Phonotactic learning without *a priori* constraints: Arabic root cooccurrence restrictions revisited

John Alderete[1], Paul Tupper[1], Stefan A. Frisch[2]
Simon Fraser University[1], University of South Florida[2]

1 Introduction

Much work in generative linguistics is nativist in the sense that the fundamental mechanisms for computing linguistic processes are assumed to be innate. In Optimality Theory (OT), for example, the building blocks of grammar, well-formedness constraints, are universal and innate ((Prince & Smolensky 1993/2004), (McCarthy & Prince 1999)). Cross-linguistic differences are accounted for by reranking these fixed universal constraints. While it is fairly certain that some aspects of language are innate in humans, it is also far from clear which aspects are innate and which simply evolve in the natural course of language development. Results from a host of different research paradigms have shown that many language processes can be learned directly from the statistical structure of experience ((Elman et al. 1996), (Spencer et al. 2009)), including nontrivial ones like dependencies between nonadjacent elements ((Gomez 2002), (Newport & Aslin 2004)). Perhaps at least some of the constraints of OT grammars can be learned from experience too.

In a sense, recent work in computational language learning in phonology anticipates this issue. Initial computational work in OT showed that, with a finite set of fixed constraints, complex linguistic systems can be learned within an OT architecture ((Tesar 1995), (Tesar & Smolensky 2000)). Related research paradigms, including the Gradual Learning Algorithm ((Boersma 1998), (Boersma & Hayes 2001)) and Harmonic Grammar ((Legendre et al. 1990), (Pater 2009)), modify how constraint-based grammars predict output forms, but they retain the assumption that the constraints themselves are given in advance of learning. More recently, however, (Hayes & Wilson 2008) call into question this assumption. In their theory, constraints can be induced from the data by search heuristics that select a small number of highly predictive constraints from a quasi-infinite constraint set. While this approach is used more as an inductive baseline to motivate the introduction of more abstract structures, it is notable in that it makes learning the constraints themselves a nontrivial part of learning.

We seek to continue this line of research by providing an additional mechanism of inducing constraints from data. In particular, we develop a connectionist architecture for learning phonotactic constraints. Below we motivate this cognitive architecture and apply it to the problem of learning root occurrence restrictions, or 'OCP effects', in Arabic. Arabic is chosen because large datasets exist, i.e., root lists and psycholinguistic experiments (Frisch et al.

2000), that enable strong tests of model performance. Also, Arabic exhibits graded phonotactic patterns that make it a good test case for any learning system designed to induce constraints. The principal result reported below is that the graded phonotactic patterns of Arabic consonant phonology can be learned as the gradual tuning of subsymbolic constraints in a connectionist network. Learning of OCP constraints in a connectionist network therefore presents a new way of inducing constraints from data.

The rest of this article is structured as follows. Section 2 provides a summary of the core empirical facts, namely the restrictions in Arabic against roots containing homorganic consonants. Section 3 provides some formal background, explaining what constraints are in connectionist networks. In section 4, a feedforward multi-layer network is developed for learning phonotactic restrictions, and it is shown to capture the core descriptive and psycholinguistic facts of Arabic consonant phonology. Section 5 develops a second connectionist network using recurrent connections to show that the learning results do not depend on the feedforward nature of the first network. The last section summarizes and makes a few concluding remarks based on our findings.

2   Root cooccurrence restrictions in Arabic

Arabic roots exhibit a phonological pattern in which there is a strong tendency against two adjacent consonants having the same place of articulation. This generalization was first clarified in (Greenberg 1950) and explored further in ((McCarthy 1988), (McCarthy 1994) and (Pierrehumbert 1993)) with different root lists. Table 1 below, from (Frisch et al. 2004), shows the consonant co-occurrence in trilateral roots (excluding glides), the most common type of root that has been the focus of most prior research. It organizes Arabic consonants into a set of homorganic natural classes typically assumed in prior work, following the autosegmental analysis of (McCarthy 1988). We refer to these classes below (excluding the uvulars) as 'same-place' classes, because they are not the same as the natural classes defined by major place features. As explained below, there are three separate coronal same-place classes, and uvulars are merged with both dorsal and pharyngeal classes. The rate of cooccurrence of two consonants in a root is quantified as a so-called O/E ratio, or the ratio of observed consonant pairs to the number of consonants that would be expected to occur by chance (Pierrehumbert 1993). The O/E ratios for sequences of adjacent consonants in a root, i.e., the first two or last two consonants, are shown in Table 1 from a dataset of 2674 triliteral verb roots compiled originally in (Pierrehumbert 1993) and based on the Hans Wehr Arabic-English Dictionary (Cowan 1979).

An O/E ratio of less than 1 indicates underrepresentation in the dataset, as shown in all the shaded cells for all same-place consonant pairs. Uvulars are also significantly underrepresented when they combine with either dorsals or pharyngeals. For this reason, uvulars are generally assumed to be in both same-place classes. While not as strong an effect, coronal stop + fricative pairs are also

underrepresented with an O/E of 0.52. Thus, after merging uvulars with dorsals and pharyngeals, there are six same-place classes in Arabic root phonotactics.

Table 1: Co-occurrence of adjacent consonants in Arabic triliteral roots

	Lab	Cor Stop	Cor Fric	Dorsal	Uvular	Phar	Cor Son
Labial [b f m]	0.00	1.37	1.31	1.15	1.35	1.17	1.18
Cor Stop [t d tˤ dˤ]		0.14	0.52	0.80	1.43	1.25	1.23
Cor Fric [θ ð s z sˤ zˤ ʃ]			0.04	1.16	1.41	1.26	1.21
Dorsal [k g q]				0.02	0.0	1.04	1.48
Uvular [χ ʁ]					0.00	0.07	1.39
Pharyngeal [ħ ʕ h ʔ]						0.06	1.26
Cor Son [l r n]							0.06

This restriction against same-place pairs is also found in non-adjacent consonants, e.g., the first and third consonant of a triliteral root, but the effect is not as strong ((Greenberg 1950), (Pierrehumbert 1993), (Frisch et al. 2004).

The above data shows that roots that contain two same-place consonants are in general prohibited. However, two identical consonants are commonly found in the second and third consonantal positions in triliteral roots, e.g., *madad* 'stretch'. Most prior work, and the table above, follow (McCarthy, 1986) in excluding roots with pairs of identical segments in counts of same-place consonant pairs because they assume an analysis in which the second and third consonants are derived in some sense (e.g., by autosegmental double-linking or reduplicative copying) from the same underlying consonant. So the two identical surface consonants do not actually constitute a consonant pair for the purpose of the restriction against homorganic consonants ((Gafos 1998), (Rose 2000), (Frisch et al. 2004)). We follow this work for the sake of concreteness, and exclude identical segments in C2C3 position from the set of patterns that our model is designed to account for.

Our simulations below are tested directly on how well they approximate Arabic speakers' responses to a wordlikeness experiment. In order to compare our results with this data, we sketch the principal questions and experimental design of (Frisch and Zawaydeh, 2001) so detailed comparisons can be made. In this study, 24 native speakers of Jordanian Arabic were given a set of nonsense words that contained triliteral roots. Subjects were asked to rate these words on a 7 point scale for the overall acceptability of the form, which was the dependent variable in all experiments. The larger finding was that the constraint against homorganic consonants, dubbed the 'OCP' for Obligatory Contour Principle, has a significant effect on subjects' ratings that cannot be attributed to certain lexical statistical effects or accidental gaps. Furthermore, subjects' ratings of these words fall on a gradient that correlates with featural similarity of the two consonants. The specific research questions, design, and results of each experiment are given below to allow for explicit comparisons in sections 4 and 5.

Experiment 1. Is the homorganic cooccurrence restriction (a.k.a., the OCP) psychologically real, and not just an effect of lexical statistics?

- independent variables: OCP violations, expected probability, neighborhood density
- results/conclusion: significant effect of OCP found on wordlikeness ratings, no other effects found and no interactions; OCP accounts for approximately 30% of subject variability

Experiment 2. Do subject ratings distinguish between systematic gaps (OCP violations) and accidental gaps (non-OCP violating, rare consonant combinations)?

- controlled variables: expected probability and neighborhood density
- variables balanced in stimuli set: bigram probability
- result/conclusion: OCP had a significant effect on wordlikeness ratings, accounting for approximately 21% of subject variability; so subjects distinguish between systematic and accidental gaps

Experiment 3. Do subject acceptability judgments exhibit different degrees of OCP violation that correlate with different degrees of featural similarity?

- variables balanced in the stimuli: expected probability, neighborhood density, and bigram probability
- independent variable: similarity of phonological features
- result/conclusion: similarity had a significant effect on wordlikeness rating (approximately 20% of subject variability); OCP is gradient

3 Constraints in connectionist grammars

Connectionist grammars are information processing models that take inputs and generate outputs, and so they can be compared with symbol-manipulating grammars as generative models. The well-known analysis of the English past tense in (McClelland & Rumelhart 1986), for example, generates past tense verbs from English present forms by computing activation patterns of inflected forms in a multi-layer node network. Connectionist networks, 'c-nets', have also been developed that capture the facts of traditional problems in phonology. For example, (Hare 1990) designed a sequential network to capture some of the core facts of Hungarian vowel harmony. Hare showed that by using a context layer, which in a sense remembers the structure of preceding elements (Jordan 1991), the c-net could explain the fact that the more similar/closer the target and trigger are, the stronger the assimilatory effect. Another example is the c-net developed in (Legendre et al. 2006) to account for the now classic OT analysis of Tashlhiyt Berber syllabification (see (Prince & Smolensky 1993/2004)). This c-net takes a sequence of segments as input, represented as input patterns of sonority classes, and the dynamics of a recurrent network assigns input segments to the proper syllable positions, peak and margin.

One can view the computation of activation patterns in a c-net as constraint satisfaction, but satisfaction of subsymbolic constraints rather than the symbolic constraints familiar from standard Optimality Theory ((Smolensky 1988), (Smolensky & Legendre 2006b)). Subsymbolic constraints are defined as the

connection weights between nodes. A subsymbolic constraint can be a single connection, or sets of connections within the larger node network. If the 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 node resembles the state of the sending node. If the connection is negative, the sending unit tries to put the receiving unit in the opposite state, so negative weights are satisfied by inducing opposite activation states downstream. Like constraint satisfaction in OT, not all constraints can be satisfied in c-nets. But as activity flows through the network, or cycles through recurrent networks, the activation values of individual units will change in a way that better satisfies these positive and negative constraints. This is the principle of Harmony Maximization of ((Legendre et al. 1990), (Smolensky & Legendre 2006a)).

Subsymbolic constraints are thus not global assessments of some property of a symbolic representation. They are the combined effect of microstructure links that can be scattered across the network. The problem of 'learning the constraints' can thus be characterized more precisely as a problem of learning the correct configuration of connection weights. We develop this approach to learning constraints from data by building two c-nets that capture OCP-Place effects.

4   A multi-layer connectionist network for learning phonotactics

Our first c-net is modeled after multi-layer networks that take linguistic forms as input and output a single score, as in the c-net developed in (Ramsey et al., 1990) to simulate learning of semantic truth values (i.e., '0' or '1') for syntactic phrase structure trees. Our multi-layer network, dubbed the Assessor Network (AN), is a deterministic feedforward network that accepts as input a triliteral root and returns an acceptability rating, i.e., a value between -1 and 1. The input to the AN, and the Recurrent Network described below in section 4, is a sequence of three segments, where each segment is a string of 17 values, either -1, 0, or 1, corresponding to the feature specifications assumed in (Frisch et al., 2004). Each triliteral root is thus a distributed representation of the featural make-up of the three consonant root, expressed as a vector of 3x17=51 elements. The AN is a three layer network with this input layer, a hidden layer of a certain number of nodes (1, 2, 5 or 10, which was varied to test the model's performance), and the output layer constituted by a single node that yields the acceptability rating. The output of the AN, or the activation state of the final output node, is comparable to the relativized acceptability score of Harmonic Grammar (Coetzee & Pater 2008) and the maxent values of MaxEnt Grammar (Hayes & Wilson 2008).

All input nodes are connected to all hidden layer nodes and all hidden layer nodes are connected to the output node. For each node in the hidden and output layer, the activation is given by:

(1) $a_i = \sigma \left(\sum_j w_{ij} a_j + b_i \right)$

where w_{ij} is the strength of the connection from node j to node i, and a_j is the activation of node j, and b_i is the bias on node i. σ is the sigmoid logistic function commonly used in connectionist modeling.

The AN only makes sensible classifications of the data when it has been trained. The AN was trained in a protocol that attempts to approximate the type of exposure that Arabic speakers confront in learning. In particular, a simplified production system was created that takes actual Arabic roots as input, and attempts to return an identical root as output. This system is noisy, however, in the sense that random variations in the activations patterns of output nodes can produce non-identical roots by replacing some or all consonants. These 'errors' are used in learning, as well as correctly reproduced output forms. In particular, if the production system gives the AN a faithfully reproduced form, all connection weights and biases in the AN are adjusted such that the AN gets closer to outputting a '1' for that form. If instead the production system gives the AN a form that differs from the actual Arabic word it used as input, the form supplied to the AN is classified as an error and all connection weights and biases in the AN are adjusted so that it outputs a '-1' for this form. In this way, the training is realistic in the sense that it is composed exclusively of actual Arabic roots or errors based on actual roots. The training regime is thus comparable to the role of production in many constraint-based learning systems (e.g., (Tesar & Smolensky 2000), (Tesar 2004)) that also use production errors as evidence in learning.

The specific training protocol for the AN was as follows. All weights and biases were initialized with small random values. Training took place over a run of 10^7 epochs, where each epoch involved presenting the AN a single output of the production system. The actual roots the AN was trained on, via the simplified production system, were a set of 3,489 triliteral roots from (Buckwalter 1997). Quadraliteral and triliterals with final geminates were excluded because they involve problems that are orthogonal to our study. The values of w_{ij} and b_i in (1) above were then computed using backpropagation (Rumelhart & McClelland 1986): these values were adjusted such that the AN gets closer to producing a '-1' for error forms and '+1' for target forms, as explained above.

We first give an overview of the network's performance by showing how it classifies all possible roots. Figure 1 below illustrates the rating results for an Assessor Network with five hidden layer units, given input trained on the Buckwalter root list. The first observation to make is that, while the ratings for the actual roots (a) overlap with the ratings for all possible roots (b) ($n = 28^3 = 21,952$), the ratings for the middle 50% of the actual words is centered around the top of the middle 50% of the ratings for all possible roots (compare the first and second boxplots), indicating that network has learned to assign higher scores to the words that it has been exposed to. Second, the opposition between the third (OCP compliant roots (c), from Frisch and Zawaydeh's Experiment 1) and fourth boxplots (OCP violating roots (d), same source) shows that the network has effectively learned the OCP. The middle 50% of the ratings for the OCP compliant roots is well above the middle 50% of the ratings for the OCP violating

roots. Thus, viewed globally, the Assessor has a strong tendency to rank roots that violate the OCP lower than those that do not.
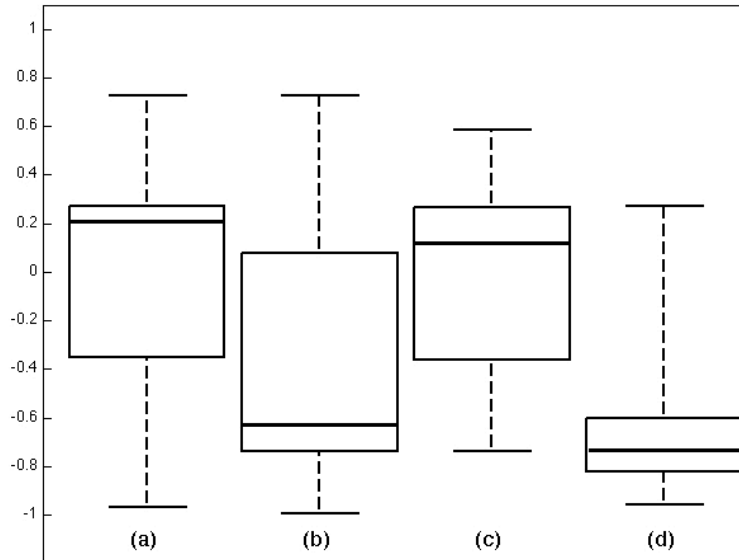

Figure 1: Acceptability scores for one trial of Assessor Network with five hidden nodes. Box plots indicate minimum, first quartile, median, third quartile and maximum scores for (a) all attested roots, (b) all possible roots, (c) OCP compliant roots, (d) OCP violating roots.

The Assessor Network assigns acceptability scores to nonactual roots, so its assessment of nonsense words can be compared directly to the native speakers' judgments of nonsense words. To do this, we conducted the same tests from (Frisch & Zawaydeh 2001), but substituted Assessor Network acceptability ratings for their wordlikeness ratings. The hidden layer of the Assessor Network can have any number of nodes, but we have investigated learning with hidden layers of between 1 and 10 hidden layer units and found that a range between 2 and 5 units produces effects parallel to the judgment data. Table 2 gives the results of a 5 unit hidden layer on three separate learning trials. All effects with significance at $p < .05$ are reported with the percentage of the variation accounted for by this effect. Under the Experiment 1 column, which used most of the stimuli, the correlation coefficients between AN outputs and Frisch and Zawaydeh's wordlikeness data (mean rating) is given as a gross measure of the correlation between the two datasets.

The results are qualitatively parallel to the Frisch and Zawaydeh's findings. In particular, in Frisch and Zawaydeh's experiment 1, OCP violation was the only statistically significant factor on wordlikeness and it still had a significant effect when bigram probability was controlled for in experiment 2. The results shown in Table 2 are therefore consistent with all these experimental findings, as OCP violation was the only significant factor in experiments 1 and 2, and similarity

was a significant factor in two of the three trials of experiment 3. We note that the percentage of the acceptability explained by the OCP is slightly higher than with the judgement data in experiments 1 and 2, but we believe that a perfect match of the two datasets is not required to demonstrate the learning of OCP-Place constraints. A c-net model and learning protocol could be constructed to produce a better quantitative match with the experimental data through manipulation of model parameters and additional training. The important finding is therefore that a relatively simple set of parameters reproduces all of the statistically significant generalizations in the behavioral data.

Table 2: Significant effects on acceptability from factors in Frisch & Zawaydeh 2001 experiments; cells show factor, percentage explained, and for experiment 1, correlation with the wordlikeness judgement data

	Experiment 1, $p<0.001$	Experiment 2, $p<0.001$	Experiment 3, $p<0.05$
Trial 1	OCP 44%; r=0.37	OCP 47%	similarity 5% (not sig.)
Trial 2	OCP 47%; r=0.48	OCP 43%	similarity 9%
Trial 3	OCP 48%, r=0.40	OCP 31%	similarity 17%

One of the common criticisms of connectionist networks is that it is difficult to understand the nature of the functions they compute because their subsymbolic structure is not easily accessible to analysis. How do we know that our c-net is working in a way parallel to OCP-Place constraints? One way to extract rule-like symbolic behavior from the rich dataset provided by our c-net is to use Classification and Regression Trees (CART). CART analysis is a commonly used statistical technique for imposing categorical structure on large and 'messy' datasets. We applied the following method to produce CART trees for the five-node network described above. For each hidden node in the trained network, we applied the `classregtree` function of Matlab in order to produce decision trees that would predict the output from the input. The inputs for CART analysis were the 21,952 possible triliteral roots (i.e., both actual and nonactual roots). The specific variables for these inputs were just the 51 feature specifications used by the c-net to describe each triliteral root (i.e., distributed representations for 3 segments × 17 features). The output variables were the activation values for a particular hidden node, rounded to -1 or 1, a '1' meaning that the node judges the form favorably, '-1' unfavorably.

The algorithm begins by identifying the single input variable, i.e., a phonological feature in a particular position, that does best at predicting the outputs. The data is then partitioned into two sets based on the value of this input feature. This procedure is then repeated recursively on each of the two sets, selecting a new input variable (=another feature in a C slot) to partition each set. We set the algorithm to stop when a set has fewer than 1,000 roots or else when all the output variables in a set are identical. At this point each terminal node is labeled either a '-1' or a '1', depending on which output variable predominates in the node. Once this is complete, the tree can be used to predict the output for a

given input by descending the tree according to the input variables and then using the label for the resulting terminal node.

We applied this procedure of generating CARTs for all five hidden layer nodes, for three different trials. With one exception, each of the hidden nodes implements an approximation of one of the six OCP-Place coocurrence restrictions active in Arabic, i.e., OCP-labial, OCP-coronal/sonorant, OCP-coronal/fricative, OCP-dorsal, and OCP-pharyngeal (the latter two overlap with uvulars, as expected). In other words, in virtually all cases, the hidden layer nodes compute symbolic-like OCP constraints. As an example, Fig. 2 shows the CART tree for the fifth hidden layer node of the 1st trial. This node approximates the OCP for [pharyngeal] specification.
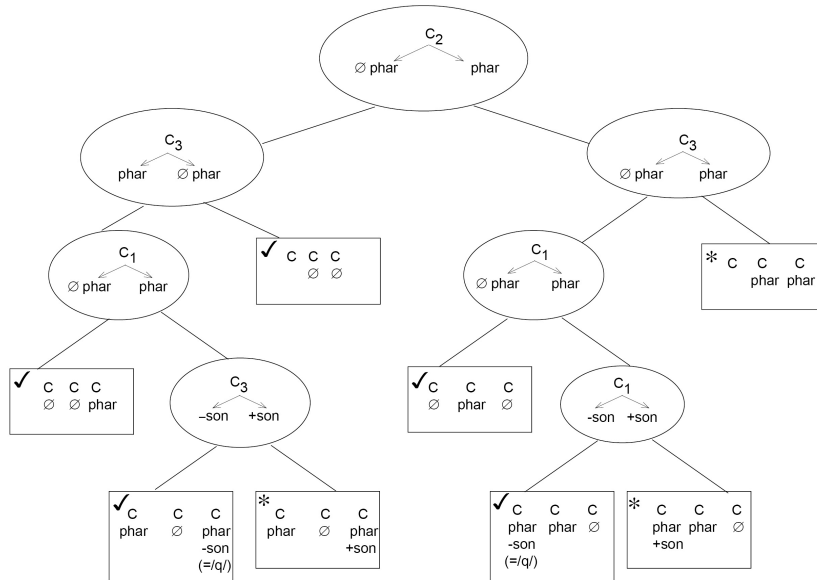
Figure 2: CART visualization of hidden layer node approximating OCP-pharyngeal. Circled nodes represent decisions about feature realization for a specified consonant, and boxed nodes represent predominant acceptable (checked) and marginal (starred) triliterals.

Two observations can be made about the CART visualization above for the OCP-pharyngeal node, and these observations are typical of the rest of the CARTs. First, this particular hidden layer node does a reasonably good job of predicting the nonoccurrence of triliterals that contain two pharyngeals. As we descend down the tree in Fig. 3, all triliterals (the boxed terminal nodes) that do not have two [phar] specifications (shown with at least two $\varnothing$) are allowed, and all but two of the schematic roots that have two [phar] are starred. These two cases involve the segment /q/ in either C1 or C3. Recall that, to be consistent with prior work, all uvulars, including /q/, have a [pharyngeal] specification. The apparent exceptions, however, reveal an important descriptive generalization in the dataset. While roots with two pharyngeals have low O/E values, most of the exceptions to OCP-pharyngeal, which again includes all uvulars, involve roots that contain /q/. There are 132 roots that contain /q/ and another pharyngeal consonant in the

Buckwalter corpus, including 15 roots fitting the exempted pattern /q + Pharyngeal + C/ and 28 matching the pattern /Pharyngeal + C + q/. This fact is why /q/ is traditionally grouped with velars in descriptive statements of consonant cooccurrence (Greenberg 1950). A more systematic study of exceptions to OCP-Place in Arabic is carried out in (Alderete et al. To Appear) and it is shown that while some exceptions evidenced by a large number of examples can be detected in a c-net, many more subtle patterns are not. To summarize the larger points, the hidden nodes are capable of approximating the functions of symbolic constraints, even when segment-level exceptions exist.

5   A recurrent connectionist network for learning phonotactics

A different approach to phonotactic learning is to encode phonotactic constraints in a model that attempts to faithfully reproduce the input. We use an architecture similar to that of (McClelland & Rumelhart 1985), namely a recurrent network, to illustrate how such a network can be used to generate acceptability values. We develop this alternative to demonstrate that success in learning phonotactics does not depend on the specific assumptions of the multi-layer network.

In the Recurrent Network (RN), a triliteral root is again represented as a distributed representation of a sequence of three consonants, with one slight adjustment. For any phonological feature, a positive value is +1 in the activation vector, but both redundant and negative values are -1. The network consists of a single layer of 3x17=51 units. The network is fed by an external input, and the output of the network is the activation vector of the network at equilibrium. Each node in the network receives an external input via connections that are not adjusted, and each node of a segment X is connected to all other nodes that encode the two other segments besides segment X. For example, the node corresponding to [nasal] for the first consonant connects to all the other nodes that encode the second and third consonants, but not to any of the other nodes encoding the first consonant. Because the network is recurrent, we require an update rule, which is defined in the following way. Let the activation of the ith node be a_i, the external output to the ith node be ext_i, the strength of the connection between node i and node j be W_{ij}, and equilibrium activations be a_i^*.

(2) Update rule for Recurrent Network
    We have
$$da_i / dt = \sigma (\textstyle\sum_j W_{ij} a_j + ext_i) - a_i$$
    Equilibrium activations satisfy:
$$a_i^* = \sigma (\textstyle\sum_j W_{ij} a_j^* + ext_i)$$

The goal in training the RN is to find W_{ij} such that $\sum_j W_{ij} a_j^* = ext_i$, or, such that the internal input to each node matches the external input. The Delta rule, a standard learning rule in c-nets (McLeod et al., 1998), is used to do this. In each epoch, where the number of epoch in training $= 10^5$, a random attested word is selected and input to the system through ext_i. The system is then allowed to

equilibriate using the current values of W_{ij}. Then the weights W_{ij} are adjusted so that $\sum_j W_{ij} a_j^*$ more closely approximates ext_i.

The mature network can then be used as an autoassociator, i.e., a system that attempts to faithfully map inputs onto identical outputs. While our RN is not a very effective autoassociator, it can be employed to measure acceptability in the following manner. The more the RN is able to reproduce an input as the external output, the more acceptable that input is. Thus, we define acceptability not in terms of an output score, as with the AN, but as a measure of the length of the distance between the external input and output vectors.

(3) Acceptability (RN): acceptability = - ‖ W a - ext ‖

We applied this model and this definition of acceptability again to the Arabic data. The results of the multi-factor tests in experiments 1-3 are given in Table 3 below, for three separate trials, again showing all effects that reach significance at level $p<0.05$. The RN gives a slightly poorer approximation of this pattern in that neighborhood density has a significant (though extremely small) effect, and the OCP accounts for a greater percentage of the variation. However, this characterization of the OCP is consistent with Frisch and Zawahdeh's results in that the OCP is the most important effect on the acceptabilty.

Table 3: Significant effects on acceptability in Recurrent Network from factors in Frisch & Zawaydeh 2001 experiments; cells show factor, percentage explained, and for experiment 1, correlation with the wordlikeness judgement data

	Experiment 1	Experiment 2	Experiment 3
Trial 1	OCP 58%, density 3%; r=0.49	OCP 49%	similarity 14%
Trial 2	OCP 58%, density 3%; r=0.48	OCP 51%	similarity 13%
Trial 3	OCP 58%, density 3%; r=0.48	OCP 50%	similarity 13%

These learning results are consistent with the results of the feed-forward Assessor Network. Therefore, it cannot be the case that the connectionist approach employed here requires the assumptions specific to that model.

5 Discussion

The simulation results above demonstrate a new way of learning gradient phonotactic generalizations that does not require the prior existence of well-formedness constraints. Our connectionist learners are able to learn complex distributional patterns in Arabic with rather simple network architectures and training regimes: a multi-layer network trained via backpropagation, and a basic recurrent network using gradient descent.

This result sets our connectionist learning system apart from many contemporary approaches to learning phonotactics. As summarized above, most constraint-ranking algorithms can find the correct ranking of constraints, given the right data and a fixed set of constraints ((Tesar 2004), (Prince & Tesar 2004); (Boersma 1998), (Boersma & Hayes 2001); (Pater 2009)). But these

investigations do not make learning the constraints themselves part of the learning problem. Furthermore, it has been conjectured that there is a close parallelism between the macro-structure of OT grammars and the micro-structure of connectionist networks (Smolensky & Legendre 2006b). However, as stated at the outset of this important work (chapter 1, section 2), the connectionist implementations of OT constraint systems have not yet shown how behavior resembling symbolic constraint interaction can be learned at this level of explanation. Our contribution to Smolensky and Legendre's research paradigm is thus to show that at least one kind of phonological pattern, place-based cooccurrence restrictions, can be learned at the micro-structure level.

The finding that constraints can be learned from data supports a comparison of our model to the MaxEnt phonotactic learning paradigm. Both approaches use principles of statistical learning to search a vast constraint space and provide the constraints that give a good approximation of the target grammar. As such, both rely heavily on a suitably large and representative data sample. Another similarity is that both approaches produce inductive baselines, or simple systems derived from data. These systems have limits, however, like the generalizations involving suprasegmentals that Hayes and Wilson document in their study. We have explored select problems in Arabic consonant phonology that extend this core phonotactic system and have likewise found certain facts that will require additional assumptions in our c-nets. For example, after extensive parameter switching, we have not found a means for our multi-layer network to learn the exceptions to the OCP in Arabic involving final geminates, i.e., roots in which the last two consonants are identical. Like Hayes and Wilson, we think that these facts demonstrate a role for constituent structure (syllables, doubly-linked segments, etc), which can be implemented in a c-net with tensor product representations (Smolensky & Legendre 2006a).

How does our c-net model differ from the MaxEnt approach generally? One empirical point is that c-nets have greater descriptive capacity than MaxEnt grammars. The c-net constraint space is uncountably infinite and so it is richer, whereas Hayes and Wilson's model makes certain modest, but restrictive, assumptions about the set of possible constraints. The math supplement to this article (available from the authors' webpages) provides a proof that compares the underlining functions of MaxEnt and c-net grammars and shows that c-nets have greater descriptive capacity. It is a non-trivial empirical issue, however, if these differences matter for describing language. A typical constraint space in MaxEnt learning is still rather large, and c-net grammars are constrained significantly by training, which must be considered.

We think that one key aspect of our approach that sets it apart from other paradigms, however, is the potential for integration with psycholinguistic models of production and perception. In the spreading interactive model of (Dell 1986), for example, selection of a word in the mental lexicon is simulated as the spreading of activation through a lexical network of many linguistic layers (i.e., morphological and phonological constituents). While there are important features

of this model that differ from our c-net, e.g., bidirectional spreading and rich linguistic representations, an important point is that lexical selection is the result of activation spreading, an output pattern predicted by parallel processing of micro-elements. Another important model is the TRACE theory of word recognition (McClelland & Elman 1986), which uses spreading activation and parallel distributed processing to work in the other direction, predicting word forms from phonetic attributes. These models have been tremendously influential and provided a set of assumptions shared with many contemporary theories of speech production and perception. We believe that the parallel-distributed processing principles at the core of these two influential theories and our c-net may allow for a more natural integration of the functions of our c-net within these models. Furthermore, this integration is highly desirable in the case of dissimilatory phenomena, like Arabic OCP-Place constraints. As shown in detail in (Frisch 1996), (Frisch et al. 2004), (Frisch 2004), and (Martin 2007), many of the properties of dissimilatory patterns can be explained as the long term diachronic effects of constraints on speech production and perception.

References

Alderete, John, Paul Tupper & Stefan A. Frisch. To Appear. A connectionist approach to phonological constraint induction: OCP[Place] in Arabic. Language Sciences.

Boersma, Paul. 1998. Functional Phonology The Hague: Holland Academic Graphics.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. Linguistic Inquiry 32.45-86.

Buckwalter, Tim. 1997. The triliteral and quadriliteral roots of Arabic. URL: http://www.angelfire.com/tx4/lisan/roots1.htm.

Coetzee, Andries & Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-ccurrence in Muna and Arabic. Natural Language and Linguistic Theory 84.289-337.

Cowan, J. Milton (ed.). 1979. Hans Wehr: A dictionary of Modern Written Arabic Wiesbaden, Germany: Otto Harrasowitz.

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

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.

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

Frisch, Stefan A. 2004. Language processing and segmental OCP effects. Phonetically-based phonology, ed. by B. Hayes, R. Kirchner & D. Steriade, 346-71. Cambridge: Cambridge University Press.

Frisch, Stefan A., Nathan R Large & David S. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. Journal of memory and language 42.481-96.

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.

Gafos, Adamantios. 1998. Eliminating long-distance consonantal spreading. Natural Language and Linguistic Theory 16.2.223-78.

Gomez, Rebecca L. 2002. Variability and detection of invariant structure. Psychological Science 13.413-36.

Greenberg, Joseph. 1950. The patterning of root morphemes in Semitic. Word 6.162-81.

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.

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

Jordan, Michael I. 1991. Serial order: A parallel distributed processing approach. Advances in connectionist theory: Speech, ed. by J. Elman & D. Rumelhard, 214-49. Hillsdale, NJ: Erlbaum.

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.

Martin, Andy. 2007. The evolving lexicon: University of California, Los Angeles.

McCarthy, John J. 1988. Feature geometry and dependency: A review. Phonetica 43.84-108.

—. 1994. The phonetics and phonology of Semitic pharyngeals. Papers in Laboratory Phonology III, ed. by P.A. Keating, 191-233. Cambridge: Cambridge University Press.

McCarthy, John J. & Alan Prince. 1999. Faithfulness and identity in Prosodic Morphology. The prosody-morphology interface, ed. by R. Kager, H.v.d. Hulst & W. Zonneveld, 218-309. Cambridge: Cambridge University Press.

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.

Newport, Elissa & Richard Aslin. 2004. Learning at a distance: I. Statistical learning of non-adjacent dependencies. Cognitive Psychology 48.127-62.

Pater, Joe. 2009. Weighted constraints in generative linguistics. Cognitive Science 33.999-1035.

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

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

Prince, Alan & Bruce Tesar. 2004. Learning phonotactic distributions. Fixing priorities: Constraints in phonological acquisition, ed. by R. Kager & J. Pater, 245-91. Cambridge: Cambridge University Press.

Rose, Sharon. 2000. Rethinking geminates, long-distance geminates and the OCP. Linguistic Inquiry 31.85-122.

Rumelhart, David & James L. McClelland. 1986. Parallel distributed processing: Explorations in the microstructure of cognition. Volumes 1-2. Cambridge, MA: MIT Press.

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

Smolensky, Paul & Géraldine Legendre. 2006a. Formalizing the principles II: Optimization and grammar. The harmonic mind, From neural computation to optimality-theoretic grammar. Vol 1: Cognitive architecture, ed. by P. Smolensky & G. Legendre. Cambridge, MA: The MIT Press.

—. 2006b. The harmonic mind. From neural computation to optimality theoretic grammar Cambridge, MA: The MIT Press.

Spencer, John P., Larissa K. Samuelson, Mark S. Blumberg, Bob McMurray, Scott R. Robinson & Bruce J. Tomblin. 2009. Seeing the world through a third eye: Developmental Systems Theory looks beyond the nativist-empiricist debate. Child Development Perspectives 3.103-05.

Tesar, Bruce. 1995. Computational Optimality Theory. Boulder: University of Colorado Doctoral dissertation.

—. 2004. Using inconsistency detection to overcome structural ambiguity in language learning. Linguistic Inquiry 35.219-53.

Tesar, Bruce & Paul Smolensky. 2000. Learnability in Optimality Theory Cambridge, MA: MIT Press.