

## Phonotactic learning without *a priori* constraints: A connectionist analysis of Arabic cooccurrence restrictions \*

John Alderete\*, Paul Tupper\*, Stefan A. Frisch†

\*Simon Fraser University, †University of South Florida

**Abstract.** In this article, we develop a connectionist model of learning phonotactics and apply it to the problem of learning root cooccurrence restrictions in Arabic. Two types of connectionist networks are developed: a multilayer network with a hidden layer and a single layer network with recurrent connections. They are both shown to classify Arabic words and nonwords in ways that are qualitatively parallel to psycholinguistic studies of Arabic. In these networks, units and connections act like soft constraints in the computation of acceptability scores. Because these constraints are malleable and can change gradually over time, the networks learn phonotactic generalizations without requiring the prior existence of the exact constraints responsible for phonotactics, a fact that sets this model apart from many phonotactic learners.

### 1. Introduction

A recent focus in research on constraint-based theories of language learning is phonotactics, or a speaker's knowledge of sound-based distributional patterns ((Tesar and Smolensky, 2000), (Boersma, 1998), (Boersma and Hayes, 2001), (Prince and Tesar, 2004), (Hayes and Wilson, 2008), (Pater, 2009)). These theories differ in many important respects, both in the nature of the constraint system to be learned and crucial assumptions about how phonotactic learning is achieved. However, a fundamental assumption shared by most contemporary theories is that the constraints for characterizing phonotactic grammars are available in advance to the learner. To learn the phonotactics of onsets, for example, the learner does not need to learn the constraints referring to syllable initial segments because they are given in advance. The learner only needs to determine the role of these constraints in grammar.

This commitment to *a priori* constraints leads to two problems for learnability theory. The 'expressiveness problem' arises because it is not known in advance that the system of constraints is capable of characterizing a given phonotactic system. This is simply because a finite set of *a priori* constraints has a finite descriptive capacity, but the set of attested phonotactic systems is not known. If it turns out that an attested phonotactic system cannot be described with the given constraints, a new set of constraints is required.

Success with the expressiveness problem is a trade off with success with a second problem. The 'search problem' is that the learner must select the correct constraint system for the data from a universe of possible constraint systems. In Optimality Theory (Prince and Smolensky, 1993/2004), the search space is significant because the space of possible constraints systems is at least as large as the factorial typology of *a priori* constraints. Recently, in part in response to the expressiveness problem, (Hayes and Wilson, 2008) have extended the search space even further by greatly enlarging the set of *a priori* constraints in phonotactic learning. They envision a search space composed of an initial set of logically possible constraints motivated by sensible assumptions in generative phonology, e.g., the available phonological features, feature

---

\*We have benefitted from the comments and questions raised by the poster session participants at the 2009 Cognitive Science Conference. This article is supported in part by a standard SSHRC research grant 410-2005-1175, and NSERC discovery grant awarded to Paul Tupper. Usual disclaimers.

underspecification, and restrictions on the terms and logical structure of well-formedness constraints. The search problem is then broken down into a selection of the operative constraints from this initial set, and the assignment of the relative importance of these selected constraints in the grammar. The significance of this approach is that the vastness of this initial search space makes the selection of the much smaller number of operative constraints an important part of learning. We approach phonotactic learning in a similar way in this article, but seek to develop alternative ways of addressing the search problem that relate to other learning problems in cognitive science.

We propose to address these two problems jointly by employing some well-known representational assumptions and learning methods in a connectionist network, or a parallel distributed processing (PDP) network (Rumelhart et al., 1986b). In particular, we model a phonotactic system as a web of neuron-like units that generates acceptability judgements of language forms using well-known algorithms in numerical computation. Following the lead of early connectionist approaches to concept learning ((McClelland, 1981), (McClelland and Rumelhart, 1985)), we do not envision an initial search space that encodes all of the generalizations relevant to phonotactic grammars. Instead, the units of a connectionist network ('c-net' henceforth) act as soft constraints that, at a particular stage in learning, have a much more limited range of computable functions. Rather than positing a vast space of fixed constraints, connectionist models posit a small number of malleable constraints that can be adjusted gradually in response to input data in order to move the learner towards a better approximation of the target grammar. Our goal is to show this assumption supports learning of a complex phonotactic system without requiring the operative constraints to exist *a priori*, making a clear contribution to the debates surrounding learnability in phonology and other areas of linguistic cognitive science.

The argument we develop below may appear to be a radical departure from constraint-based approaches to phonotactics. After all, with an almost exclusive use of numerical computation, the structure of the phonotactic system and the actual steps in learning appear to be quite different. This impression is quite mistaken, however, because our approach builds on several recent research trends in formal language learning. Like most contemporary phonotactic learners, connectionist learning is constraint-based because units (and connections) act like soft constraints in the larger computation of output forms ((McLeod et al., 1998), (Smolensky, 1988)). Second, PDP models typically employ error-corrective learning rules, like the Delta rule that adjusts connection weights to minimize the difference between the desired and error forms. Virtually all constraint-based learning models are error corrective and some even use the same learning procedures developed in the connectionist literature ((Coetzee and Pater, 2008), (Pater, 2009)). Third, while this is a learning parameter in some models, c-net learning typically involves gradual changes to the network by making small changes to connection weights. This assumption is consistent with many phonotactic learners that likewise make small changes in learning, e.g., (Boersma and Hayes, 2001). Finally, an important test of phonotactic learners is their ability to approximate gradient phonotactic patterns. PDP models share the capacity to compute such patterns with recent models, e.g., (Coetzee and Pater, 2008) and (Hayes and Wilson, 2008), because they share the assumption that phonotactic assessments can be modeled as the weighted sum of individual constraints, a point that we demonstrate in detail below.

We develop our argument by testing a connectionist learner on the phonotactics of Arabic root cooccurrence restrictions. This problem is chosen because it has already been used to test several

models of gradient phonotactics ((Anttila, 2008), (Frisch et al., 2004), (Coetzee and Pater, 2008)). Also, the Arabic dataset allows us to assess the ability to capture gradient phonotactics in a very rigorous way because the system of root cooccurrence restrictions in Arabic is one of the few cases in which phonotactic generalizations are measured in detail with behavioral data (Frisch and Zawaydeh, 2001).

The next section illustrates the core data from Arabic that are analyzed in later sections, including a summary of the three experiments of (Frisch and Zawaydeh, 2001) that documents the behavioral data we attempt to model. The next two sections present two distinct c-nets that are able to independently learn Arabic phonotactic restrictions: a multilayer feedforward network with a hidden layer (section 3) and a single layer recurrent network (section 4). The results of these sections show that a standard set of tools in connection science can model an important subsystem of Arabic phonotactics, generalize to novel forms, and capture gradient generalizations. The last section compares the results of these two c-nets with the other models mentioned above, and discusses the issues that this comparison raises.

## 2. Arabic consonant cooccurrence restrictions

Arabic verbs exhibit root-and-pattern morphology in which verb words are formed by combining discontinuous strings of consonants and discontinuous strings of vowels. The string of consonants, typically three in number, constitute the ‘root’, and they are interspersed with vowel strings in specific CV patterns to mark grammatical distinctions like tense and aspect. Arabic verb roots have attracted a lot of interest because they exhibit a phonotactic pattern in which roots have a strong tendency against having more than one consonant with the same place of articulation ((Greenberg, 1950), (McCarthy, 1988), (McCarthy, 1994), (Pierrehumbert, 1993), (Frisch et al., 2004)). A useful way of describing this consonantal cooccurrence restriction is with a ratio of observed/expected (O/E) pairs, or the number of attested pairs of same-place consonants over the number of same-place pairs that would be expected by chance (Pierrehumbert, 1993). The data below show the O/E ratios for all Arabic consonants for a dataset of 2674 verb roots compiled originally in (Pierrehumbert, 1993), based on the Hans Wehr Arabic-English Dictionary (Cowan, 1979). The groups of same-place consonants in Arabic are shown in the column on the left, which uses IPA transcription.

(1) Co-occurrence of adjacent consonants in Arabic (Frisch et al., 2004)

	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 t <sup>c</sup> d <sup>c</sup> ]		0.14	0.52	0.80	1.43	1.25	1.23
Cor Fric [ θ ð s z s <sup>c</sup> z <sup>c</sup> f ]			0.04	1.16	1.41	1.26	1.21
Dorsal [ k g q ]				0.02	0.07	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

An O/E ratio less than 1 indicates underrepresentation in the dataset, as shown in all the shaded cells above for same place consonants. The above data shows that adjacent consonants in verb roots have a strong tendency to be different in place of articulation. This tendency is also found in non-adjacent consonants, i.e., the first and third consonant of a trilateral root, but the effect is not as strong ((Frisch et al., 2004), (Pierrehumbert, 1993)). A final point is that while roots that

contain two homorganic consonants are in general prohibited, two identical consonants are allowed in the second and third consonantal positions, e.g., *madad* ‘stretch’. Most prior work, following (McCarthy, 1986), excludes these cases because they assume an analysis in which the second and third consonants in fact derive from the same underlying consonant, so the two identical surface consonants actually do not constitute a consonant pair for the purpose of the cooccurrence restrictions. We make the same assumption and therefore exclude these cases too in our simulations below.<sup>1</sup>

The consonant co-occurrence patterns above constitute gradient phonotactic patterns because they are not absolute or categorical restrictions on same-place consonant pairs. Rather, they are statistical trends that appear to fall on a gradient that relates to the overall similarity in sound structure of the two members of the pair (Frisch et al., 2004). It is gradient phonotactics that accounts, for example, for the fact that coronal stops and coronal fricatives, while not in the same homorganic groups shown above, are significantly underrepresented because they are close enough on the similarity scale to induce a phonotactic restriction (see (Padgett, 1995) for an alternative account that does not make use of gradient similarity). This generalization is a fact of the lexical database used in these studies, and it has also been documented with behavioral data in (Frisch and Zawaydeh, 2001), a study that we return to in detail below.

Gradient phonotactics as observed in Arabic roots have been used to argue for new models of linguistic competence that make gradient rather than categorical characterizations of the overall acceptability of a form (see e.g., (Frisch and Zawaydeh, 2001), (Pierrehumbert, 2003), (Treiman et al., 2000)). A nontrivial question in assessing these models is how to test their predictions for both actual words and nonsense words, a question that turns on how one actually defines phonotactic generalizations. One can think of phonotactic generalizations as descriptive generalizations that capture linguistic patterns in extant words, or one can think of them as cognitive constraints on productivity in a native speaker’s mind. Though there are cases in which phonotactic grammars that are derived from lexical O/E patterns correlate with experimental acceptability scores (see for example the discussion of English onsets in (Hayes and Wilson, 2008)), the two measures are not equivalent and so it is not clear that the correlation will hold in all cases. If one is interested in cognitive constraints, in this case the psychologically real constraints on consonant pairs, then the correct test of the phonotactic learner is to compare its results directly with the behavioral data documenting the constraint. A virtue of this approach is that it makes possible a strong test of the native speaker’s ability to generalize to new types. In the case of Arabic, psycholinguistic experimentation has documented important differences among classes of nonexistent bigrams, a fact that can be tested using behavioral measures. If on the other hand lexical statistics are used as the true characterization of a phonotactic generalization, it’s not possible to measure a model’s performance on nonwords because all nonexistent bigrams have a O/E of zero.

---

<sup>1</sup> Because our networks below are trained on the (Pierrehumbert, 1993) corpus of trilaterals, they do not systematically differentiate between roots of the form *s-m-m* and *\*s-s-m*. We therefore cannot test if our approach reproduces the experimental results of ((Berent et al., 2001), (Berent et al., 2002)), which show how similar consonant cooccurrence restrictions in Hebrew phonotactics can be generalized by native speakers to words that contain non-Hebrew sounds. While this is a limit of the current corpus materials for Arabic, we also note that the arguments of ((Marcus, 1998), (Marcus, 2001)) for why certain connectionist models are unable to generalize outside the training space do not apply in the case studied here. In particular, computing acceptability judgements does not require computing a function that maps any input onto a single unique output (an ‘uqotom’ in Marcus’ terminology).

Since our simulations below are tested directly on how well they approximate subjects' responses to the wordlikeness experiments, 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 trilateral roots, manipulated as shown below. 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,<sup>2</sup> 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 did 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 the next two sections.

**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. A multilayer connectionist network**

#### **3.1 Network architecture**

Our first c-net is modeled after multilayer 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 the

---

<sup>2</sup> The terms used in the experiment sketches below are defined as follows. The 'OCP' is a constraint used in formal phonology that states 'adjacent identical autosegments are prohibited'. In autosegmental phonology (Goldsmith, 1990), this has the effect of prohibiting two segments in the same domain from having the same specification for some feature (=autosegment), unless there is an intervening segment that has a different specification for that feature. The OCP is commonly used to account for restrictions on homorganic consonants, because they have the same specification for place features. 'Expected probability' (abbreviated exp.prob.) is the probability of independent combinations of the segments that make up a form, given their frequency in the lexicon; it is the product of monogram probabilities when more than one segment is considered. 'Neighborhood density' (density) in trilateral roots is the number of existing roots that share two of the three consonants in the appropriate serial positions. Similarity of two phonological segments is defined in (Frisch et al., 2004) as the number of shared natural classes over the sum of shared natural classes plus the non-shared natural classes.

correct truth values for syntactic phrase structure trees. This network, dubbed the Assessor Network (AN), is a deterministic feedforward network that accepts as input a trilateral 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 trilateral root is thus a distributed representation of the featural make-up of the three consonant root, expressed as a vector of  $3 \times 17 = 51$  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.

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:

$$(2) a_i = \sigma \left( \sum_j w_{ij} a_j + b_i \right)$$

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  $i$ .  $\sigma$  is the sigmoid logistic function commonly used in connectionist modeling.

The values of  $w_{ij}$  and  $b_i$  are computed with backpropagation (Rumelhart et al., 1986a). All weights and biases were initialized at small random values. Training the AN took place over a run of  $10^7$  epochs, where each epoch involved presenting a single word. At each epoch an actual Arabic root was drawn with a probability of .5. The actual Arabic word was picked at random from the word list of 2674 roots. Other inputs were generated with a probability of .5 in the following way. Three consonants were selected independently with a probability according to how often the consonant occurs in its particular serial location in the word list.<sup>3</sup> The degree of weight change produced by backpropagation was also decreased from 1 to .1 over the whole run. There is also a regularization parameter  $\alpha$  in the model that effectively decreases the connection weights and biases a bit with each epoch.

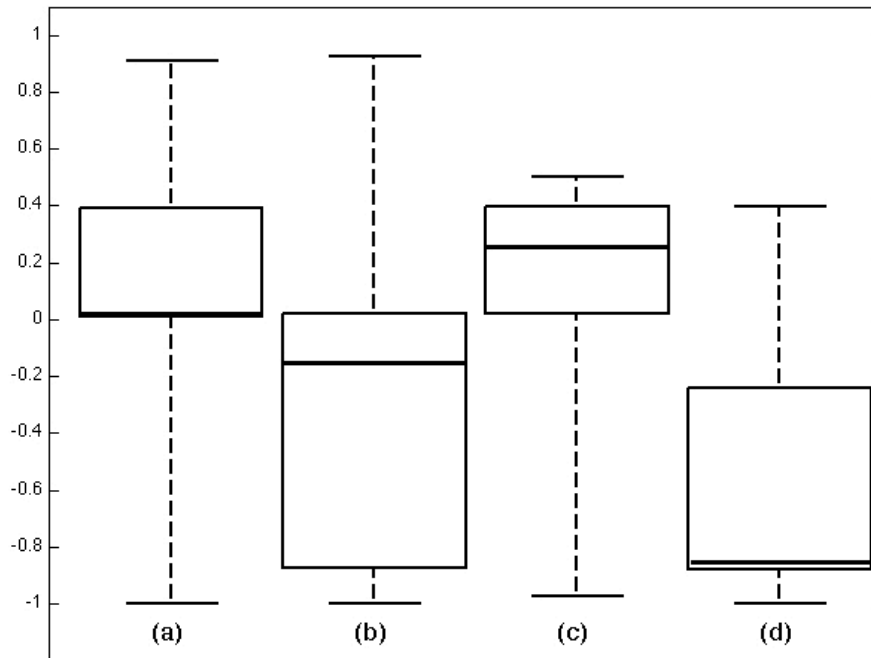
### 3.2 Results and discussion

The AN, after training, rates trilateral roots in a way that can be compared to human acceptability ratings. Figure 1 illustrates two basic rating results for one of the AN networks, namely the network with five hidden layer units. First, while the ratings for the 2674 actual roots overlap with the ratings for all possible roots ( $n = 28^3 = 21,952$ ), the ratings for the middle 50% of the actual words is above the middle 50% for all possible roots, indicating that network has learned to assign higher scores to the words to which it has been presented. Second, the opposition between the third (OCP obeyers) and fourth (OCP violators) box plots shows that the network has effectively learned the OCP, as these same two middle data populations are even further apart.

---

<sup>3</sup> One might object to this way of training the network in that it does not seem connected to any plausible cognitive process, like the generation of errors in some idealized process of language production. We have tested our model in a more complex learning model that involves error generation from an autoassociator module to train the assessor network, with very similar results to random selection of consonants matching Arabic type frequencies.

(3) Fig. 1. Acceptability scores given by one trial of the feedforward network with 5 hidden nodes. Box plots indicating minimum, first quartile, median, third quartile, and maximum scores are shown for four groups of trilateral roots: (a) all attested roots, (b) all possible roots, (c) roots from Experiment 1 with no OCP violation, (d) roots from Experiment 1 with an OCP violation.



Because the AN assigns acceptability scores to nonactual words, it is appropriate to compare the output of the AN with the judgement data of (Frisch and Zawaydeh, 2001). Recall from section 2 that Frisch and Zawaydeh conducted three experiments designed to probe the psychological reality of root cooccurrence constraints. We conducted the same tests as this work, but substituted AN acceptability ratings for their wordlikeness ratings, as shown in Tables 1-3. Table 1 illustrates the results of the three factor ANOVA on AN acceptability for all versions of the ANs, where different versions derive from differences in the number of hidden layer units. Each AN was trained on three separate trials, so the tests were conducted for each trail (each row in Tables 1-3 is a separate trail). All effects with significance at  $p < .05$  are reported and the percentage of the variation accounted for by this effect,  $r^2$ , is shown in parentheses. Finally, correlation coefficients for wordlikeness ratings ( $r_{wr}$ ) and AN acceptability ( $r_{ar}$ ) are shown for each trail, to give a general sense of the correlation between the behavioral data and the modeling results. To summarize the comparison with experiment 1, while there is some variation between modeling trials, an AN with between 2 and 5 hidden layer units provides a good fit with the judgement data from experiment 1. In these models, the OCP accounts for approximately a third of the variation in subject ratings, as found in (Frisch and Zawaydeh, 2001).

(4) Table 1. Results comparing tests for experiment 1: effect (fraction explained);  $\text{Corr}(r_{wr}, r_{ar})$

<b>1 node</b>	<b>2 nodes</b>	<b>5 nodes</b>	<b>10 nodes</b>
OCP (9%); .29	OCP (18%), exp.prob. (5%); .32	OCP (44%); .50	OCP (46%); .36
OCP (8%); .29	OCP (10%), exp.prob. (3%); .29	OCP (52%); .50	OCP (50%), exp.prob. (2%), density (2%), exp.prob. * OCP (2%); .42
exp.prob.* density (5%); -.02	OCP (15%); .32	OCP (30%); .34	OCP (50%); .39

Frisch and Zawaydeh's experiment 2 looked for an effect of the OCP after controlling for expected probability and neighborhood density, and bigram probability was balanced in the stimuli set. The modeling results of the different ANs, where again the number of hidden layer units was varied, show that ANs can approximate this finding as well. Table 2 reports the significant effects ( $p < 0.001$ ) of OCP violations on AN acceptability scores. In all cases but one trial (the single hidden node network), this effect was significant, but the AN that appears to be consistent with the rating data contains two hidden units.

(5) Table 2. Results comparing tests for experiment 2: fraction explained

<b>1 node</b>	<b>2 nodes</b>	<b>5 nodes</b>	<b>10 nodes</b>
14%	21%	61%	53%
14%	28%	74%	35%
1% (not sig.)	17%	35%	38%

Finally, experiment 3 showed that the OCP in Arabic is gradient in the sense that mean acceptability correlated with similarity. The more similar two consonants are, the stronger the effect of the OCP. Below we replace their behavioral rating data with AN acceptability, showing all effects that reach significance at  $p < 0.001$ . Again the AN's with between 2 and 5 hidden layer units are consistent with the ratings data.

(6) Table 3. Results comparing tests for experiment 3: fraction explained

<b>1 node</b>	<b>2 nodes</b>	<b>5 nodes</b>	<b>10 nodes</b>
6% (not sig.)	8%,	28%	18%
5% (not sig.)	9%	28%	6% (not sig.)
0% (not sig.)	3% (not sig.)	18%	7% (not sig.)

## 4. A recurrent network

### 4.1 Network architecture

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 and Rumelhart, 1985), namely a recurrent network, to illustrate how such a network can be used to generate acceptability values. This network, however, is not intended as a realistic model of language production.

In this Recurrent Network (RN), a trilateral 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  $3 \times 17 = 51$  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  $i^{\text{th}}$  node be  $a_i$ , the external output to the  $i^{\text{th}}$  node be  $\text{ext}_i$ , the strength of the connection between node  $i$  and node  $j$  be  $W_{ij}$ , and equilibrium activations be  $a_i^*$ .

(7) Update rule for Recurrent Network

We have

$$da_i / dt = \sigma (\sum_j W_{ij} a_j + \text{ext}_i) - a_i$$

Equilibrium activations satisfy:

$$a_i^* = \sigma (\sum_j W_{ij} a_j^* + \text{ext}_i)$$

The goal in training the RN is to find  $W_{ij}$  such that  $\sum_j W_{ij} a_j^* = \text{ext}_i$ ; in other words, so 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  $\text{ext}_i$ . 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  $\text{ext}_i$ .

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.

(8) Acceptability (RN)

$$\text{acceptability} = - \| W a - \text{ext} \|$$

This model and this definition of acceptability is again applied to the Arabic data.

## 4.2 Results and discussion

The results of the three factor test in experiment 1 are given 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 expected probability and neighborhood density both have significant (though small) effects, 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 it is the most important effect on the acceptability.

(9) Table 4. Results comparing tests for experiment 1: effect (fraction explained),  $\text{Corr}(r_{wr}, r_{ar})$

OCP (58%), exp.prob.*density (3%), exp.prob. (2%)
OCP (57%), exp.prob.*density (3%), exp.prob. (2%)
OCP (57%), exp.prob.*density (3%), exp.prob. (2%)

The next table shows that, in the RN, the OCP has significant effects on acceptability, even when expected probability and density were controlled, as with the AN.

(10) Table 5. Results comparing tests for experiment 2: fraction explained

48%
48%
49%

Finally, the similarity also correlates with RN acceptability, as shown below with the  $r^2$  values on three separate trials. This model compares with AN models with 5 hidden units.

(11) Table 6. Results comparing tests for experiment 3: fraction explained

15%
17%
16%

## 5. General discussion

The simulation results above contribute to an ecumenical understanding of phonotactic learning and the correct analysis of gradient phonotactics. In particular, these results 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: a multilayer network with hidden units and a basic recurrent network structure.

It is sometimes the case that computational modeling is used to compare two or more models, and on the basis of differences in their overall performance, a superior model is selected. We do not employ this methodology here because we do not believe modeling should have this role. It is too easy to modify model parameters to better approximate the data, and furthermore, the datasets these models are working with are still rather primitive. Rather, computational modeling is best used as a tool for exploring ideas and determining if a model in isolation effectively

addresses a problem; in this regard we agree with (McClelland, 2009). Broader conclusions about the correct model for learning phonotactics are thus likely to be established on different grounds.

A significant and complex issue that has potential to support such conclusions is the extent to which *a priori* knowledge of language is necessary. It is not the case that computational learning results require *a priori* constraints, and thereby provide such evidence, because at least two basic approaches, namely Hayes and Wilson's Maximum Entropy model and the connectionist approaches developed here, have the ability to learn gradient phonotactics without core knowledge of the operative constraints in phonotactics. Models that simply lack the ability to induce operative constraints for the input data, e.g., Harmonic Grammar learning ((Smolensky et al., 1992), (Pater, 2009)) and Optimality Theoretic learning models with error retention ((Tesar, 2004), (Prince and Tesar, 2004)) or without error retention (Tesar and Smolensky, 2000), must therefore be argued for on the basis of other kinds of evidence for this *a priori* knowledge.

The apparent asymmetries in linguistic typology are frequently employed as a form of evidence in generative linguistics, but we do not feel that the current understanding of typological patterns constitutes evidence for *a priori* knowledge. First, it is very often the case that typological generalizations are established using convenience-based samples that are difficult to interpret. Further, even when language examples are collected from a semi-random set of reference grammars, it is rarely the case that examples are sampled in a way that ensures the independence of cases, and in particular uses a procedure that controls for the influence of language affiliation (Perkins, 1989). A second issue is that typological generalizations most likely emerge from a combination of effects, some from constraints on human cognitive capacities for language, and some from historical and demographic forces that are at least in part separate from processes of human cognition. In practice it is very difficult to separate these effects, and the potential interactions among them, in a way that permits researchers to align the range of structures found in typology and the structures predicted by a model of human cognition. In sum, it could be that typological studies have identified gaps in structural typologies that *a priori* knowledge of constraints predict, but it could also be that the putative gaps reflect the incomplete nature of typological research.

Perhaps a more direct approach to comparing successful approaches to gradient phonotactics is to examine more carefully the inherent assumptions of each and see how they relate to other cognitive processes. A comparison of the multilayer AN with the two alternatives discussed above, Harmonic Grammar (HG) and Maximum Entropy (MaxEnt) models, is instructive in this regard. HG grammars actually resemble a simplified AN very closely, with the qualification that the HG models of Arabic simply lack a hidden layer. That is, our AN has to learn two weight matrices:  $W_1$  for the weight coefficients linking the input layer and the hidden layer, what might be called 'learning the constraints', and  $W_2$ , the weights linking the hidden layer nodes with the output node, or 'learning the constraint weightings'. HG models do not have a  $W_1$ ; they simply do not make learning the constraints part of the learning problem because the constraints are given in advance. MaxEnt models, on the other hand, do learn the operative constraints by induction, but they do so rather differently by positing a large universe of constraints and proposing search heuristics for selecting the operative constraints. Clearly, if learning the constraints themselves is a concern at all, the focus should be on teasing apart important differences between these two approaches to inducing the operative constraints.

One question that comes to mind immediately is how well the MaxEnt and c-net approaches compare in terms of characterizing the range of logically possible phonotactic systems, i.e., how they address the expressiveness problem. After all, c-nets use a general purpose toolbox for inducing constraints by relating hidden units to the input data, while Hayes and Wilson's model makes certain rather modest, but also potentially restrictive, assumptions about this search space by positing certain logical assumptions about the set of possible initial constraints (i.e., underspecification, reference to variables in constraints, etc.). Will both models actually be able to describe the same facts? The answer to this question, putting the issue of training aside, is 'no', because while both have a strong descriptive capacity, c-nets, which do not have the same limits on potential constraints, can describe a wider range of phonotactic effects. The appendix presents a proof that relates the underlying functions in MaxEnt and c-net models, and further shows the greater descriptive capacity of general purpose c-nets.

One aspect of the MaxEnt model that distinguishes it from our c-net is the apparent need for 'batch processing', or reference to the entire set of surface forms and probabilities of the structures in those forms. In Hayes and Wilson's approach, this need is expressed in both their algorithms responsible for selecting the operative constraints, and also the algorithms responsible for ranking the constraints. The latter requirement seems not to be a necessary assumption to MaxEnt learning, as another related model (Jäger, 2004) does not require such reference in probabilities of phonotactic patterns in the constraint weighting algorithm. However, we wonder if reference to phonotactic probabilities, in essence to an entire constellation of structures, is a realistic assumption to make on a phonotactic learner. For example, exploration in other cognitive domains like human rationality have shown that while human subjects are sensitive to frequency patterns in making rational decisions, like in tasks involving making a medical diagnosis, they are rather weak at making the same kinds of decisions when they involve manipulation of information encapsulated in a probability ((Gigerenzer, 1994), (Cosmides and Tooby, 1996)). While our training regimes described in sections 3-4 do pick out words based on their frequency distributions, the c-net learners do not actually require reference to phonotactic probabilities. Our learners could have been trained with error-corrective learning using simply the words encountered in the normal process of language acquisition. While the limits of human rationality are not in the same domain as phonotactics, and it clearly involves explicit learning, if it turns out that this constraint on human cognitive capacities carries over to language, it speaks in favor of connectionist learning because it can work exclusively with type frequencies and still successfully learn complex pattern.<sup>4</sup>

## Appendix

Both MaxEnt and c-net approaches to modeling acceptability judgements construct a function  $F$  with parameters  $p$ . The function takes a word  $x$  as input and then outputs a number  $a$  indicating the acceptability of the word.

$$a = F(x,p) = F(x).$$

---

<sup>4</sup> One thinks of the well-known studies in infant speech perception in connection with this issue, where it has been shown that infants as young as eight months can use the 'transitional probabilities' of two segments to infer word boundaries ((Saffran et al., 1999), (Saffran et al., 1996)). Importantly, the definition of a transitional probability used in this work, frequency of  $XY$ /frequency of  $X$ , is rather different than the global notion used in MaxEnt learning, as the former is a simple ratio of type frequencies. Thus, this distributional information is compatible with conversion to a frequentist representation of the kind required by the human rationality work. Indeed, connectionist sentence parsers work with a very similar kind of representation, making parsing decisions based on a local assessment of the word string to be parsed ((Elman, 1990), (Elman, 1993)).

There are two distinct aspects to the approach. One is the architecture of the network; the other is how the parameters  $p$  of the network are set by the training regime. Here we address the first aspect, and show that our c-net approach has greater expressiveness than Hayes and Wilson's MaxEnt architecture. We show that for any choice of constraints  $c_i$  and weights  $w_i$  used in the MaxEnt grammar, we can choose the number of hidden nodes and the weights on the connections so that the c-net agrees on acceptability judgements arbitrarily closely.

The MaxEnt architecture developed in (Hayes and Wilson, 2008) can be characterized as follows:

$$F(x) = \exp(-\sum_i w_i c_i(x))$$

where  $c_i$  are constraint functions and  $w_i$  are the weights. Each  $c_i$  returns either 0, 1, 2, ... for each word  $x$ , corresponding to the number of violations of a given constraint. Constraints must be given in terms of natural classes and must be translation invariant, i.e. constraints are always constraints in adjacent segments, but it does not matter where in the word adjacent segments are located. Both  $c_i$  and  $w_i$  are determined by the learning algorithm.

Our architecture is:

$$F(x) = \sigma(\delta - \sum_i w_i \sigma(b_i + (V x)_i))$$

Here  $\delta$ ,  $w_i$  and  $b_i$  are scalars.  $V$  is an  $h \times n$  matrix, where  $h$  indicates the number of hidden units and  $n$  is the length of the input  $x$ ;  $b_1$  and  $b_2$  are vectors;  $\sigma$  is a sigmoid function that gives values between 0 and 1. (In our actual simulations we used a sigmoid running between -1 and 1, but we use this equivalent alternative here to simplify discussion.)

First, considering that we are primarily interested only in relative judgements of acceptability we omit the outer function evaluations of both values for  $F$ , since  $\exp$  and  $\sigma$  are both monotonic functions. Let us call this  $G$ .

For Hayes and Wilson we have

$$G(x) = -\sum_i w_i c_i(x).$$

However, we will interpret each of their constraints as one constraint for each segment in the word, in which case  $c_i(x)$  is either 0 or 1.

For our model we have

$$G(x) = -\sum_i w_i \sigma(b_i + (V x)_i).$$

We claim that a suitable choice of  $w_i$ ,  $b_i$  and  $V$  can match our model arbitrarily close to their model. Choosing  $w_i$  to be the same as their  $w_i$  is straightforward. So it only remains to show that for each  $i$ .

$$c_i(x) = \sigma(b_i + (V x)_i)$$

for some  $b_i$  and  $V$ . Dropping the subscripts, for each constraint  $c$  of Hayes and Wilson, we need to show that there is a scalar  $b$  and a vector  $v$  such that

$$c(x) = \sigma(b + \sum_j v(j) x(j)),$$

where  $v(j)$  denotes the  $j$ th entry of the vector  $v$ .

The expression on the right gives the output of a perceptron with a smooth activation function  $\sigma$ . Let us first imagine that sigma is the step function and then we will consider our smoother case. This means that  $\sigma(b + \sum_j v(j) x(j)) = 1$  if  $b + \sum_j v(j) x(j) > 0$  and it is equal to 0 otherwise. As is well known, not all Boolean functions can be computed by a perceptron, exclusive OR being a prominent example (McLeod et al., 1998). However, they are capable of matching the limited set of Boolean functions of feature values used by Hayes and Wilson in their MaxEnt grammar, as we will show.

Hayes and Wilson use two types of constraints. The first prohibits a collection of feature values over one or more segments. For example, a constraint they propose for English onsets is

\*[+ant, +strid][-ant]

which prohibits the cluster /sr/, among others. We will consider this constraint as applying to the first two segments of a form.

This constraint can be captured by a perceptron as follows. Let  $j_1, j_2, j_3$  be the indices of the input nodes corresponding to [ant] for the first segment, [strid] for the first segment, and [ant] for the second segment, respectively. We want to find a scalar  $b$  and a vector  $v$  so that if  $x(j_1)=1, x(j_2)=1$ , and  $x(j_3)=-1$ , then  $b + \sum_j v(j) x(j) > 0$ , but otherwise  $b + \sum_j v(j) x(j) < 0$ . This is achieved by letting  $v(j_1)=1, v(j_2)=1, v(j_3)=-1$ , all other  $v(j)=0$ , and  $b=-2.5$ .

A similar idea works for any constraint of this type. Suppose a constraint is specified on  $J$  nodes with indices  $j_1$  through  $j_J$  by saying that there is a violation if for all  $k=1, 2, \dots, J$ ,  $x(j_k)=e_k$ , where each  $e_k$  is -1 or 1. To capture this with a perceptron, let  $v(j_k)=e_k$  for  $k=1, \dots, J$  and  $v(j)=0$  otherwise. Let  $b = -J + 1/2$ . If all  $J$  of the relevant nodes have their prohibited values then  $\sum_j v(j) x(j) = J$ , and  $b + \sum_j v(j) x(j) > 0$ , causing the perceptron to return 1. However, if not all of the relevant nodes are active the perceptron will return 0 as required.

The other form of constraint considered by Hayes and Wilson is implicational. An example is

\*[+lab][^+son, +cor],

which in their notation means that if a consonant is [+lab] it may only be followed by a consonant with [+son, +cor]. To capture this constraint with a perceptron, as before we let  $j_1$  correspond to [lab] on the first consonant,  $j_2$  correspond to [+son] on the second consonant, and  $j_3$  correspond to [+cor] on the second consonant. We then let  $v(j_1)=1, v(j_2)=-1/2, v(j_3)=-1/2$ , and  $b=1/4$ . Suppose  $x(j_1)=1$ . Then the only way to prevent the perceptron from being on ( $b + \sum_j v(j) x(j) > 0$ ) is if both  $x(j_2)$  and  $x(j_3)$  are both 1.

More generally, suppose we have a constraint that  $x(j_i)=e_i$  for  $i=1, \dots, K$  implies that  $x(k_m)=f_m$  for  $m=1, \dots, J$ . We assume that the indices  $j_i$  and  $k_m$  are distinct, as they are for the constraints used by Hayes and Wilson. We let  $v(j_i)=e_i$  for all  $i$ ,  $v(k_m)=f_m/K$  for all  $m$  and all other entries of  $v$  be zero. Then we let  $b = J - 1 + 1/K$ . It is straightforward to check that  $b + \sum_j v(j) x(j) > 0$  only if  $x(j_i)=e_i$  for all  $I$ , but for some  $m$ ,  $x(k_m)$  is not equal to  $f_m$ .

The arguments above were all for the case where sigma is the step function. However, a perceptron with the step function can be approximated arbitrarily well with our sigmoid function if we multiply  $v$  and  $b$  by a sufficiently large positive constant.

## References

- Anttila, Arto. 2008. Gradient phonotactics and the Complexity Hypothesis. *Natural Language and Linguistic Theory* 26:695-729.
- Berent, Iris, Everett, Daniel L, and Shimron, Joseph. 2001. Do phonological representations specify variables? Evidence from the obligatory contour principle. *Cognitive Psychology* 42:1-60.
- Berent, Iris, Marcus, Gary F, Shimron, Joseph, and Gafos, Adamantios I. 2002. The scope of linguistic generalizations: evidence from Hebrew word formation. *Cognition* 83:113-139.
- Boersma, Paul. 1998. *Functional Phonology*. The Hague: Holland Academic Graphics.
- Boersma, Paul, and Hayes, Bruce. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32:45-86.
- Coetzee, Andries, and Pater, Joe. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 84:289-337.
- Cosmides, L, and Tooby, J. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition* 58:1-73.
- Cowan, J. Milton (ed.). 1979. *Hans Wehr: A dictionary of Modern Written Arabic*. Wiesbaden, Germany: Otto Harrasowitz.
- Elman, Jeffrey. 1990. Finding structure in time: . *Cognitive Science* 14:179-211.
- Elman, Jeffrey. 1993. Learning and development in neural networks: the importance of starting small. *Cognition* 48:71-99.
- Frisch, Stefan, and Zawaydeh, Bushra. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77: 91-106.
- Frisch, Stefan A., Pierrehumbert, Janet, and Broe, Michael B. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22:179-228.
- Gigerenzer, G. 1994. Why the distinction between single-event-probabilities and frequencies is important for psychology (and vice versa). In *Subjective Probability*, eds. G Wright and P Ayton, 129-161. New York: John Wiley.
- Goldsmith, John. 1990. *Autosegmental and metrical phonology*. Oxford: Blackwell.
- Hayes, Bruce, and Wilson, Colin. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39:379-440.
- Jäger, Gerhard. 2004. Maximum entropy models and stochastic Optimality Theory. Ms. University of Potsdam.
- Marcus, Gary F. 1998. Rethinking eliminative connectionism. *Cognitive Psychology* 37:243-282.
- Marcus, Gary F. 2001. *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: The MIT Press.
- McCarthy, John J. 1986. OCP Effects: Gemination and antigemination. *Linguistic Inquiry* 17:207-263.
- McClelland, James L. 1981. Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the third annual meeting of the cognitive science society*, 170-172.
- McClelland, James L., and Rumelhart, David. 1985. Distributed memory and the representation of general and specific information. *Journal of Experimental psychology: General* 114:159-188.

- McClelland, James L. 2009. The place of modeling in cognitive science. *Topics in Cognitive Science* 1.
- McLeod, Peter, Plunkett, Kim, and Rolls, Edmund T. 1998. *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.
- Padgett, Jaye. 1995. *Structure in feature geometry*. Stanford, CA: CSLI Publications.
- Pater, Joe. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33:999-1035.
- Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13:293-315.
- Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. In *NELS* 23, 367-381.
- Pierrehumbert, Janet. 2003. Probabilistic phonology: Discrimination and robustness. In *Probability theory in linguistics*, eds. Rens Bod, Jennifer Hay and Stefanie Jannedy, 177-228. Cambridge, MA: The MIT Press.
- Prince, Alan, and Tesar, Bruce. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, eds. René Kager and Joe Pater, 245-291. Cambridge: Cambridge University Press.
- Ramsey, William, Stich, Stephen, and Garon, Joseph. 1990. Connectionism, eliminativism and the future of folk psychology. In *Connectionism: Debates on folk psychology*, eds. C. Macdonald and G. Macdonald, 311-338. Cambridge, MA: Basil Blackwell.
- Rumelhart, David, Hinton, Geoffrey E, and Williams, Ronald J. 1986a. Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 1-2*, eds. James L. McClelland, David Rumelhard and The PDP Group, 318-362. Cambridge: The MIT Press.
- Rumelhart, David, McClelland, James L., and Group, The PDP Research. 1986b. *Parallel distributed processing: Explorations in the microstructure of cognition. Volumes 1-2*. Cambridge, MA: MIT Press.
- Saffran, Jenny, Aslin, Richard, and Newport, Elissa. 1996. Statistical learning by 8-month-old infants. *Science* 274.
- Saffran, Jenny, Johnson, Elizabeth, Aslin, Richard, and Newport, Elissa. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition* 70:27-52.
- Smolensky, Paul. 1988. The proper treatment of connectionism. *The Brain and Behavioral Sciences* 11:1-23.
- Smolensky, Paul, Legendre, Géraldine, and Miyata, Yoshiro. 1992. Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition: Computer Science Department, University of Colorado at Boulder.
- Tesar, Bruce, and Smolensky, Paul. 2000. *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
- Tesar, Bruce. 2004. Using inconsistency detection to overcome structural ambiguity in language learning. *Linguistic Inquiry* 35:219-253.
- Treiman, Rebecca, Kessler, Brett, Knewasser, Stephanie, Tincoff, Ruth, and Bowman, Margo. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in laboratory phonology V: acquisition and the lexicon*, eds. Michael B Broe and Janet B Pierrehumber, 269-282. Cambridge: Cambridge University Press.