Phonological constraint induction in a connectionist network: learning OCP-Place constraints from data*

John Alderete[a], Paul Tupper[a], Stefan A. Frisch[b]

Simon Fraser University[a], University of South Florida[b]

Abstract. A significant problem in computational language learning is that of inferring the content of well-formedness constraints from input data. In this article, we approach the constraint induction problem as the gradual adjustment of subsymbolic constraints in a connectionist network. In particular, we develop a multi-layer feed-forward network that learns the constraints that underlie restrictions against homorganic consonants, or 'OCP-Place constraints', in Arabic roots. The network is trained using standard learning procedures in connection science with a representative sample of Arabic roots. The trained network is shown to classify actual and novel Arabic roots in ways that are qualitatively parallel to a psycholinguistic study of Arabic. Statistical analysis of network behavior also shows that activations of nodes in the hidden layer correspond well with violations of symbolic well-formedness constraints familiar from generative phonology. In sum, it is shown that at least some constraints operative in phonotactic grammar can be learned from data and do not have to be stipulated in advance of learning.

Highlights:

- Inferring the content of linguistic constraints, or constraint induction, is a significant part of computational language learning.
- In the micro-structure of connectionist networks, constraints are sets of connections; constraint induction therefore involves gradual adjustment of connection weights.
- A multi-layer feed-forward connectionist network trained on a representative sample of Arabic roots can rate actual and novel Arabic roots in ways qualitatively parallel to Arabic native speakers.
- Statistical analysis of the behavior of the connectionist network reveals a coarse-graining effect in which the functions computed by subsymbolic constraints correspond well to the symbolic constraints familiar from generative phonology.
- Gradient OCP-Place effects can be learned as subsymbolic constraints and do not need to be stipulated in advance of learning.

Keywords: constraint induction, subsymbolic learning, connectionism, parallel distributed processing, Optimality Theory, Arabic, co-occurrence restrictions, dissimilation, nature vs. nurture, gradience, exceptions

---

* This report has benefited from the comments and questions put to us by Joe Pater, Terry Regier, and Colin Wilson and those raised at the poster session of the 2009 Cognitive Science Conference and the 2010 Computational Modeling of Sound Pattern Acquisition conference at the University of Alberta. It is supported in part by a standard SSHRC research grant (410-2005-1175) awarded to John Alderete and an NSERC discovery grant awarded to Paul Tupper. Thanks also to Navid Alaei for help in finding and processing the Buckwalter Arabic root list and Reem Alsadoon for answering our questions about Arabic linguistics. Usual disclaimers.

1. Introduction

1.1 Constraint induction

A central challenge to linguistic theory is to predict the universal aspects of language with the same toolbox used to capture language-particular facts. A common strategy in generative linguistics is to assume that there is a universal set of well-formedness constraints available in all grammars, and that these constraints are prioritized in language-particular contexts. Classical Optimality Theory employs this strategy by assuming that universal constraints are ranked on a language-particular basis ((Prince & Smolensky 1993/2004), (McCarthy & Prince 1995)), and most theoreticians in the emerging framework of Harmonic Grammar likewise assume the universality of constraints but that constraints are weighted in order to conform to language-particular facts ((Legendre et al. 1990), (Coetzee & Pater 2008), (Farris-Trimble 2008), (Pater 2009), (Smolensky & Legendre 2006b)).

While this strategy has produced valuable results, it has also clarified certain problems with the assumption that all well-formedness constraints are universal. First, the existence of language-particular rules that seem to be completely inconsistent with universal trends have been argued to require language-particular mechanisms ((Blevins 1997), (Hayes 1999)). To retain the predictiveness of universal constraints, so the argument goes, we require specific mechanisms for inducing constraints, or even rules, from language-particular data. A similar line of reasoning emerges from apparent violations of the Continuity Hypothesis in language acquisition (Pinker 1984), according to which the formal properties of grammar are universally constrained and do not change. The differences between developing and mature phonologies, like the emergence of syllable-internal consonant harmony in child phonology, has led many researchers to argue for mechanisms of constraint induction that may produce constraints outside a universal constraint set ((Becker & Tessier 2011), (Fikkert & Levelt 2008), (Goad 2001), (Levelt & Oostendorp 2007), cf. (Inkelas & Rose 2008); see (Buckley 2003) and (Pierrehumbert 2003) for reviews).

It is not the case that all of this evidence requires language-particular constraint induction. Levelt and van Oostendorp (2007), for example, make clear that the emergence of feature co-occurrence constraints in developing consonant inventories could either be the result of constraint induction, or, alternatively, the activation/deactivation of universal constraints in response to input data. However, we believe that the weight of the evidence supports induction of language-particular constraints, and that significant interest has emerged in recent years in better understanding the nature of such a mechanism. We therefore take the existence of a constraint induction mechanism for granted in this article, and focus instead on the underlying processes of constraint induction itself.

1.2 Subsymbolic constraints

Considerable research on this problem in linguistics has assumed that induced constraints are symbolic in nature in the sense that they refer to linguistic symbols and can motivate manipulations of these symbols in outputs. Constraint induction in this approach involves either the creation or selection of constraints from a relatively large range of logically possible constraints afforded by linguistic symbol systems ((Alderete 2008), (Flack 2007), (Hayes 1999), (Hayes & Wilson 2008), (Pater 2011/To appear)). Important work has also shown how this symbolic structure can be implemented at the subsymbolic level of connectionist networks (Smolensky & Legendre 2006b). Relatively less attention, however, has been paid to the question of how symbolic structure might emerge from the induction of constraints at this

subsymbolic level (but see (Plaut & Kello 1999) for an important precedent). This question is the focus of the present work.

We argue that standard mechanisms for representing and learning linguistic structure in connectionist networks (c-net henceforth) provide a number of attractive advantages in addressing the problem of constraint induction. First, subsymbolic constraints in c-nets are understood as sets of connections within a node network. For example, the constraint ONSET (syllables must have onsets) in a recent analysis of Berber syllabification is formalized as the inhibitory links between output nodes representing syllable positions (Sorace et al. 2006). Understood as sets of connection weights, the induction of subsymbolic constraints can be modeled as the gradual adjustment of connection weights in response to natural language data. Second, c-nets have a number of properties that make them very good at capturing natural language data. Because connection weights are assigned on a continuous scale, they can readily capture gradient linguistic generalizations, generalizations that are abundant in natural language. Also, c-nets are well-suited for capturing similarity structure, like the graded generalizations correlated with featural similarity. A third consideration is that this approach is in line with a large body of work in the psycholinguistics of language production and perception that likewise models language processes as information-processing in connectionist networks. Modeling acquisition in a c-net therefore makes possible a natural integration of the induction of language-particular constraints with underlying psycholinguistic processes, a point we explore below.

A common criticism of connectionist approaches to language processes is that they have greater descriptive capacity and so they capture language-particular facts at the expense of the universality of language (see e.g., (Massaro 1988)). While it is certainly true that c-nets have rich generative and learning capacity, we believe the advantages of c-nets outlined above outweigh this objection for the following reasons. First, theoretical c-nets have rich generative capacity, but they are significantly constrained by training. Therefore, this criticism can only be assessed in the context of simulations of specific language examples, and we offer a positive assessment here from a specific phonological system. Second, given the idiosyncrasies of language-particular rules and violations of the Continuity Hypothesis outlined above, we believe that it is still rather premature to reject a highly expressive theory of constraints like that offered in connectionist models. The differences between possible and attested languages are still too poorly understood.

Finally, we do not require all constraints to be induced at the micro-structure level of c-nets. We think that it is plausible that micro-structure learning could provide a mechanism of ultimately learning symbolic macro-structure constraints. Indeed, the parallels established in Smolensky & Legendre (2006) between the subsymbolic constraints of c-nets and symbolic constraints in theories like Optimality Theory provides a framework for such an approach. It is therefore possible under this view that universal trends in linguistic typology derive from a subset of universal constraints. Or, alternatively, that the universal aspects of these constraints derive from a 'constraint hardening' that assigns symbolic structure to subsymbolic constraints learned from data (cf., (Pierrehumbert 2001)). In the latter case, the universal character arises from their integration with the psycholinguistics of production and comprehension made possible by subsymbolic constraint induction. In our analysis below of the functions computed by hidden layer nodes, we offer one statistical tool for imposing symbolic structure on subsymbolic constraints using classification and regression trees, illustrating that such a process is tractable.

1.3 Overview

We develop our argument by proposing a connectionist architecture for inducing phonotactic constraints from data. We build this architecture and apply it to the problem of learning root co-occurrence restrictions, or 'OCP-Place constraints', in Arabic. Arabic is a good case because rich datasets exist, including large root lists and psycholinguistic experiments ((Buckwalter 1997), (Dukes & Habash 2010); (Frisch & Zawaydeh 2001)), that enable strong tests of the model's performance. Also, Arabic is known to exhibit strong but graded phonotactic patterns that make it a good test case for subsymbolic constraint induction. The principal result reported below is that the graded phonotactic patterns of Arabic consonant phonology can be learned as the gradual modifications to subsymbolic constraints in a c-net. OCP constraints can be induced directly from realistic language data, and do not need to be stipulated in advance of learning.

The rest of this article is organized as follows. The next section summarizes the Arabic data that we attempt to model, including some exceptional patterns that we document for the first time in some detail. Section 3 lays out the theoretical context and some of the formal details of our model (with the rest of the details appearing in the appendix). Section 4 presents the results of our connectionist learning model, and the last section discusses some of the issues raised by the research.

2. Root co-occurrence restrictions in Arabic

2.1 Core facts

A root in Arabic is a discontinuous string of consonants that is interspersed with patterns of vowels to form stems. The number of consonants making up the root can be between two and five, but triconsonantal roots are by far the most common. Roots in a sense specify a narrow semantic field within which actual stems are realized. For example, the triconsonantal root *k-t-b* 'writing' can be interlocked with the pattern for the active participle, *CaaCiC*, to form the noun *kaatib* 'writer'. While standard reference grammars, e.g., (Ryding 2005), tend to distinguish just these roots and patterns, work in contemporary theories of morphology and phonology has further decomposed some stem patterns into grammatical morphemes of two kinds: (i) discontinuous strings of vowels and, (ii) prosodic templates to which the consonantal root and vocalic morphemes are linked up ((McCarthy 1979), (McCarthy & Prince 1990)).

There is a set of restrictions active in Arabic roots against two adjacent consonants having the same place of articulation. This generalization was first made in (Greenberg 1950) and documented further in ((McCarthy 1988), (McCarthy 1994), and (Pierrehumbert 1993)) with different datasets. Table 1 below, from (Frisch et al. 2004), organizes Arabic consonants into a set of homorganic natural classes typically assumed in prior work, following the autosegmental analysis of (McCarthy 1988).[1] We refer to these classes below (excluding the uvulars) as 'same-place' classes, because they are not the same classes as the natural classes defined by major place features. As explained below, there are three distinct coronal same-place classes, and uvulars are merged with both dorsal and pharyngeal classes.

The rate of co-occurrence 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

[1] Following prior work cited above, glides are excluded from the chart because their unusual phonology makes it difficult to establish frequency generalizations.

root, i.e., the first two or last two consonants, are shown in Table 1 from a dataset of 2,674 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 below 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 ((Pierrehumbert 1993),(McCarthy 1994)). 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 both dorsals and pharyngeals, there are six same-place classes in Arabic root phonotactics. We will use these six classes for several purposes below.

Table 1. Co-occurrence of adjacent consonants in Arabic triliteral roots (from 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 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	0.07	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

The restriction against same-place pairs is also found in non-adjacent consonants, i.e., the first and third consonant of a triliteral root, but the restriction is not as strong ((Greenberg 1950), (Pierrehumbert 1993), (Frisch et al. 2004); see also discussion below in 2.2).

Two same-place consonants are in general prohibited in Arabic roots. However, two identical consonants are commonly observed in the second and third consonantal positions in triliteral roots, e.g., *madad* 'stretch'. Most prior work in generative phonology, and the table above, follow (McCarthy, 1986) in exempting these XYY roots from place restrictions. In particular, 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 are really one underlying consonant and therefore do not constitute a consonant pair for the purpose of the restriction against same-place consonants ((Coetzee & Pater 2008), (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. In the last section, however, we discuss some ways of extending our model that can address this problem.

2.2 Documenting exceptional patterns

While the generalization banning homorganic consonants is clearly evident in Table 1, a closer look at the facts of consonant co-occurrence reveals many exceptional patterns that contain particular same-place segments in particular positions. For example, in his original 1950 article, Greenberg notes that, while pairs of pharyngeals and uvulars are in general significantly

underrepresented in Arabic, most of the exceptions to this restriction are of the form /χCʕ/, which occur at a rate approaching chance. This example, which is typical of many others, gives additional structure to the description of Arabic phonotactics. Thus, while there is an over-arching constraint banning pairs of same-place consonants, there are pockets of consonant pairs that are not as underrepresented as one would expect from a blanket restriction against homorganic consonant pairs. We describe these exceptional patterns below to document this additional layer of phonotactic structure. In section 5, we also use this description to ask if our connectionist network learner is sensitive to this level of phonotactic detail.

Our description draws on the data in the (Buckwalter 1997) root list. This list contains 4,749 roots, including both triliterals (three consonants) and quadriliterals (four consonants), but we exclude the quadriliterals for comparison with most prior work in generative phonology, which has a focus on triliterals. The triliteral section of Buckwalter's root list contains 3,823 roots, of which 334 are final-geminate roots. Since we exclude the latter, the triliteral list we use for description and modeling purposes contains 3,489 roots. This number is roughly half of the estimated 5,000-6,500 roots in the language (Ryding 2005). In addition to containing a larger number of roots than the Hans Wehr root list, it also contains both nominal and verbal roots, and so it is a more representative sample of the Arabic lexicon. This choice is important, because the larger goal of this work is to demonstrate induction of phonological constraints from realistic data. Thus, when we compare the performance of our learning model with native speaker judgments of the same roots, it is important that the training data be a good approximation of what native speakers are exposed to in language acquisition. The data supplement to this article available from the authors' websites includes the Buckwalter list transcribed in IPA, together with a set of contingency tables documenting consonant co-occurrence.

Table 2 below lists the counts of all exceptions in the Buckwalter corpus to the homorganic co-occurrence restrictions of Arabic, sorted by the six same-place classes and consonantal position. We exclude examples with identical segments, i.e. roots of the form XXY or XYX. A count is given for both the exceptional pattern and the total number of exceptions in the same position and same-place class (shown in the bottom right of each box). For example, there are 2 roots that fit the pattern /dCt/, where /d/ occurs in C1 position, /t/ in C3, and any other consonant in C2 position. This pattern accounts for 2 of the 15 total number exceptions to the OCP for coronal stops in C1C3 pairings.

Table 2. Exceptional patterns in Arabic triliteral roots, sorted by same-place class and position

	C1C2		C2C3		C1C3			
Labial	fmC	1			bCf	1		
3×3					bCm	9		
					fCm	11		
Totals		*1*		*0*				*21*
Coronal stop	dˤdC	1	Ctd	4	dCt	2	dˤCd	2
4×4			Ctˤd	1	tˤCt	3	dCtˤ	1
			Cdˤd	3	tCd	1	dˤCtˤ	3
					tˤCd	2	dCdˤ	1
Totals		*1*		*8*				*15*
Coronal fricative	sðC	2	Cʃz	1	ʃCθ	2		
7×7	ʃðC	4			ʃCð	1		
	ʃsC	1			ʃCs	3		
	ʃzC	1			ʃCsˤ	1		
	ʃsˤC	2			ʃCzˤ	1		
	ʃzˤC	2						
Totals		*12*		*1*				*8*
Dorsal	gkC	1	Cqg	1	gCk	1	ʁCq	6
5×5	χgC	1	Cgq	2	χCk	1	kCχ	3
	ʁgC	1	Cχq	1	kCg	1	gCχ	2
	χqC	1			qCg	4		
	ʁqC	1			χCg	5		
	kχC	1			ʁCg	1		
	gχC	2			gCq	2		
	kʁC	3			χCq	5		
Totals		*11*		*4*				*31*
Pharyngeal	ʔχC	4	Cχˤ	2	ʔCχ	1	ʔCh	3
6×6	ʔħC	3			ʔCħ	1	χCʔ	5
	ˤhC	4			χCˤ	9	ʁCʔ	1
	ʔhC	2			hCˤ	8	ħCʔ	3
	hʔC	1			ʔCˤ	1	ˤCʔ	1
					ˤCh	5	hCʔ	6
Totals		*14*		*2*				*44*
Coronal sonorant	nrC	1	Crl	2	rCl	14		
3×3	rnC	7	Cnl	1	nCl	22		
			Clr	1	nCr	23		
			Cnr	7	lCn	11		
			Cln	5	rCn	15		
			Crn	9				
Totals:		*8*		*25*				*85*

These patterns support and extend some of the observations made in prior work. For example, if one distinguishes these patterns by their position in the root, the number of attested pairings of non-adjacent same-place consonants (C1C3) comes to 45 (from 204 roots), which far outnumbers exceptional patterns in both initial C1C2 (= 23 patterns from 47 roots) and final C2C3 (=14 patterns from 40 roots) pairs. This fact is consistent with the observation that non-adjacent consonant pairs are less restricted than adjacent pairs ((Greenberg 1950), (McCarthy 1994), (Pierrehumbert 1993)). The exceptions to the constraint against two coronal fricatives also reveals a static generalization that Greenberg mentions in passing, namely that most of these exceptions involve /ʃ/ as an initial member of the consonant pair. Finally, these exceptional patterns also often differ in whether they are the lone examples in an otherwise general prohibition on same-place consonants, or they are instead one exceptional pattern among many. For example, there is just one exceptional pattern to the restriction against two pharyngeals in C2C3 pairs, /ʿχʿ/, but there are 5 distinct patterns in C1C2 pairs and 12 in C1C3 pairs. The facts above give richer structure to Arabic phonotactics, and we use this as a way of testing how well our learning model has learned the restrictions on homorganic consonants.

2.3 Summary

We can summarize these facts with some guiding assumptions from Autosegmental Phonology. The Obligatory Contour Principle (OCP; (Leben 1973), (Goldsmith 1976), (McCarthy 1986), (McCarthy 1988), (Yip 1989),), and extensions of it in Optimality Theory ((Myers 1997), (Suzuki 1998)), provide a means of formalizing specific consonant co-occurrence restrictions described above. OCP-Place constraints effectively ban two segments that have identical specifications for the major place features, e.g., OCP-Labial bans a form with two labial segments. In many languages, as in Arabic, additional 'subsidiary' features are needed to further specify the set of features that must also have the same specification. Thus, (Padgett 1995) argues for three sets of OCP-Coronal constraints in Arabic, OCP-[Coronal, +son], OCP[Coronal, -son, -cont], OCP[Coronal, -son, +cont], to account for the basic fact that the same-place classes of Arabic subdivide the coronals into three classes: sonorants, stops, and fricatives. To these, we add OCP[labial], OCP[dorsal], and OCP[pharyngeal], to cover the course-grained constraints exhibited by the O/E patterns in Table 1.

In addition, two other factors are important: proximity and phonological similarity. In Table 1, for example, we see that there is a stronger restriction on two coronal stops than a coronal stop and fricative, because, in the latter case, the segments differ in continuancy and are therefore less similar phonologically. In section 4, we review psycholinguistic evidence for the fact that phonological similarity is a factor in native speakers' intuitions about these restrictions. While we do not know if native speakers have robust intuitions about the exceptional patterns in Table 2, we may also ask if our analysis can discriminate between these fine-grained exceptions, essentially on a segment-by-segment basis, and the OCP constraints given above.

Finally, we acknowledge that, like most work in phonotactic learning, we have tackled just a subpart of the set of problems that need to be addressed in word learning. We have focused on root phonotactics because of its relevance to constraint induction, but Arabic learners must learn root phonotactics in the larger context of word learning, which is a different morphological domain than roots. There are known methods for extracting roots from words in languages like Arabic with root-and-pattern morphology, e.g., (Gasser 2009), which shows both sub-problems

are well-defined and tractable. But it has yet to be shown that a single learning system can solve the two problems in tandem.

3. Theoretical assumptions

3.1 Background on connectionist grammars in generative phonology

Connectionist networks (c-nets) compute input-output processes by sending information through a web of simple processing units. C-nets are often said to be neurally-inspired, but connectionist researchers generally do not claim that c-nets are entirely brainlike in the way they process information, and nor do we. The important point is that providing a model of this micro-structure constitutes a theory that makes new predictions, which we explore below in the context of the problem of constraint induction.[2]

Information is passed through a c-net by computing the activation states of simple processing units. The flow of this information is neurally-inspired in the sense that the computation of these states is done in parallel, and the activation state of one unit is affected by the activation state of other units connected to it in the network. As shown below, units are often organized into distinct layers corresponding to different types of representations. For example, input and output layers are distinguished from the hidden layer, an internal representation that restructures the input representation as a function of the first set of connection weights. The activation state of any particular unit is a function of the weighted sum of the activation states of all units sending information to it. Thus, in the enlarged fragment of unit m on the right of Fig. 1, m's activation is the weighted sum of the activation values from all units sending information to m, transformed by an activation function. A commonly used activation function is the sigmoid logistic function, which has the effect of squishing the sum of input activation states into a fixed range.

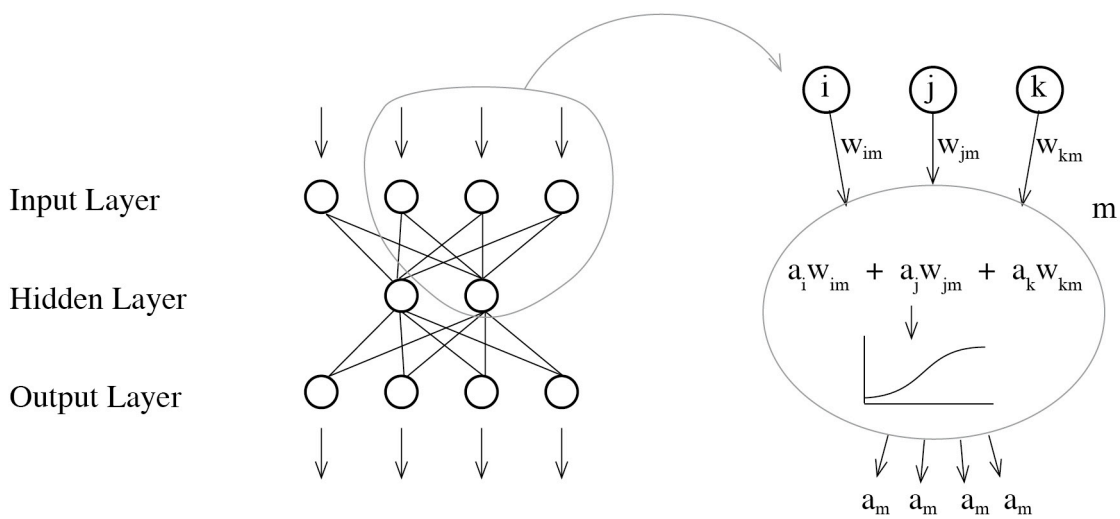

Figure 1. Information processing in a multi-layer network.

[2] See ((McLeod et al. 1998), (Mitchell 1997), and (Thomas & McClelland 2008)) for more thorough introductions to connectionist networks that go beyond this crisp overview.

C-nets and their training are characterized by a host of additional parameters (e.g., bias units that establish thresholds for activation states, plasticity parameters in adjusting connection strengths, etc.) and assumptions about the overall network architecture (number of layers, nature of representations, feed-forward vs. recurrent). We give a functional overview of these parameters for our c-net in section 3.2 and then provide more formal detail, much of which employs standard assumptions in connection science, in the appendix.

Connectionist networks are information processing systems that take inputs and produce 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 forms from English present forms in a single layer network. Problems at the heart of generative phonology have also been accounted for with connectionist networks. For example, (Hare 1990) designed a c-net to capture the core facts of Hungarian vowel harmony. Hare showed that a sequential network using a context layer, which in a sense remembers the structure of preceding elements ((Jordan 1991), (Elman 1990)), could explain the effect of similarity and proximity on vowel harmony, i.e., the fact that the more similar and closer the target and trigger are, the stronger assimilatory effect; see also (Wayment 2009) on attractor networks producing similar effects in a wider range of linguistic systems. Another example is the c-net developed in (Legendre et al. 2006) to account for the now classic OT analysis of Tashlhyt Berber syllabification (see (Prince & Smolensky 1993/2004)). This c-net takes a sequence of consonants 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.

In order to assess claims about constraint induction, it is necessary to first understand how constraints work in connectionist networks. 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)).[3] 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 spreads through the network, the activation values of individual units will change in a way that better satisfies these positive and negative constraints. This is Smolensky & Legendre's (2006) principle of Harmony Maximization.

Subsymbolic constraints are therefore not global assessments of some property of a macrostructure representation. They are the combined effect of microstructure links that can be scattered across the network. This has important consequences for constraint induction, because the problem of 'learning the constraints' can be characterized more precisely as a problem of learning the correct configuration of connection weights. Our approach to Arabic below is to

[3] We avoid the distinction between hard and soft constraints often used in comparing connectionist networks and symbolic-computational systems because hard constraints are often understood as unviolated constraints. OT uses symbolic constraints that are not hard constraints in this sense, because they can be violated if this leads to satisfaction of a higher-ranking constraint.

design a connectionist learner that can induce the correct configuration of weights from a realistic data sample.

In the context laid out above, we establish that a connectionist learning system can induce constraints from input data by learning the correct configuration of connection weights. This assumption should not be confused with the characterization frequently given to connectionist approaches to cognitive processes, namely that c-nets are complete 'blank slates' that are completely free of bias and *a priori* assumptions. Clearly, there are some assumptions that we make about the initial state of learning, described in detail below, and there are also assumptions about the model that do not change in response to data. For example, we assume a default set of initial values for connection weights, and specific plasticity and damping parameters relevant to learning. We also assume a fixed number of hidden layer units, though we vary this number in testing to find the right number of hidden layer nodes for the data. As for representations, we use a particular set of phonological features to encode input forms. This feature set constitutes a substantive hypothesis about the range of natural classes the model can be sensitive to after training. Most of these assumptions, however, are operational in nature and we think that they do not affect model performance enough to matter for our argument. The number of hidden layer nodes is crucial, however, because a particular range of units is necessary to force the network to make the right generalizations. While it is true that we do not make the number of hidden nodes a part of the learning problem in this study, we could very easily do so, because there are known protocols for pruning and sprouting hidden layer nodes in c-nets (Mitchell 1997).

The larger point, however, is that the range of possible constraints for assessing consonant combinations is largely free. For any given set of connections in our network, the set of possible constraints these connections can compute is uncountably infinite. This is because the connection weights are assigned on a continuous scale. Given the definition of subsymbolic constraints as (sets of) connections, there is an infinite set of constraints that are expressible in our network. This fact gives the connectionist approach greater descriptive capacity than other phonotactic learners, like MaxEnt grammars and Harmonic Grammar (see the math supplement of (Alderete et al. 2012) for a proof of this assertion). The open range for constraint definition in this model therefore makes the problem of learning the constraints a non-trivial one.

3.2. A connectionist model for Arabic root co-occurrence restrictions

The c-net grammar/learner for learning the OCP in Arabic is composed of two modules, an Autoassociator module and an Assessor module, as depicted in Figure 2. The Autoassociator is a single layer c-net that constitutes a simplified production module. It takes as input a triliteral root and attempts to output an identical root. Like human language production, the Autoassociator is noisy in the sense that random variation in the activation patterns may cause it to produce non-identical roots by replacing some or all of the consonants with another Arabic consonant. The output is therefore either a root identical to the input, or a non-identical root that may accidentally correspond to another attested Arabic root or is not an attested root at all.

The role of the Autoassociator in learning is like the role of production modules in many constraint-based learning systems. In constraint-ranking algorithms in OT ((Tesar & Smolensky 2000), (Tesar 2004)), for example, Production-Driven Parsing produces forms that are either correct or incorrect with respect to the target grammar. These incorrect forms, or 'errors', are important evidence in grammar learning because the learner compares these errors with target forms and uses this comparison as a way of revising the grammar. The Autoassociator plays a

parallel role: it generates forms that can be used as evidence in grammar learning, which we outline immediately below and in more detail in the appendix.

It should be said that the Autoassociator does not constitute a realistic psycholinguistic model of speech production. Its errors do not have the same structure as human speech errors, like for example exhibiting lexical or frequency biases in produced outputs (see for example (Goldrick & Daland 2009) on error structure in learning). It is rather an algorithm that transforms actual forms of the target language and supplies these outputs as evidence to a learning system. In this way, the Autoassociator plays a role that is similar to production systems in other learning systems, like computational learning in standard OT with recursive constraint demotion (see e.g., (Tesar 2004)) or error-corrective constraint weighting employed in learning Harmonic Grammars (e.g., (Coetzee & Pater 2008)), and it should be evaluated as such.[4]

The Assessor model (see Figure 2) takes a triliteral root as input and assesses it by assigning an acceptability score ranging from -1 to 1. The acceptability score is a gradient measure of the overall well-formedness of the root, and so the Assessor is a grammatical model in the sense that it assesses input forms for acceptability. The output of the Assessor, 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).

The Assessor only makes sensible classifications of the data when it has been trained, which requires the interface depicted in Figure 2 between the two modules. The training regime is described in more detail in the appendix, but in essence, the Assessor trains on the output of the Autoassociator. The network is trained on a representative data sample, namely the root list of 3,489 roots described in section 2.2. The training data was not selected from this initial large sample, or structured in any way beyond the motivated exclusion of XYY roots. If the Autoassociator gives the Assessor a form that coincides with the input to the Autoassociator (and is therefore an attested root), all connection weights and biases in the Assessor module are adjusted such that the Assessor gets closer to outputting a '1'. If instead the Autoassociator gives the Assessor a form that differs from the input to the Autoassociator (usually, but not always, an unattested root), then all connection weights and biases in the Assessor module are adjusted such that the Assessor gets closer to outputting a '-1'.

Given this overview, we can succinctly characterize the constraint induction problem with the following two questions. Will exposure to actual roots of Arabic give rise to the correct configuration of connection weights such that a trained c-net can correctly classify actual and non-actual Arabic roots? Furthermore, if the trained c-net does correctly classify the Arabic data, do the functions computed by its subsymbolic constraints resemble those employed in symbolic constraints systems, i.e., the OCP-Place constraints introduced in 2.3? This last questions is somewhat tricky to answer directly, because both the identity a constraint and its importance in a connectionist grammar is encoded as connection weights. In section 4.2, however, we give an analysis of the Assessor performance that distinguishes the connection weights responsible for constraint identity, essentially those from the input to hidden layers in Figure 2, from 'constraint

[4] However, we note that in one important aspect, this is a very conservative decision. It is well-known that speech production has a bias against sequences of similar sounds ((Dell 1984), (Dell 1986)). This avoidance of 'tongue twisters' is indeed the very structure our learning model is supposed to learn. If it can be shown that a learner can learn such patterns, even without a production system skewed toward avoiding them, then the learner will have been shown to be successful even in absence of evidence pointing to the target outcome.

weight', which are single links from the hidden layer nodes to the output node. The model therefore accounts for both constraint identity and constraint rank/weight.

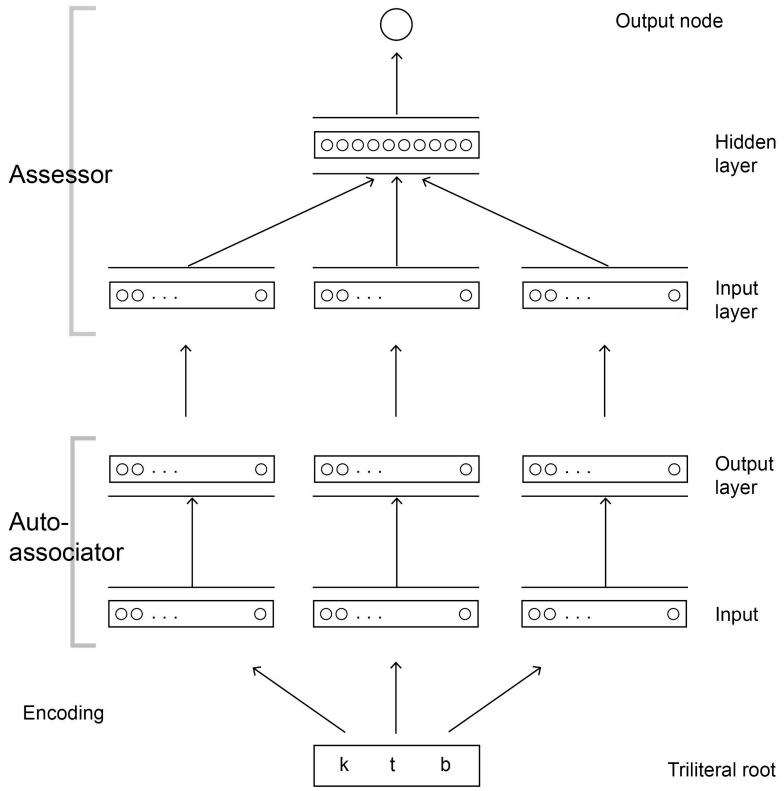

Figure 2. A feed-forward error-generating production module (Autoassociator) and a phonotactic learner (Assessor).

4. Learning results

We describe below the performance of the model after training by comparing the Assessor's output to psycholinguistic data (4.1), presenting statistical analyses of the behavior of the hidden layer nodes (4.2), and evaluating how well the Assessor module captures exceptional patterns (4.3).

4.1 Statistical effects of the OCP on acceptability

The output of the trained assessor model can be compared to human ratings of Arabic roots collected in an experiment with Jordanian Arabic speakers (Frisch & Zawaydeh 2001). However, a trained model can also be examined more broadly. Figure 3 shows the results from one version of the Assessor network with five hidden layer units. This network was tested over the entire range of possible trilateral roots ($n = 28^3 = 21{,}952$ combinations). Ratings for the actually occurring roots, shown in Figure 3a, are higher than the ratings for all possible roots, shown in Figure 3b. The range of ratings overlap, but the central tendency for the actually occurring roots is skewed toward the top of the distribution. This shows that the network has encoded patterns in the lexicon. Network ratings for novel roots from Frisch and Zawaydeh's Experiment 1 are shown in Figure 3c and 3d. OCP compliant roots in Figure 3c are rated higher than OCP violating roots in Figure 3d showing that the Accessor has encoded lexical patterns that capture the OCP.

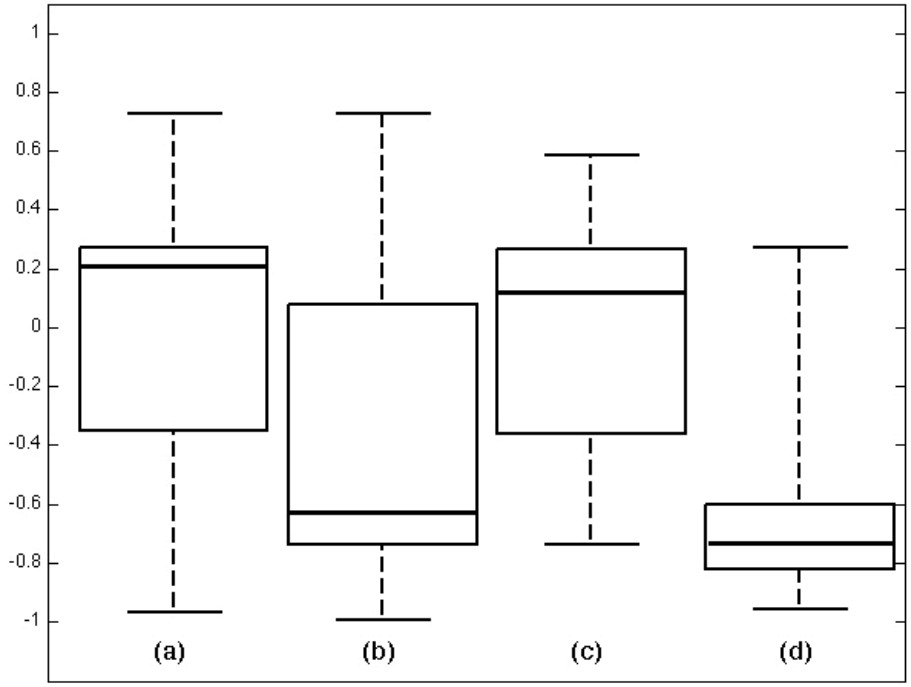

Figure 3. Acceptability scores for one trial of Assessor module with five hidden nodes. Box plots indicate minimum, first quartile, median, third quartile and maximum scores for (a) all attested roots, (b) all possible roots, (c) OCP compliant roots, (d) OCP violating roots.

A more detailed examination of the performance of the Accessor network in comparison to native speakers requires a review of the findings of Frisch and Zawaydeh (2001). They used a wordlikeness experiment to probe the psychological reality of the OCP with 24 native speakers of Jordanian Arabic. Novel word stimuli rated by native speaker participants in their experiment varied in expected probability, neighborhood density, transitional (bigram) probability, number of OCP violations, and phonological similarity of OCP-violating consonant pairs.[5] Participants rated words on a seven-point scale for acceptability as a word of Arabic. Frisch and Zawaydeh found evidence for the OCP distinct from any frequency effect and further that consonant pair similarity influenced participants' ratings, as summarized below.

Summary of results of Frisch and Zawaydeh (2001).

Experiment 1. Is the OCP psychologically real, and not just an effect of lexical statistics?

- independent variables: OCP violations, expected probability, neighborhood density

[5] The terms used in the experiment are defined as follows. '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 triliteral roots is the number of existing roots that share two of the three consonants in the appropriate serial positions. Bigram probability with respect to a given two locations in the root is the number of existing roots that have matching consonants in those locations. 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.

- 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)?

- balanced variables: expected probability, neighborhood density, bigram probability
- independent variable: OCP violation or not
- 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 of OCP violations correlate with different degrees of featural similarity?

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

To evaluate the Assessor, we examined the same questions using the Assessor module acceptability ratings. The hidden layer of the Assessor module 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 3 gives the results of a 5 unit hidden layer on three separate learning trials by way of example. All effects with significance at $p < .05$ are reported with the percentage of the variation accounted for by this effect. In addition, for the experiment 1 stimuli, the correlation between Assessor acceptability and participant mean rating is given.

Table 3. 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; network trained on Buckwalter 1997 corpus.

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

These results are qualitatively parallel to the Frisch and Zawaydeh's findings. In experiment 1 and experiment 2, the OCP violation was the only significant factor for both the model and experiment participants. Similarity was a significant factor in two of the three trials of experiment 3. Quantitatively, the OCP effects in the Assessor module are stronger than the effects found with participants, but we believe that a perfect match of the two datasets is not required to demonstrate induction of OCP constraints. Network parameters can be manipulated to produce a better quantitative match, but we believe that this sort of model tweaking is not especially insightful. The important finding is therefore that a relatively simple set of parameters reproduces all of the statistically significant generalizations in the behavioral data.

4.2 Analysis of the hidden layer units

In the network model, generalization to novel stimuli is achieved by the weighted connections through the hidden layer. Using a limited number of hidden nodes and learned connection

weights the c-net captures patterns in the lexical data without encoding each lexical item in the network. A closer analysis of the hidden layer of the Assessor network shows how the qualitative effects of symbolic constraints can be achieved through subsymbolic learning. One way to observe the nature of this learning is to extract rule-like symbolic behavior from the c-net using Classification and Regression Trees (CART). CART analysis reduces complex multidimensional data sets into a set of discrete categories using a probabilistic decision tree. The resulting tree provides a good overall fit to the data (Hastie et al. 2009). The decision tree reflects how the nodes respond to different dimensions of the data.

For each hidden node in the trained network, we applied the `classregtree` function of Matlab to produce decision trees that predict the output from the input for that node. We used all possible triliteral roots as input, based on the 51 feature specifications used by the c-net to describe each triliteral root (i.e., distributed representations for 3 segments × 17 features; see appendix). The output was an activation value for a particular hidden node, rounded to -1 or 1, where a 1 means that the node judges the form favorably, -1 unfavorably. Five trees were produced, one for each of the five hidden layer nodes.

The algorithm starts with the most predictive phonological feature and partitions the data into two sets based on that feature. Within each of those two sets, the next most predictive feature is identified and the data are further subdivided. This procedure was continued until the data was partitioned down into subsets of fewer than 1,000 roots or when partitioning the data further produced identical outputs (i.e. each subset provided the same result). At this point the tree predicts an output for any input based on the features of that input. The CART tree does not perfectly predict model performance, but it does reveal what the most important features are for determining how the hidden node reacts to different inputs.

We applied this procedure of generating CARTs for all five hidden layer nodes, for three different training 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).

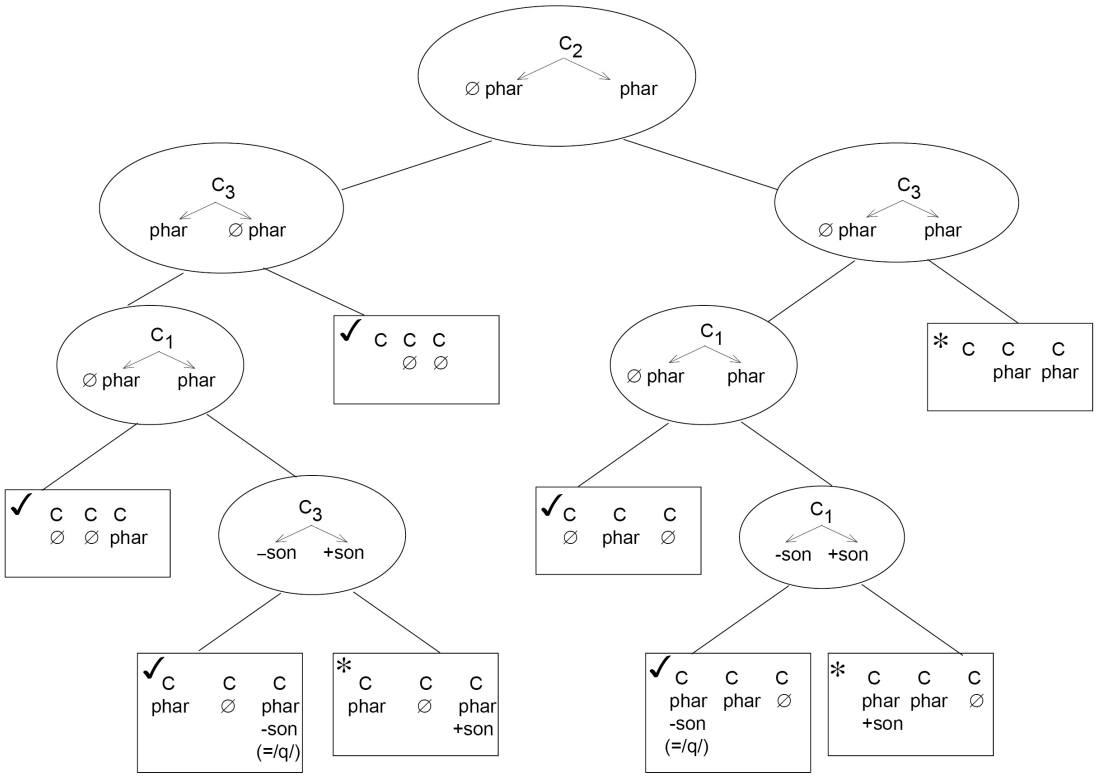

Figure 4. CART visualization of hidden layer node approximating OCP-pharyngeal. Circled nodes represent decisions about feature realization for a specified consonant, and boxed nodes represent predominant acceptable (checked) and marginal (starred) triliterals.

As an example, Figure 4 shows a CART tree capturing OCP [pharyngeal]. This particular tree is an approximation of the performance of the fifth hidden layer node in the first trial. The CART tree shows that this hidden node differentiates roots with two pharyngeal consonants from those without. The pattern */C + Pharyngeal + Pharyngeal/ emerges relatively early in the tree as unacceptable. Other forms with a single pharyngeal also emerge relatively early as acceptable. The lowest terminal nodes on the tree differentiate other combinations of pharyngeals based on the secondary [sonorant] feature, which identifies /q/ versus all other pharyngeals. The only two apparent exceptions to the OCP, patterns that are judged acceptable by the Accessor, involve the segment /q/ in either C1 or C3. However, these apparent exceptions 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 involve roots that contain /q/. There are 132 roots that contain /q/ and another pharyngeal consonant in the Buckwalter corpus. Thus, the hidden nodes are capable of approximating the functions of symbolic constraints, even when segment-level exceptions exist. In the next section, we investigate the c-net's sensitivity to a host of segment-level exceptions that are far less robust.

The CART trees above are useful for visualizing the coarse-grained nature of the functions computed by the hidden layer units. But since the CART only approximates the output of the c-net, it does not give an exact description of the behavior of a given node. To precisely quantify the relationship between the activations of hidden layer units and the OCP, the following method was used. First, we computed, for all possible triliteral roots, the violations of the specific OCP-place restrictions documented in Table 1. In particular, a violation of OCP-place requires

adjacent segments of the same class in one of the same-place classes given below in Table 5. We assign a score of -1 for that root and that class if there is an OCP violation, and a 1 otherwise. To correlate these facts with the behavior of the hidden layer, we computed the output of each hidden layer node times the weight of the connection leading out of it, for each root. This gives the acceptability of that root, as assessed by a specific hidden layer node. A positive acceptability value indicates that the node thinks the root is well-formed; a negative value corresponds to ill-formedness. We then compute the correlation coefficient of each OCP value and each hidden node's acceptability value. These correlations are shown for Trial 1 in Table 5 below (the same trial used for CART visualization in Figure 4). Results are qualitatively similar for the other trials. In Table 5 a value of 1 under a node column would indicate that a node does a perfect job of determining whether there is an OCP violation, and -1 indicates the opposite. The final column shows the correlation between the activation of the output node and the various OCP violations. This last column gives an idea of how well the entire network approximates the OCP with respect to each place of articulation. All of the cells in the last column are positive (as they are for all three trials), showing that the network assigns on average lower acceptability scores for roots that violate these specific OCP constraints than otherwise.

Table 5. Correlations between OCP violation (differentiated by place class) and weighted activations of five hidden nodes and the output node. Results from Trial 1 with training on the Buckwalter corpus. Correlations greater than 0.1 are highlighted.

Class	Node 1	Node 2	Node 3	Node 4	Node 5	Output
OCP-labial	0.0829	-0.0850	0.0583	**0.1138**	-0.0645	0.0950
OCP-cor/stop	0.0635	0.0214	0.0616	**0.1621**	-0.0859	0.1455
OCP-cor/fric	-0.1272	**0.4420**	0.0945	**0.2860**	-0.1517	0.3356
OCP-dorsal	0.0245	-0.1277	0.0948	-0.0929	**0.2544**	0.0764
OCP-phar	**0.1727**	-0.1408	**0.1160**	-0.3814	**0.7191**	0.1798
OCP-cor/son	0.0852	-0.0555	-0.0062	**0.1212**	-0.0645	0.0744

For all rows, there is a positive correlation greater than .1 between a specific OCP violation and node activation (the bold-boxed cells). In some cases, a particular node stands out as doing the work of a specific OCP-place constraint, like node 5 for OCP-dorsal. In others, the effect of the OCP for a specific place class is spread out over more than one node, for example, OCP-coronal/fricative. Furthermore, there are no strong negative correlations found in the other nodes that overpower these constraints. The highest negative correlation for the trial shown here is -0.3814 for node 4, a real outlier compared to the other negative values. But this is not strong enough to trump the effect of nodes 1, 3, and 5 in this network, which collectively approximate OCP-pharyngeal quite well.

4.3 Exceptional patterns

In section 2.2, we documented exceptions to the OCP in Arabic with a set of consonant specific patterns. The finding is that, for a pair of particular consonants, in a particular position, the distribution of the exceptions is not random. Rather, the exceptions tend to occur in specific

bigram templates. For example, out of the 21 exceptions to the rule *Labial-C-Labial, 9 are /bCm/ and 11 are /fCm/. Only 1 is /bCf/, and there are none from /mCf/, /mCb/, or /fCb/. A natural question to ask, given the network's ability to extract out /q/ from OCP-Pharyngeal constraints, is whether our c-net is sensitive to these more fine-grained exceptional patterns.

We investigate this question by comparing the Assessor's score for the exceptional patterns we have identified in Table 2 (i.e., OCP-violating patterns that occur in the lexicon), with sets of roots that violate the OCP and that do not occur in Arabic. For example, to see if the network is sensitive to exceptions of the form /fCm/, versus other similar OCP[Lab] violations, we compute the acceptability scores for all 28 roots of the form /fCm/ (28 because there are 28 consonants that could replace C), and compare them with the scores for all roots of the form XCY where X and Y are non-identical labials, and excluding all /fCm/ forms. There are 5×28 of these nonexceptional roots, because there are five other logically possible combinations of the three non-identical labials. For each consonant pair, we take the average of the score for all roots with the exceptional pair, minus the average score for all roots that violate the OCP in the same way, but with a different consonant pair. For each consonant pair and each of one of three trails of the Assessor, we get a difference in means.

The results show that there was no clear pattern to the differences between the exceptional patterns and the logically possible but unattested patterns. Thus, in each trial approximately 30 out of 81 of the exceptional patterns in Table 2 were viewed less favorably than the comparable non-exceptional pattern, contrary to the hypothesis that the network might not view (attested) exceptional patterns as violators of the OCP. Averaging the difference over all exceptional pairs yielded a mean difference of acceptability score of approximately 0.05, which is quite small relative to the difference in score between OCP-violating and OCP-compliant roots. In sum, our c-net after training does not seem to be sensitive to the fine-grained exceptional patterns in Table 2 as a whole.

Just because a pattern exists in the lexicon, however, does not mean that it is part of a native speaker's phonotactic intuitions. For example, in the English lexicon there are no instances of diphthongs before palato-alveolar fricatives, but when speakers are tested for awareness of this constraint, (for example, by comparing *foushert* with *fousert*) there is no statistically significant difference in their rankings (Hayes 2010). We cannot directly answer the question of whether native speakers of Arabic are sensitive to the patterns in Table 2 because the experiments in (Frisch & Zawaydeh 2001) were not designed to answer this question. But the pilot data available from this study does not seem to provide any support for the contention that speakers have strong intuitions of the exceptional patterns. Thus, there were 19 nonsense roots in Frisch and Zawaydeh's study that fit the templates for exceptional patterns in Table 2, and the mean of these averaged wordlikeness scores is 2.6936. The mean of mean ratings of roots (n=64) that do not fit these patterns is slightly lower at 2.4591. This is consistent with the hypothesis that native speakers rate higher the roots that fit the attested exceptional patterns, but it is impossible to tell if this difference is meaningful, given the small number of examples and inability to compare similar roots.

It is possible, however, to group classes of roots, namely certain place classes that fit the exceptional patterns, and compare their means with the means of roots that fit nonexceptional patterns. Table 6 lists the mean ratings for three non-coronal place groups (there were too few examples of roots with coronal pairs) and shows the mean ratings for exceptional vs.

nonexceptional roots and the difference of means. Except perhaps for dorsals, the data again does not show a significant trend.

Table 6. Mean wordlikeness judgments of roots with exceptional and non-exceptional OCP violating roots aggregated by selected place classes.

	Labial	Dorsal	Pharyngeal	Totals
Exceptional	2.9583	2.9275	2.455	2.7803
Non-expectional	3.1403	1.9028	2.1034	2.3822
Differences	-0.182	1.0248	0.3216	

While it is possible that native speakers are sensitive to a subset of the exceptional patterns in Table 2, we believe that the lack of evidence for a trend in this pilot data supports a conjecture that native speakers are in fact not sensitive to many of the facts at this level of phonotactic detail. This is consistent with other findings, e.g., ((Hayes 2010), (Becker et al. 2011)), and establishes a clear set of patterns that can be investigated in future experimental research. This conjecture is also consistent with our modeling results.

5. Discussion

This article has provided a cognitive architecture that makes learning the identity of grammatical constraints a significant part of learning. The web of connections in the Assessor module is a space of possible constraints, or a search space in the sense commonly used in machine learning. Different configurations of connection weights constitute different subsymbolic constraint systems. When the correct configurations are learned, the larger network can be said to have learned the target constraint system. We have shown that a two layer feed-forward network can learn the phonotactic constraints of Arabic root phonotactics by properly setting the connection weights leading into and out of a set of hidden layer units. In other words, we have shown that the functions computed by these hidden layer units after training approximate quite well the functions computed by symbolic OCP-Place constraints familiar from generative phonology. The hidden layer units do not *exactly* compute OCP-Place constraints, but this finding is consistent with the data because Arabic co-occurrence restrictions are gradient in nature and have many exceptions. The larger finding is thus that the identity of phonotactic constraints themselves can be learned from data in this case, and do not have to be stipulated in advance.

This article also makes a contribution to the issue of levels of explanation in cognitive science. 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 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 co-occurrence restrictions, can be learned at the micro-structure level and projected upward to the macro-level.

This result sets our connectionist learning system apart from many contemporary approaches to learning phonotactics. Most constraint-ranking algorithms can find the correct ranking of constraints, given the right data and a reasonable amount of time ((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. This point is self-evident for purely symbolic learning models like models that employ Tesar's recursive constraint

demotion algorithm with universal constraints, but it is less so for grammars like Harmonic Grammar that use both micro-structure (constraint weights) and macro-structure (symbolic constraints). The crucial difference between Harmonic Grammar learning and our c-net is that we make learning constraint identity a significant part of learning via backpropagation down to the input-to-hidden weight matrix. Thus, learning the first weight matrix from the output node to the hidden layer in our model (see Figure 2) is roughly analogous to the learning constraint weights in harmonic grammars. The connectionist network architecture also requires learning at a deeper level, the second weight matrix from input to hidden layer nodes. It was shown in section 4.2 that this layer roughly corresponds to the functions computed by OCP-Place constraints. There is nothing analogous to this level in Harmonic Grammar because the symbolic constraints of Optimality Theory are directly adopted and do not change in response to data.

The finding that constraints can be learned from data supports a comparison of our model to the MaxEnt phonotactic learning paradigm (Hayes & Wilson 2008). 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 (see also (Goldsmith & Riggle 2012) and (Hume & Mailhot To Appear) for recent information theoretic approaches to related problems). 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, for example, the generalizations involving suprasegmentals that Hayes and Wilson document in their study. We have explored select problems in Arabic consonant phonology that extend the core system we document here, and have found that there are also certain limits to our simple two layer system. For example, it is a basic fact of Arabic that partially similar segments are avoided, but identical segments in C2C3 position slip by the OCP constraints. This is a kind of nonlinear function that a multi-layer c-net ought to be able to learn and compute. Compare its non-linear separability (e.g., *p-b, *b-p vs p-p, b-b) with that of exclusive OR (0-1, 1-0 vs 0-0, 1-1). Yet even after extensive parameter switching with our Assessor module, pairs of identical segments are not assessed correctly if we treat them as triliterals. These simulations support the conclusions of (Berent & Shimron 1997), based on similar data from Hebrew, that linguistic constituency is necessary to express the generalization about final geminates. Similarly, the bigram templates documented in section 2 seem to be too fine-grained for our basic system to learn, though at present we do not know if native speakers also have intuitions of these templates.

We do not take these findings as insurmountable obstacles for connectionist learning models. Rather, like Hayes & Wilson, we believe they motivate additional theoretical assumptions and structure in our model. Indeed, our network is about as simple as it can get, short of a one-layer perceptron, and could be extended in a host of ways. For example, inputs could be endowed with constituency (syllables, feature classes, etc.) using tensor product representations ((Smolensky 1990), (Smolensky & Legendre 2006a)) to allow for generalization based on the role of an element in a constituent. Another option is to rely on recurrent structure to learn the constraints, like the recurrent network developed in (Alderete et al. 2012). In fact, our unpublished experimentation with such a system showed that a recurrent network trained on triliterals, including all final geminate roots, could in fact learn OCP-Place constraints. The only differences from the psycholinguistic tests documented here in Table 3 for the feed-forward network is that the recurrent network learned some small effects from neighborhood density (4%) and expected probability (1%), and had a weaker correlation between OCP violation and

similarity in experiment 3. These results and opportunities for extending the model leave us optimistic about its ability address more complex empirical patterns.

How does our c-net model differ from the MaxEnt approach generally? A basic distinction can be made by pointing out that the MaxEnt approach uses symbolic constraints and c-nets use subsymbolic constraints. In theory, this difference has empirical consequences, because the c-net constraint space is uncountably infinite and so it is richer (see (Alderete et al. 2012)). It is an open question whether this formal difference has empirical consequences that matter for the study of language. The difference between a constraint space of e.g., 300,000 symbolic constraints and an infinite set of subsymbolic constraints may not matter for most problems in generative linguistics once constraint weights are assigned to the constraints selected in a MaxEnt grammar. One might also remark that c-net learning is inherently gradual in that adjustments are made to the whole network after processing each form, while MaxEnt learning, at least as it is implemented, involves processing whole sets of language forms collectively. We do not think this is a theoretical difference of much significance, however, as there is nothing in principle that prevents an on-line version of MaxEnt learning. Indeed, Colin Wilson (personal communication) informs us that such an algorithm exists.

We think that one 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.

A step in this direction was made in (Frisch 1996) by showing a close mathematical relationship between his natural class similarity model of English and Arabic and the similarity structure predicted by spreading activation in a Dell-style production system. Indeed, the natural class model is used in this work as a closed-form approximation of a connectionist network, with the hope that future research will develop mathematically accessible connectionist models of native speaker intuitions. The present work makes a further step by showing how grammatical constraints can be represented and learned in a two-layer feed-forward network. However, problems not specific to connectionism preclude a realization of Frisch's original vision. As shown in section 3, our model has separate modules for production and assessment of phonotactic intuitions. The consequence of our decision to separate the two modules is that our

Assessor module produces acceptability scores, not linguistic representations, like the two foundational theories above. These psycholinguistic models therefore compute different functions, which must be addressed to bring our c-net closer to these models (see the recurrent network in (Alderete et al. 2012) for a start on this problem).

Another central problem is the encoding of serial order and the formalization of competitive inhibition in language processing. Our model is non-sequential in the sense that it feeds the Autoassociator and Assessor modules a preordered string of consonants. Our network is similar to the spreading activation and TRACE models, but it is distinct from other connectionist networks that model serial order as sequential output conditioned by a context layer ((Jordan 1991), (Elman 1990)); see also (Dell et al. 1997). Again, our network is non-sequential because we simply do not address the serial order problem here, but several important psycholinguistic effects depend on a competition between nodes that results from a formalization of sequences (Frisch 2004). We hope that future research can relate our findings on constraint induction to the proper characterization of serial order in respective domains of language processing.

Appendix

Assessor architecture

This appendix fleshes out the details of the connectionist network architecture. The Matlab program that we developed to implement the network is also available on the authors' webpages for further scrutiny and extension to new datasets. The Assessor is a feed-forward neural network with an input layer consisting of 51 nodes, a hidden layer consisting of 5 nodes, and an output layer consisting of 1 node (Figure 2). Every node in the input layer is connected to every node in the hidden layer, and every node in the hidden layer is connected to the output node. Each of these connections has a weight associated with it. The output node and each of the hidden nodes have a bias. These weights and biases are what is modified when the Assessor is being trained (details below). The input layer of 51 units represents a sequence of three consonants, i.e., a triliteral root, as a phonological-feature based distributed representation. We used the phonological features of (Frisch et al. 2004) for Arabic, which is essentially a variant of the widely used feature set from (Clements & Hume 1995), but adapted for Arabic. The properties of this system relevant to Arabic consonant phonology are (i) it uses primary place features, [labial], [coronal], [dorsal], and [pharyngeal], to designate the places of articulation in Arabic, and, like most prior work, (ii) uvulars are both primary [dorsal] and [pharyngeal], because they are restricted in both of these major place classes. In order to avoid the criticism that the natural class of uvulars has been constructed to fit our model results, we also assume that all uvulars, including /q/, are specified for both [dorsal] and [pharyngeal], despite the traditional grouping of /q/ with dorsals (Greenberg 1950). There are 17 phonological features in total, so, since each node encodes a single feature value, a sequence of three consonants can be represented with 51 nodes.

Unit activation states correspond to traditional feature values in the following way: '+' = +1, '-' = -1, and all trivially redundant features, i.e., features not specified for a particular segment, receive a 0. These input activations, the weights on the connections between the input layer and the hidden layer, and the biases on the hidden nodes, determine the activation of the hidden nodes. Then, the activation of the hidden nodes, together with the weights on the connections between the hidden layer and the output node, and the bias of the output node, determine the activation of the output node. The computation of activation states through the network is calculated as show below.

(A.1) Activation in the Assessor module

Let inp_i indicate the activation of the input nodes for $i = 1,...,51$, h_i indicate the activation of the hidden node for $i = 1,...,5$, and out indicate the activation of the output node. The relation between these activations is:

$$h_i = \sigma\left(\sum_j W_{1,ij} inp_j + b_i\right)$$

$$out = \sigma\left(\sum_i W_{2,i} h_i + b_{out}\right)$$

where

$W_{1,ij}$ is the weight on the connection between the jth input node and the ith hidden node

$W_{2,i}$ is the weight on the connection between the ith hidden node and the output node

b_i is the bias on the ith hidden node

b_{out} is the bias on the output node

σ is a sigmoid logistic function with $\sigma(-\infty) = -1,\ \sigma(\infty) = 1$.

Autoassociator architecture

The Autoassociator is a feed-forward network with no hidden layer (Figure 2). There are 84 input nodes and 84 output nodes (=3 slots × 28 consonants). The roots are represented in the two modules using two different representational schemes. The Autoassociator uses a so-called 'localist' representation, meaning that, for each consonantal slot, there is a single active unit that represents that consonant of Arabic in the input representation. For example, when C1 is /b/ the first unit of a sequence of 28 units is 1 and all others are 0. Though there are arguments for using localist representations for problems like the representation of concepts (see e.g., (Bowers 2009)), our motivation is purely operational. The output of the Autoassociator needs to provide unattested but logically possible roots in Arabic. Local encoding gives the required output control because the output layer can only represent roots with Arabic consonants. Alternative representation schemes do not give us this control.

The Assessor module, on the other hand, necessarily uses features-based distributed representations, rather like distinctive feature representations commonplace in generative phonology. Distributed representations are simply nonlocal representations. This means that information may be distributed across the representation (= the string of units for a consonant), and the one-to-one relationship between unit activation and consonant type in localist encoding is not guaranteed. Distributed representations are required in the Assessor module because the goal of this module is to capture feature-based generalizations about consonant cooccurrence restrictions. It is simply impossible to achieve this goal unless the activation states of units correspond to phonological feature values, values that are shared among consonants in a natural class. These assumptions therefore require that the interface between the Autoassociator and the Assessor have a conversion process that takes the locally encoded Autoassociator root and converts it to the equivalent distributed representation. This is a simple conversion process and has no function other than making it possible for the Autoassociator to 'talk to' the Assessor.

Returning to the details of the Autoassociator, each input node is connected to all output nodes with some weight. Each output node has a bias, and also receives a random Gaussian input that is regenerated every time a new root is input to the network. Because of this noise, the Autoassociator does not give the same output each time a fixed input is given, even when the weights are held constant. When a root is input to the Autoassociator, the output nodes are activated. Due to the noise in the system, the output activations are not all 1 or 0, as would be required for the output to be a root itself. The output is therefore converted to a root by taking the largest activation value for each of the three consonant slots and choosing the consonant corresponding to highest activation value.

The output activation of the ith output node of the Autoassociator is given below.

(A.2) Output activation patterns in the Autoassociator

$$out_i = \sum_j W_{a,ij} inp_j + b_i + \eta \times rand_i$$

where

$W_{a,ij}$ is the weight on the connection between the jth input node and the ith output node,

b_i is the bias on the ith output node

$rand_i$ is a random number drawn from a standard normal distribution each time the output is computed

η is a fixed parameter which specifies the amount of noise in the network. We chose η=0.325. This value yielded a mature network that gave an attested root about half the time and an error root about half the time. We have chosen an error rate of 50% for the Autoassociator for reasons of convenience only. Similar results can be obtained for a much lower error rate at the cost of running for more epochs and making the rate of Assessor training greater for errors than for non-errors. This latter modification would reasonable assuming that the fewer the errors the learner is exposed to, the more salient they are.

The vector *out* is not in the same format as the input vectors, since its elements will not typically be either 0 or 1. The output is converted to the input form by, for each group of 28 nodes, selecting the most active and setting its activation to 1, and then setting all other nodes in that group to zero activation. This procedure could be implemented by introducing inhibitory connections between the output nodes belonging to the same consonant slot, but we have chosen this simpler idealization. This result of this rounding procedure is dubbed $roundout$ below.

Training

We gave our larger network a representative sample of triliteral Arabic roots in order to give a realistic approximation of the type of data an Arabic learner would be exposed to. In particular, the Autoassociator was fed the entire 3,489 triliteral root list described in 2.2. This is all of the triliterals from the Buckwalter root list, except roots with final geminates, and represents approximately half of all roots in Arabic. As explained in 2.2, these are excluded to be consistent with prior research that treats them as biliterals.

The Autoassociator is trained using the Delta rule, a standard method in machine learning for simple one-layer networks (McLeod et al. 1998). The input is randomly selected from the list of attested roots. The network computes an output from the input activations and the randomly generated values $rand_i$ as described above. The output is then compared to the input. The weights and biases are modified based on the difference between the input and the output, as shown below.

(A.3) Delta rule for updating the Autoassociator weights and biases

$$b_i = b_i + \delta \times \left(roundout_i - inp_i\right)$$

$$W_{a,ij} = W_{a,ij} - \delta \times (out_i - inp_i) \times inp_j$$

The effect of the Delta rule is to change the weights and biases so that the actual output given the particular input is closer to the target output. In our simulations, the variables in b and W_a were

all initialized to 0. The training consisted of 10^5 epochs. In each epoch an attested root was randomly selected from the root list and input to the network. $rand_i$ was generated for each node and out and $roundout$ were computed. The expressions above were used to update b and W_a. Additionally, the weights were then reduced by a factor of $1 - \alpha\delta$ with every epoch, where α= 0.001. This is a standard technique to prevent overfitting by the network (Hastie et al., 2009). Finally, δ was chosen to vary with time so that δ was 0.1 at the beginning of the training and 0.0001 at the end of the training.

Once the Autoassociator was trained, the Autoassociator weights and biases were fixed and this module was then used to generate training data for the Assessor. The Assessor weights and biases were initialized to small random values before the training. Training consisted of 10^7 epochs. At each epoch of the training the Assessor, a root was randomly selected from the list of attested roots. The root was then passed through the Autoassociator to generate an output root. If the input root and the output root were identical, the target was chosen to be 1. If the input root and the output root were different, the target was chosen to be -1. The output root was then input to the Assessor. Based on the difference between the output from the Assessor and the target, the weights and biases of the Assessor were updated using one step of backpropagation. Backpropagation is a method for training feed-forward neural networks with one or more hidden layers (Hastie et al., 2009). For our network, backpropagation first updates W_2 and b_{out} exactly as in learning by the Delta rule, with the goal of making the output closer to the desired target. Additionally, the activations of the hidden nodes are modified, also with the goal of bringing the value of the output closer to the target. Finally, W_1 and b are modified as with the Delta rule, so that the actual activation of the hidden nodes with the given input is closer to the new hidden node activations. As with training the Autoassociator, there is a parameter δ that controls the rate of learning, which was varied from δ=1 at the beginning of the training to δ=0.1 at the end. To reduce overfitting, weights and biases were decreased by a factor $1 - \alpha\delta$ with every step, where $\alpha = 10^{-5}$.

References

Alderete, John. 2008. Using learnability as a filter on computable functions: a new approach to Anderson and Browne's generalization. Lingua 118.1177-220.

Alderete, John, Paul Tupper & Stefan Frisch. 2012. Phonotactic learning without a priori constraints: Arabic root cooccurrence restrictions revisited. In press: Proceedings of the 48th annual meeting of the Chicago Linguistics Society. Chicago: University of Chicago.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. Language 87.84-125.

Becker, Michael & Anne-Michelle Tessier. 2011. Trajectories of faithfulness in child-specific phonology. Phonology 28.163-96.

Berent, Iris & Joseph Shimron. 1997. The representation of Hebrew words: Evidence from the obligatory contour principle. Cognition 64.39-72.

Blevins, Juliette. 1997. Rules in Optimality Theory: two case studies. Derivations and constraints in phonology, ed. by I. Roca, 227-60. Oxford: Clarendon.

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.

Bowers, J. S. 2009. On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. Psychological Review 116.220-51.

Buckley, Eugene. 2003. Children's unnatural phonology. Proceedings of the 29th annual meeting of the Berkeley Linguistics Society, 523-34.

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

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

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. 1984. Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. Journal of Experimental Psychology: Learning, Memory and Cognition 10.222-33.

—. 1986. A spreading interactive theory of retrieval in sentence production. Psychological Review 93.283-321.

Dell, Gary S., L. K. Burger & W. R Svec. 1997. Language production and serial order: A functional analysis and a model. Psychological Review 104.123-47.

Dukes, Kais & Nizar Habash. 2010. Morphological annotation of Quranic Arabic. Language Resources and Evaluation Conference (LREC). Valletta, Malta.

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

Farris-Trimble, Ashley. 2008. Cumulative faithfulness effects in phonology: Indiana University.

Fikkert, Paula & Clara C Levelt. 2008. How does Place fall into place? The lexicon and emergent constraints in children's developing phonological grammar. Contrast in phonology: theory, perception, and acquisition, ed. by B.E. Dresher & K. Rice, 231-68. Berlin & New York: Mouton de Gruyter.

Flack, Kathryn. 2007. The sources of phonological markedness: University of Massachusetts, Amherst.
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., 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.
Gasser, Michael. 2009. Semitic Morphological Analysis and Generation, Using Finite State Transducers with Feature Structures. Paper presented at the Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.
Goad, Heather. 2001. Assimilation phenomena and initial constraint ranking in early grammars. Proceedings of the 25th annual Boston University Conference on Language Development, ed. by H.-J.A. Do, L. Dominguez & A. Johansen, 307-18. Somerville, MA: Cascadilla Press.
Goldrick, Matthew & Robert Daland. 2009. Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. Phonology 26.147-85.
Goldsmith, John. 1976. Autosegmental phonology: MIT Doctoral dissertation.
Goldsmith, John & Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language and Linguistic Theory 30.859-96.
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.
Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2009. The elements of statistical learning. New York: Springer.
Hayes, Bruce. 1999. Phonological restructuring in Yidin and its theoretical consequences. The derivational residue in phonological Optimality Theory, ed. by B. Hermans & M. van Oostendorp, 175-205. Amsterdam: John Benjamins.
—. 2010. Learning-theoretic linguistics: Some examples from phonology. Presentation given at 2010 Cognitive Science Conference, session in honor of Rumelhart Prize winner James McClelland. UCLA.
Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39.379-440.
Hume, Elizabeth & Frédéric Mailhot. To Appear. The role of entropy and surprisal in phonologization and language change. Origins of sound patters: Approaches to phonologization, ed. by A. Yu. Oxford: Oxford University Press.
Inkelas, Sharon & Yvan Rose. 2008. Positional neutralization: A case study from child language. Language 83.707-36.

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.
Leben, Will. 1973. Suprasegmental Phonology: MIT Doctoral dissertation.
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.
Levelt, Claartje & Marc van Oostendorp. 2007. Feature co-occurence constraints in L1 acquisition. Linguistics in the Netherlands.162-72.
Martin, Andy. 2007. The evolving lexicon: University of California, Los Angeles.
Massaro, Dominic W 1988. Some criticisms of connectionist models of human performance. Journal of memory and language 27.214-34.
McCarthy, John J. 1979. Formal problems in Semitic phonology and morphology: MIT Doctoral dissertation.
—. 1986. OCP Effects: Gemination and antigemination. Linguistic Inquiry 17.207-63.
—. 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. 1990. Prosodic morphology and templatic morphology. Perspectives on Arabic linguistics II: Papers from the Second Annual Symposium on Arabic Linguistics, ed. by M. Eid & J. McCarthy, 1-54. Amsterdam: John Benjamins.
—. 1995. Faithfulness and reduplicative identity. University of Massachusetts Occasional Papers 18, Papers in Optimality Theory, ed. by J. Beckman, S. Urbanczyk & L. Walsh, 249-384. Amherst, MA: Graduate Linguistic Student Association.
McClelland, James L. & Jeffrey Elman. 1986. The TRACE model of speech perception. Cognitive Psychology 18.1-86.
McClelland, James L. & David E. Rumelhart. 1986. On learning the past tenses of English verbs. Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume 2: Psychological and biological models, ed. by J.L. McClelland, D.E. Rumelhart & T.P.R. Group, 216-71 Cambridge, MA: The MIT Press.
McLeod, Peter, Kim Plunkett & Edmund T. Rolls. 1998. Introduction to connectionist modelling of cognitive processes. Oxford: Oxford University Press.
Mitchell, Tom M. 1997. Machine learning. Boston, MA: McGaw Hill.
Myers, Scott. 1997. OCP effects in Optimality Theory. Natural Language and Linguistic Theory 15.847-92.
Padgett, Jaye. 1995. Stricture in feature geometry. Stanford, CA: CSLI Publications.
Pater, Joe. 2009. Weighted constraints in generative linguistics. Cognitive Science 33.999-1035.
—. 2011/To appear. Canadian raising with language-specific weighted constraints. Language.
Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. NELS 23, 367-81.
Pierrehumbert, Janet B. 2001. Why phonological constraints are so coarse-grained. Language and Cognitive Processes 16.691-98.

—. 2003. Phonetic diversity, statistical learning, and acquisition of phonology. Language and Speech 42.115-54.
Pinker, Steven. 1984. Language learnability and language development. Cambridge, MA: Harvard University Press.
Plaut, David C. & Christopher T Kello. 1999. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. The emergence of language, ed. by B. MacWhinney. Mahwah, NJ: Lawrence Erlbaum Associates, Ltd.
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.
Ryding, Karin C. 2005. A reference grammar of Modern Standard Arabic. Cambridge: Cambridge University Press.
Smolensky, Paul. 1988. On the proper treatment of connectionism. The Brain and Behavioral Sciences 11.1-23.
—. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. Artificial Intelligence 46.156-216.
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.
Suzuki, Keiichiro. 1998. A typological investigation of dissimilation: University of Arizona Doctoral dissertation.
Tesar, Bruce. 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.
Thomas, Michael S. C. & James L. McClelland. 2008. Connectionist models of cognition. Cambridge handbook of computational psychology, ed. by R. Sun, 23-58. Cambridge: Cambridge University Press.
Wayment, Adam. 2009. Assimilation as attraction: Computing distance, similarity, and locality in phonology: Johns Hopkins University.
Yip, Moira. 1989. Feature geometry and cooccurrence restrictions. Phonology 6.349-74.