

# Learning Phonotactic Distributions

Alan Prince & Bruce Tesar<sup>1</sup>  
Department of Linguistics  
Rutgers Center for Cognitive Science  
Rutgers University, New Brunswick

## 1 The problem

All languages have distributional regularities: patterns which restrict *what* sounds can appear *where*, including *nowhere*, as determined by local syntagmatic factors independent of any particular morphemic alternations. Early generative phonology tended to slight the study of distributional relations in favor of morphophonemics, perhaps because word-relatedness phonology was thought to be more productive of theoretical depth, reliably leading the analyst beyond the merely observable. But over the last few decades it has become clear that much morphophonemics can be understood as accommodation to phonotactic requirements, e.g. Kisseberth (1970), Sommerstein (1974), Kiparsky (1980), Goldsmith (1993), etc. A German-like voice-neutralizing alternation system resolves rapidly when the phonotactics of obstruent voicing is recognized. And even as celebrated a problem in abstractness and opacity as Yawelmani Yokuts vocalic phonology turns on a surface-visible asymmetry in height-contrasts between long and short vowels.<sup>2</sup>

Distributions require nontrivial learning: the data itself does not explicitly indicate the nature, or even the presence, of distributional regularities, and every distributional statement goes beyond what can be observed as fact, the ‘positive evidence’. From seeing X in this or that environment the learner must somehow conclude ‘X can *only* appear under these conditions and *never* anywhere else’ – when such a conclusion is warranted.

A familiar learning hazard is immediately encountered. Multiple grammars can be consistent with the same data, grammars which are empirically distinct in that they make different predictions about other forms not represented in the data. If learning is based upon only positive evidence, then the simple *consistency* of a grammatical hypothesis with all the observed data will not guarantee that the hypothesis is correct, even when adequate data are provided. This problem is typically characterized in terms of relationships among grammars. If one grammar generates all the observable forms licensed by another grammar along with yet other forms, then it is tricky to tell them apart, though crucial to do so. All positive data that support the less-inclusive grammar also support the broader grammar. More precisely, no fact consistent with narrower grammar can ever be *inconsistent* with the predictions of its more-inclusive competitor, even though the broader grammar also allows forms not generable – ruled out (negative evidence!) – by the less-inclusive grammar. This is *the subset problem*, familiar from the work of Angluin (1980) and Baker (1979). If a learner mistakenly adopts the superset grammar when the subset grammar is correct, no possible positive evidence can contradict the learner’s adopted but incorrect hypothesis. The misstep of choosing a superset grammar makes the subset grammar unlearnable (from positive evidence). By contrast, if on the basis of the same ambiguous evidence the subset grammar is chosen, positive evidence – observable forms licensed by one grammar but not the other – will exist to move the learner along when the superset grammar is the actual target.

As an abstract example of the problem, consider two languages:  $A = \{a^n, n \geq 1\}$  which consists of every string of *a*’s and  $AB = \{a^n b^m, n \geq 1, m \geq 0\}$ , which consists of any strings of *a*’s possibly followed by a string of *b*’s. A is a proper subset of AB. Suppose that the target language for

learning is A. If, upon encountering data from the target language (strings of *a*'s), the learner should guess that the language is AB, then there will be no positive evidence contradicting this hypothesis, and the learner is stuck in error. Now consider the situation in which the target of learning is AB. If the learner happens to encounter data consisting only of strings of *a*'s, and guesses A, an error has also been made. But here positive evidence exists which contradicts this hypothesis – namely strings containing *b*'s – and when any such are encountered, the grammar A can be rejected as inconsistent with observation.

To see the relevance to phonotactic learning, consider the case of complementary distribution. Imagine a language in which *s* and *š* both appear: what is the learner to conclude if the first encounter with *š* occurs in a word like *šite*? Suppose that as in Nupe and (Yamato) Japanese and many other languages, *š* can in fact only occur before *i*. The overhasty generalizer who decides that *š* occurs freely can never be liberated from this notion, because every further *š*-word that comes is entirely consistent with it. The negative fact that *š* fails to occur before *a, e, o, u* simply doesn't register on the positive-minded learner. If the learner had opted for a narrower grammar – say one in which *š* is derivable by palatalization of *s* – not only would the irrecoverable overgeneralization have been avoided, but there would be no difficulty in correcting the hypothesis, should the target language turn out upon further sampling to be more like English: observation of words like *šem* and *šam* would refute the subset grammar.

Within the tradition of language learnability work, the standard response to the subset problem is *the subset principle* of Berwick (1982): whenever more than one language is consistent with the data, and there is a subset relation between the languages, always select the grammar for the subset language. That way, a grammar isn't selected which *overgenerates*, that is, generates forms which are illicit when the target language is the subset language.

It might seem that the subset principle is the last word to be had on the issue: include the step in any learning algorithm, and the subset problem will be avoided. However, implementation is far from trivial. Empirical linguistic reality forces linguistic theories to permit large numbers of possible grammars, and this makes it impractical to equip a learner with an exhaustive lookup-table indicating subset relations between the languages generated by each and every pair of grammars. Given the combinatorial structure of modern linguistic theories, such a table would threaten to eclipse in size the linguistic system giving rise to the grammars. Linguistic theory aims to ease the learning problem, not to magnify it into something that requires a huge paralinguistic life-support system.

A reasonable response would be to try to compute subset relations on the basis of the individual analytical atoms of linguistic systems. For instance, different settings of a particular parameter might result in languages with subset relations. This approach has been commonly assumed when the subset principle is discussed in the context of the principles and parameters framework (see, for example, Jacobowitz 1984, Wexler & Manzini 1987). But it does not provide a complete and principled solution. Clark (1992), using what he calls “shifted languages”, shows that a learner can obey the subset principle on a parameter-by-parameter basis, yet may still end up (incorrectly) in a superset language, when subset relations depend on correlated parameter settings. Even more radically, when the elements of a linguistic theory *interact*, in violation of the Independence Principle of Wexler & Manzini, the possible subset (and non-subset) relationships between different settings of one parameter may depend crucially upon how other parameters of the system are set.

Consider the following schematic example. Suppose a subsystem in the principles and parameters framework has three binary parameters, as illustrated in Table 1 below. The subset/superset relations of languages distinguished by the settings of P1 depend upon how the other parameters, P2 and P3, are set.

**Table 1 Hypothetical parametric system: the subset/superset implications of P1 depend upon the settings of P2 and P3 (Note: *a-f* are observable forms, not entire structural descriptions).**

P1	P2	P3	Language	Subset Relations
–	–	–	{ <i>a,b,c,d</i> }	<i>Neither subset of other</i>
+	–	–	{ <i>a,b,e,f</i> }	
–	+	–	{ <i>a,b</i> }	–P1 ⊆ +P1
+	+	–	{ <i>a,b,e</i> }	
–	–	+	{ <i>a,c,d</i> }	+P1 ⊆ –P1
+	–	+	{ <i>a,d</i> }	

Observe the relationship between the languages defined for the opposing values of P1. When, as in the first pair of rows, the other parameters are set –P2 /–P3, the –P1 and +P1 languages have no subset relations. When, as in the middle pair of rows, the other parameters are set +P2/–P3, the language with –P1 is a subset of the language with +P1. Finally, when the setting is –P2/+P3, the language with –P1 is now a *superset* of the +P1 language. The learner cannot assume in advance that a particular value for parameter P1 is *the* subset value.<sup>3</sup> Positing interaction is an almost inevitable consequence of the drive for generality in theories of linguistic form, and in section 6 below, we point to a realistic analog within Optimality Theory.

The subset principle, then, is really more of a *goal* than a computational means for achieving that goal. It is a property possessed by a satisfactory theory of learning-from-positive evidence: not a mechanism within the theory, but a desired consequence of its mechanisms. *Given a set of data and a set of grammars which are consistent with that data, the learning device functions to pick the grammar generating the language which is a subset of the languages generated by the other consistent grammars, when the subset/superset relation exists.* There is no guarantee as to the computational ease of determining which of the grammars is the subset; indeed difficult theorem-proving may be required.

The subset principle can be seen as an instance of the general occamite analytical strategy: an inquirer should always select the ‘most restrictive’ hypothesis consistent with the data. But the term ‘restrictive’ has a number of different applications in the grammatical context, and it is worthwhile to sort out the one that is relevant here. Under certain interpretations, it applies *only* to candidate theories of Universal Grammar and not at all to individual grammars. In one such sense, perhaps the primary sense of the term, a grammatical *theory* is ‘restrictive’ to the degree that it provides few grammars consistent with determinative data (cf. the concept of VC-dimension in formal learning theory, as in Vapnik and Chervonenkis 1971); a ‘restrictive’ theory eases the learning task by limiting ambiguity in the interpretation of the data. In another sense, also applied at the same level, a theory of grammar is sometimes said to be ‘restrictive’ if it allows for a relatively limited typology of distinct grammars. A theory of UG that predicts only the basic word orders {SOV, SVO} is sometimes said to be more restrictive than one that predicts, say, {SOV, SVO, VSO}. This second sense is independent of the first and need have no relevance to learnability, so long as the evidence for each of the options is readily available and

unambiguous. In contrast to these, the sense that interests us applies to individual grammars within a *fixed* theory of grammar: if grammar G1 generates a language which is a subset of the language of grammar G2, then we will say that G1 is *more restrictive* than G2. Yet further uses of ‘restrictive’ at this level of analysis are occasionally found, as the term comes to be associated with any property presumed desirable that involves having less of something. But criteria of relative grammatical desirability where no subset relationship exists – for example, having fewer rules – have been considered to be of limited concern for language learnability. Competing grammar-hypotheses that do not lead to subset languages can be distinguished by data: each generated language has forms not found in the other, which can appear as positive evidence.

The subset issue appears in full force in the learning of distributional regularities. Indeed, the descriptive language used to depict phonotactics usually takes the form of stating what *cannot* occur in the language, even though the data available as evidence depict only those things which *can* occur. For the learning of phonotactic distributions, then, the goal is to always select the most restrictive grammar consistent with the data. The computational challenge is to efficiently determine, for any given set of data, which grammar is the most restrictive. Since psychological realism demands efficient computability, solving this problem is crucial to the generative enterprise.

## 2 The target grammar

Under OT, the restrictiveness of a grammar depends upon the relative ranking of the constraints: given several alternative grammars that are equivalent over the observed data, we need to be able to choose the ranking that projects the most limited language. The possible interactions among constraints can be quite complex, but a first cut can be made by observing the interactions between markedness constraints and faithfulness constraints in the core theory. Markedness constraints, violated by structures in the output, can function to prevent such structures from appearing in grammatical forms. Faithfulness constraints, on the other hand, are violated by input-output disparities, and can be ranked so as to require that input structures be retained even when they are marked or entail the presence of other marked structure. Broadly speaking, increased domination of markedness constraints over faithfulness constraints will lead to a reduced language consisting of relatively unmarked forms. This suggests that subset/superset configurations among observable language data can be managed by attention to markedness/faithfulness relationships within the grammar.

Important steps in this direction have been taken by theorists of child acquisition of phonology such as Gnanadesikan (1995), Levelt (1995), Demuth (1995), Smolensky (1996a), Bernhardt & Stemberger (1997), who have shown how the limited early repertory, and progress away from it, can be understood in terms of an initial stage where all markedness constraints dominate all faithfulness constraints ( $M \gg F$ , for short).<sup>4</sup> Sherer (1994), van Oostendorp (1995), Smolensky (1996b), and Hayes (1999, this volume) have also observed the relation between markedness dominance and the subset principle. It is important, however, to keep the subset issue notionally distinct from issues in the analysis of early acquisition patterns. The proposals we will entertain are not intended to provide a direct account for child language data, although we expect that they ought to bear on the problem in various ways. Speech relates to grammar as behavior to knowledge, with all the attendant complexities of interpretation.

To see the differences between markedness and faithfulness explanations for the same ambiguous data, let us reconsider the complementary distribution problem mentioned above. Suppose a language allows *ʃ* only before *i*, and bans *s* from that environment. When presented

with a form ...*ši*... , why doesn't the learner simply conclude that the segment *š* is generally and freely admitted in the language? UG must provide a constraint that is effectively antagonistic to *š*, call it  $M(*š)$ , and a constraint with the effect of banning the sequence *si*,  $M(*si)$ .<sup>5</sup> Restricting discussion to the faithfulness constraint  $F(\text{ant})$  demanding the preservation of *anteriority*, we have two distinct accounts of the observed datum:

- (a)  $F(\text{ant}) \gg \{M(*š), M(*si)\}$  • the faithfulness explanation.
- (b)  $M(*si) \gg M(*š) \gg F(\text{ant})$  • the markedness explanation.

The first, via dominant faithfulness, admits *all š*. The second enforces the  $M \gg F$  structure and admits *š* only in *ši*, eliminating it elsewhere: this is the most restrictive way to admit some *š* under the assumptions we have made about the constraint system. The  $M \gg F$  solution obtains a  $ši \rightarrow si$  map by specifying a ranking *entirely within* the  $M$  bloc.

Several properties of the account are worth noting.

- In=Out. No special assumptions about distinctions between underlying and surface form are required, and consequently no techniques for hypothesizing, testing, or constraining underlying forms need be called on. The learner encounters a datum and tries to get the grammar to reproduce it exactly, taking input and output to be identical.<sup>6</sup> This involves active learning, because the  $M \gg F$  condition generally encourages *nonidentity* between input and output.

- Richness of the Base. Because the learner manipulates constraints on the elements of structure rather than simply listing encountered forms, the grammar that is learned will have consequences beyond the motivating data. In the case at hand, the learner must also have grasped  $M(*š) \gg M(*s)$ , for those constraints  $M(*s)$  that penalize *s*. This ranking could come from whatever cognitive mechanisms lead to universal markedness hierarchies; or it could come from direct observation of the distribution of *s* in the language. For example, given the occurrence of ...*sa*..., the mapping  $/sa/ \rightarrow sa$  will be posited, and under the  $M \gg F$  regime, this requires  $M(*š) \gg M(*s)$ ; the reverse ranking would need Faithfulness to preserve *s*. Consequently, when the learner opts for the markedness solution to the appearance of ...*ši*... , a grammar banning *\*si* is automatically obtained; crucially, there is no contemplation of the unobserved  $/si/$  as a *possible input*. The distinction between *s* and *š* has been contextually neutralized in the effort to reproduce observed ...*ši*... against the background supposition that *š* should not exist. By configuring a grammar to perform the identity map under  $M \gg F$  bias, the learner will end up with a grammar that deals with all possible inputs (“richness of the base”), without having to review each of them.

- Durability of  $M \gg F$  structure. Learning takes place over time as more data accumulate. The choice between faithfulness and markedness solutions recurs at each stage of the process. It is not enough to set up an initial state in which  $M \gg F$ ; rather, this must be enforced throughout learning, at each step. This point figures in Itô & Mester (1999b) and is also explicitly recognized by Hayes (1999, this volume). Further justification is provided in section 4.4 below.

The basic strategy, then, is to seek the most restrictive grammar that maps each observationally-validated form *to itself*. This grammar will give us  $G(d_i) = d_i$  for every datum  $d_i$  – the  $d_i$  are the ‘fixed points’ of the mapping – and should give us as few other output forms as possible, given arbitrary input. Although we seek restrictiveness, we do not attempt to directly monitor the contents of the output language, *i.e.* to compute the entirety of  $G(U)$ , where  $U$  is the set of universally possible inputs. In essence, we have set ourselves a different but related goal: to find a grammar that *minimizes* the number of fixed points in the mapping from input to output. Each faithfulness constraint demands that some piece of linguistic structure be fixed in the grammatical mapping, and faithfulness constraints combine to force fixed-point mappings of whole composite forms. By keeping the faithfulness constraints at bay, we will in general discourage fixed-point mappings in favor of descent to the lowest markedness state possible. The

goal of fixed point minimization is not abstractly identical to language-based restrictiveness – they are the same only when the generated language is exactly the set of fixed points – but they appear to be quite close under standard assumptions about the nature of markedness and faithfulness constraints.<sup>7</sup>

Minimization of fixed points is, again, not something to be directly computed. We need a measure that can be applied to the grammar itself, without reviewing its behavior over the set of possible inputs. The idea, of course, as in the literature cited above, is that markedness constraints should be ranked ‘as high as possible’ and faithfulness constraints ‘as low as possible’. We can move in the direction of greater precision by specifying that the faithfulness constraints should be dominated by as many markedness constraints as possible. This is precise enough to generate a numeric metric on constraint hierarchies, which we will call *the r-measure*.

- (1) **R-measure.** The r-measure for a constraint hierarchy is determined by adding, for each faithfulness constraint in the hierarchy, the number of markedness constraints that dominate that faithfulness constraint.

The measure is coarse, and it does not aim to reckon just the crucial M/F interactions; but the larger the r-measure, the more restrictive the grammar ought to be, in the usual case. Even so, the r-measure is not the end-all, be-all characterization of restrictiveness. It says nothing of the possible restrictiveness consequences of different rankings among the markedness constraints or among the faithfulness constraints. Below we will see that the intra-markedness issue is solved automatically by the class of learning algorithms we deal with; the intra-faithfulness problems are distinctly thornier and will be the focus of much of our discussion.

It is also not obvious (and perhaps not even likely) that for every data-set there is a unique hierarchy with the largest r-value.<sup>8</sup> But the r-measure is a promising place to start. Given this metric, we can offer a concrete, computable goal: any learning algorithm should return a grammar that, among all those consistent with the given data, has the largest r-measure.

**A conflict avoided: high vs. low.** Before asking how to achieve this goal, it is worth noting a couple of the r-measure’s properties. First, the most restrictive possible grammar, by this measure, will have all of the faithfulness constraints dominated by all of the markedness constraints, with an r-measure that is the product of the number of faithfulness constraints and the number of markedness constraints. Thus, for a given OT system, the maximum possible r-measure is that product.

Second, selecting the hierarchy with the largest r-measure is consistent with the spirit of ranking each of the markedness constraints ‘as high as possible’, i.e., with respect to all other constraints, both markedness and faithfulness. This distinguishes the r-measure from another possible measure: that of adding together, for each faithfulness constraint, the number of constraint *strata* above it – call it the s-measure. (A stratum is a collection of constraints undifferentiated in rank, produced by the learning algorithms we employ.) Consider the stratified hierarchy in (2): the r-measure is 2+3=5, while the s-measure is 1+3=4.

$$(2) \{M_1 M_2\} \gg \{F_1\} \gg \{M_3\} \gg \{F_2\} \quad r=5, s=4$$

Now consider the stratified hierarchy in (3), which is just like (2) save that the top stratum has been split, creating a dominance relationship between  $M_1$  and  $M_2$ .

(3)  $\{M_1\} \gg \{M_2\} \gg \{F_1\} \gg \{M_3\} \gg \{F_2\}$   $r=5, s=6$

Adding this new intra-markedness relationship doesn't directly change the relationship between any of the faithfulness constraints and any of the markedness constraints. Correspondingly, the r-measure of (3) is 5, the same as for (2). However, the s-measure of (3) is 6, larger than the s-measure for (2). Maximizing the s-measure will have the effect of demoting markedness constraints in ways not directly motivated by the data. By maximizing the r-measure instead, learning can be biased toward having faithfulness constraints low in the ranking without directly impacting the relative ranking of the markedness constraints. This turns out to be a good thing because, as discussed below, it permits the continued use, when ranking the markedness constraints, of the principle of constraint demotion (CD), which mandates that constraints be ranked as high as possible. Using the r-measure instead of the s-measure to characterize 'as low as possible' for faithfulness constraints avoids a possible conflict between such low-ranking and the goal of ranking markedness constraints maximally high.

### 3 The learning situation

For the purposes of this paper, we assume a learner not (yet) interested in unifying underlying forms for morphemes. Such a learner is responsive only to syntagmatic factors, and may simply treat prosodic words as morphologically opaque entities. If some surface morphological segmentation has been achieved, the learner could also detect distributions sensitive to whatever categories can be posited as present in observed forms; similarly for the analysis of prosodic structure. We further assume that the learner always adopts, as the underlying form, precisely the surface analysis of the overt form that has been heard, i.e. whatever analysis the learner is currently capable of. The correct mapping will always be an identity mapping for the structures in the surface form; these must be fixed points for the grammar. This does not mean that the correct mapping is identity over all inputs; in the vast majority of cases there are other inputs not observed or contemplated that are mapped unfaithfully to lower-markedness outputs. All overt forms actually observed as positive data must be generable by the grammar, and their corresponding underlying forms must, we assume, be mapped faithfully.

These assumptions are not adopted solely to keep the problem simple. Ecologically, it is reasonable to start the learning process with observations that can be made without extensive comparative analysis of words and their meanings; hence the overlap in assumptions with empirical work on early phonological acquisition in children, such as Gnanadesikan (1995), Levelt (1995), Demuth (1995), and others. Linguistically, it is clear from studies, such as those by Kisseberth (1970), Sommerstein (1974) and by now many others, that phonotactic knowledge is of significant benefit when the task of learning the morphophonology is taken up.

Distributional evidence has some inherent limitations: from data with only CV syllables, one cannot distinguish whether epenthesis or deletion is responsible for dealing with problematic inputs. In general, if UG allows a range of solutions to a markedness constraint, distributional evidence will only assure the learner that at least one of them is in use, without specifying which one. The more tightly the range is specified, either by UG or by other informative constraint interactions, the more the learner will grasp. But even in the case of ambiguity, the analysis of distributional data will give excellent evidence about relations within the markedness subcomponent of the grammar, and very good evidence as well about the niches in the hierarchy where faithfulness constraints fit. The stage investigated here will typically not fix the complete constraint hierarchy for the entire phonology of the language, but it will aim to provide a crucial

infrastructure that limits the range of options that the learner must consider. According to this program of explanation, the increasingly sophisticated learner will have the luxury of revising the constraint hierarchy developed through phonotactic learning as knowledge of morphophonemic relations develops, using further techniques for hypothesizing and testing nontrivial underlying forms, ultimately arriving at the correct adult grammar.

Having taken *via* the r-measure a global stance on the restrictiveness of grammars, the challenge we face is to devise an efficient and local step-by-step algorithm that produces grammars in accord with it. The transition between global goal and local implementation is not entirely trivial, as we will see.

Our point of departure is an existing procedure for constructing a constraint hierarchy from data: Recursive Constraint Demotion. RCD generates a grammar consistent with a set of *mark-data pairs*. A mark-data pair (mdp) is an elementary ranking argument: a competition between two candidates from the same input, one of them optimal, the other not, along with the information about their constraint-violation profiles. The member of the pair that is (desired to be) optimal will be termed *the winner*; the other, suboptimal competitor will be *the loser*. The important action takes place comparatively: the raw quantitative violation data must be analyzed to determine, for each constraint, which member of each pair does better on that constraint, or whether the constraint is neutral between them; mere violation is, of course, meaningless by itself. (This step is known as *mark cancellation*.) We need a technical term to describe the basic comparative relationship that emerges after mark cancellation: let us say that a constraint *prefers* one candidate to another when the first candidate has fewer violations than the second. Note that a candidate can satisfy a constraint completely, yet still not be ‘preferred’; it all depends on how its mdp competitor fares.

A single md-pair contains information about how the constraints must be ranked so that the desired winner actually wins. At least one of the constraints preferring the winner must dominate *all* of the constraints preferring the loser. This, rephrased, is the Cancellation/Domination Lemma of Prince & Smolensky (1993:148). We know from Cancellation/ Domination that the loser-preferring constraints must *all* be subordinated in the ranking, and subordinated to constraints that prefer the associated winners. Turning this around, we see that the complementary set of constraints – those that *do not prefer any losers* among the mdp list, those that only prefer winners when they prefer anything – can all be ranked at the top. The algorithm develops from this observation.

RCD applies to a constraint set coupled with a list of elementary ranking arguments (mdp’s, possibly arising from different inputs), and recursively assembles a constraint hierarchy consistent with the information in the mdp-list. RCD does not aim to construct all possible totally-ordered rankings consistent with an mdp-list. The goal of RCD is to produce a single *stratified hierarchy*, where the strata consist of nonconflicting constraints. Various total orderings can be produced by ranking the constraints within the strata, retaining the domination order between strata. The stratified hierarchy produced by RCD has a unique property: each constraint is placed in the highest stratum that it can possibly occupy.

The goal of grammar-building is to organize the constraints so that all known ranking arguments are simultaneously satisfied. If any such constraint hierarchies exist, the stratified hierarchy produced by RCD algorithm will be among them.

RCD starts by collecting all the constraints that prefer only winners when they prefer anything (they *prefer no losers*). These are placed at the top of the hierarchy, forming a stratum of non-conflicting winner-preferring and winner-loser-neutral constraints. The learner then dismisses from further consideration all those md-pairs  $W_i \sim L_i$  on which some constraint just ranked *prefers* winner  $W_i$  to loser  $L_i$ . This move discards those competitors that lose on some constraint in this stratum, and leaves behind those md-pairs where both members have done equally well. In this way, the list of md-pairs shrinks just as the list of constraints to be ranked is shrinking.

The learner then faces exactly the same problem again – ranking a set of constraints so as to be consistent with a list of ranking arguments, but with fewer constraints and fewer md-pairs. This is a canonical recursive situation, and RCD simply repeats its ranking-and-elimination procedure on the depleted collection of md-pairs and constraints. The second stratum of the hierarchy is now filled with those constraints that prefer only winners among the md-pairs left over from the first round. If the ranking arguments in the mdp-list are mutually consistent, the RCD process will repeat until all md-pairs have been dismissed, and all constraints have been placed in the hierarchy. A grammar will have been found.

Should the data be inconsistent, a point will be reached where none of the remaining constraints prefers only winners – each yet-to-be-ranked constraint will prefer at least one loser on the remaining mdp list. These constraints cannot be ranked, and it follows that no grammar exists for the original mdp list over the constraint set. Our focus here will be on finding grammars for consistent data, but we note that detection of inconsistency via RCD can play an important role whenever generation and pruning of tentative, fallible interpretations of data is required, as shown in Tesar (1997, 1998).

Here’s a brief example that illustrates the use of RCD. The mdp list is presented in the form of a comparative tableau (Prince 2000; 2001a,b), the appropriate representation for ranking arguments. Competing members of a mark-data pair are shown as ‘x~y’, where ‘x’ is the desired optimum or winner, and ‘y’ is a competitor and desired loser. As always, we say that a constraint *prefers* candidate x to candidate y if x has fewer violations of the constraint than y does; and vice versa. If a constraint prefers one candidate of the pair to the other, then a token that indexes the preferred candidate is placed in the appropriate cell: ‘W’ if it prefers the winner, ‘L’ if it prefers the loser. If a constraint prefers neither candidate (because both violate it equally), the cell is left blank. A constraint hierarchy (arrayed left-to-right in domination order, as usual) is successful over an mdp list if the first nonblank cell encountered in each row, going left to right, contains W. This indicates that the highest ranking constraint that distinguishes the members of the pair prefers the winner. In terms of the comparative representation, then, the action of RCD is to take a tableau of W’s and L’s and find an arrangement of the columns in which all L’s are preceded by at least one W. Let us perform RCD.

MDP: Winner ~ Loser	C1	C2	C3
a. W1 ~ L1	W	L	
b. W2 ~ L2	L	W	W

Of the three constraints, C3 alone *prefers only winners* (prefers no losers). The first round of RCD will therefore place C3 in the top stratum. This eliminates mdp (b) by virtue of the ‘W’ awarded by C3, leaving the following:

MDP: winner ~ loser	C1	C2
a. W1 ~ L1	W	L

Now C1 *prefers only winners* (and more, albeit rather trivially: lacking neutral cells, it prefers *all* winners). C1 goes into the second stratum and eliminates mdp (a), exhausting the mdp-list and relegating C2 to the lowest stratum.

Notice that C1 did not at first *prefer-only-winners*, prior to the elimination of mdp (b), because C1 prefers L2. Thus, we can say that C1 was *freed up for ranking* when C3 entered the first stratum. The resulting hierarchy,  $C3 \gg C1 \gg C2$ , leaves the tableau looking like this:

MDP: winner ~ loser	C3	C1	C2
a. W1 ~ L1		W	L
b. W2 ~ L2	W	L	W

As promised, each row begins with a ‘W’ that settles the competition in favor of the winner.

A full presentation of RCD, along with a formal analysis of its basic properties, can be found in Tesar (1995), Tesar & Smolensky (2000); for an informal account, see Prince (2000). Samek-Lodovici & Prince (1999) explores the relationship between harmonic bounding and RCD, and provide an order-theoretic perspective on both. Prince (2001b) further explores the relationship between RCD and the logic of ranking arguments.

RCD applies not to raw data but to mark-data pairs: observable winners paired with unobserved and ungrammatical forms (losers). The learner must therefore be equipped to construct an appropriate mdp list from observed data, algorithmically supplying the losing competitors. Two criteria must be satisfied by any competitor-generating mechanism. First, the generated loser must in fact be ungrammatical as an output for the given input. Second, the comparison thus constructed should be *informative*: of all the multitudes of ungrammatical outputs, the one chosen to compete should lead to the imposition of new ranking relations through RCD; otherwise there is no point to computing the comparison.

The method of error-driven learning discussed in Tesar & Smolensky (2000) satisfies both criteria. Suppose a hypothesized grammar at some stage of learning produces an *error*, an incorrect mapping from input to output. Rectifying the error requires re-ranking so that the desired candidate fares better on the constraint hierarchy than the currently-obtained but spurious optimum. RCD will accomplish this if the new md-pair  $\langle \textit{desired winner} \rangle \sim \langle \textit{currently obtained winner} \rangle$  is added to the mdp-list. The grammar itself, through its errors, thus supplies the meaningful competitors that are needed to improve its performance.<sup>9</sup> Other methods of finding useful competitors can be imagined – for example, conducting some kind of search of the ‘nearby’ phonological or grammatical space – but we will adhere to the error-driven method in the interests of focus, well-definition, and gaining a more precise understanding of its strengths and weaknesses.

It is important to note that intermediate grammar hypotheses, based on partial data, will often have strata containing constraints that turn out to conflict when more mdp’s are considered. By consequence, such grammars will tend to produce multiple outputs for each input, due to undecided conflicts, in cases where the final grammar would produce only one. Following Tesar & Smolensky (2000), we regard each output that is distinct from the observed datum as an error

to be corrected. We abstract away from issues of variation, then, and work under the idealization that each input corresponds to only one legitimate output.

The course of learning is assumed to proceed over time as in the Multi-Recursive Constraint Demotion (MRCD) of Tesar (1998). Rather than manipulating a grammar-hypothesis *per se* by switching rankings around, the learner develops and retains a permanent data-base of mdp's. A grammar hypothesis can be generated easily from the current data-base by the application of RCD. From this point of view, the contribution of the phonotactic/identity-map learning stage is precisely the accumulation of a valuable mdp data-base. MRCD allows RCD to be utilized with 'online' learning, that is, with progressive augmentation of the learner's data-base. The learner uses the current grammar to apply error-driven learning to each form as it is observed. When an error occurs on a form, the learner constructs a new mdp and adds it to the list (which was initially empty). A new constraint hierarchy is then generated by applying RCD to the entire accumulated mdp list, and the learner is ready to assess the next piece of data when it arrives. In this way, the learner responds to incoming data one form at a time, constructing and storing mark-data pairs only when necessary.

## 4 Postponing the placement of faithfulness constraints

### 4.1 Biasing RCD

While any given set of mark-data pairs is typically consistent with quite a number of stratified hierarchies, RCD returns the unique stratified hierarchy in which each constraint is ranked as high as it will go. Unmodified, the algorithm will give exactly the opposite of what we want for faithfulness constraints. In the complementary distribution case, for example, the presence of ...*ši* ... in the data is entirely compatible with top-ranking of F(ant), and so up it goes. Indeed, this ranking is compatible with the *absence* of ...*ši*... from the data, and the absence of *š* generally. Under simple RCD, all constraints – and hence all faithfulness constraints – will reside at the top until there is evidence to contrary. No such evidence will ever arise for faithfulness constraints when the identity map is optimal. Faithfulness will be demoted only when non-identity mappings are desired; that is, when input and output differ crucially, a situation that for us requires knowledge of morphophonemic alternations. A strictly RCD-based learner cannot grasp distributions from positive evidence alone.

RCD must therefore be reconfigured so as to impose high-ranking *except for faithfulness constraints*. The modified algorithm should place a faithfulness constraint into a hierarchy only when absolutely required. If successful, this strategy will yield a hierarchy with a maximal r-measure. Let us call the new algorithm – or rather *class of algorithms*, since a number of variants are possible – Biased Constraint Demotion (BCD): the ranking procedure will be biased against faithfulness and in favor of markedness constraints as candidates for membership in the stratum under construction.

On each recursive pass, RCD places into the next stratum those constraints that prefer only winners among the mdp-list. Our basic principle takes the form of the following modification to RCD, given in (4):

- (4) **Faithfulness Delay.** On each pass, among those constraints suitable for membership in the next stratum, if possible place only *markedness constraints*. Only place faithfulness constraints if no markedness constraints are available to be placed in the hierarchy.

In this way, the placement of faithfulness constraints into the hierarchy is delayed until there are no markedness constraints that can do the job. If all the md-pairs are consistent with a hierarchy having the entire set of faithfulness constraints at the bottom, that is what will be returned.

However, the data may require at least one faithfulness constraint to dominate some markedness constraint. This will be happen only when all of the constraints available for placement into the next stratum are faithfulness constraints. At this point, the algorithm must place at least one faithfulness constraint into the next stratum in order to continue.

What is not obvious is what, precisely, to do at this point. At least one faithfulness constraint must be ranked, but which one?

#### 4.2 Don't Place Inactive Faithfulness Constraints

All unranked faithfulness constraints will be available for ranking at each step of BCD. The identity map satisfies every criterion of faithfulness, and no faithfulness constraint can ever *prefer a loser* under such circumstances. All faithfulness constraints prefer only winners from the outset, and any F constraint not already ranked will be in a state of permanent readiness.

With markedness constraints, the situation is quite different; they have no affinity for the identity map or any other and are as likely to prefer losers as winners. Assuming that no markedness constraints are rankable at some stage, the crucial question is this: which faithfulness constraints *need* to be ranked? From the point of view of unbiased RCD, any F constraint at all will do. But if F constraints are to be ranked as low as possible, in accord with the r-measure, the answer must be: just those faithfulness constraints whose ranking will free up markedness constraints for ranking in the next round.

'Freeing up' comes about in general when a just-ranked constraint dismisses an md-pair in which the loser is preferred by certain constraints. With this md-pair gone, such once-problematic constraints may now fall into the prefers-only-winners category, and therefore be rankable. The following tableau illustrates this property:

MDP: Winner ~ Loser	M1	M2	F1	F2
a. W1 ~ L1	W	L		
b. W2 ~ L2	L	W	W	

At this stage, neither M1 nor M2 can be ranked – both prefer some loser (marked by L in the cell). If we rank F2, we accomplish nothing, since the set of md-pairs remains the same. If we rank F1, however, the problematic mdp (b) will be eliminated, taking with it the L-mark that holds M1 in check. With mdp (b) gone, M1 is 'freed up' and may be ranked next.

As a first step toward identifying the best F constraints to rank, we can set aside a constraint-type that it is definitely **not necessary** to rank: the F that prefers no winners, being neutral over the md-pairs at hand. Any such constraint, when ranked, will never eliminate any md-pairs and cannot free up markedness constraints. This gives us our second principle:

- (5) **Avoid the Inactive.** When placing faithfulness constraints into the hierarchy, if possible only place those that *prefer some winner*.<sup>10</sup> If the only available faithfulness constraints *prefer no remaining winners*, then place all of them into the hierarchy.

Observe that the second clause only applies when the bottom-most stratum is being constructed. Following the principle will ensure that completely inactive faithfulness constraints always end up at the very bottom of the hierarchy. (Completely inactive markedness constraints will, of course, end up at the top.) With this much of a selection-criterion at hand, we can give a version of the Biased Constraint Demotion algorithm. This is not the final version of our proposal, but it expresses the ideas presented thus far. We write it out as generic pseudocode, with italicized comments on the side.

#### 4.2.1 Version 1 of The BCD Ranking Algorithm

Repeat the procedure until all constraints have been ranked.

Given: a list of mark-data pairs, and a set of constraints distinguished as markedness/faithfulness:

	<i>[Comment:</i>
Identify the constraints NoL that prefer no losers ( <i>prefer only winners</i> )	<i>[Collect the rankable constraints;</i>
If at least one of NoL is a markedness constraint	<i>[Place only M constraints</i>
place only the markedness constraints of NoL in the next stratum	<i>whenever possible;</i>
<b>else</b>	<b><i>[but if NoL consists entirely of F</i></b>
<b>if at least one of NoL prefers a winner</b>	<b><i>Place only active F;</i></b>
<b>place [only those members of NoL that prefer a winner] in the next stratum</b>	
else	<i>[but if no F in NoL are active, form</i>
place all of NoL in the next stratum	<i>bottom-most stratum from NoL.]</i>
end-if	
end-if	

Remove all mark-data pairs where the winner is preferred by one of the just-ranked constraints

#### 4.2.2 The Logic of the Algorithm (Version 1)

The initial *If*-clause ensures that faithfulness constraints are not ranked when markedness constraints are available. Markedness constraints will continue to be ranked as high as possible, because they are placed into the ranking as *soon* as they prefer-only-winners among the current mdp-list. Faithfulness constraints are placed into the hierarchy when they are the only ones available for ranking. In such a circumstance, all remaining markedness constraints must be dominated by at least one of the rankable faithfulness constraints.

The bracketed, bolded section of the algorithm contains the conditions determining which among the available faithfulness constraints will be ranked; it is this section that we must scrutinize and sharpen. Currently, it places a necessary condition, one that will persist, but needs further elaboration. Looking only at the isolated behavior of individual F constraints, it picks up the weakest form of activity. All that is asked for is that an F constraint *prefer* some winner; the behavior of other constraints on the relevant md-pair is not considered.

Remark. Because discussion in the literature often concentrates only on M/F conflict, it is important to recall that for a markedness constraint M1 to be rendered inactive, it is not always necessary that it be dominated by every faithfulness constraint whose violation could satisfy M1. Suppose that a (suboptimal) form satisfying M1 via the violation of faithfulness constraint F1 necessarily also violates another *markedness* constraint M2. It would be adequate if M2 were to dominate M1, which in turn dominates F1. This would also be preferable due to the lower ranking of F1. (In the complementary distribution case mentioned above, M1 = \*š, M2= \*si; the failure of

$\check{s}i \rightarrow si$  is guaranteed not by F(ant) but by M2.) This is the kind of competition between markedness and faithfulness explanations which the algorithm will handle correctly.

### 4.3 Locating the Competition

Optimality Theory does not run on positive evidence alone: each elementary ranking argument turns on the contrast between an optimal form and another form presumed to be *suboptimal* – ungrammatical, in a word, as the output for the given input. The desired optima come from observation, but selecting their failing competitors asserts a kind of negative evidence. Where do they come from?

Above we sketched the notion of error-driven learning (as in Tesar 1997), by means of which the grammar itself is used as the source of informative contrasts, those which compel modification of the current hierarchy. Let us consider how this applies in the phonotactic learning context. Given a stratified hierarchy current at some stage of learning, one can compute the optima that it predicts as output for some input. When an observationally-supported *output* form is used as *input*, the current grammar may produce unwanted output from it. It may fail to reproduce the input in the output, or it may produce other forms as co-optima, because the strata contain constraint-ties that have not yet been worked out. On the assumption that there is a unique output for each input, all such nonmatching forms are guaranteed to be suboptimal, and they can be used as fodder for ranking arguments (mdp's). They are errors, useful because they can be identified and corrected.

Because the identity map is the sought-for optimum, the losing competitors produced by this method do not earn their erroneous superiority by being more faithful, and must instead be *less marked*, in terms of the current stratified hierarchy. Expanding the mdp-list to include these implied suboptima leads to a new grammar when BCD is applied, either through new rankings within the markedness group, or through high-ranking of a faithfulness constraint, if necessary.

Negative evidence acquired in this fashion does not provide the learner with the tools to resolve the subset problem. The driving assumption – that each input gives rise to one output – is no help at all. Having spotted  $\check{s}i$  in the environment, the learner can be sure that  $/\check{s}i/$  must come out only as  $\check{s}i$ , not as  $si$  or  $\check{s}a$ . But no evidentiary hint is offered as to what an input like  $/\check{s}a/$  should lead to, when  $\check{s}a$  has never been observed. The information gained from errors is not evidence about what's missing from the overt, observable language. This is the key question, and the one that cannot be asked: is  $\check{s}$  free or restricted in distribution? is  $\check{s}a$  in the language *or not*? In the absence of positive evidence to the contrary, BCD aims to resolve such questions in favor of the restrictive pole of the disjunction.

### 4.4 The Initial State: One among Many

Linguistic theorists and language acquisitionists alike often focus on the *initial state* as having special properties. Under BCD, the initial state is not arbitrary, nor does it require special stipulation. Assume a state of complete ignorance in the learner; apply the algorithm to an *empty* list of md-pairs. All of the markedness constraints are placed into the hierarchy first, followed by the faithfulness constraints. The initial hierarchy is useful in the present context because it provides a starting point for error-driven learning.

This hierarchy is the familiar  $M \gg F$  configuration: it is the same initial state that has been proposed by Smolensky (1996a,b) and others cited above. Smolensky proposes that the learner

starts with this initial hierarchy, and constraint demotion subsequently mixes the faithfulness and markedness constraints, with the expectation that markedness constraints would only be demoted below faithfulness constraints to the extent necessary. (Independent of constraint demotion, Itô & Mester (1999a) propose a principle of ‘ranking conservatism’ that aims to have the effect of maximally preserving initial-state ranking relations.) The present proposal is another, stronger version of the same theme: instead of confining the asymmetry between markedness and faithfulness to the initial state, we posit a constant markedness-favoring force throughout learning. Under the new proposal, at any given time, the subordination of markedness to faithfulness constraints will always aim to be the minimum necessary to accommodate the data. The initial state is but one consequence of this constant force.

The two approaches are not equivalent. Consideration of their differences shows that  $M \gg F$  cannot just be an *initial state*, subject to minimal perturbation by any version of CD that fails to take direct account of the M/F distinction. As new data arrive, on the Smolensky and Itô & Mester views, learning involves a revision of the currently-maintained hierarchy or hierarchies, and under standard ‘on-line’ CD theory, any such revision can upset the general  $M \gg F$  relation quite massively.

Imagine the two initial strata:

$$\begin{array}{c} \{M_1, M_2, \dots, M_n\} \gg \{F_1, F_2, \dots, F_k\}. \\ \leftarrow M \quad \rightarrow \quad \leftarrow F \quad \rightarrow \end{array}$$

Suppose a datum comes in which requires  $F_1 \gg M_1$ . Standard ‘on-line’ CD, operating on the current hierarchy, demotes  $M_1$  to the highest possible stratum, a new one in this case, retaining as much as possible of the current hierarchy:

$$\{M_2, \dots, M_n\} \gg \{F_1, F_2, \dots, F_k\} \gg \{M_1\}$$

The measure  $r$  diminishes by  $k$ , the number of F constraints that  $M_1$  no longer dominates, which is in fact the entire list of faithfulness constraints! CD has grossly upset the  $M \gg F$  relation, with potentially great and unwanted consequences for the violation of  $M_1$ . Standard CD is conservative of the hierarchy’s stratum structure, but allows very significant shifts in domination relations. Under BCD, the resulting hierarchy would have more strata, but minimal dominance of F over M:

$$\{M_2, \dots, M_n\} \gg \{F_1\} \gg \{M_1\} \gg \{F_2, \dots, F_k\}$$

Here  $r$  descends by just 1, and  $F_1$  dominates only the single M constraint that the data compels it to. Recall that BCD does not directly transform one ranking into another, but re-derives an entire hierarchy from the enhanced list of md-pairs.

Itô & Mester’s notion of ranking conservatism seeks to regulate transition from one state of the grammar to the next by avoiding, to the extent possible, the imposition of new rankings that contradict those already established. This leads to a preference for within-stratum refinement, which establishes rankings where none had existed before, as opposed to reversing dominance relations that are currently present. The expectation, then, is that the initial two stratum  $M \gg F$  structure will be carried forward as an  $M \gg F$  preference, via conservatism, into the hierarchies that descend from it. The idea is plausible and attractive, but it encounters at least two serious

difficulties. First, the dynamics of ranking change under this proposal can be rather complex, due to the effects of transitivity, and the original relations (and stratum-internal nonrelations) can get lost after a few steps, so that it is no longer possible to distinguish correctly on the basis of conservatism between markedness and faithfulness solutions, when both are possible. Second, it runs into a problem diagnosed by Broihier (1995) for a modification of the CD algorithm that incorporates a kind of ranking conservatism. It turns out that a conservative strategy can stumble badly over the rankings it preserves from previous states, in ways that depend pathologically on the sequencing of the data. In the worst case, the strategy will interpret consistent data as if it were inconsistent, leading to a nonterminating loop of demotion and re-demotion of the same pair of constraints. (Details of these problems are presented in Appendix 1.) Until such issues are resolved, we feel compelled to maintain the hypothesis that the  $M \gg F$  bias is an intrinsic, persistent property of the learning algorithm.

Other articulations can be added to the initial state, e.g., markedness scales enforced through fixed domination relationships, etc. As with the  $M \gg F$  bias, any such additional fixed relationships must be actively enforced throughout learning. Ongoing respect for fixed constraint relationships is not an additional burden created by the RCD-based approach; any theory of learning constraint rankings must achieve it. Under the approach proposed here, based on RCD, all such relationships must be continuously active, and the initial state follows as a consequence.

## 5 Freeing up markedness constraints

### 5.1 The Problem

The overarching goal in placing F constraints is to free up markedness constraints for ranking in such a way as to maximize the r-measure: typically, this will mean freeing up the maximal number of markedness constraints, and freeing them up as soon as possible. A markedness constraint is fully rankable when it reaches the prefers-winners-only state, a purely constraint-internal matter. The F constraints, by contrast, are in that state from the start, and it is their interaction with other constraints that determines when they should be ranked. F constraints should go into the hierarchy only when no markedness constraints are rankable; but when several F constraints are available, the correct choice among them need not be straightforward. Here we review the kind of situations that can arise and develop a strategy that leads to a sharpened version of the BCD algorithm.

**Role of Conflict.** Consider first the configuration portrayed below. No M constraints are rankable.

MDP: Winner ~ Loser	F1	F2	M1
a. W1 ~ L1	W		L
b. W2 ~ L2		W	

Should we rank F1, F2, or both? Observe that F2 is quite useless, even though it prefers a winner and is therefore active in the sense used above. By contrast, F1 *conflicts* with M1 and *must* be ranked above it. (Constraint conflict means that ‘W’ and ‘L’ both occur in the same row.) Even better, when F1 is ranked, it frees up M1. Because F2 conflicts with nothing, it makes no difference at all where it is ranked. By the r-measure, it should be ranked at the bottom. This would also be true if M1 preferred the winner for mdp (b). The first moral, then, is that F-constraints should be ranked only if they *conflict* with other constraints.

The present approach already achieves this result without the need for further intervention. In cases of *nonconflict* where F is active, the loser is harmonically bounded by the winner. (Harmonic bounding by the winner in comparative tableaux is signaled by the occurrence of a tableau row containing only W's and possibly blanks.) The error-driven competitor-selection mechanism will not find harmonically-bounded losers, because they never appear as winners for any provisional hierarchy. Consequently, no such competitions will arise under the error-driven regime, and no special clauses will be required to interpret them. In addition, error-driving aside, it should be noted that some logically possible patterns may simply not occur: winner/loser pairs distinguished by faithfulness but not by markedness (as in mdp (b) above).

**Freeing up.** Conflict is a necessary precondition for freeing up, because every removal of an L-containing row involves conflict resolution. But it is not sufficient. *All* L's in a constraint column must go, if the (M) constraint is to become rankable. In general, there may be many paths to freedom. Let us examine two fundamental complications: gang-effects, and cascading consequences.

**F-Gangs.** It can easily be the case that no single F constraint working by itself can free up a markedness constraint, yet a set of them can combine to do so. Here's an illustration:

MDP: Winner ~ Loser	F1	F2	F3	M1
a. W1 ~ L1	W		W	L
b. W2 ~ L2		W		L
c. W3 ~ L3	W			L

M1 will be freed up if both F1 and F2 are ranked. F3 can also be thrown into the mix, but it merely duplicates the work already done. For maximal r-measure, F3 should not be ranked at this point. (Note that this also means that a procedure of randomly ranking active F constraints one-at-a-time is sure to be r-wise defective.) The algorithm must not only be able to identify effective sets of F constraints, it must be able to adjudicate between different such 'gangs' in order to determine which is most efficient.

There are two dimensions of efficiency to be measured: the size of the F-gang, and the size of the set of M constraints that it frees up (on which more immediately below). These two measures can trade-off under special circumstances. Although the general expectation in ranking F is surely that fewer is better, it is possible to construct cases in which a greater initial investment in F yields a better r-measure in the relative ranking of M and F overall.<sup>11</sup> Pursuing this potential trade-off with full vigor is computationally unattractive, and therefore offensive to mental realism. Until there is evidence that it is meaningful or even likely to occur with realistic F and M constraints, we prefer to develop a version of BCD that simply favors smaller F-gangs. This leads to the following principle:

- (6) **Smallest effective F sets.** When placing faithfulness constraints into the hierarchy, place the *smallest set* of F constraints that *frees up some markedness constraint*.

This strategy leaves open the action to be taken when more than one F-gang is of minimal size. To see what hangs on the choice between them, we turn to our second focus, the cascade of distant consequences that can flow from an F-ranking decision.

**M-Cascades.** Placing an F constraint may not only free up *some* markedness constraint – it may free up a markedness constraint that itself frees up yet more markedness constraints. Setting loose such a cascade can result in higher ranking of M and lower ranking of F than would otherwise occur.

Suppose that at a certain stage there are two F constraints freeing up distinct markedness constraints. Consider the following situation:

**Table 2 When should F1 and F2 be ranked?**

MDP: Winner ~ Loser	F1	F2	M1	M2
a. W1 ~ L1	W		L	
b. W2 ~ L2		W	W	L

Neither M constraint can be placed at this point, since each prefers at least one loser (cells marked L). We must look to the F constraints. Both F1 and F2 are equally suitable by the freeing-up criterion, and each forms a smallest set of rankable F. If F2 is ranked first, the following hierarchy results:

$$(7) F2 \gg M2 \gg F1 \gg M1 \quad r = 1$$

If F1 is ranked first, though, the md-pair (a) is eliminated, and with (a) gone, the constraint M1 suddenly becomes available for ranking. Ranking M1 next then frees up M2 for ranking, and F2 drops to the bottom. The result is hierarchy (8):

$$(8) F1 \gg M1 \gg M2 \gg F2 \quad r = 2$$

These considerations indicate that the r-measure can respond to effects that are not immediately visible in the next round of ranking, but are consequences of that next round, or consequences of consequences, and so on. Here again a tactical issue arises: to pursue the global r-measure unrelentingly would probably require, in the worst case, sorting through something close to all possible rankings. It is doubtful that that there is much pay-off in this mad pursuit. We conjecture that it will suffice for linguistic purposes to consider only the M-freeing consequences up to the next point where an F constraint must be ranked, i.e. until the cascade of M constraints arising from F-placement runs out. This leads to our last principle:

- (9) **Richest Markedness Cascade.** When placing faithfulness constraints into the hierarchy, if more than one F-set *freeing up some markedness constraint* is of smallest size, then place the F-set that yields the largest set of M constraints in contiguous subsequent strata, i.e. until another F constraint must be ranked. If there is more than one of these, pick one at random.

As with the smallest-F-set criterion, it is possible with a sufficient number of suitably interacting M and F constraints to craft a case where going for a smaller immediate M-cascade is offset by gains further on.<sup>12</sup> More investigation is needed to see what role, if any, this might play in linguistic systems, and whether such distinctions actually lead to discernible differences in substantive restrictiveness. If, as we suspect, they do not, then the suggested criterion may be more than just a reasonable heuristic compromise between the r-measure and computational efficiency. On the other hand, if distinctions of richness between immediate cascades fail to lead to discernible differences in restrictiveness of the resulting grammars, then it might even be preferable to forgo the effort of computing them as well.

## 5.2 BCD Revised

Implementing the principles developed above requires further articulation of the F-placement clause given in the preliminary version of the algorithm. We need to install a subroutine that seeks the smallest effective set of F constraints – the smallest set that collectively frees up some markedness constraint – and in case of ties on size, looks for the one that releases the biggest markedness cascade.

The subroutine we propose starts out by looking for single effective F constraints, and if this fails, advances to looking for two-member sets, and so on. Although the combinatorics are unpleasant in the worst case, experience to date suggests that the routine will never have to look very far. If the result of the sizing subroutine is unique, it is turned over to the main routine for ranking. If nonunique, it is turned over to a further subroutine that pursues the markedness cascade released by each contender, looking ahead, as it were, via calls to the main BCD routine. The most successful F-set is returned to the main routine for ranking; ties for success are decided randomly.

We note that the algorithm could be pushed in the direction of greater complexity by being restructured to chase down ever more distant consequences of ranking decisions. It could also be scaled back, if its level of sensitivity to the ultimate  $M \gg F$  structure turns out to be excessive when the limitations of realistic constraints are taken into account. At present, it stands as a reasonable, if in part conjectural, compromise between the global demands of the r-measure and the need for local computability.

### 5.2.1 The full BCD algorithm

Given: a list of mark-data pairs, a set of constraints distinguished as markedness/faithfulness

#### Main Routine

**BCD**( ):

Repeat (Until all constraints have been ranked)

Set **NoL** to be the constraints not yet ranked that prefer no losers

If at least one of **NoL** is a markedness constraint

    Set **RankNext** to be the set of markedness constraints in **NoL**

Else

    If at least one of **NoL** prefers a winner

        Set **RankNext** to be the constraint set returned by **Find-minimal-faith-subset**(**NoL**)

    Else

        Set **RankNext** to be all the constraints of **NoL**

    End-if

End-if

Place **RankNext** in the next stratum

Delete all mark-data pairs where the winner is preferred by a constraint in **RankNext**

End-Repeat

### Faithfulness Constraint Set Selection

**Find\_minimal\_faith\_subset**(NoL):

Set `ActiveFaith` to be those members of NoL that prefer a winner in one of the remaining md-pairs

Set `FsetSize` to start at 0

Repeat (until a constraint set freeing a markedness constraint is found)

    Increase `FsetSize` by one

    Generate all subsets of `ActiveFaith` that are of size `FsetSize`

    For each such subset `FaithSet-x`

        If `FaithSet-x` frees up a markedness constraint, add it to `FreeingFaithSets`

    End-for

End-Repeat

If `FreeingFaithSets` contains only one faithfulness constraint set

    Set `BestFaithSet` to be the constraint set in `FreeingFaithSets`

Else

    Set `BestFaithSet` to the constraint set returned by **Select-best-faith-set**(`FreeingFaithSets`)

End-if

Return(`BestFaithSet`)

### Choosing Between Same-sized Minimal Faithfulness Constraint Subsets

**Select-best-faith-set**(`FreeingFaithSets`):

For each set `FaithSet-y` in `FreeingFaithSets`

    Place `FaithSet-y` in the next stratum of the constraint hierarchy under construction

    Continue BCD forward until another faithfulness constraint must be placed in the hierarchy

    Set `Value-y` to the number of markedness constraints ranked after the placement of `FaithSet-y`

End-for

Set `BestFaithSet` to the member `FaithSet-y` of `FreeingFaithSets` with the largest `Value-y`

If there is a tie for the largest `Value-y`, pick one of them at random

Return(`BestFaithSet`)

## 5.3 Alternative Criteria for Selecting Faithfulness Constraints

In BCD, the selection of what faithfulness constraints to rank at a given point during learning is driven primarily by the freeing up of markedness constraints. This is a consequence of the use of the *r*-measure to characterize restrictiveness: BCD strives to optimize the *r*-measure, which values the ranking of *M* constraints above *F* constraints, and the algorithmic decisions are determined largely by that imperative.

Hayes (1999:§7.6.3, this volume) proposes a different basis for the selection of faithfulness constraints, and in fact the major differences between our proposal and Hayes's reside in the approach to faithfulness selection. Rather than trying to enforce the  $M \gg F$  bias within the *F* selection routine, Hayes's Low Faithfulness CD algorithm aims to identify which faithfulness constraints are least likely to be *redundant* with other constraints. Consider the situation in which a certain faithfulness constraint  $F_k$  can account for a mark-data pair, and no other constraint can. So long as the loser of the pair is not harmonically bounded, i.e. so long as there is at least one *L* in its tableau row, then the constraint  $F_k$  cannot be relegated to the bottom of the hierarchy: it must dominate at least one other constraint. Hayes proposes that because such constraints are clearly needed to explain at least some data, they are good candidates to place into the ranking. The general "(data)-effectiveness" of a faithfulness constraint is characterized in terms of its ability to eliminate some mark-data pair which could be eliminated by *as few* other constraints as

possible. (The ability to eliminate a pair that no other constraints can eliminate indicates the maximum possible degree of data-effectiveness.) The Low-Faithfulness CD algorithm, when called upon to choose among faithfulness constraints to be ranked, selects (all of) those that have a maximum effectiveness measure. The hope is that greater restrictiveness will be achieved by postponing the ranking of less effective F constraints.

The difference between the two proposals can be cast (roughly) as a difference in focus on the problem of F selection. BCD focuses on relationships between constraints in the hierarchy under construction, aiming directly to maximize the collective domination of M over F. Low-Faithfulness CD focuses on the relationship between constraints and data, working within individual mark-data pairs, seeking to deal with each pair according to a criterion of constraint efficiency. A complete evaluation of the consequences of pursuing these disparate foci is, of course, quite non-trivial, and will doubtless be the subject of future investigation.

## 6 Special/general relationships: F vs. M

Faithfulness constraints may stand in a kind of special/general relationship, called ‘stringency’ in Prince (1997): one may be more stringent in its requirements than the other, in that violation of the one (the special case) entails violation of the other, but not vice versa. Canonical examples are provided by the positional faithfulness constraints of Beckman (1998): since F:IDENT/ONS( $\phi$ ) is violable only when the bearer of feature  $\phi$  appears in onset position, the constraint F:IDENT( $\phi$ ) is more general and more stringent, being violable everywhere. Gnanadesikan (1997) notes the same relationship among her scale-based faithfulness constraints.<sup>13</sup>

As demonstrated in Smith (1999) and Hayes (1999), special/general relationships among F constraints can cause significant difficulty for distributional learning. The general F constraint will always be available to do any work that its specialized cognate can. But if the learner mistakenly high-ranks the general version, no positive evidence can contradict the error. To see this, let us review a Lombardi (1998)-style treatment of voicing. The relevant constraints are F:IDENT(voi), F:IDENT/ONS(voi), and M: \*+voi, where by “voi” is meant the feature of *obstruent* voicing. Working from distributional evidence only, consider the observation that *da* occurs in the language. What should the learner deduce?

/da/ →	F:IDENT/ONS(voi)	F:IDENT(voi)	M: *+voi
a. da ~ ta	W	W	L

This is the canonical situation in which some F must be ranked, since the only available M prefers the loser. The tableau puts both F constraints on an equal footing, and BCD would pick randomly between them. But the choice is consequential. If the general constraint is placed, then no amount of *rat* and *tat* will inspire the positive-only German learner to imagine that *rad* and *tad* are forbidden. But if the special constraint is placed, the German learner is secure, and the English learner can be disabused of overrestrictiveness by subsequent encounters with positive data.

A plausible move at this point would be to add an injunction to the F-selection clause of BCD, along the lines of Hayes (1999:19, this volume) and Smith (1999), giving priority of place to the special constraint (or to the most special, in case of a more extended hierarchy of stringency). We are reluctant to take this step, because it does not solve the general problem. There are at least two areas of shortfall, both arising from the effects of constraint interaction. Both illustrate that the relation of stringency is by no means limited to what can be computed

pairwise on the constraints of CON. First, two constraints that have only partial overlap in their domain of application can, under the right circumstances, end up in a special to general relationship. Second, the special/general relation that is relevant to restrictiveness can hold between constraints that seem quite independent of each other. Let us review each in turn.

**Derived subset relations.** The first effect is akin to the one reviewed above in section 1 for parameter-setting theories, where dependencies among the parameters can reverse or moot subset relations. Two constraints which do not stand as special to general *everywhere* can end up in this relation within a hierarchy where higher-ranking constraints re-shape the candidate set in the right way. Because the set of candidates shrinks as evaluation proceeds down the hierarchy, the general Venn diagram for overlap between domains of relevance of two constraints can lose one of its ‘ears’ at a certain point, converting overlap into mere subsetting.

Consider two such generally-overlapping constraints: F:Ident/ $\sigma_1(\varphi)$ , which demands identity when the  $\varphi$ -bearer is in the first syllable, and F:Ident/ $\sigma'(\varphi)$ , which demands identity when the  $\varphi$ -bearer lies in a *stressed* syllable. Neither environment is a subset of the other over the universal set of forms, but if there are higher-ranked constraints that ensure stress on initial syllables as well as other stresses elsewhere in the word, then suddenly any violation of F:Ident/ $\sigma_1(\varphi)$  is also interpretable as a violation of F:Ident/ $\sigma'(\varphi)$ , though not vice versa. Contrapositively put, faithfulness in  $\sigma'$  implies faithfulness in  $\sigma_1$ , but it is possible to be narrowly faithful in  $\sigma_1$  without being faithful in all  $\sigma'$ . In short, the  $\sigma_1$  constraint has become a special case of the  $\sigma'$  constraint, contingently. Suppose  $\varphi$  is vowel-length, and the relevant markedness constraint is M:\*V: , banning long vowels generally. Upon observation of a word like *pá:to* , we have a situation like this:

/pá:to/	M:InitialStress	F:Ident/ $\sigma_1(\varphi)$	F:Ident/ $\sigma'(\varphi)$	M:*V
pá:to ~ páto		W	W	L

The learner who picks the stressed-based condition will have made an irrecoverable overgeneralization error if the target language only allows long vowels in the first syllable. (As always, if the learner chooses  $\sigma_1$  incorrectly, when the target language allows long vowels in all stressed syllables, or freely distributed without regard to stress, then positive evidence will be forthcoming to force correction.)

The direction of implication will be reversed in a language which allows at most one stressed syllable, always initial, so long as words with *no* stressed syllables also occur. Here faithfulness in  $\sigma_1$  implies faithfulness in  $\sigma'$ , but not vice versa. (As for the plausibility of this case, we note that although stress is usually required in all content words, in languages where stress behaves more like pitch accent, stressless words may indeed occur; see for example, the discussion of Seneca in Chafe (1977:178-180) and Michelson (1988:112-113).)

This result effectively rules out a solution to the general problem that is based on direct comparison of the internal syntactic structures of constraints, in the manner of Koutsoudas, Sanders, & Noll (1974) or Kiparsky (1982). The constraint-defining expression simply doesn't contain the required information, which is contextually determined by the operation of individual constraint hierarchies on individual candidate sets. We tend to regard this tactic as distinctly implausible anyway, since it requires a meta-knowledge of the structure of formal expressions which is otherwise quite irrelevant to the working of the theory. Nor is it plausible to imagine that the learner is somehow equipped to characterize surviving sets of candidates in such a way that subset relations could be computed. Here again, an appeal would have to be made to a kind of

meta-knowledge which the ordinary functioning of the theory quite happily does without, and which will be in general quite difficult to obtain.

**Third party effects.** To see how independent-seeming constraints may be tied together by the action of a third constraint, let us examine a constructed voicing-agreement pattern of a type that can be legitimately abstracted from present theoretical and empirical understanding of such systems. It is similar to that of Ancient Greek and, as Bruce Hayes reminds us, also resembles the aspects of Russian used by Halle (1959) in his celebrated anti-phoneme argument.

Suppose a language has a contrast between voiced and voiceless stops, which we will write as *p* and *b*, but no contrast among the fricatives, with voiceless *s* as the default and *z* arising only by regressive assimilation in clusters. Let us assume simple syllable structure, no clusters of stops, and agreement of voicing in clusters. We then have the following distribution of forms in the output:

pa	ap	sa	*za	apsa	aspa
ba	ab	as	*az	*abza	azba

What we are aiming for is a grammar in which the mappings *asba*, *azba* → *azba* provide (by regressive voicing) the *only* source of *z*, while *z,s* → *s* prevails everywhere else. The key datum is *azba*, which has multiple interpretations. It can be nonrestrictively interpreted as evidencing general faithfulness to *b*, implying that *\*abza* ought to be admitted, since clusters must agree in voicing. Even worse, the existence of *azba* might be taken to evidence general faithfulness to *z*, implying *\*za* and *\*az* in addition to *\*abza*. We actually want *azba* to be interpreted in terms of special onset-sensitive faithfulness to the voicing of *b*: this admits *azba* but no other instances of output *z*.

Even at this coarse level of description, the problem is already clear: the datum *azba* admits of three distinct faithfulness explanations, each with different restrictiveness consequences: how do we choose the *most* restrictive placement of these F? As we will see, the r-measure-driven BCD conflicts with the special-over-general criterion, which itself has nothing to say about the choice between faithfulness to b-voicing and faithfulness to z-voicing.

For concreteness, let us assume the following constraints, which are like those in Lombardi (1998):

M:AGREE(voi)	Adjacent obstruents agree in voicing.
M:*b	No b (voiced stops)
M:*z	No z (voiced fricatives)
F:STOP-VOI/ONS	Preserve stop-voicing when output correspondent is in Onset.
F:STOP-VOI	Preserve stop voicing.
F:FR-VOI/ONS	Preserve fricative voicing when output correspondent is in Onset.
F:FR-VOI	Preserve fricative voicing.

The error-driven method of selecting suboptimal candidates for comparison, which collects false optima from the immediately prior grammar (here, the initial  $M \gg F$  stage), will find competitors for *ba* (*pa* is less marked no matter what the ranking of the M constraints is), *ab* (*ap* is less marked) and *azba* (*aspa* is less marked). This yields the following tableau:

W~L	M:AGREE	M:*b	M:*z	F:stp-voi/O	F:stp-voi	F:fr-voi/O	F:fr-voi
ba ~ pa		L		W	W		
ab ~ ap		L			W		
azba ~ aspa		L	L	W	W		W

Recall that the left-hand member of each pair  $x \sim y$  is not only the desired winner, but also an exact match with its assumed input.

BCD puts the constraint M:AGREE at the top, as shown, but after that, no other M constraint is immediately rankable. Each remaining active F frees up an M constraint; which should be chosen?

- BCD as currently formulated unambiguously chooses the *general* F:STOP-VOI, because it frees up both remaining M constraints. The other two F constraints free up only one M constraint.
- The doctrine of preferring special to general favors the placement of F:STOP-VOI/ONS over placing its general cognate, regardless of the diminished freeing-up capacity of the special constraint. But this doctrine *makes no distinction* between F:STOP-VOI/ONS and F:FR-VOI, which are not members of the same stringency family.

Each choice determines a distinct grammar with different predictions about what additional forms are admitted (see Appendix 3 for details). High-ranking of F:STOP-VOI leads to this:

(10) General F:STOP-VOI high. **r=10**  
 {M:AGREE} >> {**F:STOP-VOI**} >> {M:\*b, M:\*z} >> {F:STOP-VOI/O, F:FR-VOI/O, F:FR-VOI}

► This grammar predicts /**absa**/ → **abza**, introducing  $z$  in a nonobserved context.

High-ranking of the special version of the stop-voicing faithfulness leads to this:

(11) Special F:STOP-VOI/ONS high. **r=9**  
 {M:AGREE} >> {**F:STOP-VOI/O**} >> {M:\*z} >> {F:STP-VOI} >> {M:\*b} >> {F:FR-VOI/O, F:FR-VOI}

► This grammar introduces no further  $z$  than are observed: /**abza**, **absa**/ → **apsa**.

High-ranking of fricative-voicing faithfulness leads to this:

(12) F:FR-VOI high. **r=9**  
 {M:AGREE} >> {**F:FR-VOI**} >> {M:\*z} >> {F:STP-VOI} >> {M:\*b} >> {F:FR-VOI/O, F:STP-VOI/O}

► This grammar preserves all / $z$ /, and therefore predicts the widest range of nonobserved forms through new identity maps: /**abza**/ → **abza**, /**za**/ → **za**, /**az**/ → **az**.

It is worth noting that BCD has successfully found the hierarchy with the optimal  $r$ -measure. It is the  $r$ -measure's characterization of restrictiveness that has failed.

Stepping back from the details, we see that the path to greatest restrictiveness will be obtained by ranking those F constraint with, roughly speaking, *fewest additional consequences*. In the case of an F: $\alpha$  vs. F: $\alpha$ /Ons pair, it is clear that the restricted version will have this property.

But here we face an additional choice, from outside a single stringency family. F:STOP-VOI/ONS is by no means a special case of F:FR-VOI, and indeed one might naively imagine them to be completely independent – what should *az* have to do with *ba*? In isolation, nothing: yet they are connected by the force of M:AGREE in *azba*.

It is tempting to imagine that the notion of *fewest additional consequences* might be reducible to some computable characteristic of performance over the mdp-list. In the case of the special-general stringency pair operative here, observe that the special F constraint deploys fewer W's than the general version. This is precisely because the special constraint is relevant to fewer structural positions and will be vacuously satisfied in regions of candidate space where the general version assesses violations. Since in the current learning context the optimal form satisfies all faithfulness constraints, it either ties or beats the suboptimum on each of them. One might then attempt to capitalize on this observable property, and add a bias to the algorithm favoring an M-freeing F constraint that assesses fewest W's over the mdp-list.<sup>14</sup> (A similar effect would be obtained if the algorithm monitored and favored vacuous satisfaction.) But the data that are available to the learner under error-driven learning will not always include the relevant comparisons.<sup>15</sup> In the case at hand, it is the dangerously general F:FR-VOI that is observably the most parsimonious of W's, using just one while successfully freeing up M:\*z.

It appears that the algorithm should choose the F constraint that is relevant to the narrowest range of structural positions. By this, any F/P for some position P should be favored over any F, given the choice, even if they constrain different elements. But in general this property need not be obviously marked in the constraint formulation itself. Between F/P<sub>1</sub> and F/P<sub>2</sub> the choice will be more subtle, as we saw in our first example, and subject to conditioning by other constraints.

We must leave the issue unresolved, save for the certainty that managing the special/general relationship cannot be reduced to statements about the formulations of constraints in CON as they are presently understood. A natural further question arises immediately: what of the special/general relations among *markedness* constraints? Somewhat surprisingly, perhaps, it turns out that absolutely nothing need be done about these. The RCD method of ranking as high as possible handles these relations automatically and correctly. The learner can be completely oblivious to them.

To see this, let us reconstruct the German-like situation along markedness lines, following Itô & Mester (1997). In addition to a single monolithic F:IDENT(voi), we need the following constraints:

- (13) M:\*+voi                    Obstruents must not be voiced.  
 (14) M:\*+voi/Coda            Coda obstruents must not be voiced.

Suppose now that the learner must deal with the form *da*.

/da/ →	M:*+voi/Coda	F:IDENT(voi)	M:*+voi
a. da ~ ta		W	L

BCD applied here will yield the following hierarchy:

M:\*+voi/Coda ≫ F:IDENT(voi) ≫ M:\*+voi

Codas are predicted to be voiceless, but outside of codas, the voiced-voiceless contrast is preserved.

With M-constraints ranked as high as possible, both special and general remain in force at the top until forms are observed that violate them (positive evidence). Under the markedness account, the special constraint directly forbids what is to be forbidden. (By contrast, the faithfulness approach derives the forbidden as the complement of what must be preserved.) The learner will be motivated to *demote* a special constraint like M: \*+voi/Coda only by a form that violates it, one that actually contains a voiced obstruent coda, like *ad*. (The faithfulness account reaches its crux as soon as it encounters *da*, with a voiced obstruent in the onset.). Absent such forms, the learner stays with the most restrictive grammar. The affinity for restrictiveness falls directly out of the core constraint demotion idea.

In a world without positional faithfulness, some of the pressure on BCD to decide among F constraints would go away; in particular, all special/general issues associated with positional specification would evaporate. We do not presume to adjudicate the issue here, but it is worth noting that complete elimination of the special/general problem would require developments in the theory of faithfulness that would handle the kinds of emergent relationships illustrated above. It is at least intriguing that relationships of arbitrary complexity between markedness constraints are always handled with ease by constraint demotion. This provides strong learning-theoretic motivation for attention to the character of the faithfulness system. Short of complete elimination of positional faithfulness, one might nevertheless hope to minimize the problem by minimizing dependence on positional faithfulness in favor of positional markedness, as in Zoll (1998).

Another case of constraints with complicating relationships has been discussed by McCarthy (1998), that of output-output faithfulness. He argues that the default in learning must be to favor hierarchies in which output-output faithfulness constraints, like markedness constraints, dominate ordinary faithfulness constraints. In the present context, output-output faithfulness constraints receive no special recognition, and require no extra apparatus; they are treated just like markedness constraints, and ranked as high as possible by the learner. The bias exerted by BCD is reserved for input-output faithfulness constraints. The default learner behavior of ranking constraints as high as possible treats output-output faithfulness constraints properly. The justification for this derives from the fact that the learner has access, in the form of positive data, to the base output form: whenever OOF is violated, there will be positive evidence, in the visible lack of base/derived-form identity, that will indicate the need for domination of OOF. The analogous structure in input-output faithfulness, the input, is precisely what the learner does *not* have direct access to. (On this, see also Hayes:1999/this volume, who envisions an important role for these in recovery from mistaken conclusions drawn from distributional learning.)

## 7 “As low as possible”: The disjunction threat

The chief problem addressed in this paper – choosing which minimal set of faithfulness constraints to place next into the developing hierarchy – is symptomatic of a fundamental danger lurking in the effort to rank constraints *as low as possible*. The formal problem addressed by the original Constraint Demotion principle had to do with the intrinsic computational difficulties involved in working out which constraints *must* dominate which others, described by Tesar & Smolensky (1998). Consider the scheme given in (15), which describes the information contained in a mark-data pair. Here, W1, W2, ..., are constraints preferring the winner of the pair, while L1, L2, ..., are constraints preferring the loser. By the Cancellation/Domination Lemma, it must be the case that if the winner is to beat the loser, the full hierarchy must meet the following condition:

$$(15) \quad (W1 \text{ or } W2 \text{ or } W3 \text{ or } \dots) \gg (L1 \text{ and } L2 \text{ and } L3 \text{ and } \dots)$$

For a given md-pair, **at least one** constraint preferring the winner, ( $W1$  **or**  $W2$  **or**  $W3$  **or** ...), must dominate **all** of the constraints preferring the loser, ( $L1$  **and**  $L2$  **and**  $L3$  **and** ...). The formula condenses a potentially very large number of distinct hypotheses: one for each non-empty subset of  $W_i$ 's. And this statement gives the hypothesis set for just one md-pair. Multiplying them all out is sure to produce an impressive tangle of ranking arguments. Constraint demotion avoids the problem of “detangling the disjunction” of the constraints favoring the winner, by instead focusing on demoting the constraints in the *conjunction*, those preferring the loser. When rendered into an algorithm, the effect of focusing on the conjunction is to leave constraints as high as possible, because constraints are only demoted when necessary, and then only as far down as necessary. A correct hierarchy is guaranteed to emerge, even though the algorithm has never attempted to determine precisely which constraints *must dominate* which others, satisfying itself with finding out instead which constraints *can* dominate others.

But when we attempt to determine which faithfulness constraints to rank next, we are precisely trying to ‘detangle the disjunction’, by picking from among the (faithfulness) constraints preferring the winner(s). Attempting to select from among the faithfulness constraints preferring winners steers a course straight into the formal problem that constraint demotion avoids. This does not inspire great confidence in the existence of an *efficient* fully general solution to the formal problem.

Fortunately, a fully general solution is probably not required: the choice is only made from among faithfulness constraints, and only under certain circumstances. Recall the simple example of indeterminacy:  $F1$  must dominate  $M1$ , and  $F2$  must dominate  $M2$ , with no dependencies indicated by the md-pairs between ( $F1$  and  $F2$ ) or between ( $M1$  and  $M2$ ). If there are in fact no dependencies at all (indicated by the mark-data pairs or not) between ( $F1$  and  $F2$ ), ( $M1$  and  $M2$ ), ( $F1$  and  $M2$ ), or ( $F2$  and  $M1$ ), then the choice won't matter. The same grammar should result from either of the hierarchies in (16) and (17).

(16)  $F1 \gg M1 \gg F2 \gg M2$

(17)  $F2 \gg M2 \gg F1 \gg M1$

Indeed, ranking (18) should also give the same grammar, although it would fare worse on the r-measure than the other two.

(18)  $\{F1, F2\} \gg \{M1, M2\}$

In such a circumstance, there really would be no harm in randomly picking one of  $\{F1, F2\}$  to rank first. So, the fact that there is no obvious basis for a choice does not automatically mean that there is “wrong choice” with attendant overgeneration.

More generally, how much the learner needs to care about such choices will depend upon the structure of the faithfulness component of UG. This leads to a request for a more informed theory of faithfulness, in the hopes that a more sophisticated understanding will yield a better sense of how much computational effort the learner ought to exert in making these ranking decisions during learning.

## Appendix 1: Problems with Ranking Conservatism

### Problem 1: Losing track of initial $M \gg F$ structure

The hypothesis of ranking conservatism of Itô & Mester (1999a) is that a new hierarchy *preserves* as much as possible of the already determined ranking relations. Thus, if  $C_1$  and  $C_2$  belong to the same stratum, imposing  $C_1 \gg C_2$  is cost-free, because it doesn't contradict anything already established; but if  $C_1 \gg C_2$  in the current state, it is undesirable to reverse it to  $C_2 \gg C_1$ . Given a hierarchy  $M \gg F$  for sets of constraints  $M$ ,  $F$ , this strategy will promote solutions *within* the  $M$  and  $F$  blocs. However, once the  $M$  and  $F$  constraints have become entangled, the key distinction can be lost. The essential problem is that transitivity of ranking entails further relations that obscure the original state, and preserving them leads to an inability to distinguish  $M$  from  $F$  solutions, or even a preference for  $F$  over  $M$ . Here is an example of the first.

Suppose we have a initial  $M \gg F$  structure on three  $M$  constraints and one  $F$  constraint:

$$(19) \quad \{M_1 M_2 M_3\} \gg \{F\}$$

For conciseness and legibility, we will write this with bars separating the strata:

$$(20) \quad M_1 M_2 M_3 | F$$

Now imagine two pieces of input, encountered sequentially, the first entailing  $F \gg M_3$ , the second  $M_1 \gg M_2$ .

$$(21) \quad M_1 M_2 M_3 | F \rightarrow M_1 M_2 | F | M_3 \rightarrow M_1 | M_2 | F | M_3$$

Now we encounter a datum entailing  $F \gg M_2$ . The result is a simple switch of neighbors:

$$(22) \quad M_1 | M_2 | F | M_3 \rightarrow M_1 | F | M_2 | M_3$$

By this point we have thoroughly lost track of the fact that there's no substantive reason for  $M_2 \gg M_3$ , a ranking that popped up as a consequence of  $F \gg M_3$ . But we must try to preserve it nonetheless.

Now suppose we encounter a datum that admits of either a faithfulness or a markedness solution: say, either  $F \gg M_1$  or  $M_3 \gg M_2$ . Notice that both are entirely consistent with the ranking conditions that have come before. The problem is that each involves only one shift in ranking relations.

$$\begin{array}{l} \text{F solution:} \quad \mathbf{M}_1 | \mathbf{F} | M_2 | M_3 \quad \rightarrow \quad \mathbf{F} | \mathbf{M}_1 | M_2 | M_3 \\ \text{M solution:} \quad M_1 | F | \mathbf{M}_2 | \mathbf{M}_3 \quad \rightarrow \quad M_1 | F | \mathbf{M}_3 | \mathbf{M}_2 \end{array}$$

Ranking conservatism has failed to diagnose the difference between  $M$  and  $F$  solutions.

### Problem 2: Broihier's Cycle

Online CD works by modifying an old ranking into a new one, working one md-pair at a time. (RCD constructs an entire ranking from a set of md-pairs.) Broihier (1995:52-54) considers a variant of online CD that places each demoted constraint in its own stratum: this will have the effect of preserving as many ranking relations as possible, in the manner of ranking conservatism. Thus, starting from  $AB|C$ , and given data entailing  $A \gg B$ , we go to  $A|B|C$ , preserving the relation

$B \gg C$ . By contrast, standard online CD would go to  $A|BC$ , keeping both B and C as high as possible. The problem is that  $B \gg C$  is not secure as a relation worth preserving. In the typical stratified hierarchy  $AB|C$  means that *either* A or B dominates C in the final grammar.

The result of this method of demotion is that the order in which forms are encountered can impose phony domination relations, and preserving these can run afoul of the treatment of disjunctions in ranking arguments, leading to the interpretation of consistent data as contradictory. Here we present Broihier’s example.

Imagine that we are targeting  $A|B|C|D$ , but we preserve ranking relations maximally. Suppose we see data motivating the rankings  $(A \gg B)$ ,  $(A \gg C)$ , and  $(A \gg D)$ , in that order. The following sequence of hierarchies emerges:

$$(23) \quad ABCD \rightarrow ACD|B \rightarrow AD|C|B \rightarrow A|D|C|B$$

Observe that standard CD would come up with  $A|BCD$ , regardless of order of data presentation.

Now we encounter data entailing  $C \gg D$ :

$$(24) \quad A|D|C|B \rightarrow \mathbf{A|C|D|B}$$

Suppose now we are faced with data leading to the disjunctive choice ‘*either*  $(D \gg C)$  *or*  $(B \gg C)$ ’. This is perfectly consistent with everything seen so far, because even though the first disjunct must be false, the second disjunct can be true. (The analyst can see this, but the algorithm, working one datum at a time, cannot.) The desideratum of keeping C as high as possible and putting it in its own stratum (maximally preserving its previous ranking relations) forces us to act on the  $D \gg C$  disjunct – the one that contradicts previous data!

$$(25) \quad A|C|D|B \rightarrow \mathbf{A|D|C|B}$$

But this is exactly the ranking we arrived at after the first sequence of demotions. So we’re in a cycle, and we must persist in it interminably, bouncing between the last two pieces of data considered.

## Appendix 2: Getting Greediness to Fail

Here we show a contrived situation in which ranking a smaller set of faithfulness constraints at a given stage ultimately results in a lower r-measure than does the ranking of a larger set of faithfulness constraints at the same stage. We believe that this is the smallest case that can be constructed.

	<b>M1</b>	<b>M2</b>	<b>M3</b>	<b>M4</b>	<b>F1</b>	<b>F2</b>	<b>F3</b>	<b>F4</b>
<b>P1</b>	L				W	W		
<b>P2</b>	L				W		W	
<b>P3</b>	L				W			W
<b>P4</b>		L	L	L		W	W	
<b>P5</b>		L	L	L		W		W
<b>P6</b>		L	L	L			W	W

The example has 8 constraints, 4 markedness constraints **M1-M4**, and 4 faithfulness constraints **F1-F4**. It has six mark-data pairs, **P1-P6**. Initially, each of the markedness constraints prefers a loser in at least one of the pairs, so none of them may be ranked at the top of the hierarchy. Thus, at least one faithfulness constraint must be ranked first.

The set {F1}, consisting of only a single faithfulness constraint, is adequate to free up a markedness constraint, specifically, M1. Once those two constraints have been ranked, all of the other faithfulness constraints must be ranked before any more markedness constraints are freed up. The result is the following hierarchy, with an r-measure of 3.

$$(26) \quad \{F1\} \gg \{M1\} \gg \{F2\ F3\ F4\} \gg \{M2\ M3\ M4\} \quad r = 3$$

If, instead, the learner were to rank the other three faithfulness constraints first, then all four markedness constraints would be freed up, permitting F1 to be dropped all the way to the bottom of the hierarchy. Perhaps surprisingly, the result is a hierarchy with a larger r-measure, 4.

$$(27) \quad \{F2\ F3\ F4\} \gg \{M1\ M2\ M3\ M4\} \gg \{F1\} \quad r = 4$$

Finally, an even higher r-measure results from initially ranking any two of F2-F4.

$$(28) \quad \{F2\ F3\} \gg \{M2\ M3\ M4\} \gg \{F1\} \gg \{M1\} \gg \{F4\} \quad r = 7$$

This third hierarchy also illustrates that the strategy of selecting that subset of the faithfulness constraints that frees up the most markedness constraints also can fail to produce the highest r-measure: initially, {F2 F3 F4} frees up 4 markedness constraints, while {F2 F3} only frees up 3 markedness constraints, yet it is the latter that leads to the higher r-measure.

Observe that the pattern of relative violation of the constraints in this example is quite odd, from a linguistic point of view. First of all, markedness constraints M2-M4 are entirely redundant. For these mark-data pairs, at least, one would do just as much work as the three. This is not by accident; if one of M2-M4 didn't prefer the loser in one of P4-P6, then it would be possible to free that markedness constraint with only one faithfulness constraint, and the contrary result would not occur. Pairs P4-P6 cannot be resolved by the ranking of any single faithfulness

constraint, but neither is any single faithfulness constraint required: any subset of size 2 or greater of F2-F4 will do the job.

Observe also that the pairs suggest that faithfulness constraint F1 is “equivalent” to the combination of {F2,F3,F4} with respect to justifying violation of M1. However, it must be so without any of F2-F4 constituting special cases of F1, or else the relevant facts about pairs P4-P6 would change, with F1 single-handedly freeing all markedness constraints.

Of particular interest is the extent to which the cases where the algorithm fails to optimize the r-measure are also cases where the r-measure fails to correlate with actual restrictiveness. If the differences in r-measure are among hierarchies generating identical grammars, then failure to find the hierarchy with the best r-measure is not a problem, so long as the hierarchy actually learned is one generating the restrictive language.

### Appendix 3: Unexpected Special/General Relations

More details of the p/b/s example. The grammars given are constructed by BCD, with different choices of F for the second stratum.

#### Grammar #1. General F:STOP-VOI high. r=10

{M:AGREE} >> {F:STOP-VOI} >> {M:\*b, M:\*z} >> {F:STOP-VOI/O, F:FR-VOI/O, F:FR-VOI}

In the rows below the heavy line, inputs are tracked that do not come from observation of positive evidence, but that give information about the further predictions of the induced grammar. The key result is /absa, abza/ → abza, introducing z in a nonobserved context. Stop-voicing is always preserved and therefore ‘dominant’ in clusters.

	M	F	M		F		
	M:Agr	F:stp-voi	M:*b	M:*z	F:stp-voi/O	F:fr-voi	F:fr-voi/O
ba ~ pa		W	L		W		
ab ~ ap		W	L				
azba ~ aspa		W	L	L	W	W	
★ /abza/ → abza ~ apsa		W	L	L		W	W
★ /absa/ → abza ~ apsa		W	L	L		L	L
/za/ → sa ~ za				W		L	L
/az/ → as ~ az				W		L	

#### Grammar #2. Special F:STOP-VOI/ONS high. r=9

{M:AGREE} >> {F:STOP-VOI/O} >> {M:\*z} >> {F:STOP-VOI} >> {M:\*b} >> {F:FR-VOI/O, F:FR-VOI}

This is the most restrictive grammar and introduces no further z than are observed. Stop-voicing spreads regressively in clusters, which are otherwise voiceless: /abza, absa/ → apsa.

	M	F	M	F	M	F	
	M:Agr	F:stp-voi/O	M:*z	F:stp-voi	M:*b	F:fr-voi	F:fr-voi/O
ba ~ pa		W		W	L		
ab ~ ap				W	L		
azba ~ aspa		W	L	W	L	W	
★ /abza/ → apsa ~ abza			W	L	W	L	L
★ /absa/ → apsa ~ abza			W	L	W	W	W
/za/ → sa ~ za			W			L	L
/az/ → as ~ az			W			L	

**Grammar #3. F:FR-VOI high. r=9.**

{M:AGREE} >> {F:FR-VOI} >> {M:\*z} >> {F:STOP-VOI} >> {M:\*b} >> {F:FR-VOI/O, F:STOP-VOI/O}

This grammar preserves all /z/, and therefore predicts the widest range of nonobserved forms through new identity maps: /abza/ → abza, /za/ → za, /az/ → az. Fricatives retain voicing everywhere and therefore fricative voicing is dominant in clusters.

	M	F	M	F	M	F	
	M: Agr	F: fr-voi	M:*z	F:stp-voi	M:*b	F:fr-voi/O	F:stp-voi/O
ba ~ pa				W	L		W
ab ~ ap				W	L		
azba ~ aspa		W	L	W	L		W
★/abza/ → abza ~ apsa		W	L	W	L	W	
★/apza/ → abza ~ apsa		W	L	L	L	W	
/absa/ → apsa ~ abza		W	W	L	W	W	
/za/ → za ~ sa		W	L			W	
/az/ → az ~ as		W	L				

## References

- Angluin, Dana. 1980. Inductive inference of formal languages from positive data. *Information and Control* **45**. 117-135.
- Baker, C. L. 1979. Syntactic theory and the projection problem. *LI* **10:4**. 533-581.
- Beckman, Jill. 1998. *Positional faithfulness*. PhD dissertation, University of Massachusetts, Amherst. ROA-234.
- Bernhardt, Barbara & Joseph Stemberger. 1997. *Handbook of phonological development: From the perspective of constraint-based nonlinear phonology*. Academic Press.
- Berwick, Robert. 1982. *Locality principles and the acquisition of syntactic knowledge*. PhD dissertation, MIT.
- Broihier, Kevin. 1995. Optimality Theoretic Rankings with Tied Constraints: Slavic Relatives, Resumptive Pronouns and Learnability. Ms, MIT. ROA-46.
- Chafe, Wallace. 1977. Accent and related phenomena in the Five Nations Iroquoian languages. In Charles N. Li and Larry M. Hyman (eds.) *Studies in Stress and Accent (Southern California Occasional Papers in Linguistics, 4)*. 169-81.
- Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* **2:2**. 83-149.
- Demuth, Katherine. 1995. Markedness and the development of prosodic structure. *NELS* **25**. 13-25. ROA-50.
- Dresher, B. Elan. 1999. Charting the learning path: Cues to parameter setting. *LI* **30:1**. 27-67.
- Dresher, B. Elan & Jonathan Kaye. 1990. A computational learning model for metrical phonology. *Cognition* **34**. 137-195.
- Gnanadesikan, Amalia. 1995. Markedness and faithfulness constraints in child phonology. Ms, University of Massachusetts, Amherst. ROA-67.
- Gnanadesikan, Amalia. 1997. *Phonology with ternary scales*. PhD dissertation, University of Massachusetts, Amherst. ROA-195.
- Goldsmith, John. 1993. Phonology as an intelligent system. In D. J. Napoli and J. Kegl (eds.) *Bridges between psychology and linguistics: A Swarthmore Festschrift for Lila Gleitman*. Hillsdale, NY: Lawrence Erlbaum Associates. 247-267.
- Hale, Mark and Charles Reiss. 1997. Grammar optimization: The simultaneous acquisition of constraint ranking and a lexicon. Ms, Concordia University. ROA-231.
- Halle, Morris. 1959. Questions of Linguistics. *Supplemento a Il Nuovo Cimento* **13**, ser. **10**. 494-517.
- Hayes, Bruce. 1999. Phonological acquisition in Optimality Theory: The Early Stages. Ms, UCLA. ROA-327.
- Itô, Junko and Armin Mester. 1997. Sympathy Theory and German Truncations. In Viola Miglio and Bruce Morén (eds.) *University of Maryland Working Papers in Linguistics 5: Selected phonology papers from Hopkins Optimality Theory Workshop 1997/University of Maryland Mayfest 1997*. 117-139. ROA-211.
- Itô, Junko and Armin Mester. 1999a. The Structure of the Phonological Lexicon. In Tsujimura, Natsuko, (ed.) *The Handbook of Japanese Linguistics*. Malden, MA, & Oxford, U.K.: Blackwell Publishers. 62-100.
- Itô, Junko, and Armin Mester. 1999b. On the sources of Opacity in OT: coda processes in German. To appear in Caroline Féry and Ruben van der Vijver (eds.) *The optimal syllable*. Cambridge University Press. ROA-347.
- Jacobowitz, Celia. 1984. On markedness and binding principles. *NELS* **14**.
- Keer, Edward. In prep. *Geminates, the OCP, and faithfulness*. PhD dissertation, Rutgers University, New Brunswick.
- Kiparsky, Paul. 1980. Vowel harmony. Ms, MIT.

- Kiparsky, Paul. 1982. Lexical phonology and morphology. In I. S. Yang (ed.) *Linguistics in the Morning Calm*. Seoul: Hanshin. 3-91.
- Kisseberth, Charles. 1970. On the functional unity of phonological rules. *LI* 1. 291-306.
- Koutsoudas, A., G. Sanders & C. Noll. 1974. The application of phonological rules. *Lg* 50, 1-28. (Drafted September, 1971: fn. 2).
- Levelt, Claartje. 1995. Unfaithful kids: Place of articulation patterns in early child language. Paper presented at the Department of Cognitive Science, The Johns Hopkins University, Baltimore, MD.
- Lombardi, Linda. 1998. Restrictions on the direction of voicing. Ms, University of Maryland. ROA-246.
- McCarthy, John. 1998. Morpheme structure constraints and paradigm occultation. Ms, University of Massachusetts. To appear in Catherine Gruber, Derrick Higgins, Kenneth Olson, and Tamra Wysocki (eds.) *CLS 32, vol. II: The Panels*. Chicago: Chicago Linguistic Society.
- Michelson, Karin. 1988. *A comparative study of Lake-Iroquoian accent*. Dordrecht: Kluwer Academic Publishers.
- Oostendorp, Marc van. 1995. *Vowel quality and phonological projection*. PhD dissertation, Tilburg University. ROA-84.
- Prince, Alan. 1997. Paninian Relations. Lecture given at LSA Summer Institute, Cornell University.
- Prince, Alan. 2000. Comparative tableaux. Ms, Rutgers University, New Brunswick. ROA-376.
- Prince, Alan. 2001a. Entailed Ranking Arguments. *RuLing Papers* 2:117-140. Rutgers University, New Brunswick
- Prince, Alan. 2001b. Entailed Ranking Arguments (extended version). Ms, Rutgers University, New Brunswick.
- Prince, Alan & Paul Smolensky. 1993. Optimality Theory. Technical Report RuCCS-TR-2, Rutgers Center for Cognitive Science. To appear MIT Press.
- Samek-Lodovici, Vieri & Alan Prince. 1999. Optima. Ms, University of London & Rutgers University, New Brunswick.
- Sherer, Tim. 1994. *Prosodic phonotactics*. PhD dissertation, University of Massachusetts, Amherst. ROA-54.
- Sommerstein, Alan H. 1974. On phonotactically motivated rules. *JL* 19. 71-94.
- Smith, Jennifer. 1999. Special-general relations among faithfulness constraints and learning. Talk presented at the Rutgers-UMass Joint Class Meeting.
- Smolensky, Paul. 1996a. On the comprehension/production dilemma in child language. *LI* 27. 720-731. ROA-118.
- Smolensky, Paul. 1996b. The initial state and 'Richness of the Base.' Technical Report JHU-CogSci-96-4. ROA-154.
- Tesar, Bruce. 1995. *Computational Optimality Theory*. PhD dissertation, University of Colorado, Boulder. ROA-90.
- Tesar, Bruce. 1997. Multi-Recursive Constraint Demotion. Ms, Rutgers University. ROA-197.
- Tesar, Bruce. 1998. Using the mutual inconsistency of structural descriptions to overcome ambiguity in language learning. *NELS* 28. 469-483.
- Tesar, Bruce. 2000. Using inconsistency detection to overcome structural ambiguity in language learning. Ms, Rutgers University, New Brunswick. ROA-426.
- Tesar, Bruce & Paul Smolensky. 1998. The learnability of Optimality Theory. *LI* 29:2. 229-268.
- Tesar, Bruce & Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.
- Vapnik, V. N. & A. Y. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16. 264-280.

- Wexler, Ken & Rita Manzini. 1987. Parameters and learnability in binding theory. In Thomas Roeper and Edwin Williams (eds.) *Parameter setting*. 41-76. Dordrecht: Reidel.
- Zoll, Cheryl. 1998. Positional Asymmetries and Licensing. Ms, MIT. ROA-282.

---

<sup>1</sup> The authors names are given alphabetically. Prince would like to thank the John Simon Guggenheim Memorial Foundation for support. Thanks to Jane Grimshaw, John McCarthy, Paul Smolensky, and the Rutgers Optimality Research Group for helpful discussion.

Hayes (1999, this volume) independently arrives at the same strategy for approaching the learning problems dealt with here, and he offers similar arguments for the utility and interest of the approach. We have tried to indicate by references in the text where the identities and similarities lie. Along with the overlap in basics, there are significant divergences in emphasis, analytical procedure, and development of the shared ideas; we suggest that readers interested in the issues discussed below should consult both papers.

<sup>2</sup> The nonexistence of long high vowels on the surface in Yawelmani Yokuts, coupled with their underlying necessity, is the centerpiece of generations of commentary. Bruce Hayes reminds us, however, that the allomorphs of the reflexive/reciprocal suffix contain the otherwise unobserved high long vowels, indicating the need for a more nuanced view of their still highly restricted distribution in the language.

<sup>3</sup> Different kinds of interactions have led Dresher & Kaye (1990) and Dresher (1999) to hypothesize elaborate paralinguistic strategies such as pre-establishing, for each subsystem, the order in which its parameters must be set. This is meant to replace learning from error. Ambiguity of analysis, rather than the subset relation, provides the central difficulty they focus on.

<sup>4</sup> Swimming against the tide, Hale & Reiss (1997) insist on  $F \gg M$  as the default. For them, then, word learning yields no learning of phonology. To replace  $M \gg F$  learning, they offer an “algorithm” that regrettably sets out to generate and sort the entirety of an infinite set in its preliminary stages.

<sup>5</sup> For ease of exposition, we write as if there was only one constraint with the effect \*š and one with the effect \*si. In a fuller treatment of the example, we would see that *all* such constraints would be subordinated in (a) and that *some*  $M(*si)$  would have to dominate all  $M(*š)$  in (b). The thrust of the argument – distinguishing between markedness and faithfulness solutions to ambiguous data – is preserved amidst any such refinements.

<sup>6</sup> Hayes (1999, this volume) operates under the same  $I=O$  assumption, which he attributes to a suggestion from Daniel Albro.

<sup>7</sup> The two diverge when a language admits forms which are not fixed points. Such forms do not add to the number of fixed points, but do increase the number of forms overall. A simple illustration is a language with a chain-shift. Consider the following two mappings:  $[x \rightarrow y, y \rightarrow z, z \rightarrow z]$  and  $[x \rightarrow z, y \rightarrow z, z \rightarrow z]$ . The two mappings have the same number of fixed points (one), but differ in the languages produced: the first, with the chain-shift, produces  $\{y, z\}$  while the second produces the more restrictive  $\{z\}$ . We are assuming that the learning of all non-identity mapping relations, including chain-shifts, is done at the stage of morphophonemic learning.

<sup>8</sup> This correlates with the fact that there is typically going to be a number of different grammars, with different patterns of faithfulness violation, that produce the same repertory of output forms. For example, whether syllable canons are enforced by deletion or epenthesis cannot be determined by distributional evidence. Similarly, in the complementary distribution case, we simply assumed a set of mappings among the allophones – but the same pattern could be gotten by deleting the restricted allophone from environments where it doesn’t appear. The expectation is that improved theories of constraints will restrict such options, and that even when no improvement is possible, morphophonemic learning will be able to profit from the partial success obtained by identifying the high-ranked markedness constraints.

<sup>9</sup> In order to deal with such intermediate grammars, a definition of ‘optimum’ must be offered for cases where the (mutually unranked) constraints within a stratum turn out to conflict. Tesar (1997) proposes the computationally efficient heuristic of merely lumping together the violations. Errors produced under this definition will always be informative, but some useful suboptima will be missed. Tesar (2000) offers further discussion of this point, along with a proposal for reducing the number of missed suboptima.

---

<sup>10</sup> In this regard, our proposal is identical to Hayes (1999:§7.6.3). Beyond this, the proposals diverge in their handling of cases where more than one F constraint prefers some winner.

<sup>11</sup> See Appendix 2 for the details of such an example.

<sup>12</sup> See Appendix 2 for the details of such a case.

<sup>13</sup> We gloss over the distinction noted by Prince (1997 *et seq.*) between stringency on the elements of structure and stringency on whole candidate forms. The constraints we discuss guarantee only the former, but only the latter leads to true special-case general-case behavior. The two converge when competing forms contain the same number of instances of the focal configuration, as in all the cases considered here.

<sup>14</sup> A tension has developed here between the r-measure, which will in general favor those F constraints that set out the most W's, giving them the greatest chance at freeing up M constraints, and considerations of specificity, which will have the opposite effect.

<sup>15</sup> The key issue is the (un)availability of competitors that will 'turn on' the F constraint in every region of candidate space. For example, the true generality of F:FR-VOI will only be seen when competitions like *sa~za* and *as~az* are evaluated, leading to W's that encode the implicit declaration "relevant to onsets", "relevant to codas", etc. But in such cases the required competitors are *more marked* than the desired optimum, and they will never present themselves as false optima from an earlier grammar hypothesis. Hence, they simply will not be found by the error-driven method of competitor-generation. Additional, more inclusive methods of competitor generation might also be considered, which could increase the informativeness of the distribution of W's. But it bears emphasizing that exhaustive inspection of all competitors is computationally intractable.