Constraint Interaction: A Lingua Franca for Stochastic Theories of Language*

Matthew Goldrick
Northwestern University

Psycholinguistic and stochastic grammatical theories of sound structure are typically stated in quite different formal vocabularies (i.e., connectionist networks vs. stochastic extensions of Optimality Theory). This commentary argues that this difference obscures a common set of core principles: in each framework, the generation of behavior can be conceptualized as the consequence of the interaction of two broad types of constraints. To support this claim, Warker and Dell's connectionist proposal is re-cast as a type of constraint-based theory. This reconceptualization not only increases our understanding of how their proposal accounts for empirical data but also allows us to better understand its relationship to stochastic grammatical theories. The connection between these cross-disciplinary perspectives suggests that the substance of theoretical disagreements reflects not principles of processing vs. grammar but rather conflicting claims regarding the precise nature of constraints and their interactions.

1. Introduction

There are two prominent theoretical traditions that aim to understand how the cognitive representations and processes underlying our knowledge of sound structure generate probabilistic behavior (e.g., the focus of this volume, speech errors). Stochastic generative grammatical theories typically take as their goal the specification of the non-deterministic function that generates output structures. Prominent formal frameworks for specifying this function include various stochastic extensions of Optimality Theory (Prince and Smolensky 1993); in these approaches, the grammar is specified by the interaction of a set of ranked, violable constraints (for a review of other stochastic formalisms in phonology, see Pierrehumbert 2001). Psycholinguistic theories (such as that presented by Dell and Warker (2007)) aim to specify the cognitive mechanisms that use long-term memory representations of phonological form to construct context-specific utterances plans (e.g., by associating segments to positions in prosodic structure). The predominant formalism for expressing such theories is connectionist networks; the cognitive mechanisms are specified in terms of spreading activa-

* Thanks to the workshop participants for helpful discussions and especially to Brady Clark for conceptual and editorial advice.

MIT Working Papers in Linguistics 53, 95–114
The State of the Art in Speech Error Research: Proceedings of the LSA Institute Workshop
Carson T. Schütze and Victor S. Ferreira (Eds.)

tion between simple processing units (see, e.g., Goldrick in press for an overview in the domain of speech production).

Both stochastic Optimality Theoretic grammars and connectionist psycholinguistic theories seek to characterize how cognitive representations and processes are structured so as to yield probabilistic behavior. Despite this shared goal, each framework focuses on different means of achieving it (i.e., characterizing functions vs. cognitive mechanisms) and makes use of distinct formal vocabularies (i.e., ranked constraints vs. spreading activation networks). These differences make it difficult to recognize the dimensions along which theories stated in one framework conflict or agree with theories stated in another. This commentary examines the formal properties of these theories to demonstrate that they can both be conceptualized as utilizing constraint interaction to generate behavior. More specifically, both types of theory can be seen as relying on the interaction of two broad types of constraints: one preferring an output structure associated with the specific input and another type expressing more general preferences over output structures. Recognizing these similarities allows one to better understand the relationship between stochastic grammatical and processing theories of sound structure. Specifically, this commentary claims that theoretical disputes do not reflect contrasting grammatical vs. processing principles but rather concern disagreements over the precise nature of phonological constraints and their interactions.

To argue for this perspective, the second section of the commentary presents an analysis of Warker and Dell's (2006) connectionist account of phonological processing that characterizes network processing in terms of constraint interaction. The ability of their connectionist theory to account for four empirical findings from speech errors is then discussed in terms of constraint interaction. The third section compares this constraint-based psycholinguistic account to constraint-based grammatical accounts (stated in various stochastic extensions of Optimality Theory). The generality and implications of this parallelism between psycholinguistic and stochastic grammatical theories is then discussed.

This commentary is situated within two broader research programs: one that analyzes connectionist computation in terms of constraint satisfaction (e.g., Ackley, Hinton and Sejnowski 1985; Hopfield 1982; Smolensky 1986, 2006b) and another that establishes parallels between connectionist approaches to cognition and constraint-based grammatical theories (e.g., Bernhardt and Stemberger 1998; Legendre, Sorace and Smolensky 2006; Prince and Smolensky 1993). This rich body of work has established important general principles for constraint-based theories of cognition. The analyses reported here demonstrate that these can be fruitfully applied to a specific, extant psycholinguistic proposal, revealing the connections that exist between it and existing grammatical frameworks.

Note that this discussion focuses on the relationship between stochastic generative grammars and connectionist psycholinguistic theories. Many important issues fall outside this scope, including the relationship between stochastic and non-stochastic generative grammars (see Newmeyer 2003 and associated commentaries in *Language* 81(1) for a recent discussion) and the relationship between connectionist and non-connectionist theories of cognition (for recent discussions, see Anderson and Lebiere 2003; Smolensky 2006b).

2. A constraint-based analysis of Warker and Dell's connectionist theory

Warker and Dell (2006; see also Dell and Warker 2007) present a connectionist theory of metrical encoding, a core aspect of phonological processing. Many psycholinguistic theories of speech production assume a distinct stage of processing where the melodic content of an utterance (e.g., segments) is associated with metrical structure (e.g., syllables, feet, and prosodic words; Dell 1986; Garrett 1984; Shattuck-Hufnagel 1992; Levelt, Roelofs and Meyer 1999). Warker and Dell focus on the assignment of segments to syllable positions. Following the general structure of other proposals, they assume that the input to this syllabification process consists of a set of segments along with a specification of their serial order. The output is a representation in which these segments have been associated to syllable positions.

Warker and Dell make two critical theoretical claims that augment this standard framework. First, they propose that the syllabification process is implemented by a three-layer connectionist network; the mapping between input and output representations is mediated by an intermediate representation. Second, they assume that an implicit learning mechanism continually modifies how activation spreads between these representations (based on statistical relationships between input and output representations).

To examine the predictions of this theory, they constructed a simulation focusing on the syllabification of CVC monosyllables. The simulation's input representation has two components: an unordered set of segmental units (e.g., /t/, /k/, /æ/…); and a set of syllable identity units (e.g., CAT, TACK, ACT…) that serve to identify the serial order of segmental units. The output representation consists of three pools of segmental units representing the onset, nucleus, and coda of the CVC syllables. Each pool of output units is identical; there are three distinct copies of each segmental unit, one for each position. A set of twenty intermediate (hidden) units mediates between these representations; these are fully connected to both the input and output units. The mechanism that implicitly learns statistical relationships between input and output representations is instantiated by the backpropagation algorithm (Rumelhart, Durbin, Golden and Chauvin 1996; Rumelhart, Hinton and Williams 1986). As discussed by Warker and Dell (2006; Dell and Warker 2007), the behavior of this simulation is broadly consistent with empirical studies of speech errors.

2.1 Analysis of Warker and Dell's proposal

Warker and Dell's theory consists of two broad principles for the organization of the process of syllabification (spreading activation in a multilayer network; implicit learning of statistical relationships between input and output representations via modification of the spreading of activation). They argue that these principles constitute a viable account because a simulation instantiating these principles exhibits behavior broadly consistent with the empirical data. One problematic aspect of such a research strategy is that data and theory are always mediated by simulations whose structure is poorly understood (McCloskey 1991). As noted by Mozer and Smolensky (1989: 3),

> One thing that connectionist networks have in common with brains is that if you open them up and peer inside, all you can see is a big pile of goo. Internal organization is obscured by the sheer number of units and connections.

An alternative approach (Plaut, McClelland, Seidenberg and Patterson 1996) is to consider a simplified instantiation of the broad theoretical principles which shares the core properties of the complex simulations. While such an instantiation is too simplistic to provide a full account of the empirical data, it is much more amenable to detailed analysis. This allows a tighter connection to be established between theoretical principles and the empirical predictions of the proposal.

2.1.1 Covariant learning algorithms

Warker and Dell's proposal is one specific version of the *covariant learning hypothesis* (VanOrden, Pennington and Stone 1990) which claims that learning involves strengthening connections between covarying representational units or sets of representational units. Warker and Dell's simulation instantiates this using the backpropagation algorithm. This method is powerful enough to allow connectionist networks to learn complex input-output mappings (such as that required by the syllabification process; see below for further discussion) but it is not amenable to closed form analysis. Such an analysis is possible, however, for a simpler covariant learning algorithm known as Hebbian learning. Consider a two-layer, fully connected feed-forward network incorporating Warker and Dell's input and output representations. Upon presentation of training pattern p, the weight from input unit i to output j (w_{ji}) is modified according to the rule shown in (1).

$$\Delta w_{ji}^{[p]} = \varepsilon \bar{s}_j^{[p]} \bar{s}_i^{[p]} \tag{1}$$

Where $\bar{s}_x^{[p]}$ is the specified activation of unit x for training pattern p and ε is the learning rate.

This simple algorithm is clearly an instantiation of the covarying learning hypothesis; the spreading of activation is directly proportional to the covariance of individual input and output units.

2.1.2. Analysis: Simplified Hebbian network

As shown by Plaut et al. (1996), at the end of training the behavior of a Hebbian network can be described by the following equation.

$$s_j^{[t]} = \sigma\left(\varepsilon \sum_p \bar{s}_j^{[p]} F^{[p]} O^{[p][t]}\right) \tag{2}$$

When some pattern t is clamped on the input units of the network, the activation of some particular output unit s_j ($s_j^{[t]}$) reflects a weighted sum, over all training patterns p, of each training pattern's specified output ($\bar{s}_j^{[p]}$). The

contribution of each training pattern is weighted by its frequency ($F^{[p]}$) and its similarity to the current input ($O^{[p][t]}$—the dot product of the training pattern's specified input vector *p* and the current input pattern *t*). (The sum is scaled by a learning rate ε and "squashed" by a nonlinear activation function σ (e.g., the logistic) to determine the precise activation value.) Critically, the output of this covariant learning algorithm reflects the ensemble of training patterns, weighted by two factors: *frequency* and *similarity* (Plaut et al. 1996).

To better understand frequency and similarity effects in Warker and Dell's specific proposal, assume that all training patterns have target outputs of +/−1 over a set of localist representational units (e.g., /t/-onset). This is appropriate for the dichotomous empirical data we will consider below (e.g., preservation vs. violation of target syllable position for consonants). Following this assumption, (3) rewrites (2) to show the network's output for some (trained) target pattern α.

$$(3) \qquad s_j^{[\alpha]} = \sigma\left(\varepsilon \bar{s}_j^{[\alpha]} \sum \left(\begin{array}{l} +F^{[\alpha]} \\ +\sum_{p:\bar{s}_j^{[p]}=\bar{s}_j^{[\alpha]}} F^{[p]} O^{[p][\alpha]} \\ -\sum_{p:\bar{s}_j^{[p]}=-\bar{s}_j^{[\alpha]}} F^{[p]} O^{[p][\alpha]} \end{array} \right)\right)$$

To highlight the relationship between network behavior and the correct (trained) output for target pattern α ($\bar{s}_j^{[\alpha_T]}$), the target output value has been pulled out of the summation from (2). The sign and magnitude of the summation will therefore determine how close the network output is to the trained output. If the summation is negative, the network output will take on the sign opposite the target value; if positive, the network output will be the same sign. The magnitude of the summation will influence how close the output will be to one of the activation extrema (i.e., +/− 1; due to the squashing effect of the sigmoid function σ, the activation values are confined to this region).

After the target output value has been factored out, the summation can be decomposed into three factors. These factors are either positive (for training patterns that share the same output value as the target pattern) or negative (for training patterns associated with the opposite value). With the output values factored out, the only terms left from the summation in (2) are frequency and similarity; each factor is some combination of these two terms. First, the target pattern α contributes positively to the summation based on its frequency (as its similarity to itself is 1). The second positive factor is other patterns whose trained output is the same as α ("friends" of the target; Stemberger 2004); their contribution is weighted by their frequency and similarity to the target pattern. The final, negative, factor is the frequency- and similarity-weighted contribution of other patterns whose trained output is the opposite of α (the target's "enemies"; Stemberger 2004).

2.1.3 Constraint satisfaction in the Hebbian network

Instead of focusing on how each factor influences the sign and magnitude of the summation, we can focus directly on its influence on the output. To the extent that the positive factors are greater in magnitude than the negative one, the summation will be positive and the output will be closer to that of the trained target (i.e., of the same sign). If the negative factor is larger in magnitude, the summation will be negative and the output will take on the opposite sign. The push/pull effect of these different factors can be thought of as a type of constraint interaction. Positive factors are constraints that "prefer" that the output take on the trained target value; negative factors are constraints that prefer the opposite. For example, the positive constraints express their preference by making the summation more positive (which, all else being equal, will cause the output activation to be the same sign as the target). The influence that each factor's preference exerts on the output—the weight of each constraint—is determined by each factor's magnitude. The output reflects the value that best satisfies these weighted constraints. For example, if the positive constraints are stronger (i.e., the positive factors are larger in magnitude), the output will be closer to that of the trained target.

Following this analogy, each factor can be recast as being one of two types of constraint on output activation. The first factor is a very specific constraint; it prefers that the output reflect the trained value for this particular target pattern. Its weight is the frequency of the training pattern. The second and third factors can be seen as more general constraints on output values. The second is a constraint reflecting a general preference for the output value specified by the target; it reflects the sum total preference, over all trained patterns, for the target's output value. Similarly, the third factor reflects the general preference (across all training patterns) for the opposite output value. The weighting of these two general constraints is a reflection of two factors: one, the frequency with which each constraint is obeyed in the training set (i.e., the frequency of the supporting training patterns); two, the degree to which each constraint is obeyed in this particular context (i.e., the similarity of the supporting training patterns' inputs to the target pattern).

It is important to note that although this analysis serves to highlight the contribution of various types of training patterns to network output, the distinction between constraint types does not carry over into network processing mechanisms. Each constraint type is realized over the same representational units and connection weights. Furthermore, although there are some similarities, there are also many critical differences between this conception of constraint interaction and that utilized in grammatical formalisms. This relationship is explicated in greater detail in Section 3 below.

2.2 How do constraints allow the network to account for the empirical data?

Re-conceptualizing the Hebbian network's behavior in this manner allows us to understand its performance using the language of constraint interaction. In particular, error types that satisfy these constraints can be seen as more likely than constraints that violate them. Following Warker and Dell, let us assume that the probability of various error types is determined by the relative activation of

segment–syllable position bindings in the output (e.g., /t/-onset, /k/-coda). Since activation values reflect the interaction of the three network constraints, segment–syllable position bindings will be active when such an output satisfies network constraints—and inactive to the degree that such an output violates them. Consider the first network constraint, which prefers that the output reflect the trained values for the specific input pattern. For input pattern "map," this constraint will push the activation of /m/-onset towards +1; the presence of an /m/ in onset is consistent with the preferences of this constraint. In contrast, the presence of /m/ in coda would be inconsistent with its preferences. This first network constraint will therefore push the activation of the corresponding /m/-coda unit towards –1.

The following subsections examine how Warker and Dell's theory accounts for four empirical phenomena from speech errors. Following Dell and Warker (2007), the discussion is primarily focused on how such errors reflect the influence of acquired phonotactic restrictions. As noted above, Warker and Dell have constructed a simulation of their proposal and its behavior is consistent with the first three empirical findings discussed below (for discussion, see Warker and Dell 2006; Dell and Warker 2007). This section relies on the constraint analysis above to discuss how the theory—independent of its simulation implementation—can account for these findings.

2.2.1 Syllable position preservation

Many studies have observed that segments tend to preserve their target syllable position in errors. For example, *napkin* is more likely to be erroneously produced as *kapkin* (target onset intruding into onset position) than *papkin* (target coda intruding into onset position; see Vousden, Brown and Harley 2000, for a recent review). Warker and Dell's theory would attribute this effect to the first network constraint, which compels similarity to the target pattern. An error which places a segment into the same syllable position as the target (e.g., onset → onset) will better satisfy this constraint than an error placing it into another syllable position (e.g., onset → coda). All else being equal, then, the former error type will be more active, making it more probable.

Note that the strength of this constraint will vary with pattern frequency. Therefore, this theory predicts that syllable position preservation should be stronger for higher frequency targets (I am unaware of any study that has examined errors for such an effect). However, in Warker and Dell's simulations, all legal patterns had the same frequency; any frequency modulation was therefore obscured.

2.2.2 First order phonotactic learning

Dell, Reed, Adams and Meyer (2000) reported that participants could acquire first-order phonotactic restrictions in an implicit learning paradigm. For example, following exposure to a set of syllables where /f/ appeared in onset (but never in coda), participant's speech errors involving /f/ were overwhelmingly restricted to onset position (approximately 3% of such errors appeared in coda position in Dell et al.'s experiment 1). In contrast, segments that appeared equally often in the onset and coda of syllables in the set often violated their target syllable position (roughly 30% of such errors in the same experiment).

Warker and Dell's theory accounts for this finding by changes in the relative strength of the second and third network constraints. Consider the activation of the /f/-onset unit in the experimental condition described above. Exposure to the set of syllables in the experiment will strengthen the second constraint; the relative frequency of friends of /f/-onset will be increased. Conversely, it will weaken the third constraint; as the relative frequency of friends increases, the relative frequency of enemies (dispreferring /f/-onset) will decrease. The acquired asymmetry in the weighting of the second and third constraints will push the activation of /f/-onset towards +1. In contrast, for segments appearing equally often in onset and coda the frequency of friends and enemies is equal. Under this frequency distribution, the second and third constraints will have roughly equal weightings and the second constraint will be unable to push the activation of particular segment-syllable position bindings towards +1.

2.2.3 Second order phonotactic learning

Warker and Dell (2006) examined the acquisition of second order phonotactic restrictions. Whereas first order restrictions on segments refer only to syllable position, their experiments considered restrictions that referred to other phonological structures in the word (e.g., adjacent vowel, non-adjacent medial consonants). For example, in one condition they exposed participants to syllables in which /f/ occurred in onset only when the vowel was /æ/; when the vowel was /ɪ/, /f/ was associated to coda. Participants in their study were able to acquire such restrictions; however, they required substantially more exposure than for first order phonotactics.

In Warker and Dell's theory, this learning delay arises because second order training strengthens not only the second but also the third network constraint. When learning first order phonotactics, participants are exposed only to friends of the target; whenever /f/ appears in the input representation, it is always mapped onto /f/ in the output (e.g., all /f/ syllables resemble /fæm/). In contrast, second order phonotactic learning involves exposure to enemies. In the condition outlined above, participants would be exposed not only to syllables like /fæm/ but also syllables like /mɪf/. Since Warker and Dell claim that the input to the syllabification process contains an unordered set of segments, syllables that share segments in any position have non-zero overlap. This makes /mɪf/ an enemy of /fæm/; it overlaps with /fæm/'s input representation (e.g., both have the segment /f/) and it specifies an inconsistent output (e.g., specifying /f/-onset should be –1). Because both friends and enemies are presented during training, the second network constraint's weight is not significantly stronger than that of the third—blocking the network from acquiring behavior consistent with the second order phonotactic.

This type of interference from similar enemies is known more generally in the connectionist literature as *crosstalk*. In fact, the simplified Hebbian system that forms the basis of the analysis here is incapable of overcoming this effect given the input and output representations Warker and Dell assume. Solving this problem requires the presence of hidden units and a more powerful learning algorithm—for example, the backpropagation algorithm used in Warker and Dell's simulations (see Rumelhart et al. 1986 for discussion). The influence of crosstalk not only reveals the limits of our simplified analysis but also provides

a principled reason for Warker and Dell's more complex three-layer architecture.

2.2.4 The content of phonotactics

The results above concern learnability of phonotactics of varying complexity (i.e., first vs. second order). Another set of studies have examined the content of phonotactics. Clear evidence of learning has been obtained when phonotactic restrictions refer strictly to aspects of phonological structure. The studies discussed above show that participants can learn restrictions on consonants involving syllable (or perhaps word) position and can also learn to combine this with information about other segments present in the word (e.g., adjacent vowels). Goldrick (2004) found that participants can also acquire first order (sound-syllable position) phonotactics at the featural level (e.g., labiodental fricatives /v/ and /f/ occurred in onset position 75% of the time, coda 25%).

In contrast, other results suggest that participants have great difficulty acquiring constraints that do not refer solely to phonological structure. Although speech rate clearly influences the realization of phonological structure (for recent examples, see Davidson 2006; de Jong 2001), rate in and of itself is not a structural component of phonological representations. Consistent with the non-structural characterization of speech rate, Dell and Warker (2007) report that speakers are unable to acquire rate-dependent phonotactics.

Warker and Dell's theory accounts for these contrasting results by virtue of the composition of input representations. Recall that under their proposal acquiring a phonotactic restriction involves strengthening the second network constraint relative to that of the third. In order for this to occur, the relative frequency of patterns which are friends of the target must be increased. Importantly, friends are not simply patterns which share the same output as the target—they must also be similar to the target's input representation. Since only phonological structure (e.g., segment identity) is represented within Warker and Dell's theory, the prosodification process will fail to recognize the similarity of patterns sharing non-phonological structure. In this theory, "friendship" can exist only along purely structural dimensions. Thus, phonotactic restrictions that make reference to phonological structure (syllable position, segmental or featural structure) are learnable, while those that make reference to non-structural aspects of sound (speech rate) are not.

Note that the particular input representations that Warker and Dell used in their simulations did not specify featural structure (i.e., they were composed of atomic segmental identity units). The simulation would therefore be unable to account for Goldrick's (2004) results, which document that humans can learn patterns at this level of structure. However, this is not an in-principle problem for Warker and Dell's theory. Although the particular input representations their simulations utilized lacked features, including other aspects of phonological structure in the input representation would accord with the broad principles of their proposal. This point reveals another virtue of the constraint-based analysis; by abstracting away from the particular simulation implementation, it is possible to understand the degree to which data contradict core versus peripheral assumptions of the connectionist theory.

Such analysis can reveal not only the virtues but also the faults of a proposal. A large body of sociolinguistic research suggests that individuals are sensitive to variation in phonological structure cued by non-structural social variables (see the papers in Chambers, Trudgill and Schilling-Estes 2002 for recent reviews). Such findings suggests that Warker and Dell's theory places too great a restriction on the capacities of the cognitive processes encoding phonotactic regularities. Interestingly, in a perceptually-based implicit learning paradigm (similar to that used in the production-based studies discussed above), Onishi, Chambers and Fischer (2002) found that adult participants could *not* acquire phonotactic restrictions that were gender-specific (e.g., male speaker restricted /f/ to onset, female speaker restricted /f/ to coda). This appears to contradict sociolinguistic findings showing that speakers acquire gender-dependent variation (see Cheshire 2002 for a review). Resolving these conflicting findings is an important area for further development of theories such as Warker and Dell's.

3. Comparison with constraint-based stochastic grammars

Reconceptualizing Warker and Dell's proposal in terms of constraint interaction not only allows us to better understand how the theory accounts for behavioral data. It also allows us to recognize how this theory compares and contrasts with those stated in the other dominant formalism for characterizing the cognitive mechanisms responsible for generating probabilistic behavior—stochastic extensions of Optimality Theory (OT).

3.1 Constraints in the network vs. Optimality Theory

As originally formulated (Prince and Smolensky 1993), constraints in OT come in two basic types. Faithfulness constraints prefer that output structures preserve input structure along various dimensions (e.g., McCarthy and Prince's (1995) MAX constraint prefers output structures that include the segments present in the input). Markedness constraints express preferences over different output structures irrespective of the input (e.g., Prince and Smolensky's constraint ONSET disprefers syllables without onsets). These two basic OT constraint types have clear similarities to the constraints in Warker and Dell's network. The first network constraint, like Faithfulness, prefers the output structure specific to the particular input pattern. In contrast, the second and third network constraints, like Markedness, reflect general preferences—they are not specific to the particular input pattern.

Although there are important similarities between the two approaches, there are also notable differences. Some are merely notational. The network constraints are defined positively; for example, the first network constraint increases the summation, pushing the output towards the target value. In contrast, OT constraints are often defined negatively; MAX punishes outputs lacking target segments. (See Legendre et al. 2006, for further discussion of the interrelationship between positive and negative constraint-based analyses.)

Other differences between OT and network constraints are more substantive. For example, the first network constraint is specific to each particular trained input pattern; its strength varies with the frequency of the pattern. In contrast, OT Faithfulness constraints (such as MAX) often apply equally to all

input patterns (see section 4 for further discussion). Another critical distinction between frameworks is due to the architectural assumptions of Warker and Dell's proposal. In OT, Markedness and Faithfulness constraints contrast in terms of the representations they can refer to (output only vs. input/output relations). In contrast, due to its strictly feed-forward architecture, all constraints in Warker and Dell's theory are conditioned by the structure of *input* patterns, not by outputs.[1] This is because the strength of each network constraint is conditioned not by output structure but by input structure. As shown in (3), the weight of each constraint in the connectionist theory is a function of two factors: the frequency of its participating patterns ($F^{[p]}$) as well as their overlap with the *input* representation of the target ($O^{[p][\alpha]}$). This property of the network restricts the types of constraints it can specify. For example, when the network acquires the phonotactic restriction "/f/ should occur in onset position," the network constraint could be glossed as "output unit /f/-onset should be active when input unit /f/ is active." The parallels between Warker and Dell's theory and OT are therefore limited to the claim that constraint types contrast in scope (specific to the input vs. more general); OT's formal distinction between constraint types has no correspondent in this particular connectionist proposal.

Those familiar with OT may wonder how the network can encode phonotactic restrictions without using purely output-based constraints. Like many generative phonological theories, Warker and Dell's assumes an input representation that encodes a subset of the phonological structures present in the output. For example, both input and output representations for the syllable /fæm/ specify that an /f/ occurs in the string; only the output specifies its prosodic position. Due to this redundancy, the network constraints can encode phonotactic restrictions that are conditioned by structure within the shared subset. For example, the implicit learning studies reviewed above examined restrictions on the association of segments to particular prosodic positions (e.g., /f/ was restricted to onset). Warker and Dell's proposal can encode these restrictions because segmental identity is specified in the input representation (in addition to the output).

Two additional qualifying points are in order regarding the relationship of the second and third network constraints to Markedness constraints. First, this relationship is complicated by the agnosticism of the OT formalism regarding the source of such constraints. Some OT proponents assume Markedness constrains are innate (e.g., Jusczyk, Smolensky and Alloco 2002) while other assume they arise from the learner's experience producing and perceiving phonological structure (e.g., Hayes 1999). The network formalism bears a close relationship to this latter perspective; the constraints reflect the frequency with which the network has encountered a given structure during training. The second point is the relationship of these network constraints to an extension of OT that proposes constraints compelling identity among morphologically related output forms (i.e., Output-Output Faithfulness/OO-F; Benua 1997, et seq.). Unlike OO-F constraints, which privilege particular output forms, the second and third network constraints reflect the summed influence of all non-target patterns. These network constraints do not therefore privilege the properties of *one particular* output form; all forms contribute in proportion to their frequency in the training

[1] Note this is not a necessary property of connectionist networks; see Smolensky (2006a) for discussion of constraint satisfaction in a much broader class of networks.

set. The general scope of these network constraints therefore makes them most comparable to Markedness constraints in OT.

3.2 Constraint interaction in the network vs. Maximum Entropy formulations of Stochastic Optimality Theory

As originally formulated, OT grammars were deterministic; a given input would always be mapped to the same output form(s). Such a formalism is clearly inappropriate for modeling stochastic behavioral data such as speech errors. This section and the following review two broad types of extensions to OT have been proposed to allow it to capture such data. The first extension embeds OT constraints within a Maximum Entropy (ME) framework (Goldwater and Johnson 2003; Hayes and Wilson 2006; Jäger 2004; M. Johnson 2002; Wilson 2006). In common with other formal characterizations (e.g., Prince and Smolensky 1993: Chapter 5; Tesar and Smolensky 1998) each OT constraint is instantiated by a function $c_p(j_\alpha,[t_\alpha])$ which returns the number of violations that the pth constraint assigns to output candidate j_α given input t_α. (Note for functions defining Markedness constraints, t_α is irrelevant.) Each constraint is assigned a weight λ_p which determines its relative strength in influencing the output. Consistent with the common OT formulation of constraints as negative, all weights here are negative (such that increasing numbers of constraint violations decrease probability). ME models define the conditional probability of an output form j_α given some input t_α as:

$$\text{(4)}\quad \Pr(J = j_\alpha \mid T = [t_\alpha]) = \frac{1}{Z^{[t_\alpha]}} e^{\sum_p -\lambda_p c_p(j_\alpha,[t_\alpha])}$$

The $Z^{[t_\alpha]}$ term normalizes the sum of weighted constraint violations by summing the constraint violations over the set of all possible output structures J. Thus, the log probability of any particular output structure is proportional to the weighted sum of its constraint violations.

As in the network, ME constraint interaction is numerical; the relative probability of a form is therefore related to the relative value of constraints preferring vs. dispreferring that form. To illustrate the close parallelism of these two frameworks, consider the relative probability of two candidate structures α (representing the target output) and β (representing a competing non-target output). This is analogous to our network analysis, which considered a single output unit which took on one of two values (+/–1). Given (4), the log conditional probability of the target form (given the target's input t_α) relative to the probability of the competitor will be equal to the difference in the sum of their weighted constraint violations.

$$\text{(5)}\quad \ln\left(\frac{\Pr(J = j_\alpha \mid T = [t_\alpha])}{\Pr(J = j_\beta \mid T = [t_\alpha])}\right) = \sum_p -\lambda_p c_p(j_\alpha,[t_\alpha]) - \sum_p -\lambda_p c_p(j_\beta,[t_\alpha])$$

Suppose following the discussion above that there are two general types of OT constraints. Assume a Faithfulness constraint penalizes deviation from the target output, such that following holds:

$$(6)\quad \begin{aligned} &c_{Faith}(j_\alpha,[t_\alpha]) = 0 \\ &c_{Faith}(j_\beta,[t_\alpha]) > 0 \end{aligned}$$

For these two output forms, there are two corresponding Markedness constraints, which penalize certain structures regardless of the input.

$$(7)\quad \begin{aligned} &c_{Mark-\alpha}(j_\alpha) > 0 \\ &c_{Mark-\alpha}(j_\beta) = 0 \\ &c_{Mark-\beta}(j_\alpha) = 0 \\ &c_{Mark-\beta}(j_\beta) > 0 \end{aligned}$$

Substituting (6) and (7) into (5) and eliminating zero terms allow us to re-express the relative probability of the target form as the interaction of three forces.

$$(8)\quad \ln\left(\frac{\Pr(J = j_\alpha \mid T = [t_\alpha]}{\Pr(J = j_\beta \mid T = [t_\alpha]}\right) = \begin{pmatrix} +\lambda_{Faith} c_{\text{Faith}}(j_\beta,[t_\alpha]) \\ +\lambda_{Mark-\beta} c_{Mark-\beta}(j_\beta) \\ -\lambda_{Mark-\alpha} c_{Mark-\alpha}(j_\alpha) \end{pmatrix}$$

Comparison of (8) with (3) reveals the close connection between this simplified ME model and Warker and Dell's connectionist proposal. Here, the relative probability of the target form is related positively to two constraints. The first is a Faithfulness constraint; since the ME model is formulated using negative constraints, this constraint is not a positive preference for the target form α but a dispreference for the competitor structure β. The second positive influence on the target structure's probability comes from a more general Markedness constraint that prefers the target structure (in the negative formulation, expressed as a dispreference for the competitor structure). The final constraint exerts an inhibitory influence on target output probability; it reflects a general preference for the competitor structure (by dispreferring the target structure). These three forces closely parallel those identified in the analysis of the connectionist network above. This close connection is not surprising, given that ME models are part of a broad class of statistical models (see M. Johnson 2002 for discussion) that includes many connectionist networks as well as the connectionist precursor to OT, Harmonic Grammar (Legendre, Miyata and Smolensky 1990).

3.3 Constraint interaction in the network vs. variable ranking formalizations of Stochastic Optimality Theory

The original formulation of OT differs from ME models not only in its determinism, but also in its reliance on non-numerical constraint interaction. Rather than associate each constraint with a weight, OT as proposed by Prince and Smolensky (1993) ranks constraints in a strict preference order. Regardless of the number of violations assigned by a lower-ranked constraint, the preferences of higher-ranked constraints always dominate. Some versions of Stochastic OT maintain this method of constraint interaction. In these systems, stochastic behavior is produced by allowing multiple constraint rankings to be associated to a single speaker. For example, if some set of rankings X results in α being optimal, while some other set of rankings Y results in competitor structure β being optimal, stochastic selection of rankings from either X or Y will produce output variation between α and β.

3.3.1 The Gradual Learning Algorithm (GLA)

Boersma (1997) proposed a variable ranking extension to OT coupled with a learning algorithm (the GLA). In this proposal, multiple sets of strict constraint rankings are generated by associating each constraint with a probability distribution along an (arbitrary) continuous scale. Several variants of this have been proposed; the discussion below focuses on Boersma and Hayes (2001), in which each constraint C is associated with an independent Gaussian random variable with a standard deviation of 2.0 and a specified mean μ_c (its ranking value). To produce an output, each constraint is associated with a particular value (a selection point) drawn from this probability distribution. The rank order of these selection points (highest-lowest) is then used to determine a strict ranking of the constraints (highest ranked–lowest ranked). If two or more constraints' ranking values lie close enough together, the rank order of their selection points will switch with non-zero probability—yielding multiple possible constraint rankings and potentially producing variation.

Under the assumptions of strict ranking, a target output will be more harmonic than a competitor only if at least one constraint preferring the target dominates all constraints preferring the competitor (c.f. Prince's 2002 "elementary ranking condition"). The GLA's predicted probability of a target output structure α in a two-candidate competition with β is therefore the probability that the selection point of at least one constraint that prefers α to β (the set W_α) will be greater than that of all constraints preferring β to α (the set L_α). (The selection point for constraints that do not distinguish the pairs are irrelevant.) If there is only a single constraint δ in W_α, the outcome of the two candidate competition reduces to the probability that this constraint's selection point will be greater than that of all constraints in L_α:

$$(9) \qquad \Pr(J = j_\alpha \mid T = [t_\alpha]) = \int_{-\infty}^{\infty} \phi(m;\mu_\delta) \prod_{\rho \in L_\alpha} \Phi(m;\mu_\rho) dm$$

where $\phi(m;\mu)$ is the probability density function for a normal distribution with mean μ evaluated at point m and $\Phi(m;\mu)$ is the corresponding cumulative probability distribution function (as variance is constant in Boersma and Hayes's (2001) proposal, it is omitted here). (See Maslova (2004) for an alternative derivation of these results and further discussion of GLA output probabilities.)

Assume as above there are three constraints for the two candidate competition (Faithfulness, Markedness-α, and Markedness-β). Only Markedness-α prefers the competitor (by dispreferring the target); the other constraints prefer the target. Since W_β contains only one constraint, (9) is most easily expressed in terms of the output probability of the competitor β.

$$(10)\ \Pr(J = j_\beta \mid T = [t_\alpha]) = \int_{-\infty}^{\infty} \phi(m;\mu_{Mark-\alpha})\Phi(m;\mu_{Faith})\Phi(m;\mu_{Mark-\beta})dm$$

As with the analysis of the connectionist and ME proposals, (10) reveals that the relative strength of the various constraints determines the relative probability of competitor and target. The integral will take on its greatest value when the probability distribution function for the constraint favoring the competitor (Markedness-α) takes on large values at the same points as the cumulative distribution functions for constraints favoring the target (Faithfulness, Markedness-β). In other words, the probability of the competitor will be highest when the ranking value of the constraint favoring it (i.e., the peak of the ranking value's probability distribution) is significantly greater than the ranking value of constraints favoring the target (such that the cumulative distributions of these constraints' ranking values have reached their peak). The relative strength of Markedness-α—reflected by its ranking value—will therefore determine the relative probability of the competitor being produced.

3.3.2 Partial orderings

Anttila (1997) proposed to produce multiple sets of strict rankings by allowing grammars to specify partial orders of constraints.[2] Let O be the set of all total orderings consistent with a partial order P (i.e., all linear extensions of the partial order). Anttila defined the probability of a given output α as the number of rankings in O where α is optimal divided by the total number of rankings in O. Following the section above, note that for a two-candidate competition the probability of competitor β can be calculated as the probability that the single constraint violated by the target α outranks the other two constraints. Given a partial order P and a corresponding set of total orderings O Anttila's proposal therefore predicts the output probability of the competitor to be the number of rankings in O in which Markedness-α outranks both Faithfulness and Markedness-β divided by the total number of rankings in O.

As with the other analyses it can be shown that the probability of producing the target vs. the competitor is related to the relative strength of constraints

[2] A set of related proposals specifies partial orders by allowing the ranking of one or more constraints to "float" or vary over a subset of the hierarchy of fully-ranked constraints (Reynolds 1994 et seq.).

favoring each outcome. Suppose the competitor is realized as an output under some partial order. If we alter the partial order by specifying that one of the constraints favoring the target strictly dominates the constraint favoring the competitor, the constraint favoring the competitor will be weakened and the competitor's output probability will decrease. For example, consider again a partial order P with a set of total orderings O as above. Let O_β be a non-null subset of O where the competitor β is optimal (i.e., Markedness-α outranks both Faithfulness and Markedness-β). Define a new partial order $P' = P \cup \{\text{Faithfulness} >> \text{Markedness-}\alpha\}$. As summarized in (11), the size of the corresponding sets O' and O'_β will be at least one total ordering smaller than that of O and O_β, decreasing the output probability of the competitor β.

(11)

$$\Pr(J = j_\beta \mid T = [t_\alpha], P) = \frac{|O_\beta|}{|O|} > \frac{|O_\beta| - 1}{|O| - 1} \geq \frac{|O'_\beta|}{|O'|} = \Pr(J = j_\beta \mid T = [t_\alpha], P')$$

In sum, although the strict domination method of constraint interaction makes these systems farther removed from the connectionist theory than the ME model, they are all guided by the general principles of constraint interaction. The relative strength of constraints preferring the target vs. those preferring a competitor determines the relative probability of different outputs.

4. Discussion: Constraint interaction and sound structure

Analysis of a psycholinguistic theory (stated in the connectionist formalism) revealed that it could be characterized by the interaction of two types of weighted constraints: one type specific to the input (preferring the target representation) and the second a more general class of constraint preferring some output structure (either the target output or an alternative). This was used to understand how Warker and Dell's theory could account for four empirical patterns from speech error data. The analysis also provided a motivation for the full complexity of their computational approach (intermediate hidden representations; backpropagation learning) and allowed us to distinguish data that violated critical vs. non-critical assumptions of their proposal. The third section discussed how aspects of both the content of the network constraints as well as their interaction resonated with the general characteristics of linguistic proposals stated within various stochastic extensions of OT.

The parallels between these two approaches suggest that contrasts between stochastic generative grammars and psycholinguistic theories are not arguments between two incompatible architectures, but are instead similar to contrasts between theories within each discipline. Since constraint interaction is neither inherently connectionist nor inherently grammatical, stating proposals in this framework allows us to focus on the substantive issues that divide different theoretical proposals—both within and across disciplines.

One particular theoretical issue noted above concerns the specificity of constraints preferring that the output structure match that of the input (i.e.,

Faithfulness). As discussed above, Warker and Dell's proposal claims that this constraint is pattern-specific (e.g., it is sensitive to the frequency of the specific input pattern). In contrast, as noted above, many OT proposals assume that Faithfulness constraints apply equally to all input patterns (e.g., MAX). This disagreement, however, is not limited to cross-discipline disputes. Connectionist psycholinguistic theories have also assumed constraints whose influence is constant across all inputs. For example, Wheeler and Touretzky's (1997) theory of prosodification assumes that all lexical representations are equally capable of activating a given feature representation in the output. Similarly, within the linguistics literature, exemplar theories (e.g., K. Johnson 1997; Pierrehumbert 2002) have proposed that idiosyncratic properties of individual stored forms can exert an influence on perception or production. The existence of similar claims across disciplines suggest that contrasting grammars and networks is not the proper formulation of theoretical disputes. Rather, in this domain disagreements concern a much more specific, substantive claim: whether item-specific or more general Faithfulness constraints provide a better account of phonological patterns and behavior.

Another contrast noted above concerns the nature of constraint interaction. Warker and Dell's proposal, unlike standard OT, assumes that constraint interaction is numerical. However, as shown by the contrasting formulations of stochastic OT, this is not simply a psycholinguistic vs. grammatical dispute; within stochastic grammatical theories, numerical constraint interaction has its proponents (e.g., Jäger and Rosenbach 2006) as well as its detractors (e.g., Kiparsky in press).

The general framework of constraint interaction has proven itself to be a useful tool for understanding the cognitive representations and processes underlying our knowledge of sound structure—both implicitly within psycholinguistic theories and explicitly within stochastic generative linguistic proposals. Bringing this principle to the fore and stating theories in a common vocabulary can allow theorists to focus on the substantive issues that distinguish theories. This cannot help but further the common cross-disciplinary goal—a more refined understanding of phonological cognition.

References

Ackley, David H., Geoffrey E. Hinton and Terrence J. Sejnowski (1985). A learning algorithm for Boltzmann machines. *Cognitive Science* 9, 147–169.

Anderson, John R. and Christian Lebiere (2003). The Newell Test for a theory of cognition. *Behavioral and Brain Sciences* 26, 587–640.

Anttila, Arto (1997). Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory*, ed. Frans Hinskens, Roeland van Hout and Leo Wetzels, 35–68. Amsterdam: John Benjamins. Rutgers Optimality Archive ROA-63, http://roa.rutgers.edu/.

Bernhardt, Barbara H. and Joseph P. Stemberger (1998). *Handbook of phonological development from the perspective of constraint-based nonlinear phonology*. San Diego: Academic Press.

Benua, Laura (1997). Transderivational identity: Phonological relations between words. Doctoral dissertation, University of Massachusetts, Amherst. Rutgers Optimality Archive ROA-259, http://roa.rutgers.edu.

Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences Amsterdam* 21, 43–58.
Boersma, Paul and Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45–86.
Chambers, J. K., Peter Trudgill and Natalie Schilling-Estes (2002). *Handbook of language variation and change*. Oxford: Blackwell.
Cheshire, Jenny (2002). Sex and gender in variationist research. In Chambers et al. 2002, 423–443.
Davidson, Lisa (2006). Schwa elision in fast speech: Segmental deletion or gestural overlap? *Phonetica* 63, 79–112.
de Jong, Kenneth J. (2001). Rate-induced resyllabification revisited. *Language and Speech* 44, 197–216.
Dell, Gary. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review* 93, 283–321.
Dell, Gary. S., Kristopher D. Reed, David R. Adams and Antje S. Meyer (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26, 1355–1367.
Dell, Gary S. and Jill A. Warker (2007). Using slips to study phonotactic learning in the laboratory. This volume.
Garrett, Merrill F. (1984). The organization of processing structure for language production: Applications to aphasic speech. In *Biological perspectives on language*, ed. David Caplan, Andre R. Lecours and Allan Smith, 172–193. Cambridge, MA: MIT Press.
Goldrick, Matthew (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language* 51, 586–603.
Goldrick, Matthew (in press). Connectionist principles in theories of speech production. In *Oxford handbook of psycholinguistics*, ed. M. Gareth Gaskell. Oxford: Oxford University Press.
Goldwater, Sharon and Mark Johnson (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm Workshop on Variation within Optimality Theory,* ed. Jennifer Spenader, Anders Eriksson and Östen Dahl, 111–120. Stockholm: Stockholm University.
Hayes, Bruce (1999). Phonetically-driven phonology: The role of Optimality Theory and inductive grounding. In *Functionalism and formalism in linguistics (Vol. I, General papers)*, ed. Michael Darnell, Edith Moravscik, Michael Noonan, Frederick Newmeyer and Kathleen Wheatly, 243–285. Amsterdam: John Benjamins.
Hayes, Bruce and Colin Wilson (2006). A maximum entropy model of phonotactics and phonotactic learning. Unpublished manuscript, University of California, Los Angeles. Rutgers Optimality Archive ROA-858, http://roa.rutgers.edu/.
Hopfield, John J. (1982). Neural networks as physical systems with emergent computational abilities. *Proceedings of the National Academy of Sciences USA* 79, 2554–2558.
Jäger, Gerhard (2004). Maximum entropy models and stochastic Optimality Theory. Unpublished manuscript, University of Postdam.
Jäger, Gerhard and Anette Rosenbach (2006). The winner takes it all—almost: Cumulativity in grammatical variation. *Linguistics* 44, 937–971.
Johnson, Keith (1997). Speech perception without speaker normalization: An exemplar model. In *Talker variability in speech processing*, ed. Keith Johnson and John W. Mullennix, 145–165. San Diego: Academic Press.
Johnson, Mark (2002). Optimality-theoretic lexical functional grammar. In *The lexical basis of sentence processing: Formal, computational and experimental issues*, ed. Paola Merlo and Suzanne Stevenson, 59–74. Amsterdam: John Benjamins.

Jusczyk, Peter W., Paul Smolensky and Theresa Allocco (2002). How English-learning infants respond to markedness and faithfulness constraints. *Language Acquisition* 10, 31–74.

Kiparsky, Paul (in press). Where stochastic OT fails: A discrete model of metrical variation. To appear in *Proceedings of Annual Meeting of the Berkeley Linguistics Society*.

Legendre, Géraldine, Yoshiro Miyata and Paul Smolensky (1990). Harmonic Grammar—A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 388–395. Hillsdale, NJ: Lawrence Erlbaum.

Legendre, Géraldine, Antonella Sorace and Paul Smolensky (2006). The Optimality Theory–Harmonic Grammar connection. In *The harmonic mind: From neural computation to Optimality-Theoretic grammar (Vol. 2, Linguistic and philosophical implications)*, ed. Géraldine Legendre and Paul Smolensky, 339–402. Cambridge, MA: MIT Press.

Levelt, Willem J. M., Ardi Roelofs and Antje S. Meyer (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22, 1–75.

Maslova, Elena (2004). Stochastic OT as a model of constraint interaction. Unpublished manuscript, Stanford University. Rutgers Optimality Archive ROA-694, http://roa.rutgers.edu.

McCarthy, John J. and Alan Prince (1995). Faithfulness and reduplicative identity. *University of Massachusetts Occasional Papers 18: Papers in Optimality Theory*, 249–384. Amherst, MA: GLSA, University of Massachusetts, Amherst. Rutgers Optimality Archive ROA-60, http://roa.rutgers.edu.

McCloskey, Michael (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science* 2, 387–395.

Mozer, Michael and Paul Smolensky (1989). Using relevance to reduce network size automatically. *Connection Science* 1, 3–16.

Newmeyer, Frederick (2003). Grammar is grammar and usage is usage. *Language* 79, 682–707.

Onishi, Kristine H., Kyle E. Chambers and Cynthia L. Fisher (2002). Learning phonotactic constraints from brief auditory exposure. *Cognition* 83, B13–B23.

Pierrehumbert, Janet B. (2001). Stochastic phonology. *Glot International* 5, 195–207.

Pierrehumbert, Janet B. (2002). Word-specific phonetics. In *Papers in Laboratory Phonology 7*, ed. Carlos Gussenhoven and Natasha Warner, 101–140. Berlin: Mouton de Gruyter.

Plaut, David C., James L. McClelland, Mark S. Seidenberg and Karalyn Patterson (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103, 56–115.

Prince, Alan (2002). Arguing optimality. In *Papers in Optimality Theory II*, ed. Andries Coetzee, Angela Carpenter and Paul de Lacy, 269–304. GLSA: Amherst, MA. Rutgers Optimality Archive ROA-536, http://roa.rutgers.edu.

Prince, Alan and Paul Smolensky (1993/2002/2004). *Optimality Theory: Constraint interaction in generative grammar*. Technical report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ. Technical report CU-CS-696-93, Department of Computer Science, University of Colorado, Boulder. Revised version, 2002: Rutgers Optimality Archive ROA-537, http://roa.rutgers.edu/. Published 2004, Oxford: Blackwell.

Reynolds, William (1994). Variation and phonological theory. Ph.D. dissertation, University of Pennsylvania.

Rumelhart, David E., Richard Durbin, Richard Golden and Yves Chauvin (1996). Backpropagation: The basic theory. In *Mathematical perspectives on neural networks*, ed. Paul Smolensky, Michael C. Mozer and David E. Rumelhart, 533–566. Mahwah, NJ: Lawrence Erlbaum.

Rumelhart, David E., Geoffrey E. Hinton and Ronald J. Williams (1986). Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, Foundations)*, David E. Rumelhart, James L. McClelland and the PDP Research Group, 318–362. Cambridge, MA: MIT Press.

Shattuck-Hufnagel, Stefanie (1992). The role of word structure in segmental serial ordering. *Cognition* 42, 213–259.

Smolensky, Paul (1986). Information processing in dynamical systems: Foundations of harmony theory. In *Parallel distributed processing: Explorations in the microstructure of cognition (Vol. 1, Foundations)*, David E. Rumelhart, James L. McClelland and the PDP Research Group, 194–281. Cambridge, MA: MIT Press.

Smolensky, Paul (2006a). Optimization in neural networks: Harmony maximization. In *The harmonic mind: From neural computation to Optimality-Theoretic grammar (Vol. 1, Cognitive architecture)*, ed. Géraldine Legendre and Paul Smolensky, 345–392. Cambridge, MA: MIT Press.

Smolensky, Paul (2006b). Computational levels and integrated connectionist/symbolic explanation. In *The harmonic mind: From neural computation to Optimality-Theoretic grammar (Vol. 2, Linguistic and philosophical implications)*, ed. Géraldine Legendre and Paul Smolensky, 503–592. Cambridge, MA: MIT Press.

Stemberger, Joseph P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language* 90, 413–422.

Tesar, Bruce and Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry* 29, 229–268.

VanOrden, Guy C., Bruce F. Pennington and Gregory O. Stone (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review* 97, 488–522.

Vousden, Janet I., Gordon D. A. Brown and Trevor A. Harley (2000). Serial control of phonology in speech production: A hierarchical model. *Cognitive Psychology* 41, 101–175.

Warker, Jill A. and Gary S. Dell (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology: Learning, Memory and Cognition* 32, 387–398.

Wheeler, Deirdre W. and David S. Touretzky (1997). A parallel licensing model of normal slips and phonemic paraphasias. *Brain and Language* 59, 147–201.

Wilson, Colin (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30, 945–982.

Department of Linguistics
Northwestern University
2016 Sheridan Rd.
Evanston, IL 60208
USA

goldrick@ling.northwestern.edu
http://ling.northwestern.edu/~goldrick