To appear in A. Bosch and S.J. Hannahs (eds.), The Routledge handbook of phonological theory.

Connectionist approaches to generative phonology[*]

John Alderete, Paul Tupper
Simon Fraser University

Abstract. While connectionist models are ubiquitous in psycholinguistic approaches to language processing, they are less well-known as generative models of grammar. This work surveys a body of literature in which connectionist models have been developed to address problems central to generative phonology. The focus is on explaining to the newcomer precisely how these models work, and in particular how they grapple with locality, gradience, opacity, and learnability in phonology. An understanding of connectionist phonology both gives a deeper understanding of past developments in phonological theory and a glimpse into its future.

1. Introduction

For many, generative phonology is a kind of symbolic logic for sound patterns. Phonological analysis involves the assignment of contrastive sounds to discrete symbols, or defining the phonemic inventory of a language. Further analysis proposes morpho-phonemic and allophonic processes that transform these symbols to new ones in environments that are likewise defined with symbolic sound structure. Phonological structure may also be organized into syllables and other higher level units like prosodic feet that enrich the symbolic phonology with constituency. This article is about looking a layer beneath this symbolic structure and examining how phonological processes can be accounted for at the micro-structure level in connectionist networks. A connectionist network is a web of interconnected micro-processors that is organized and tuned in such a way to implement input-output processes. Individually, the micro-processors do not do anything close to what we think of when we think of phonology. However, when these micro-processors are organized in the appropriate layers, and when the links between them are properly adjusted, the larger network can produce outcomes that have linguistic interpretations and indeed mimic symbol-manipulating phonology. It is possible, therefore, to construct a generative grammar out of a web of interconnected micro-processors, but the resulting model works rather differently than the macro-structure models some have come to take for granted. The goal of this chapter is to illustrate how these connectionist models work precisely and how they have been applied to some of the problems central to phonological theory.

Connectionist phonology did not develop in a vacuum, and so to understand its motivation, it will be useful to understand a bit of its history. Emerging fully in the 1980s, connectionism is an interdisciplinary research program that pursued the idea that cognitive processes arise from brain-like computation and development. Connectionist models are characterized by networks of micro-processors, called units, which are linked together by connections. Information-processing in connectionist networks is intended to be a simplified model of how information is processed in human neural networks, where units correspond to individual neurons, and connections correspond to the synapses linking them together. Information in connectionist networks is often distributed across multiple units and passed between layers in parallel, as many researchers believe to be true of brain-like computation. This comparison with the human brain has its limits (see below), but these rather general assumptions

[*] We are grateful to Sara Finley, S. J. Hannahs, and anonymous reviewers for helpful comments and suggestions. We alone are responsible for any errors that remain.

have made possible principled explanations of certain key facts about human cognition (see Elman et al. (1996) and McLeod et al. (1998) for useful introductions). The parallel-distributed processing nature of connectionist networks means that they are fault tolerant (resistant to minor damage), content-addressable (memory can be accessed from part of a memory), and well-suited for capturing graded categories and similarity structure. The cognitive architecture of connectionist networks supports the formalization of mature cognitive abilities, as well as the development of those abilities in learning through adjustment of the connection weights. This chapter explores how these assumptions apply to phonological systems.

Another important distinction that can be made between connectionist and traditional generative approaches to language is that connectionist approaches are often task-oriented (Joanisse 2000). Knowledge of language in connectionist approaches can be built up from the acquisition of several distinct cognitive processes, e.g., the development of sound categories in infants, word learning from age 1 onward, and word production and recognition processes. On this view, knowledge of sound patterns is in a sense emergent from the knowledge acquired in these distinct processes (Plaut & Kello 1999), and so language acquisition in connection science has a focus on the communicative function of language. This approach can be contrasted with the more formal approach to acquisition in generative traditions where language learning is typically modeled as a selection of a particular formal grammar from an inventory of possible grammars. Thus, traditional generative grammar posits a notion of linguistic competence that accounts for all and only the attested linguistic systems, typically vetted by typological study. In learning, children are conjectured to select their grammars from this universal inventory. In contrast, connectionist phonology is often linked to models of linguistic behaviors like speech production (Dell 1986; Dell et al. 1993; Goldrick & Daland 2009), speech perception (Gaskell et al. 1995; McClelland & Elman 1986), and language acquisition (Stemberger 1992). Children do not really select a grammar. Rather, their grammatical knowledge is seen as a by-product of these distinct linguistic behaviors.

While much of connectionist phonology is task-oriented and rooted in specific theories of psycholinguistics (see Goldrick (2007) and Stemberger *this volume* for review), there is a significant literature that uses connectionist models to tackle classic problems in generative phonology. Early conceptualization of connectionist phonology in Lakoff (1988, 1993) viewed phonological processes as the work of so-called "constructions", or constraints that hold within or between certain levels of word and sound structure. Lakoff's constructions, which in a sense anticipates later theories of constraint-based phonology, were not explicitly implemented, but his illustrations sketch how the simultaneous application of these constraints could give structure to complex systems of rule interaction and iteration; see Wheeler and Touretzky (1993) for concrete phonological systems using Lakoff's constructions. Another line of research initiated by John Goldsmith (see Goldsmith 1993 for an introduction) grapples with syllable and metrical structure, and illustrates how connectionist networks can address problems of locality and the analysis of graded phonological categories like sonority. We also review connectionist models that were developed to account for other kinds of phonological and morpho-phonological phenomena, including vowel and consonant harmony (Hare 1990; Wayment 2009), disharmony (Alderete et al. 2013), prosodic morphology (Corina 1994), and general models of morpho-phonemic alternations (Gasser & Lee 1990; Hare 1992). Our goal is to explain to the newcomer how these models work, and, importantly, how they address problems central to phonology, like the analysis of locality, gradience, opacity, and learnability.

As successful as they are in addressing these problems, it is fair to say that connectionist approaches to generative phonology have been eclipsed by other research trends in phonology, including Optimality Theory (McCarthy & Prince 1995; Prince & Smolensky 1993/2004), Articulatory Phonology (Browman & Goldstein 1989; Browman & Goldstein 1992), and the larger field of laboratory phonology. However, connectionist approaches in fact have considerable common ground with many of these approaches, and other more recent theoretical developments. For example, connectionist phonology shares with Optimality Theory the idea that phonological systems are constraint-based and information is processed in parallel. Indeed, explicit formal parallels are made between the macro-structure of OT grammars and the micro-structure of connectionist networks in Smolensky and Legendre (2006). Connectionism also shares with exemplar phonology (e.g., Pierrehumbert (2003), Wedel (2006)) the assumption that phonological knowledge is built up gradually from experience, and that phonological models must have a robust account of continuous and graded structure (see e.g., Bybee and McClelland (2005)). Connectionism also provides some of the foundational assumptions of new theories of linguistic knowledge currently in development, including dynamic field theory (Spencer et al. 2009), gradient symbol processing (Smolensky et al. 2014), and information-based theories (Currie Hall 2009). An understanding of connectionist phonology therefore both helps one understand past developments in phonological theory, and it gives a glimpse into the future of phonology.

The rest of this chapter is organized as follows. The next section provides the formal background necessary for understanding how connectionist phonology works. Section 3 reviews some core problems in generative phonology and then goes on to survey a range of connectionist models for addressing some of these problems in segmental and prosodic phonology. Section 4 goes back over the results of these distinct models and examines the nature of the explanations they offer, as well as pointing out some of the known problems for connectionism. Section 5 looks ahead to the future of connectionist phonology.

2. Background

Connectionist models differ from most generative models in that they use numerical computation rather than symbol manipulation. Connectionist networks work on vectors and matrices of real numbers and use principles of linear algebra and calculus to produce outcomes of a desired type. But connectionist models of language are still capable of computing the same kinds of functions that generative models do, as shown by many of the examples we survey in section 3. At the computational level (Marr 1982), therefore, connectionist models can work on the same kinds of inputs and map them to the same kinds of outputs as symbolic computational models.

Information is processed in connectionist networks by computing the activation states of simple micro-processors called units (see McLeod et al. (1998), McMurray (2000), Smolensky (2006a), and Thomas and McClelland (2008) for more detailed introduction to connectionism). The activation of a particular unit is a real number that, together with other activation values, can be given a linguistic or psychological interpretation. In other words, the values of a collection of units can represent a linguistic structure, like a syllable or a word. The flow of this information is neurally-inspired in the sense that the computation of activation states is assumed to be processed in parallel, and the activation state of a particular unit is affected by the activation states of other units connected to it. Figure 1 illustrates some of the core ideas of information processing in connectionist networks. In this illustration, units are organized into distinct layers that correspond to different types of representations. For example, the input and output layers encode the structures coming into and out of the network, and these are distinguished from a so-called hidden layer that mediates between the two. Hidden layers are internal representations that can restructure the input in a way that makes possible certain mappings that would not otherwise be possible.

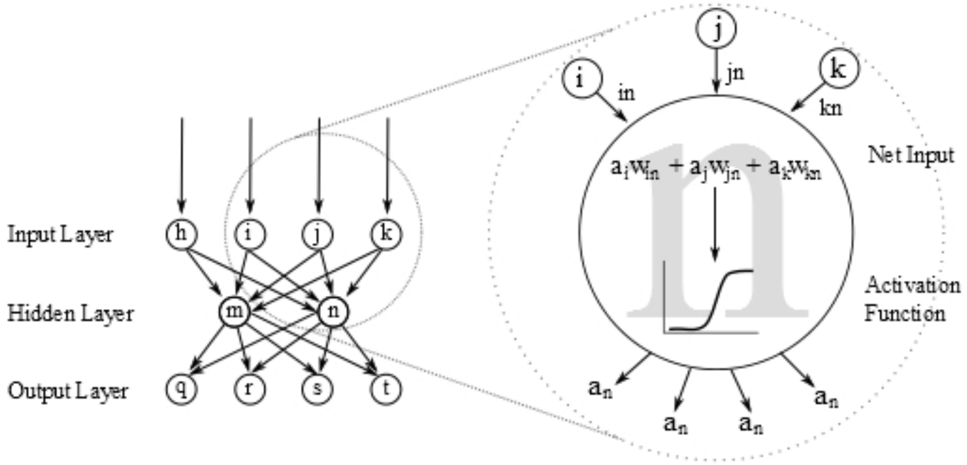


Figure 1. Spread of activation in a simple three layer feedforward network.

The activation state of any given unit is a function of the weighted sum of the activation states of all units sending information to it. Thus, in the enlarged fragment on the right of Figure 1, the activation value of unit *n* is the weighted sum of all activation values feeding into it, i.e., units *i, j, k*, transformed by a so-called activation function (see the side bar for more definitions and formulas). Concretely, if unit *i* has the activation value of 1.0 and it feeds into *n,* and there is a connection weight of .5 linking the two units, then the *input* from *i* to *n*, that is, the specific contribution from this unit, is .5. The so-called *netinput* into unit *n* is simply the sum of all these inputs. The result of this summation is then transformed by the activation function. The activation function may introduce nonlinearity into the system, increasing the expressive power of the network considerably, and it also forces outputs into a numerical range that supports an interpretation suitable for linguistic analysis. Activation functions are often tailored to the problem at hand, but a common one is a sigmoid logistic function. Summarizing, activation values are the output of the composed function: $f_{\text{activation}}$(netinput).

To make all of this a bit more concrete, you can compare the flow of information in a connectionist network to the flow of energy on a light board. Imagine, for example, that the units of Figure 1 are individual light bulbs that can receive energy to produce light. The greater the energy, the stronger the luminance of the light bulb's output, with potentially infinite shades luminance. The spread of activation values from the input to hidden, and then hidden to output, can be thought of a spreading of light patterns generated by the energy surging across the light board, with different light patterns at different states in the spreading action. A particular pattern of light in the input or output can be decoded as a particular representation; it has a meaning in the system.

Panel A. Definitions

Units: micro-processors that receive and pass on information from connected units (artificial neurons)

Weights: real numbers that are co-efficients modulating the spread of activation (artificial synapses)

$input_{ij}$ = $a_i w_{ij}$, or the contribution from a sending unit to a receiving unit

$netinput_i$ = $\sum_j a_j w_{ij}$, or the total contribution to a unit from sending units

Activation function: a function that transforms the output of *netinput* into a bounded range

This metaphor underscores some fundamental differences between connectionist models and symbolic computational models. First, the changing activation patterns involve rather low-level processing of a single currency, activity. Complex patterns and categories are not classified by using different privative elements, but rather with combinations of the activations of simple processing units. Second, since activation values are real numbers, their values can fall anywhere on a continuous dimension, which enables connectionist models to capture graded patterns. This property of connectionist networks makes them well-suited for analyzing graded concepts like sonority and similarity in phonology.

As far as giving the input and output layer a coherent interpretation in linguistics, it is typical to distinguish two kinds of representations, local and distributed representations. Local representations are simply vectors of activation values in which a single element in the vector is used to code a specific category. Thus, to encode a particular structure, one unit is active (canonically represented with "1") and all other units are inactive ("0"). For example, the phoneme [b] can be represented with a vector in which the first element is "1" and all others are "0", so there is a one-to-one relationship between specific elements in the vector and categories in the system. Distributed representations are defined simply as non-local representations: the information coding a particular category is the total combination of activation values in the row vector and there is no limitation that only one is active and all others are not. Illustrations of these two types of representations are given below.

(1)	Category	Local representation	Distributed representation
	"phoneme [b]"	[1 0 0 0 0 0 0 0]	[1 0 1 1 1 0 1 0]
	"phoneme [d]"	[0 1 0 0 0 0 0 0]	[1 0 0 0 1 0 1 1]
	"phoneme [k]"	[0 0 0 0 0 0 1 0]	[1 1 1 0 0 1 0 1]

While both local and distributed representations are used in language processing, distributed representations are more common in connectionist phonology because they are consistent with feature-based generalization. For example, individual elements in the row vector can correspond to the feature values of distinctive features in phonology, supporting generalizations based on these features. Many networks that try to capture generalizations with phonological features encode feature values in units in some way, e.g., [+voice] might be "1" for a specific unit, and [-voice] "-1" for the same unit.

Another important feature of connectionist networks is how time unfolds. Time is characterized in these models as a series of discrete states or ticks of a connectionist clock that meters the passing of information from unit to unit. Some psycholinguistic models are actually sensitive to time in this way. For example, in Dell's (1986) spreading interactive model of language production, speech errors occur with a greater frequency in short time intervals (e.g., 4 ticks in the connectionist clock) as opposed to long time intervals (say, 8 time ticks) because, in longer intervals, the network has a chance to settle into the right activation values of intended speech units before they are selected for speech. However, connectionist models of generative grammar in general do not exploit connectionist time in this way. For example, in Goldsmith and Larson's (1990) model of syllabification (illustrated below in section 3.2), syllabic role nodes are initialized at a certain level of activation, then activation is sent back and forth from neighboring nodes until the nodes settle into an equilibrium state where they do not change very much. Though the intermediate states of these units do influence the outcome, the model is mainly interested in giving interpretations of the inputs and outputs of the system. In this way, most connectionist generative grammars resemble symbolic generative grammars, as intermediate representations do not have any special interpretation other than they are interim representations that can lead to the correct outcomes.

A different aspect of time, the serial order of elements, has been modeled in connectionism rather differently than generative grammar through the use of a context or state layer. Many of the networks we discuss below are non-sequential in the sense that they do not process sequences as a series of discrete units, but rather an entire structure as a whole. The model of English vowel alternations discussed in section 3, for example, processes an entire string of consonants and vowels in tandem. There are important theoretical reasons for this non-sequential property. For example, in psycholinguistic models of speech production (see Stemberger, *this volume*), the processing of a particular segment at a particular point in time can look ahead in the speech stream, or even look back, as evidenced by anticipatory speech errors like *[l]eading* <u>*l*</u>*ist* for *reading list*. This fact means that the language processing requires simultaneous access to both future and past speech units, as embodied in a non-sequential network.

However, simultaneous processing of entire strings can be rather limiting for the modeler because it essentially commits the model's representation of processing units to fixed structures. To address this problem, and others, so-called sequential networks were developed. Jordan (1986) developed such a model for sequences of speech segments in order to account for well-known effects of coarticulation (see Lathroum (1989) for a nontechnical explanation of this model), and Elman (1990) uses a similar kind of model for predicting words in full sentences. In these sequential networks, an entire string is not processed as a whole. Rather, the elements of the string are processed one by one, and each element is processed in the context of what has come before. Another aspect of this kind of model that makes it rather different from non-sequential models is that it tends to involve learning an arbitrary association between two structures. Thus, in Jordan's original sequential model for coarticulation, the function computed by the network was to associate a so-called plan representation with a specific sequence of speech sounds. This association is comparable to the Saussurian sign, where the plan can be thought of as a concept (the signified) and the sequence of segments is the phonological word associated with that concept (the signifier). These associations are learned by allowing a feedback loop between the output and input layers. In essence, the hidden layer on each successful pass through the network is allowed to "see" the network's previous output in the form of the state layer. Thus, the first element in a sequence is processed without any prior structure, but the input to subsequent elements is both the plan layer and the context layer. This type of network has been employed in the analysis of harmony phenomena, which is illustrated in section 3.5.

Another important component of connectionist networks is how memories are learned. For the most part, when we speak of learning in connectionist models, we mean adjustments to the weight matrix that encodes the associations between two layers. Memories of language are stored in these matrices, and so learning these memories involves changing the values of the specific connection weights in response to experience. In Figure 1, for example, the mapping from the hidden to output layers is encoded in an *n* by *m* matrix, where *n* is the number of hidden layer nodes and *m* is the number of output nodes, so 2 by 4, or eight connection weights. Learning the correct mapping from the hidden to output layers involves learning specific values for the eight cells in this matrix. Typically, the values of this matrix are set to 0 or random numbers in the beginning of learning, and then adjusted gradually in response to data. The Delta rule, given below, is a common type of supervised learning rule that makes weight changes in a way that tends to push the network in a direction of making better approximations of the desired mapping. The Delta rule assumes that the learner knows the target activation value of the node *i* and also retains the actual activation value of this node. The difference between the target value and the actual value then is the error. The weight change is the product of the error and the activation value of the input node, scaled by a learning rate ε. Applying this rule iteratively, to all connections in the weight matrix, will gradually reduce the error and therefore improve the network's performance.

(2) Delta rule:
$$\begin{array}{lll} \Delta w_{ij} = [\,a_i \text{ (desired)} - a_i \text{ (observed)}] & a_j & \varepsilon \\ \phantom{\Delta w_{ij} =} [\qquad\qquad \textit{error} \qquad\qquad] & a_{input} & \textit{learning rate} \end{array}$$

Networks with hidden layers require more complex learning rules because we assume that the learner of such a network is not provided with the target activation values of the hidden layer nodes. Backpropagation learning (Rumelhart et al. 1986) involves sending the error signal of an accessible node backward into the network to the inaccessible nodes, in proportion to the weights connecting those nodes, and then updating the deeper connection weights this way (see e.g., Mitchell (1997) for a good introduction to backpropagation learning). In sum, learning in connectionist networks is error-corrective and modeled as gradual adjustments to weight matrices in response to data.

Finally, connectionist networks can be characterized by their overall organization, and therefore how activity flows through the network. The network illustrated in Fig. 1 is a feed-forward network in which activity passes successively through a set of layers, similar to the feedforward nature of phonological derivations in classic generative phonology (though of course the representations are rather different). Connectionist networks sometimes also have feedback loops in which the activation states of some units feed into the units of another layer that has already received input. Such networks are sometimes called recurrent networks. The eight unit recurrent network developed in McClelland and Rumelhart (1985) to solve certain problems in concept learning is a good example of such a network. Sequential networks, like Jordan's (1986) network developed for coarticulation, are recurrent networks because the output of one pass through the network feeds back into the processing of later units through the context layer. Networks can also be constituted by a single layer of units that interact with each other, as in the single layer models discussed in sections 3.2 and 3.3. These networks are designed to model the creation of prosodic structure through competitive inhibition of adjacent units on a single layer of information processing. Thomas and McClelland (2008) is a helpful resource that reviews the historical evolution of these and other neural network architectures.

To summarize the network structures discussed above, the properties below illustrate some of the main ways connectionist networks can be tailored to specific problems, as exemplified in the next section.

(3) Some properties that characterize connectionist networks

a. Encoding categories in representations: local representations have a one-to-one correspondence between individual categories and units; distributed representations do not

b. Existence and number of hidden layers: some networks just have an input and an output, while others have one or more hidden layers

c. Organization of layers: feed-forward networks pass information from one layer to the next, without feedback to prior layers; recurrent networks, like sequential networks, allow feedback

d. Hidden layer units: the number of units in hidden layers can be important in achieving the correct outcome; generalization of the network to novel stimuli is usually forced by having a fewer number of units than the layer feeding into the hidden layer; memorization typically requires an equal or larger number of hidden layer units

e. Conception of sequences: non-sequential networks process an entire sequence simultaneously, sequential networks process individual elements of a sequence in their linear order

3. Connectionist models of phonology

3.1 Classic problems in generative phonology

Let us start with some of the core problems that have been the focus of phonological analysis for some years.

> *Locality*. Phonological processes tend to be local, that is, the target and trigger are generally "close" to each other in some formal sense; even apparent non-local phenomena like stress and vowel harmony, can be viewed as local with the right representations.
>
> *Gradient and scalar phonology*. Many phonological phenomena cannot easily be characterized by binary oppositions, and instead generalizations need to be made with respect to scales or continuous dimensions, e.g., sonority, metrical prominence, and similarity.
>
> *Opacity*. Many phonological effects are not apparent from the surface phonological form. Models of phonology must therefore contend with phonological effects of structure that is hidden from view.
>
> *Learning*. Phonological inventories and processes are learned effortlessly by small children. Phonological analyses can be evaluated by considering if and how they can be learned.

Few would dispute the importance of having a theory that can give natural solutions to these problems. Below, we flesh out how these problems have been addressed in connectionist models of phonology.

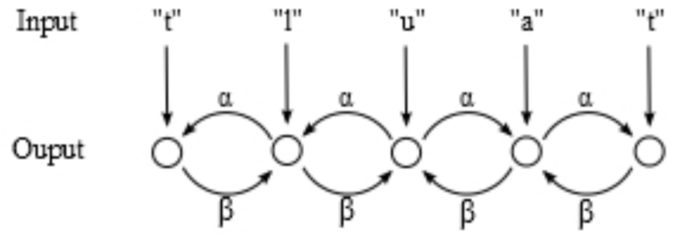
3.2 Syllables

Syllable structure is a good place to start because it is a crucial aspect of any phonological system, and it is a good point of departure for studying connectionist phonology. There are many distinct algorithms for building syllables (Blevins 1995; Itô 1989; Steriade 1982), but, at their heart, syllabification algorithms implement the rather simple idea that syllables are centered over sonority peaks. Thus, syllables are canonically built up in three concrete steps. First, a syllable nucleus is built over high-sonority segments, typically vowels or other sonorants. Syllable onsets are then formed by grabbing a string of rising-sonority consonants and placing them in syllable-initial position. Finally, the residue is dumped into the syllable coda, a process that is subject to certain constraints. Residual material that cannot be put in coda position, for example, may trigger repair rules like epenthesis or deletion.

For many languages, the job of pin-pointing the center of the syllable is a simple matter of finding vowels, which, in turn, triggers the above cascade of operations that fill the onset and coda positions. There are a number of languages, however, for which the job of assigning syllable roles is rather non-trivial, and following the standard protocol sketched above leads to significant loss of generalization. Dell and Elmedlaoui (1985, 1988, 2002)   document such a problem in Tashlhiyt Berber (Afro-Asiatic). Syllables are built in this language by seeking out high-sonority nuclei, while, at the same time, requiring all non-initial syllables to have an onset. This sonority-seeking algorithm, however, is sensitive to eight different levels of sonority, distinguishing two classes of vocoids (low vs. non-low vowels) and six classes of consonants (liquids, nasals, voiced fricatives, voiceless fricatives, voiced stops, voiceless stops). If we follow standard practice of first selecting the correct segment for the nucleus position, this results in eight distinct subroutines, interspersed with onset formation and other principles dictating how already-syllabified segments must be organized (Goldsmith and Larson 1990, Prince and Smolensky 1993/2004).

An alternative to this approach is to find some natural way of capturing the fact that syllabification is sensitive to the distinct sonority levels. Goldsmith and Larson (1990) provide such an alternative by analyzing sonority as continuous activation values (see also (Goldsmith 1992a; Goldsmith 1993b)). In particular, they model the basic competition for the nucleus position in Tashlhiyt Berber as competitive inhibition between adjacent segments. The syllabifications computed by their model are the same as Dell and Elmedlaoui's generative account, and indeed all syllabification systems: it takes a string of segments as input and then returns an assignment of this string to syllabic roles. However, their analysis avoids the need for a myriad of subroutines because their connectionist network captures the effect of sonority with a continuous variable, activity.

Goldsmith and Larson's model is a single layer network with lateral inhibition of adjacent segments. Figure 2 sketches the model and illustrates how it works for the syllabification of the Berber word *tL.wAt* (a Berber place name, capitals are nuclei). Each unit in the model represents a single segment. Different from feed-forward networks, the syllabification algorithm works by changing the states of each unit in the output layer as a function of the influence from its neighbors. As units pass though successive states, they settle into an equilibrium state where their activation values do not change very much (state *n* in Figure 2). The resulting output pattern is one with an alternating pattern of "low-high-low-high" activation values that is interpreted linguistically as syllable peaks (high) and margins (low).



State 0	0	5	7	8	0
State 1	-3.00	0.80	1.70	7.30	-0.80
State 2	-0.48	4.28	2.54	8.31	-0.73
State 3	-2.57	3.52	1.58	8.18	-0.83
State 4	-2.11	4.30	1.73	8.34	-0.82
⋮					
State n (equilibrium)	-2.59	4.32	1.56	8.34	-0.83

Figure 2. Single layer local network for Berber syllabication, based on Goldsmith and Larson (1990); (α,β)=(.6, .10)

Let's flesh out how the specific activation values are calculated in Goldsmith and Larson's model to see how the numerical computation produces this result. Segments are assigned a numerical value for their inherent sonority ranging between 0 and 8 (excluding 6). This value is based on their membership in the eight basic sonority classes and is in principle learnable. Thus, voiced stops like *t* have an inherent sonority of 0, the low vowel *a*, 8, etc. The network is initialized by feeding it these inherent sonority values, in the order they appear in a Berber word, shown in Figure 2 at State 0. The activation value of unit *i* at the next state, or the next tick in the connectionist clock, is then calculated as the inherent sonority of *i*, minus the weighted activation value of its neighbor to the left (which it is connected to by weight β) and its right (connected to by weight α) at State *i*. In this illustration (α,β) = (.6,.1), so we can figure out State 1 for *u* by simply subtracting from *u*'s inherent sonority (=7) the inhibitive value of *l* on its left (= 5 *.1) and *a* on its right (= 8 * .6). Thus, the activation value for the state representing *u* is 7 at state 0, but at state 1 it is: 7 – (.5 + 4.8) = 1.7. The influence from its neighbors has drastically cut the activation value of *u* here, which means its chances of appearing in a syllable peak have just been severely reduced. This same competitive inhibition takes place for all segments simultaneously, with a stronger push downward in activation contributed by the segment on the right, until each unit reaches its equilibrium state.

Goldsmith and Larson's analysis of Berber syllabification gives a very direct analysis of the claim that phonology is local. In their system, which they call "local modeling", segments can only interact with immediately adjacent segments, and yet a global pattern of alternating sonority rises and falls emerges from a series of local struggles for the nucleus position. Also, their analysis captures the graded nature of the sonority scale by assuming sonority is a continuous variable and that all segments compete for a high-sonority position.

Subsequent work has pursued these insights and developed a conceptual framework for relating constraint interaction in connectionist networks to the role of constraints in symbolic theories of grammar like Optimality Theory. In Legendre et al. (2006), a model similar to Goldsmith and Larson's single layer model is developed and tested more thoroughly with computational simulations. The Legendre et al. model, dubbed BrbrNet, is similar in having a single layer with competitive inhibition, but with two main differences. First, the inputs to the network are different. Instead of a linear relationship between sonority and activation where both rise by increments of 1, this relationship is exponential in BrbrNet, in particular $2^{son(x)}$-1. Concretely, the initial activation value in Goldsmith and Larson's model grows with increments of 1, as in: 1, 2, 3 … 8, but it grows exponentially in BrbrNet, i.e., 1, 3, 7, … 255. This exponential growth is argued to be necessary to capture the strict domination structure of OT grammars in the Berber words tested. However, Tupper and Fry (2012) examined the implications of this claim for certain problematic forms and found that in order to give the correct outcomes, this relation needs to be superexponential, a fact that raises problems for the biological plausibility of the model (see below).

Second, the connection weights between output units are assumed to be symmetric, so the competitive inhibition from the neighbor on the left is the same as the neighbor on the right. In Goldsmith and Larson's model these are two independent parameters, α and β, but Legendre et al. argue that symmetry is necessary in order to implement the principle of harmony maximization. This principle is the analogue in OT grammars to the notion that the winner best satisfies the constraint hierarchy. To understand this point, it is necessary to explain the nature of constraints in connectionism. In connectionist networks, individual connections, or sets of connections, encode constraints (Smolensky 1988). Connectionist constraints may not have as transparent an interpretation as well-known constraints in symbolic phonology, like the ones used in Optimality Theory. For connectionist constraints, if a connection between two units is positive, the unit sending information tries to put the unit immediately downstream into the same positive state it is in. The constraint is in a sense satisfied if the state of the receiving unit resembles the state of the sending unit. If, on the other hand, the connection is negative, the sending unit tries to put the receiving unit in the opposite state, so negative weights are satisfied by inducing the opposite activation states downstream.

With this definition, we can make parallels between the constraints of connectionist networks and OT constraints. For example, Legendre et al. (2006) show how the negative connections between output nodes in BrbrNet can be compared to Onset in an OT grammar. These negative connections mean that every unit is doing its best to push down the activation value of the segment on its left (and right as well). Interpreted linguistically, this "push down your neighbor on the left" means that segments try to make their neighbors into syllable margins, which is effectively the function of Onset. Getting back to harmony maximization, the gradual changes in the states of the output, as illustrated in Figure 2 above, can be seen as a gradual process of working toward a state that better satisfies this "push your neighbor into a margin" goal. Over time, then, the network maximizes harmony (better achievement of constraints), just like OT grammars pick a winner that best satisfies a language particular constraint hierarchy. While the number of OT analyses for which such parallels have been made is small in number, Legendre et al.'s demonstrations are quite compelling and insightful. A final question raised by the use of symmetric weights is whether onsets should behave just like codas in competitive inhibition. We return to this question in the analysis of French syllabification immediately below.

These analyses of Tashlhiyt Berber illustrate how cumbersome subroutines in a derivational analysis can be avoided, but what does this approach say about other, perhaps more common, syllabification systems? Laks (1995) applied the Goldsmith and Larson model to French syllabification, and showed how subtle facts of French could be modeled with such an account. Additionally, Laks showed that the parameters of the network could be learned through normal processes of error-corrective learning, and the learning results have interesting implications for the nature of the Onset constraints in the model.

Laks constructed a sample of 832 input strings of French and matched the segmental strings with syllabifications based on the intuitions of six native speakers. Laks used a model similar to Larson and Goldsmith's, and modified the parameters of this model in a training phase using an error-corrective technique suggested in Larson (1992). The key differences with the Berber network are that Laks used a different sonority scale tailored to French in the initial parameter settings, and allowed the inherent sonority to be changed in learning. In particular, the initial three point scale was: consonants = -5, glides = 0, vowels, = 5, but after training, a larger set of contrasts was learned that distinguished six important sonority classes necessary for French. Also, Laks distinguished the output links (α,β) for connecting vowels with their neighbors (.5, .5) from those connecting non-vowels (-.5, -.5), and also allowed these to be modified in learning. The use of negative inherent sonority values and connections means that low sonority segments can actually positively contribute to the activation of high-sonority segments like vowels, which is not possible in either BrbrNet or Larson and Goldsmith's original analysis.

Laks presented 20% of the corpus to this algorithm, and then trained the network by allowing the offending segments' inherent sonority to be changed, and the links of this segment and its neighbors to be changed. After training, the mature network was then tested against the rest of the dataset, with near perfect syllabification (99.87% accuracy). While one might object to the "brute force" nature of the learning algorithm, some of the achievements of the learning system are remarkable and worth considering for further research. In particular, the network started with a rather coarse three-way sonority scale, but learned a very natural six-way sonority scale common to many phonological systems. Also, the simulation results show that some segments can have different derived sonority levels based on context. For example, *s* behaves rather differently in /str/ contexts like *apɔstrɔfə* 'apostrophe', from /rst/ contexts, as in *karstə* 'karst', with a much higher derived sonority in coda position. Finally, Laks points out that the network can be naturally extended to account for the gradient intuitions that native speakers seem to have about certain classes, like ambisyllabic consonants, because the derived sonorities in these contexts are less clear-cut than in other contexts.

One important theoretical implication of Laks' learning results is that its accuracy seems to depend on asymmetric parameters for output links. All of the (α,β) values for the six different sonority classes have higher values for the segment on the right than on the left, and these six classes differ in degree of difference between α and β. For example, vowel links are set for (.282,.481) after training, while liquids end up as (-.36, -.3). These results do not show conclusively that the output link parameters must be asymmetric, because it might be possible to just modify inherent sonority in learning. However, they do seem to challenge the claim made in Legendre et al. (2006) that these links are symmetric; see also Touretzky and Wang (1992) on asymmetric connections and directionality in phonology.

3.3 Stress

The previous account of syllables showed how connectionist networks can account for global patterns of alternating sonority peaks and falls with simple local interactions. This account made use of a common denominator for assigning syllabic roles, sonority, which is realized as a continuous variable, activity. Goldsmith (1992b) extends this approach to stress systems, using essentially the same type of connectionist network, but modeling local competitive inhibition of stress prominence.

Goldsmith's model for stress is again a single layer network, but instead of representing a sequence of segments, the output layer represents a sequence of syllables. In other words, the network is structured just like the network in Figure 2, but with different parameters. Initial activations are assigned based on the language particular properties of the stress system (e.g., initial, penultimate, final stress), and then adjacent syllables compete with each other for prominent syllable status. The table in (4) illustrates how this competition accounts for a stress system with initial stress and alternating secondary stresses. At state 1, the first unit, representing the first syllable, is assigned an initial jolt of activation, but all other units (=syllables) have no activation. At the next state, the activation of the second unit is 0.0 + (1 * -0.7) = -.70. This in turn leads to a positive contribution of .14 to the first unit at state 3 because the weight connecting a unit to its neighbor on its left is -.02, so a_1 = 1 + (-0.2 * -.70) = 1.14. The local influences on neighboring syllables spreads through the word until again the layer settles into an equilibrium state where the unit activation values do not change very much. The resulting pattern, shown at the bottom of (4), is then interpreted as the surface stress pattern, perhaps after transforming the numbers to more discrete values.

(4) State changes for stress system with initial main stress, alternating secondary
Parameters: K(1) = 1.0, K(*i*)= 0.0, (α,β) = (-0.2, -0.7)

	σ_1	σ_2	σ_3	σ_4	σ_5
State 1	1				
State 2	1	-.70			
State 3	1.14	-.70	.49		
State 4	1.14	-.90	.49	-.34	
State 5	1.18	-.98	.70	-.34	.24
State 6	1.18	-.98	.70	-.54	.24
State 7	1.19	-.98	.78	-.54	.37
…					
State *n*	1.20	-1.01	.84	-.69	.48

The single layer model illustrated here has the appearance of a single linguistic representation, either a single string of segments or syllables. As a result, it may seem to lack the derivational steps that are necessary for some kinds of phonological effects. In stress systems, for example, stress assigned on a prior cycle can account for a stress lapse that is retained in a later stage of the word, even though the stress that led to the lapse has been lost. This kind of opaque interaction is argued to require derivational stages in Cohn's (1989) analysis of Indonesian stress, cf. Cohn and McCarthy (1998), as illustrated below for six syllable words. The lack of a secondary stress on the third syllable in (5b) is due to the presence of a penultimate stress at the stage when stress is assigned to the five syllable stem, which is later de-stressed because of word-level penultimate stress in the suffixed form.

(5) Six syllable words in Indonesian: monomorphemic vs. polymorphemic

a. ò o ò o ó o

b. ò o o̲ o ó + o

Goldsmith (1992) shows how this kind of "hidden structure" can be accounted for more directly with the subtle dynamics of his model by simply assigning initial activation values of 1.0 to the penultimate syllable of both the stem and the word. In particular, he shows how the initial high activation of the fourth syllable in (5b) pushes down the activation of the third syllable, accounting for its lack of stress. But, at the same time, the fifth syllable pushes down the activation of the fourth syllable, resulting in it ultimately having activation value consistent with the lack of stress. In a sense, the initial states of the fourth syllable act like an intermediate representation, triggering an effect but ultimately fading away as the layer reaches equilibrium. However, the analysis does not require distinct levels of representation, as the cyclic account does, and the opacity effect is produced with surprisingly little theoretical machinery.

Prince (1993) gives a formal analysis of Goldsmith's connectionist model for stress, and probes its typological consequences. In general, it is treated as a predictive model, like other generative models, and it is evaluated based on how well it accounts for all and only known stress systems. On the positive side, Prince points out that the model accounts for all the stress patterns in Prince (1983), an authoritative reference on certain kinds of stress patterns. It is also successful in accounting for stress window effects. Because of built-in rhythmic alternation, stress is forced to fall within 3 units of the end of a string, which accounts for well-known cases like Spanish that limit stress to three syllables from the end of a word. However, it appears that that the model also over-generates, as it predicts many patterns that do not appear to exist in the world's languages. For example, if an input has two prominent syllables, the model can output an alternating string of stress that begins at one prominent input and ends at the other end, with the rest of the word unstressed (cf. Indonesian polymorphemic words above). Another example is a system where medial syllables are stressed as a rule, rather than the universal pattern of main stress aligning with an edge. It is clear from Prince's investigation that some "pathological systems" are predicted, but one might also reply that Goldsmith's core model is largely unconstrained and predicts a number of well-attested systems with just a handful of free parameters. Perhaps limitations on the range of initial activations (as argued in Idsardi (1992) for lexical stress systems) would produce a better goodness of fit between predicted and attested cases.

3.4 Segmental mappings in morpho-phonemics

Moving down to the segmental level, any model of phonology will have to contend with morpho-phonemics. Morpho-phonemic processes can instantiate automatic phonological changes, e.g., devoicing in English plurals, or non-automatic changes in the sense that they are associated with particular constructions or lexical strata, like the ablaut alternations in English strong verbs. We illustrate a model of morpho-phonemics using a non-automatic process in English, because, as a result of the legacy of the English past tense debate (see McClelland and Rumelhart (1986) and Pinker and Prince (1988) *et seq.*), fully-implemented models of non-automatic processes are far more prevalent. The underlying mechanisms of spreading activation are the same as those used for automatic morpho-phonemic processes, so the analyses of these non-automatic processes extend to simpler automatic processes. The model developed in Plunkett and Marchman (1991, 1993) for the vowel changes in English past tense provides a representative example of this kind of system. Like Plunkett and Marchman's model, a number of connectionist models have been developed to largely address problems in morphology, but in the process, account for non-trivial phonological alternations (Gasser & Lee 1990; Hahn & Nakisa 2000; Hare et al. 1995a; Plunkett & Nakisa 1997); see Stemberger *this volume* for a review of these models and other models, and Anttila (2002) and Inkelas et al. (1997) for discussion of the nature of these problems in symbolic phonology.

Plunkett and Marchman's model is a feed-forward, multi-layered, non-sequential network that uses distributed representations to encode the phonological structure of present and past tense stems. Figure 3 illustrates these basic properties. The input to the system is a distributed representation of a three-segment present tense form. Each segment is encoded as a sequence of 0's and 1's in a six node sequence, and the values for these nodes correspond to values of phonological features necessary to uniquely distinguish segments (i.e., features coding major class, voicing, place and manner). Thus, the sound *p* is represented as [0 1 1 1 1 1], which contrasts with *b* in the second slot reserved for voicing information: [0 0 1 1 1 1]. Three segment inputs and outputs therefore have 18 nodes for the stem (3 * 6) and they also have a final two node sequence for encoding the allomorphs of the past tense suffix, i.e., *-t, -d, -əd,* and -∅ (i.e, the absence of a suffix, as in strong verbs). The function computed by the network is therefore one that maps present tense verbs to their correct past tense forms, including modifying the vowels in irregulars that exhibit ablaut alternations. This mapping is achieved by spreading activation from the input to hidden layer consisting of 20 nodes, and then from the hidden layer to the output layer. It is thus feed-forward because activation values spreading from one layer to the next in a uniform direction. It is also non-sequential because the network has no conception of how the pronunciation of segments unfolds in time. The first segment is simply the one represented with the first six nodes, the second the next six nodes, etc., and all segments are presented to the network simultaneously as the input. Finally, the network has three layers, including a hidden layer that can restructure the input in a way that makes possible certain associations and generalizations. Plunkett and Marchman (1991) compared this three-layer network to a simpler one with just two layers (after McClelland and Rumelhart's original (1986) model for English) and found that this hidden layer was indeed necessary to capture the facts of English.

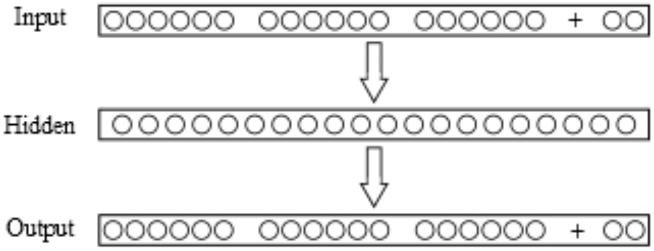


Figure 3. Three layer feed-forward network for English past tense, based on Plunkett and Marchman (1991, 1993).

Examining the activation dynamics in an actual word is useful to explain how the network works. To do English morpho-phonemics, the network must learn both the vowel changes in irregular verbs and the lack of vowel change with regulars. Concretely, it must learn that the past of *meet* [mit] is *met* [mɛt], but also that a suffix is required and the vowels do not change in *cheat* → *cheated*. The input-output pairing in (6) illustrates what this means concretely for *meet-met*. Thus, in the environment *m__t*, the 9th, 10th, and 12th node must change from a "1" to a "0", and everything else must stay the same, but this vowel change must not happen in the *ch__t* environment. The specific model parameters are not given in Plunkett and Marchman (1991), but we know that the input will be restructured in the hidden layer in such a way that it can be classified as an irregular verb and that the combined input-to-hidden and hidden-to-output mappings will change the “1”s to “0”s in the right slots.

(6)	Input:	[m]	[i]	[t]	∅
		010011	111111	001110	+ 00
	Output:	[m]	[ɛ]	[t]	∅
		010011	$11\underline{00}1\underline{0}$	001110	+ 00

An aside about this network is that its task-orientation makes it a little different than typical generative models that map abstract underlying representations onto surface forms. The network simply learns to associate actual present tense forms to actual past tense forms. Though the network does use a hidden layer, which might be compared to something like an intermediate representation (with important differences), the main point here is that the model does not assume a native speaker has abstract information about the input of the present and past tense forms. Learning English morphology is about learning the association between two existing words (see recent discussion in the generative literature, e.g., Albright (2002), also casting doubt on the role of underlying representations).

Another aspect of the hidden layer worth commenting on is the number of hidden layer nodes. Plunkett and Marchman (1991) varied the number of hidden layer units from 10 to 120 units and found that 20 units was a good compromise between an attempt to optimize performance and maximize the generalization properties of the network. This is likely due the fact that the English past tense exhibits both sound based generalizations, i.e., the family resemblances within strong verbs, and many exceptions. The network therefore needs a sufficient number of units for coding the exceptional sound patterns. Simpler more systematic phonology, like final devoicing, however, can be coded with far fewer units because the associations between the natural classes of the input and output are cleaner. Hidden layers with far fewer nodes than the nodes of the inputs and outputs are often used as bottlenecks that force generalizations, whereas a large number of nodes permits item-level associations akin to rote memorization. Connectionist modelers therefore sometimes have to experiment with the number of hidden layer nodes to find the right range suitable for their data. While it is sometimes argued that language particular and phenomenon-specific hidden layers is descriptive in nature and challenges universal conceptions of the cognitive architecture, the specific number of hidden layer nodes is in principle learnable through mechanisms of sprouting and pruning nodes (Fahlman & Lebiere 1990; LeCun et al. 1990), so this argument is more complex and requires further investigation.

3.5 Assimilation and dissimilation

In sections 3.2 and 3.3, we have examined how continua like sonority and stress prominence are captured in connectionist models. Another kind of continuous structure that is an important factor in phonological processes is phonological similarity. The similarity of two segments is rather important in the analysis of segmental phonological processes. For example, many studies of dissimilation have shown how graded categories of similarity are necessary for capturing place cooccurrence restrictions (Frisch 1996; Pierrehumbert 1993). Similarity is also crucial to the analysis of harmony rules, as the activity of a harmony rule is often predicated on some kind of shared feature structure. Connectionist networks are good at capturing graded categories of similarity because distributed representations of activation patterns are sensitive to similarity structure that is not easily captured in symbolic models. We review below some connectionist analyses of harmony and disharmony phenomena that capitalize on these features of connectionist networks.

Many vowel harmony rules only apply when the target and trigger are sufficiently similar in the phonological feature space. Building on the insights of Jordan (1986) for coarticulation (see section 2), Hare (1992) builds a connectionist model of Hungarian vowel harmony specifically designed to address this problem. The analysis requires two key assumptions: (i) that certain nodes can be unspecified for an activation value and thus acquire its value from the nodes representing neighboring segments, and (ii) activation of a current layer is influenced by the output on a prior cycle. Hare's model accounts for the second assumption with a sequential model in which the output of the model, a distributed representation of a vowel feature matrix, cycles back to a state layer, which is then fed as input for the processing of the next vowel (Figure 4).

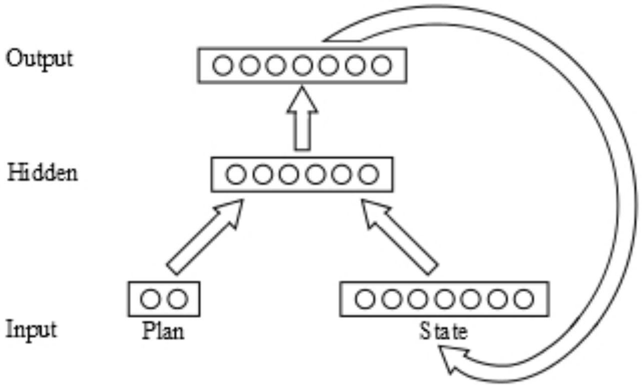


Figure 4. Sequential network of Hare (1992) for vowel harmony.

Let's first flesh out how exactly the model works as a model of vowel harmony, and then return to the issue of capturing the similarity effect. The larger function computed by the network is a mapping of a plan input to a sequence of vowel outputs that together constitute a string of vowels. The plan is an arbitrary vector of activation values that functions something like the linguistic concept that the vowel sequence corresponds to. In other words, if we are interested generating the surface word *CaCi-Ce*, the units associated with the plan layer are just an arbitrary set of activation values that triggers that sequence of vowels. This model is really different, therefore, from feedforward models like Plunkett and Marchman's model of English because it is a mapping from a linguistic concept to a phonological form, not one phonological form to another form. The output at each step is a seven unit distributed representation of a vowel where the activation values of each node correspond to traditional phonological features for vowels (i.e., backness, height, roundness, sonority). As a sequential model, the complete phonological form is generated by successive cycles through the network, where each distributed representation of a vowel is both the output representation of a vowel and the context for the next vowel. Thus, the activation vector of each vowel output is fed back into the network as the state layer. For example, the input for the second cycle, which generates the second vowel, is both the plan input and the state input, which is a kind of memory buffer of the vowel pattern just generated, i.e., the first vowel. The associations between the plan and the sequence of vowels is learned though error-corrective backpropagation learning (see section 2).

The simulation results sketched below in (7) show how Hare's model captures the similarity effect for some key examples. The seven element vectors on the right show how the network encodes ideal vowel features. These are target activation values that the network is trained on—the actual values produced by the network are close approximations of these. The network is designed to model [back] harmony, so the first element in the vector under the [back] column below is unspecified in the last vowel of the sequence. In this case, the final output determines the a/e value of the suffix *–nAk*, which in Hungarian marks the dative. There is no target value for [back] in the last cycle through the network, so it gets its value (underlined below) from the state layer. Which prior vowel colors this final vowel is a matter of phonological similarity, computed as a function of the shared non-back features (the shared features are boxed below). When the closest vowel, V2, is more similar than other vowels, its values are carried over in the final run through the network, as shown in (7a). Here, the actual value for the unit associated with [back] in V3 is .86, but a rounding-up procedure enables this to be interpreted as a "1", which is the value for [+back]. If, on the other hand, a vowel in a non-adjacent syllable is more similar than the local vowel in V2 position, the suffix vowel harmonizes with the more similar but distant vowel. In (7b), *a* shares all its features with the suffix vowel, while *i* shares virtually no features, so *a* is the trigger. Hare's analysis of the similarity effect thus accounts for a basic fact of the system, which is that *i* is transparent, i.e., not a trigger in vowel harmony. A curious fact of Hungarian, however, is that if the suffix vowel is preceded by two transparent *i* vowels, a front vowel does trigger harmony. This fact is also accounted for in Hare's analysis, because the network can look back two syllables for a similar non-adjacent trigger, but no further than this, as demonstrated (7c).

(7) Similarity effects in Hungarian vowel harmony

			target vowel vectors						
	Position	V	back	height			rd		son
a. Local trigger V2 more	V1	ü	0	1	1	1	1	0	0
similar than V1	V2	o	1	0	0	1	1	0	1
e.g., *pügo-nak*	V3	a/e	.86	0	0	0	0	1	1
		→a							
b. Nonlocal trigger V1 more	V1	a	1	0	0	0	1	1	1
similar than V2	V2	i	0	1	1	0	0	0	0
e.g., *taxi-nak*	V3	a/e	.89	0	0	0	1	1	1
		→a							
c. Two syllable threshold	V1	a	1	0	0	0	0	1	1
e.g., *anali:zis-nek*	V2	a	1	0	0	0	0	1	1
	V3	i	0	1	1	1	0	0	0
	V4	i	0	1	1	1	0	0	0
	V5	a/e	.08	0	0	0	0	1	1
		→e							

In sum, the sequential network captures the similarity effect, which is both the result of the activation patterns of prior vowel outputs and the formal limits on retaining the memory of these prior activation patterns.

A somewhat different approach is taken in Wayment (2009) to the similarity effect on phonological processes, which illustrates some additional theoretical assumptions. Consonant harmony, while relatively rare in the world's languages, exhibits the same kind of phonological similarity requirement on the target and trigger as vowel harmony. Thus, in Ineseño Chumash, consonants that share the features [+continuant, coronal] agree in the feature [anterior], e.g., /k-su-ʃojin/ → *k-ʃu-ʃojin* 'I darken it' (Hansson 2001). Like Hare's approach, Wayment (2009) captures the similarity structure of target and trigger in a connectionist network, but the specific mechanism is rather different. Instead of implementing time as a sequence of cycles through a recurrent network, time is captured in a time vector, a specific string of units dedicated to representing the position of a segment in a string that is distinct from the units used to code features. The vector for feature values and the time vector are then combined though a process of filler-role binding that makes use of tensor product representations (Smolensky 2006b). The feature vector encodes the filler of a segment, i.e., what segment it is, and the time vector encodes its role, or where it appears in the string. The entire segment, in a particular position, is encoded with the tensor product of the two vectors, whose dimension is the product of the two vectors. Wayment convincingly shows how phonological similarity can be captured in a single layer network with these filler/role representations (in particular, a Hopfield network), and harmony can be predicated on this similarity together with locality. Wayment further shows how the constraints of his network resemble the properties of a set of attraction constraints in Optimality Theory (Burzio 2002a; Burzio 2002b), illustrating another parallel between the micro-structure of connectionist networks and the macro-structure of OT.

Similarity and gradience have also been the focus of many investigations of dissimilatory phenomena, i.e., dissimilation processes where two similar sounds become less alike, or static root co-occurrence restrictions that achieve the same effect. While some patterns of dissimilation are nearly categorical, dissimilation tends to be a statistical fact of the lexicon and its strength scales with phonological similarity of segments. For example, in Arabic, restrictions against homorganic consonants are stronger for consonants that share more features. As shown in Frisch et al. (2004), phonological similarity between two consonants, established through a metric of shared phonological features, negatively correlates with the representation of the pair in the lexicon. This statistical effect is clearly psychologically real, because native speakers are sensitive to the similarity avoidance effect when they rate nonsense words (Frisch & Zawaydeh 2001). Capitalizing on the ability of connectionist networks to capture gradient effects such as this, Alderete et al. (2013) constructed a connectionist grammar for assessing Arabic roots and analyzed its properties.

Alderete et al.'s model is a non-sequential multi-layer network that takes distributed representations of Arabic triliterals as inputs and outputs a value that assesses the triliteral on a continuous scale from -1 to 1. This network functions differently than other networks, as it does not transform one string into another (Plunkett and Marchman's model of English), and it does not associate a plan with a phonological form (Hare's model of Hungarian). Rather, it functions something like a grammar that takes inputs and assesses them for their overall grammaticality (see Ramsey et al. (1990) for a similar use of output nodes in syntax). In particular, the input has 51 units, or three sets of 17 units, where units encode the feature specifications of the three consonants. The network uses the feature assumptions of Frisch et al. (2004), which is essentially the feature system of Clements and Hume (1995), adapted for Arabic. The activation values of the input spread to a hidden layer of five nodes and then onto the output node responsible for assessing the input. The connection weights between the output node and the hidden layer, and the hidden layer and the input, were trained on a comprehensive sample of Arabic roots, where the connections were gradually adjusted so that attested roots caused the network to produce a value close to “1”, and unattested roots a “-1”. The trained network was shown to capture the effects of similarity, both in a comprehensive test of nonsense words and judgement data from native speakers. In particular, the values for the output node were compared with the human judgement data classifying Arabic roots in Frisch and Zawaydeh (2001), and the network accounted for the same effect of phonological similarity on grammaticality found in this study.

Alderete et al. also scrutinized the internal workings of the network to see how it relates to macro-structure analyses of dissimilation. In particular, using certain statistical techniques, they examined the behavior of each hidden layer node in the trained network to see how it classified the data. In each of three trials, they found that the functions computed by the hidden layer nodes corresponded to a known set of constraints in constraint-based phonology, Obligatory Contour Constraints for place features (Myers 1997; Suzuki 1998). For example, one hidden node functions like the OCP for [pharyngeal] specifications, another hidden layer node for OCP[dorsal], etc. Moreover, the network was shown to capture the differing magnitudes of these distinct constraints and their idiosyncratic exceptions. This example therefore shows, like BrbrNet, that connectionist networks can closely parallel the effects of known phonological constraints.

3.6 Other phenomena

The above survey is by no means exhaustive of the types of phonological processes that connectionist networks have been designed to capture, but it is a good overview of the types of models employed. We summarize briefly some additional phenomena that connectionist models have been built for, and also sketch a few of the problems that have not been solved yet.

From the discussion of connectionist approaches to syllabification and stress, one might form the impression that prosodic constituents themselves are not necessary. Segments and syllables are organized into larger groups centered over peaks of different kinds, but there is no need to invoke the category of a syllable or metrical foot. Many phonological and morphological phenomena do seem to require reference to prosodic structure, and one well-known case is prosodic morphology (McCarthy & Prince 1986; McCarthy & Prince 1993). In prosodic morphology, morphemes are shaped by language particular prosodic units, and part of the analysis has to determine just how the shape facts are predicted. Corina (1994) investigated this problem in Ilokano reduplication, testing to see if a particular type of connectionist network could induce the CVC shape of Ilokano reduplicative prefixes. Corina built a sequential network (as in Hare’s model above) that produced a segment-by-segment output of the desired phonological form. The input to the model was a local representation that combined semantic information and a distributed representation of either a plain form or reduplicated form. After training, the network was found to make many errors, and so it cannot be said to fully account for the facts. However, the network did learn the gross CVC pattern, which the author attributes to the network’s sensitivity to the sonority of the input segments (a structure that was implicitly encoded in the segments) to infer larger prosodic structure. One limitation of the model, shared with Hare’s model of vowel harmony, is that the network only has memory of the two previous segments it has generated. This is a general problem of the sequential networks based on Jordan’s (1986) original design, so perhaps the deeper memory into prior structure allowed in newer models like Elman’s (1990) simple recurrent networks would help improve performance.

Connectionist models have also been developed to account for other phonological processes like epenthesis, deletion, and devoicing (Gasser & Lee 1990; Hare 1992; Hare et al. 1989). It is fair to say, however, that the difficulties of implementing connectionist networks have hampered progress in covering more phonological ground. Well-known segmental processes like palatalization and laryngeal alternations have not really been studied, and tone has also been largely ignored. While initial conceptions of connectionist phonology had a broad vision of grappling with complex rule systems and interaction among various linguistic levels (Lakoff 1993; Wheeler & Touretzky 1993), and while some progress has been made on focused problems (Touretzky & Wheeler 1990a; Touretzky & Wheeler 1990b; Touretzky & Wheeler 1990c; Touretzky & Wheeler 1991), we do not know of any implemented connectionist analyses that approach anything like the rich rule complexity found in cases like Mohawk phonology (Halle & Clements 1983). Another lacuna seems to be the Elsewhere Principle, the idea that specific processes take precedence over more general ones (Kiparsky 1973), though see Tabor et al. (2013) for an analysis of the emergence of the Elsewhere Principle in a classification task. Perhaps one avenue of future exploration is to model the gradual phonological processing of harmonic serialism (McCarthy 2000) at the micro-structure level. Indeed, the basic conception of harmonic serialism, that phonological processes proceed step-by-step and incrementally maximize harmony with respect to a constraint ranking, is rather parallel to the workings of recurrent networks (see discussion in 3.2). In sum, connectionist approaches have not fully addressed some of the problems that phonologists are interested in, but there are some tractable ideas that may help progress towards meeting this goal.

4. Explanations, and challenges to them

To put the models reviewed above in a broader perspective, we reexamine some of the explanations they give to problems in phonology, and also flesh out some of the challenges still faced by this approach.

4.1 Biological plausibility

Some of the initial impetus for connectionist research is the idea that it implements cognitive processes with brain-like computation. Surely, the principles of parallel processing and distributed representation has brought the program a big leap forward in this regard, but many issues remain with the biological plausibility of connectionist networks. The first is based on the analogy between connectionist nodes (or units) and human neurons. Connectionist units have continuous activations, but actual neurons are different from these units in that they are either firing or not. The firing of a neuron occurs at effectively a single instant in time, and then the neuron goes into a refractory period before it can fire again. Some psycholinguistic models actually include a refractory period, like the resetting of activation values to zero in spreading-interactive models of speech production (Dell 1986; Stemberger 2009). But even in these models, firing-refractory states are not broadly invoked across the board, and most linguistic models do not employ such a mechanism. Another problem is that neurons are very sparsely connected, but connectionist models tend to have a rich set of interconnections.

In order to interpret connectionist simulations as neurological in nature, it is necessary to interpret activation as firing rate, i.e., the number of fires per second, and each unit as representing the aggregation of many neurons. Thus, the activation of a single unit can be thought of as corresponding to the average firing rates of many neurons (Dayan & Abbott 2001). This interpretation puts strong constraints on dynamic connectionist networks if we want them to be biologically plausible. For example, the shortest time interval between two firings of the same neuron is on the order of one millisecond. It is therefore unreasonable to expect that significant changes in the firing rate of a neuron (i.e., an activation of a unit) can significantly change over a shorter time interval than that. For many linguistic tasks that take place on the time scale of seconds, the ratio of time scales between the fastest process in the network to the slowest process can be at most 10^5. Any network that utilizes a greater range of time scales is not biologically plausible. As an example, Tupper and Fry (2012) show that connectionist networks implementing OT-style constraint systems, such as BrbrNet (see section 3.2), require a greater range of time scales than this to function properly. Furthermore, this syllabification system was rather simple, involving only a handful of well-formedness constraints. The time scale problem becomes even more difficult as the number of constraints increases.

Another difficulty for connectionist modeling of language has to do with training connectionist networks. Hebbian learning as described above and used explicitly by some linguistic models (e.g., Waymant's (2009) model for consonant harmony), has broad empirical support as a neural-level model of learning. However, Hebbian learning cannot effectively train connectionist networks with hidden layers, and hidden layers have been shown to be crucial to the success of many connectionist models, like Plunkett and Marchman's model of English morpho-phonology (see section 3.4). As explained in section 2, models with hidden layers require backpropagation of the error signal. However, backpropagation as it is usually implemented is unlikely to occur in human neural networks (but see O'Reilly (1996) and Hinton (2016) for some possibilities). In sum, before the connectionist analyses fleshed out here can cash out on the biological plausibility argument, a serious reexamination of the relation between units and neurons, as well as learning, must take place. It should be emphasized, however, that the time scale problem and learning issues are not unique to connectionist models. Connectionism simply makes specific assumptions, some of which directly address the mind-brain problem, and these assumptions lead to difficult questions about how to interpret signal processing in explicit biological models of human neural networks.

4.2 Learning

Another attractive aspect of connectionist modeling is its basic approach to learning. There are well-studied algorithms for training connectionist networks that, once set to initial random weights, do surprisingly well at both modeling known patterns and generalizing to new ones, as illustrated with many of the above examples in section 3. Furthermore, the gradual adjustment of connection weights achieved by these algorithms can be linked in natural ways to the processes that underlie language development. For example, the gradual accumulation of phonological patterns can be seen as a natural by-product of word learning. Some of the models reviewed in section 3 seem successful in this kind of task-oriented approach to learning phonology. For example, Hare's model of learning vowel harmony learns associations between plans and pronunciations, i.e., the relation between concepts and phonological structure, which is essentially word learning. Wayment's model of learning consonant harmony is likewise consistent with word learning, and has the added bonus that it relies only on Hebbian learning. Finally, recent work has also shown how the identity of phonological well-formedness constraints can be learned in connectionist networks (Alderete et al. 2013, cf. Hayes and Wilson 2008).

However, some problems addressed by connectionist models have not really been completely solved. Laks' model of learning French syllables has incredible accuracy, but one might object that the learning algorithm it uses is too brute force and allows adjustment of too many free parameters (i.e., the initial activation, and connection weights both to and from all neighbors). This in turn means that it can learn unattested linguistic patterns. Alderete et al.'s approach to learning the OCP also performs well, but the mappings achieved by the model cannot as yet be thought of as a model of production or perception, so the network behavior does not yet have a natural psycholinguistic interpretation. Examination of the errors produced by connectionist networks can also weigh in on how well the network parallels language development (see e.g., Plunkett (1995)), and further study of phonological development in connectionist networks must attend to this.

4.3 Gradience and scales

One of the clear advantages of connectionist approaches to phonology is that they are naturally gradient, so they give direct analyses of graded phonology and scales. The units that make up layers and the connection weights themselves have continuous values, which permits infinite shades of grey in terms of capturing points on a linguistic dimension. The examples in section 3 illustrated the importance of graded categories in many domains, from suprasegmentals (sonority based syllabification and stress) to segmental phonology (assimilation and dissimilation). These analyses lead to two new questions. The suprasegmental analyses are of interest because they seem to obviate the need for phonological constituency, at least for these select phenomena. We ascribe "peak" and "margin" categories to the higher and lower sonority elements in Berber syllabification models, but this does not mean the segments should be interpreted as forming syllables. These analyses can thus account for the variant realizations of segments, like the difference between a glide and vowel in Tashlhiyt Berber, but the analyses themselves do not involve constituents. One might reasonably ask, then, if phonology needs these constituents at all? It seems unlikely that all of the phenomena traditionally ascribed to prosodic units can be modeled with strictly local interaction. There are just too many phonological processes that depend on the foot and the syllable, and they do not seem easily accounted for with a kind of alignment of high or low activation values. How would laryngeal neutralization in codas be approached or spreading rules that make reference to foot structure? It seems therefore that some mechanism for positing prosodic constituency, and even feature geometry, seems necessary, and the tensor product representations developed in Smolensky (2006b) (see section 3.5) are suitable to this task.

Another issue is how other known scalar phenomena might be treated in connectionist networks. While modern phonology tends to break sound structure into a set of binary oppositions, a number of phonological processes seem to be sensitive to intrinsic scales that are not easily captured by this simple system of contrast (Foley 1970; Gnanadesikan 1997), like Gnanadesikan's inherent voicing scale: voiceless obstruent < voiced obstruent < sonorant. Perhaps these scales, which are often ternary in nature, could be captured in connectionist grammars. A fundamental distinction is made between "adjacent" and "non-adjacent" elements on these scales, and if a natural linguistic dimension could be established that ties all elements on the scale together, then continuous activation values would be suitable to this kind of problem. In other words, the approach to scales is not limited to phonological similarity and sonority.

4.4 Algebraic phonology

One problem that plagues many connectionist networks is that they cannot easily instantiate variables. To make this problem clear, it is necessary to distinguish a certain type of connectionism, namely associationism, from other types, like the models found in Smolensky and Legendre (2006) and Eliasmith (2013), which are in general quite close to the assumptions of mainstream phonology. In associationist connectionism, of the kind represented in McClelland and Rumelhart (1986) *et seq*., and extended to some extent by the models of Hare (1992) and Plunkett and Marchman (1991), the networks themselves have very little *a priori* assumptions, and cognitive processes are built up from data using general-purpose learning procedures. While some assumptions, like the featural basis for unit activation, and the number of nodes, are necessary assumptions to account for the data, the basic idea of these models is that phonological knowledge is built up directly from experience, and very little information is precompiled in it for phonology.

This style of associationism has a problem with implementing variables of the type commonly found in just about every domain of linguistics (see Berent (2013) for extended argumentation). For example, suppose a network is told that AA, DD, ZZ and GG are grammatical expressions in a language, but that AZ, EG, FE, and SP are not. Suppose then we query the network to see if the novel form EE is grammatical. Most associationist networks, with extensive training, will conclude that EE is not grammatical because EE bears no resemblance to any of the grammatical examples, but bears some similarity to the examples EG and FE. Many phonological systems require this kind of generalization, the most obvious of which is the representation of a geminate, and experimental investigations have shown that humans form this kind of generalization (Berent et al. 2001; Gallagher 2013).

The problem with the XX generalization, where X is some atomic structure in a grammatical expression, is not that connectionist networks cannot represent them. The problem is that they do not induce the pattern from limited data. They cannot generalize the pattern to segments that are not in the training set (Marcus 2001; Pinker 1999; Pinker & Prince 1988; Tupper 2016). In order to handle this kind of "generalization across the board", a network has to have such behavior built into it, such as a mechanism that checks the identity of two segments or that copies material. This has been proposed in some connectionist models (Hare et al. 1995b; Shultz 1999); see also Gallagher's (2013) proposal to remedy this problem in Maximum Entropy grammars (Hayes & Wilson 2008).

5. Directions for future research

5.1 Connectionism and Optimality Theory

In a sense, part of the roots of Optimality Theory comes from connectionism. The most direct connection is Harmonic Grammar (Legendre et al. 1990), which represents a kind of "half way point" between connectionist networks and OT grammars because of its use of weighted constraints. Digging deeper, though, is the basic idea that grammar can be constituted by a set of constraints. This is a fundamental idea of connectionism because connections serve as constraints on the possible activations of two nodes (Smolensky 1988), and it is also fundamental to OT. Finally, the idea that outcomes are produced by simultaneous evaluation of multiple factors, and assessed for overall harmony, is again central to both models. It is true that most of OT involves symbolic computation, and connectionist networks use numerical computation, but the focus on constraints and parallelism gives the two approaches significant common ground (McCarthy 2002; Smolensky & Legendre 2006).

The examples above that establish parallels between connectionist networks and OT grammars, like the role of the Onset constraint in symbolic and connectionist phonology, are fascinating in their own right, and they bring to the fore the shared principles in the different approaches. These examples are currently few in number, however, and they are based on small fragments of phonological systems. Whether a rich set of parallels exists between the two types of analysis remains to be seen. For example, in the case of Arabic consonant phonology, Alderete et al. (2013) show how the hidden layer nodes "act sort of like" OCP-Place constraints, but the resemblance is not at all exact, and connectionist constraints are in fact laden with context-sensitive effects and exceptions. OT constraints like the OCP are cleaner and do not generally have detailed context-sensitivity. The French syllabification and harmony examples present similar segment- and feature-level intricacies that also seem to defy projection to the macro-structure level. On the other hand, other aspects of OT models seem worthy of examination, like the connection between harmonic serialism and recurrence mentioned in section 3.6.

5.2 Connectionism and exemplar phonology

Another research domain that connectionism can shed some light on is exemplar phonology, or the use of exemplar models of classification and production in phonological analysis (see Wedel (2006) for a review). Connectionist models actually start with very different theoretical assumptions from exemplar models (though some hybrid models do exist, e.g., Kruschke (1992)). As we have seen, connectionist networks involve layers of units linked by connections with varying weights. Connectionist networks implement processes by sending information through a web of nodes and outputting an activation pattern that has an interpretation of some kind. On the other hand, the fundamental unit in exemplar models is the exemplar itself, a detailed memory of some linguistic token. Each exemplar has a location in a space of representations, a label indicating what it is an exemplar of, and a weight indicating how strong the memory of the token is. Linguistic processes include classifying new tokens by their similarity to exemplars already labeled in the representational space, and generating a new token by sampling exemplars according to their weight and reproducing one. Despite these rather different assumptions, there is important common ground between the two models that is useful to understanding how the models work (Bybee & McClelland 2005). First, both are naturally gradient because of their use of continuous variables, i.e., unit activations and connection weights for connectionism, the representational space and exemplar weights for exemplar models. Second, both have an emphasis on learning: information about phonological patterning is built up over stored observations. Third, related to the emphasis on learning, no effort is made to minimize the role of long-term memory. Both models are also task-oriented in the sense that knowledge of sound structure is built from normal processes of word learning, which contrasts with more formal models of language learning that deemphasize the role of memory. Finally, both models have difficulty with making human-like generalization, especially generalization to inputs unlike those in the training data (see Berent's (2013) points on generalization of phonological rules to non-native segments).

Besides these similarities, recent trends in cognitive science are now blurring the lines between exemplar models and connectionist models. A comparatively new development is dynamic field theory, a modeling paradigm built upon connectionist foundations (Erlhagen & Schöner 2002; Johnson et al. 2008). Besides the standard units and connections of connectionism, dynamic field theory introduces neural fields, dense arrays of units interconnected with each other, as well as with other fields and units. Whereas in connectionism each unit has an activation that depends just on time, a field has activation that depends both on time and on the particular location in the field. Neural fields have a combination of excitatory and inhibitory connections that allow the formation of stable peaks of activation, and the location of these peaks of activation can be used to represent the values of continuous variables, such as physical location or color. Furthermore, dynamic fields can be seen as a way of neutrally implementing exemplar models. In particular, the activation of a field at a particular location can be used to represent the weight and number of exemplars at a given location in representational space (Jenkins & Tupper 2016; Tupper 2014). See Spencer et al. (2009) for an overview of dynamic field theory and its rich interaction with connectionism.

6. References

Albright, Adam. 2002. The identification of bases in morphological paradigms: University of California, Los Angeles.

Alderete, John, Paul Tupper & Stefan A. Frisch. 2013. Phonological constraint induction in a connectionist network: learning OCP-place constraints from data. Language Sciences 37.52-69.

Anttila, Arto. 2002. Morphological conditioned phonological alternations. Natural Language and Linguistic Theory 20.1-42.

Berent, Iris. 2013. The phonological mind Cambridge: Cambridge University Press.

Berent, Iris, Daniel L Everett & Joseph Shimron. 2001. Do phonological representations specify variables? Evidence from the obligatory contour principle. Cognitive Psychology 42.1-60.

Blevins, Juliette. 1995. Syllable in phonological theory. The handbook of phonological theory, ed. by J. Goldsmith, 206-44. Cambridge, MA: Blackwell.

Browman, Carol & Louis Goldstein. 1989. Articulatory gestures as phonological units. Phonology 6.201-51.

—. 1992. Articulatory phonology: An overview. Phonetica 49.155-80.

Burzio, Luigi. 2002a. Missing players: Phonology and the past-tense debate. Lingua 112.157-99.

—. 2002b. Surface-to-surface morphology: When your representations turn into constraints. Many morphologies, ed. by P. Boucher, 142-77: Cascadilla Press.

Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. The Linguistic Review 22.381-410.

Clements, G.N. & Elizabeth V. Hume. 1995. The internal organization of speech sounds. The handbook of phonological theory, ed. by J.A. Goldsmith, 245-306. Cambridge, MA: Blackwell.

Cohn, Abigail. 1989. Stress in Indonesian and bracketing paradoxes. Natural Language and Linguistic Theory 7.167-216.

Cohn, Abigail & John J. McCarthy. 1998. Alignment and parallelism in Indonesian phonology. Working Papers of the Cornell Phonetics Laboratory 12, 53-137. Ithaca, NY: Cornell University.

Corina, David. 1994. The induction of prosodic constraints: Implications for phonological theory and mental representations. The reality of linguistic rules, ed. by S.D. Lima, R.L. Corrigan & G.K. Iverson, 115-45. Amsterdam: John Benjamins.

Currie Hall, Kathleen. 2009. A probabilistic model of phonological relationships from contrast to allophony: Doctoral dissertation.

Dayan, Peter & L. F. Abbott. 2001. Theoretical neuroscience: Computational and mathematical modeling of neural systems. The MIT Press, 2001. Cambridge, MA: The MIT Press.

Dell, François & Mohamed Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. Journal of African Languages and Linguistics 7.105-30.

—. 1988. Syllabic consonants in Berber: some new evidence. Journal of African Languages and Linguistics 10.1-17.

—. 2002. Syllables in Tashlhiyt Berber and in Moroccan Arabic Dordrecht: Kluwer.

Dell, Gary S. 1986. A spreading-activation theory of retrieval in sentence production. Psychological Review 93.283-321.

Dell, Gary S., Cornell Juliano & Anita Govindjee. 1993. Structure and content in language production: A theory of frame constraints in phonological speech errors. Cognitive Science 17.149-95.

Eliasmith, Chris. 2013. How to build a brain: A neural architecture for biological cognition Oxford: Oxford University Press.

Elman, Jeffrey. 1990. Finding structure in time. Cognitive Science 14.179-211.

Elman, Jeffrey, Elizabeth Bates, Mark Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. Rethinking innateness: A connectionist perspective on development Cambridge MA: MIT Press.

Erlhagen, Wolfram & Gregor Schöner. 2002. Dynamic field theory of movement preparation. Psychological Review 109.545-72.

Fahlman, S & C Lebiere. 1990. The CASCADE-CORRELATION learning architecture. Technical Report CMU-CS-90-100. Pittsburgh: Computer Science Department, Cargnie Mellon University.

Foley, James. 1970. Phonological Distinctive Features. Folia Linguistica IV 1/2.87–92.

Frisch, Stefan. 1996. Similarity and frequency in phonology: Northwestern University Doctoral dissertation.

Frisch, Stefan A., Janet Pierrehumbert & Michael B. Broe. 2004. Similarity avoidance and the OCP. Natural Language and Linguistic Theory 22.179-228.

Frisch, Stefan & Bushra Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. Language 77. 91-106.

Gallagher, Jillian. 2013. Learning the identity effect as an artificial language: Bias and generalization. Phonology 30.1-43.

Gaskell, M. Gareth, Mary Hare & William Marslen-Wilson. 1995. A connectionist model of phonological representation in speech perception. Cognitive Science 19.407-39.

Gasser, Michael & Chan-Do Lee. 1990. Networks that learn about phonological feature persistence. Connectionist natural language processing, ed. by N. Sharkey, 349-62. Oxford: Intellect.

Gnanadesikan, Amalia. 1997. Phonology with ternary scales: University of Massachusetts, Amherst Doctoral dissertation.

Goldrick, Matt. 2007. Connectionist principles in theories of speech production. The Oxford handbook of psycholinguistics, ed. by G. Gaskell, 515-30. Oxford: Oxford University Press.

Goldrick, Matthew & Robert Daland. 2009. Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. Phonology 26.147-85.

Goldsmith, John. 1992a. Local modeling in phonology. Connectionism: Theory and Practice, ed. by S. Davis, 229-46. Oxford: Oxrford University Press.

—. 1992b. Local modelling in phonology. Connectionism: Theory and Practice, ed. by S. Davis, 229-46. New York: Oxford University Press.

—. 1993a. Harmonic phonology. The Last Phonological Rule: Reflections on Constraints and Derivations, ed. by J. Goldsmith, 21-60. Chicago: University of Chicago Press.

— (ed.) 1993b. *The last phonological rule: Reflections on constraints and derivations*. Chicago: University of Chicago Press.

Goldsmith, John & Gary Larson. 1990. Local modeling and syllabification. Proceedings of the 26th annual meeting of the Chicago Linguistics Society, Part 2., ed. by K. Deaton, M. Noske & M. Ziolkowski. Chicago: Chicago Linguistics Society.

Hahn, Ulrike & Ramin Charles Nakisa. 2000. German inflection: Single route or dual route? Cognitive Psychology 41.313-60.

Halle, Morris & George N. Clements. 1983. Problem book in phonology : a workbook for introductory courses in linguistics and modern phonology Cambridge, Mass.: MIT Press.

Hansson, Gunnar. 2001. Theoretical and typological issues in consonant harmony: University of California, Berkeley Doctoral dissertation.

Hare, Mary. 1990. The role of similarity in Hungarian vowel harmony: A connectionist account. Connectionist natural language processing, ed. by N. Sharkey, 295-322. Oxford: Intellect.

—. 1992. Phonological representation and processing in connectionist networks: UC San Diego.

Hare, Mary, David Corina & Garrison Cottrell. 1989. A connectionist perspective on prosodic structure. Proceedings of the fifteenth annual meeting of the Berkeley Linguistics Society, 114-25. Berkeley: UC Berkeley.

Hare, Mary, Jeffrey Elman & Kim G Daugherty. 1995a. Default generalisation in connectionist networks. Language and Cognitive Processes 10.601-30.

—. 1995b. Default generalization in connectionist networks. Language and Cognitive Processes 10.601-30.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39.379-440.

Hinton, Geoffrey E. 2016. Can sensory cortex do backpropagation? Paper presented at the Information, Inference, and Energy. A symposium to celebrate the work of Professor Sir David MacKay FRS., University of Cambridge.

Idsardi, William. 1992. The computation of prosody: MIT Doctoral dissertation.

Inkelas, Sharon, Orhan Orgun & Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of grammar. Derivations and constraints in phonology, ed. by I. Roca, 393-418. Oxford: Oxford University Press.

Itô, Junko. 1989. A prosodic theory of epenthesis. Natural Language and Linguistic Theory 7.217-59.

Jenkins, Gavin & Paul Tupper. 2016. A dynamic neural field model of speech cue compensation. Proceedings of the 28th annual meeting of the Cognitive Science Society.

Joanisse, Marc F. 2000. Connectionist phonology: University of Southern California Doctoral.

Johnson, Jeffrey S, John P Spencer & Gregor Schöner. 2008. Moving to higher ground: The dynamic field theory and the dynamics of visual cognition. New Ideas in Psychology 26.227-51.

Jordan, Michael I. 1986. Attractor dynamics and parallelism in a connectionist sequential machine. Proceedings of the Eight Annual Conference of the Cognitive Science Society, 531-46. Hillsdale, NJ: Lawrence Erlbaum.

Kiparsky, Paul. 1973. "Elsewhere" in phonology. A Festschrift for Morris Halle, ed. by S. Anderson & P. Kiparsky, 93-106. New York: Holt, Rinehart and Winston.

Kruschke, John K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review 99.22-44.

Lakoff, George. 1988. A suggestion for a linguistics with connectionist foundations. Proceedings of the 1988 Connectionist Models Summer School, ed. by D. Touretzky, G.E. Hinton & T.J. Sejnowski, 301-14. San Mateo, CA: Morgan Kaufmann.

—. 1993. Cognitive Phonology. The last phonological rule: Reflections on constraints and derivations., ed. by J. Goldsmith, 117-45. Chicago: University of Chicago Press.

Laks, Bernard. 1995. A connectionist account of French syllabification. Lingua 95.51-76.

Larson, Gary. 1992. Automatic learning in a dynamic computational network. Proceedings of the Interdisciplinary Workshop on Compositionality in Cognition and Neural Networks I, ed. by D. Andler, E. Bienenstock & B. Laks. Paris: CREA. Ecole Polytechnique.

Lathroum, Amanda. 1989. Feature encoding by neural nets. Phonology 6.305-16.

LeCun, Y, J. S Denker & S. A Solla. 1990. Optimal brain damage. Advances in neural information processing systems, Volume 2, ed. by D. Touretzky, 598-605. San Mateo, CA: Morgan Kaufmann.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society, ed. by M. Ziolkowski, M. Noske & K. Deaton, 237-52. Chicago: Chicago Linguistic Society.

Legendre, Géraldine, Antonella Sorace & Paul Smolensky. 2006. The Optimality Theory-Harmonic Grammar connection. The harmonic mind: From neural computation to Optimality Theoretic grammar, ed. by P. Smolensky & G. Legendre, 339-402. Cambridge, MA: The MIT Press.

Marcus, Gary F. 2001. The algebraic mind: Integrating connectionism and cognitive science Cambridge, MA: The MIT Press.

Marr, David C. 1982. Vision: A computational investigation into the human representation and processing of visual information San Francisco: W. H. Freeman.

McCarthy, John J. 2000. Harmonic serialism and parallelism. Proceedings of the North East Linguistics Society 30, ed. by M. Hirotani, 501-24. Amherst, MA: CLSA Publications.

—. 2002. A thematic guide to Optimality Theory Cambridge: Cambridge University Press.

McCarthy, John J. & Alan Prince. 1986. *Prosodic Morphology*. University of Massachusetts at Amherst and Brandeis University Rutgers University Center for Cognitive Science, Technical Report No. 32.

—. 1993. *Prosodic Morphology I: Constraint interaction and satisfaction* Rutgers Center for Cognitive Science, Technical Report No. 3.

—. 1995. Faithfulness and reduplicative identity. University of Massachusetts Occasional Papers 18, Papers in Optimality Theory, ed. by J. Beckman, S. Urbanczyk & L. Walsh, 249-384. Amherst, MA: Graduate Linguistic Student Association.

McClelland, James L. & Jeffrey Elman. 1986. The TRACE model of speech perception. Cognitive Psychology 18.1-86.

McClelland, James L. & David Rumelhart. 1985. Distributed memory and the representation of general and specific information. Journal of Experimental psychology: General 114.159-88.

McClelland, James L. & David E. Rumelhart. 1986. On learning the past tenses of English verbs. Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume 2: Psychological and biological models, ed. by J.L. McClelland, D.E. Rumelhart & T.P.R. Group, 216-71 Cambridge, MA: The MIT Press.

McLeod, Peter, Kim Plunkett & Edmund T. Rolls. 1998. Introduction to connectionist modelling of cognitive processes Oxford: Oxford University Press.

McMurray, Bob. 2000. Connectionism for ...er... linguists. The University of Rochester Working Papers in the Language Sciences, 2000(1), ed. by K. Crosswhite & J. McDonough, 72-96. Rochester: University of Rochester.

Mitchell, Tom M. 1997. Machine learning Boston, MA: McGaw Hill.

Myers, Scott. 1997. OCP effects in Optimality Theory. Natural Language and Linguistic Theory 15.847-92.

O'Reilly, R. C. 1996. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. Neural Computation 8.895-938.

Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. NELS 23, 367-81.

—. 2003. Probabilistic phonology: Discrimation and robustness. Probability theory in linguistics, ed. by R. Bod, J. Hay & S. Jannedy, 177-228. Cambridge, MA: The MIT Press.

Pinker, Stephen. 1999. Words and rules New York: Harper Collins.

Pinker, Steven & Alan Prince. 1988. On Language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Cognition 28.73-193.

Plaut, David C. & Christopher T Kello. 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. The emergence of language, ed. by B. MacWhinney. Mahwah, NJ: Lawrence Erlbaum Associates, Ltd.

Plunkett, Kim. 1995. Connectionist approaches to language acquisition. The handbook of child language, ed. by P. Fletcher & B. MacWhinney, 36-72: Blackwell.

Plunkett, Kim & Virginia Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. Cognition 38.43-102.

—. 1993. From rote learning to system building: acquiring verb morphology in children and connectionist nets. Cognition 48.21-69.

Plunkett, Kim & Ramin Charles Nakisa. 1997. A Connectionist Model of the Arabic Plural System. Language and Cognitive Processes 12.807-36.

Prince, Alan. 1983. Relating to the grid. Linguistic Inquiry 14.19-100.

—. 1993. *In defense of the number i. Anatomy of a linear dynamic model of linguistic generlizations*. Rutgers University, Center for Cognitive Science.

Prince, Alan & Paul Smolensky. 1993/2004. Optimality theory: Constraint interaction in generative grammar Malden, MA: Blackwell.

Ramsey, William, Stephen Stich & Joseph Garon. 1990. Connectionism, eliminativism and the future of folk psychology. Connectionism: Debates on folk psychology, ed. by C. Macdonald & G. Macdonald, 311-38. Cambridge, MA: Basil Blackwell.

Rumelhart, David, Geoffrey E Hinton & Ronald J Williams. 1986. Learning internal representations by error propagation. Parallel distributed processing: Explorations in the microstructure of cognition. Vol 1-2, ed. by J.L. McClelland, D. Rumelhard & T.P.R. Group, 318-62. Cambridge: The MIT Press.

Shultz, Thomas R. 1999. Rule learning by habituation can be simulated in neural networks. Proceedings of the twenty first annual conference of the Cognitive Science Society.

Smolensky, Paul. 1988. On the proper treatment of connectionism. The Brain and Behavioral Sciences 11.1-23.

—. 2006a. Formalizing the principles I: Representation and processing in the mind/brain, 147-205.

—. 2006b. Tensor product representations: Formal foundations. The harmonic mind: From neural computation to Optimality-Theoretic grammar, ed. by P. Smolensky & G. Legendre, 271-344. Cambridge, MA: The MIT Press.

Smolensky, Paul, Matt Goldrick & Donald Mathis. 2014. Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. Cognitive Science 38.1107-38.

Smolensky, Paul & Géraldine Legendre. 2006. The harmonic mind. From neural computation to optimality theoretic grammar Cambridge, MA: The MIT Press.

Spencer, John P., Michael S. C Thomas & James L. McClelland (eds) 2009. *Toward a unified theory of development: Connectionism and dynamic field theory re-considered*. New York: Oxford.

Stemberger, Joseph P. 1992. A connectionist view of child phonology: Phonological processing without phonological processes. Phonological development: Models, research, implications, ed. by C.A. Ferguson, L. Menn & C. Stoel-Gammon, 165-89. Timonium, MD: York Press.

—. 2009. Preventing perseveration in language production. Language and Cognitive Processes 24.1431-70.

—. this volume. Connectionist phonology: Interfaces. The Routledge Handbook of Phonological Theory, ed. by A. Bosch & S.J. Hannahs.

Steriade, Donca. 1982. Greek Prosodies and the Nature of Syllabification: MIT Doctoral dissertation.

Suzuki, Keiichiro. 1998. A typological investigation of dissimilation: University of Arizona Doctoral dissertation.

Tabor, Whitney, Pyeong W Cho & Harry Dankowicz. 2013. Birth of an abstraction: A dynamical systems account of the discovery of an Elsewhere Principle in a category learning task. Cognitive Science 37.1193-227.

Thomas, Michael S. C. & James L. McClelland. 2008. Connectionist models of cognition. Cambridge handbook of computational psychology, ed. by R. Sun, 23-58. Cambridge: Cambridge University Press.

Touretzky, David S & Xuemei Wang. 1992. Energy minimization and directionality in phonological theories. Proceedings of the 14th annual conference of the Cognitive Science Society.248-52.

Touretzky, David S & Deirdre W Wheeler. 1990a. A computational basis for phonology. Technical Report AIP 113. Carnegie Mellon University. Pittsburgh, PA.

—. 1990b. From syllables to stress: A cognitively plausible model. Technical Report AIP 117. Carnegie Mellon University. Pittsburgh, PA.

—. 1990c. Two derivations suffice: The role of syllabification in cognitive phonology. Technical Report AIP 116. Carnegie Mellon University. Pittsburgh, PA.

—. 1991. Exploiting syllable structure in a connectionist phonology model. Advances in neural information processing systems, ed. by R.P. Lippmann, J.E. Moody & D.S. Touretzky, 612-18. San Mateo, CA: Morgan Kaufmann Publishers.

Tupper, Paul. 2014. Exemplar dynamics models of the stability of phonological categories. Proceedings of the 26th Annual Meeting of the Cognitive Science Society.

—. 2016. Which learning algorithms can generalize identity-based rules to novel inputs? Proceedings of the 28th Annual Meeting of the Cognitive Science Society.

Tupper, Paul & Michael Fry. 2012. Sonority and syllabification in a connectionist network: An analysis of BrbrNet. The sonority controversy, ed. by S. Parker, 385-409.

Wayment, Adam. 2009. Assimilation as attraction: Computing distance, similarity, and locality in phonology. Doctoral dissertation: Johns Hopkins University.

Wedel, Andrew B. 2006. Exemplar models, evolution and language change. The Linguistic Review 23.247-74.

Wheeler, Deirdre & David Touretzky. 1993. A connectionist implementation of cognitive phonology. The Last Phonological Rule: Reflections on Constraints and Derivations, ed. by J. Goldsmith, 146-72. Chicago: University of Chicago Press.