Emergent systemic simplicity (and complexity)*

Joe Pater
University of Massachusetts Amherst

SUMMARY

Across phonology and syntax, the typological probability of one structure being present in a linguistic system often depends on other related aspects of that system. For example, voiced [g] is more probable in a language if it contains voiced [b] than if it does not, and a left-headed PP is more probable in a language that contains left-headed VPs than in one that has right-headed VPs. These dependencies can be seen as preferences for *systemic simplicity*, for uniform expression of laryngeal contrasts across place, and for uniform syntactic headedness. Both the systemic and the probabilistic nature of these generalizations pose deep challenges for linguistic theory. I provide an account of these and some other instances of systemic simplicity in terms of an emergent learning bias, which has a probabilistic effect on typology. Given an initial state with free variation, interacting learners with the right constraint sets tend to create systems that display systemic simplicity. I also show that the same learning theory leads to emergent complexity, in that under the right conditions, contrast spontaneously emerges from variation.

RÉSUMÉ

En phonologie et en syntaxe, la probabilité typologique d'une structure étant présente dans un système linguistique repose souvent sur d'autres aspects connexes de ce système. Par exemple, un [g] voisé est plus probable dans une langue si elle contient le [b] que si elle n'en a pas, et un PP dont la tête est à gauche est plus probable dans une langue qui contient des VP dont la tête est à gauche que dans une langue avec des VP dont la tête est à droite. Ces dépendances peuvent être vues comme une préférence pour la *simplicité systémique*, pour l'expression uniforme de contrastes de voisement, et pour l'uniformité syntaxique. La nature systémique et probabiliste de ces généralisations posent des défis profonds pour la théorie linguistique. Je les analyse en termes d'apprentissage avec un biais émergeant, qui a un effet probabiliste sur la typologie. L'interaction des apprenants a tendance à créer de la simplicité systémique, si on postule les bonnes contraintes. La même théorie de l'apprentissage produit aussi de la complexité émergente : sous certaines conditions, du contraste apparaît spontanément à partir de la variation.

* It's a great pleasure to present to Glyne Piggott this paper, which I hope in some way pays tribute to his inimitable fashion of synthesizing insights from diverse frameworks, which has always deeply influenced me. I would also like to thank participants in the 2009 KNAW Academy Colloquium on Language Acquisition and Optimality Theory, the 2011 MOT Workshop at McGill, and in Ling 730, Fall 2011 UMass Amherst. Special thanks to Daniel Currie-Hall, Minta Elsman, Matt Goldrick, Alice Harris, Clint Hartzell, John Kingston, Claire Moore-Cantwell, Elliott Moreton, Paul Smolensky, Robert Staubs and Lisa Travis for particularly useful discussion, and to Magda Oiry for help with the *résumé*. This research was supported by NSF grant BCS-0813829 to the University of Massachusetts Amherst.

McGill Working Papers in Linguistics, 22.1, Winter 2012

1 SYSTEMIC SIMPLICITY: THE CHALLENGE

I use the term "systemic simplicity" to refer to linguistic generalizations that have the following form:

(1) *Structure S is more probable in context X if S also appears in context Y.*

I will first discuss typological generalizations, where "probable" refers to distributions across languages, but the definition can also be applied within a language, where probability refers to the chance of a structure being added to the language, in acquisition or in historical change; this extension will become important shortly.

Table 1 provides a typological example of systemic simplicity, due to John Kingston (p.c.). It shows the number of languages in the UPSID-92 database (Maddieson and Precoda 1992) that have both [b] and [g] (244), neither (153), just [b] (43) and just [g] (11) (all of the languages have [p] and [k]). As the Observed/Expected (O/E) values indicate, languages that have just one of the voiced stops are underrepresented; a chi-square test shows that this distribution is highly unlikely to have arisen by chance ($\chi^2 = 260$, d.f. = 1, $p < 0.01$). Although this sample has not been controlled for genetic relatedness, the skew towards languages having both [b] and [g] or neither seems likely to be strong enough to persist in further analysis.

Table 1: Distribution of voicing contrasts by place in UPSID-92

	[b]	no [b]
[g]	244 (O/E = 1.52)	11 (O/E = 0.12)
no [g]	43 (O/E = 0.34)	153 (O/E = 2.13)

This case fills out the schema for systemic simplicity as follows: voicing is more probable on a labial stop if it also appears on a dorsal stop (and vice versa). The generalization is "systemic" because we cannot just say that voiced labial stops (or dorsals) are rare: it is only the systems that have them without a voiced stop at the other place of articulation that are underrepresented.

The "simplicity" of systemic simplicity cannot be illustrated with this table. It is simple to have a voicing contrast across places of articulation insofar as this minimizes the number of features that need to be used to maintain contrasts between words in a language. The cross-linguistic tendency to use consonantal features in this economical fashion is documented statistically in Clements (2003). I will return to inventory economy in section 5.

This small example can be used to illustrate the deep general challenge that systemic simplicity poses for linguistic theory. I am unaware of any discussion of these problems with respect to phonological feature economy, but the issues are the same as those discussed for the largely parallel case of cross-linguistic tendencies toward uniform phrasal headedness across syntactic categories (see e.g. Aristar 1991, Dryer 1992). First, the systemic nature of these generalizations means that they cannot be expressed in standard theories of grammar, which evaluate a single representation at a time for well-formedness, or which derive a single

representation from another. For this example, such theories do not provide the formal means to state a constraint that allows a voiced dorsal only if the inventory also contains a voiced labial, or to write a rule that devoices a voiced dorsal if the inventory fails to have a voiced labial. Second, if a theory did permit these sorts of systemic constraints or rules, it would also have to deal with the probabilistic nature of the generalizations: languages that have only the voiced labial or dorsal are only relatively rare, not unattested. Standard theories of generative grammar only deal with categorical distinctions in cross-linguistic attestedness, whether language types exist or not, and not with finer degrees of probability (see also Coetzee 2002 and Bane and Riggle 2008 on probabilistic typological generalizations in phonology).

The problem of systemic simplicity might seem to be resolved by including in one's theory of grammar constraints or rules that state a general prohibition against, or preference for, structure *S* across contexts *X* and *Y*. For example, in Optimality Theory (OT: Prince and Smolensky 1993/2004), it is standard to invoke a general *Voice constraint that penalizes both [b] and [g], and a general Ident-[Voice] constraint that penalizes their devoicing. In fact, no one has suggested that these constraints account for feature economy, and it is easy to see why. The constraint set with just these two constraints is incomplete for the full typology, since they would only generate the two more common language types that either have or lack voicing across both places of articulation. Once we add labial- and dorsal-specific versions of *Voice or Ident-Voice to our constraint set, we can generate the full typology, but we have no account of the typological skew away from voicing contrasts at a single place of articulation. The insufficiency of general constraints as an account of systemic simplicity is not specific to Optimality Theory, nor is it specific to phonology; see e.g. Travis (2011) for parallel discussion with respect to micro- and macro-parameters in syntax.

In what follows, I show that general constraints can have the desired effect of producing skews toward systemic simplicity if we adopt a probabilistic grammar formalism and a broadly used learning algorithm (see also Martin to appear for related results). In this account systemic simplicity is "emergent" in that there are no systemic constraints of the type "*S* in *X* only if *S* in *Y*". It is also emergent in that general cross-context constraints are standardly used in linguistic theory, and in that the learning algorithm I adopt is widely used in cognitive modeling and machine learning; neither was proposed to deal specifically with systemic simplicity.

In the next section I introduce the grammar and learning models, and show that a voiced stop is learned more quickly if the inventory also contains voiced stops at other places of articulation. Section 3 turns to the modeling of regularization in language learning and historical change, with an example from lexical stress. Section 4 shows how a tendency toward uniform syntactic headedness emerges from learner interaction. Section 5 returns to phonological feature economy, showing that featural contrast and economy also emerge from learner interaction.

2 SYSTEMIC GENERALIZATION AND LEARNING RATE

The present account of systemic simplicity is essentially a formalization of Martinet's (1968: 483) explanation for phonological feature economy:

(2) ...each of [the features] being more frequent in speech, speakers will have more occasions to perceive and produce them, and they will get anchored sooner in the speech of children. A phoneme that is integrated into one of those bundles of proportional oppositions which we call 'correlations' will in principle be more stable than a non-

> integrated phoneme.../ɵ/ and /ð/ in English have been preserving for centuries their practically worthless opposition simply because they are perfectly integrated in the powerful correlation of voice....

Here, I will provide a learning simulation that shows that [g] is "anchored sooner" when the inventory contains [b] and [d] than when it lacks voiced consonants at the other places of articulation. The two languages are shown in (3):

(3)

bdg	*ptg*
Word 1 [pi]	Word 1 [pi]
Word 2 [ti]	Word 2 [ti]
Word 3 [ki]	Word 3 [ki]
Word 4 [bu]	Word 4 [pu]
Word 5 [du]	Word 5 [tu]
Word 6 [gu]	Word 6 [gu]

Both languages contain six words with initial consonants at three places of articulation, labial, coronal and dorsal, and both have a single voiced dorsal [g]. The difference between them lies in whether they have voiced [b] and [d] in Words 4 and 5: *bdg* does, while *ptg* does not.

I'll first describe the model of grammar assumed in this experiment, and with minor modifications, in all the subsequent experiments below. It is a probabilistic variant of Harmonic Grammar (HG; see Smolensky and Legendre 2006 and Pater 2009, to appear for introductions and overviews), in that it uses weighted constraints to define a probability distribution over candidates. The specific model I adopt here is generally referred to as Maximum Entropy Grammar (MaxEnt; Goldwater and Johnson 2003, Wilson 2006, Jäger 2007, Hayes *et al*. 2009; see Coetzee and Pater 2011 for a comparison with other models of phonological variation). A MaxEnt grammar defines the probability of a candidate as proportional to the exponential of its Harmony, that is, of the weighted sum of constraint violations. I will explain further using the constraint set for this experiment.

There are two types of constraint. The constraints in (4a.) are output constraints (usually called markedness constraints in OT). *Voice penalizes voicing on all stops, while the three constraints *[b], *[d] and *[g] penalize voicing on each of the three stops. "Word-X-Feature-F" in (4b.) represents a set of constraints that demand that each word bear a particular set of features in the output. These constraints collapse the roles of underlying representations and faithfulness constraints in OT (see also Hare and Elman 1995, as well as Aronoff and Xu 2010 and the references cited therein). Following Aronoff and Xu (2010), I will refer to these as Realization constraints. For each of the 6 words, there are two constraints, each one demanding the features that make up the phonetic forms in (3), as well as the minimally different form that differs in voicing (e.g. for Word 1, there is W1-[bi] and W1-[pi]).

(4) a. *Voice, *[b], *[d], *[g]
b. Word-X-Feature-F

The candidate sets for each Word are similarly limited to the voiced and voiceless forms of each stop. The tableau in (5) shows the candidate set for Word 4, and all of the relevant constraints. Violations are indicated with negative integers.

(5) *Candidate set for Word 4*

	p	*H*	W4-[bu] 0	W4-[pu] 0	*Voice 10	*[b] 10
a. Word 4, [pu]	1	0		–1		
b. Word 4, [bu]	0	–20	–1		–1	–1

This tableau also shows the initial constraint weights under each constraint name. Output constraints were given an initial weight of 10, while the Realization constraints were given an initial weight of 0. The weighted sum of violations, or Harmony, of each candidate is shown under the column headed by *H*, while the probabilities of the candidates (which are proportional to exp(*H*)) are indicated under the heading *p*. As the Harmony of one candidate is lowered relative to another, so is its probability: with these initial weights, the probability of the candidate with the initial voiced obstruent approaches 0. This instantiates the standard OT approach to the initial state (see Jesney and Tessier 2011 for references and an HG implementation), in which distinctions are collapsed in favour of predictability and/or articulatory ease.

I'll now describe the learning procedure, which again is essentially the same for all of the experiments reported here. I use an update rule that would likely be referred to as a variant of the Delta Rule in connectionist modelling (Rumelhart and McClelland 1986a *et seq*.), and as the Perceptron update rule in the natural language processing literature (see Johnson 2007). This update rule has also been broadly used in the linguistic literature. For all of the cases here, it is identical to Boersma's (1997) Gradual Learning Algorithm for Stochastic OT (see Pater 2008, Jäger 2007, and Boersma and Pater to appear on the difference). Jäger (2007) in fact presents an earlier phonological implementation of exactly this grammar and learning model, whose update rule he refers to as a sampling version of Stochastic Gradient Ascent,[1] while Boersma and Pater (to appear) and others use this learning algorithm with Noisy Harmonic Grammar. The present choice of MaxEnt as the grammar model was one of convenience: it allows calculation of probabilities without the sampling or integration needed for Noisy Harmonic Grammar.

Each learning trial begins with the random selection of one of the 6 correct (Word, SR) pairs from (3). In OT learning terminology (Tesar and Smolensky 2000), this is the Winner. The learner then samples from the probability distribution defined by the grammar over candidates for the same Word. If the form that the learner's current grammar generates does not match the learning datum (and is hence a "Loser" in OT terms), the constraint weights are updated. This is done by taking the difference between the violation scores of the Winner and Loser, scaling this difference vector by the learning rate, and adding the resulting values to the constraint weights. The construction of a difference vector $W - L$ is illustrated in (6) with Winner (Word 4, [bu]) and

[1] It is not clear why Jäger (2007) used the sampling version (AKA Perceptron); see Johnson (2007) for the formulation of Stochastic Gradient Ascent in the context of a general discussion of learning of MaxEnt grammars. Stochastic Gradient Ascent could have been used for the simulations here, though in initial explorations I found that it has somewhat weaker tendency toward regularization, which would have led to less striking experimental results (but not necessarily less accurate ones, given the highly idealized nature of these simulations).

Loser (Word 4, [pu]). The weight of the constraint W4-[bu] would be increased, while the weights of W4-[pu], *Voice, and *[b] would be decreased.

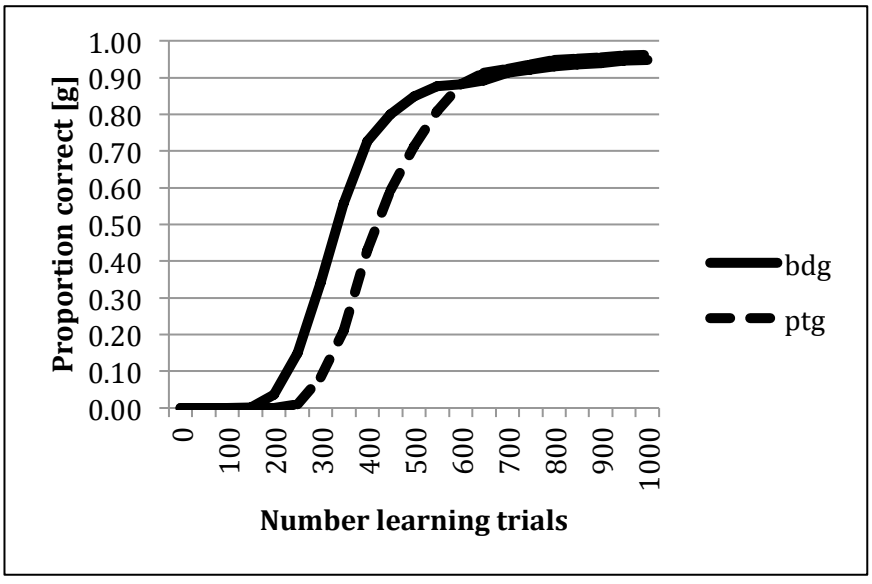
(6) *A Winner, Loser, and difference vector for datum Word 4, [bu]*

	W4-[bu]	W4-[pu]	*Voice	*[b]
Winner Word 4, [bu]		–1	–1	–1
Loser Word 4, [pu]	–1			
W – L	+1	–1	–1	–1

In this and all following experiments, the learning rate was set at 0.1, so 0.1 would be added or subtracted from each constraint in this example. A zero minimum was imposed on the output of the learning rule so that constraints would retain their intended meaning (e.g. negatively weighted *Voice would penalize voiceless rather than voiced stops).

The graph in Figure 1 plots the proportion of [g] for Word 6 that the learners' grammars generate across the first 1000 learning trials, sampled at 50 trial intervals, averaging across ten runs for each of the *ptg* and *bdg* datasets. Between trials 200 and 500, the *bdg* learners on average assign more probability to correct [g] than do the *ptg* learners. After that point, the *ptg* learners do very slightly better. I have not subjected these results to statistical analysis, but based on my experience with many other similar simulations, it seems that the early *bdg* advantage is extremely reliable. It isn't clear whether the later slight *ptg* advantage does generalize, but see section 5 for related discussion.

Figure 1: Proportion [g] for Word 6 across the first 1000 trials



The advantage of a *bdg* learner derives from the structure of the constraint set and the nature of the update rule. To see this, consider what happened when a learner received the learning datum (Word 4, [bu]), and generated (Word 4, [pu]) as a Loser. Note that on the basis of a datum containing a voiced labial, the learner will not only decrease the weight of the constraint specific to that structure, *[b], but it will also decrease the weight of the more general constraint that penalizes all voiced stops, *Voice. Because the probability of (Word 6, [gu]) depends not only on the weight of *[g], but also on that of *Voice, the weight update based on (Word 4, [bu]) will increase the probability of correct [gu] as well as correct [bu]. That is, with this update rule, the presence of the general constraints in the constraint set leads to generalization in learning.

Moreton and Pater (2011) survey a range of laboratory learning studies that show that formal simplicity leads to faster learning, and Pater and Moreton (in prep.) discuss the modelling of these results in terms very similar to those here. I don't know of any direct evidence that [g] is acquired more quickly when [b] is also present in a language, as this learning simulation predicts; the focus here is on showing how the learning bias could have an impact on linguistic typology, as Martinet (1968) speculated.

3 STRESS REGULARIZATION

I'll now turn to a quite different case of systemic simplicity, to illustrate the broad scope of this approach, and to begin to show how the learning bias could have an effect on the synchronic shape of languages. Words that disobey an otherwise general pattern of stress placement are often regularized in acquisition (e.g. Hochberg 1988), and in diachrony (e.g. Phillips 1984, Sonderegger and Niyogi in press). To see how this instantiates systemic simplicity, consider the toy languages in (7).

(7) *Two hypothetical lexical stress systems*
a. 9 words with stress on the rightmost syllable
b. 9 words with stress on the leftmost syllable

The question is which one would be more likely to add a tenth word with stress on the rightmost syllable. Insofar as languages tend towards regularity, language (7a.) would be the more likely to add a finally stressed word. This is systemic simplicity in that the probability of a word with rightmost stress depends on the presence of other words with that same stress pattern.

The following experiments show how this result follows from the same types of assumptions about learning and constraints as in the last section. The constraint set is as in (8):

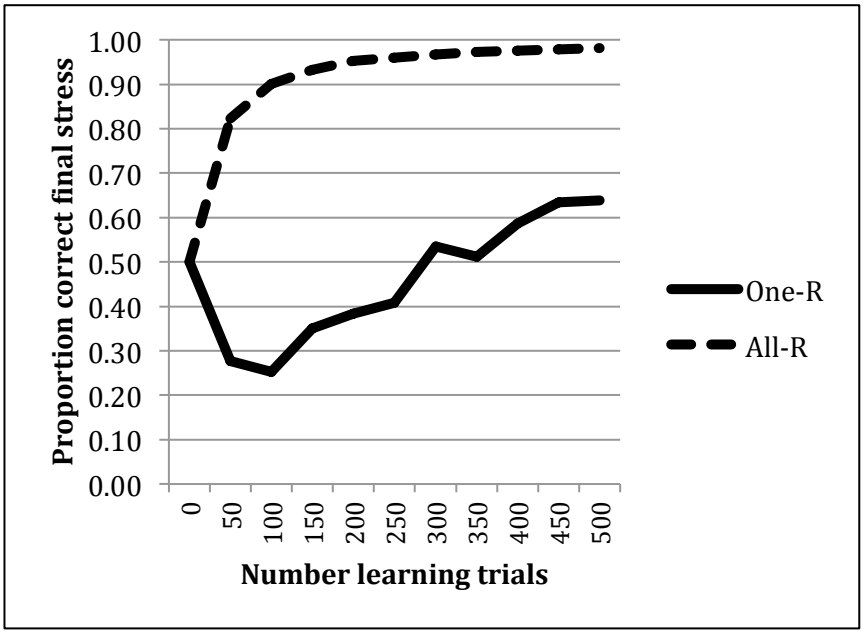
(8) a. Stress-R/L
Assign a violation if stress is not on the right-/leftmost syllable
b. Word-X-Stress-R/L
Assign a violation if stress on Word X is not on the right-/leftmost syllable

As in the last experiment, there is a general constraint that applies across contexts, as well as constraints that apply in the specific contexts, though here the contexts are words, rather than places of articulation. This constraint set also differs in that it is not biased toward one pole of the opposition: instead of the general *Voice constraint, we have general constraints preferring each of the two options for stress placement.

The languages are as just described: one language has ten words with stress on the rightmost syllable, and the other has nine words with leftmost stress and one word with stress on the rightmost syllable. Each word has a pair of associated specific stress constraints, demanding stress on each side, and the candidates are the two stressings of each word. The first experiment tests how quickly final stress is acquired in each instance. The learning parameters were the same as in the last experiment, except that all constraints started out with zero weight.

Figure 2 shows the result over ten runs for each set of data, in terms of the proportion of the probability assigned to the rightmost syllable of words that are correctly stressed that way. "All-R" is the language with ten words with rightmost stress, and the proportion is the average across all ten words, averaged again over the ten runs. "One-R" is the language with just one word with final stress, and the proportion is for that word averaged over ten runs. After 500 trials, the proportion of correct final stress is already getting very close to 1 for the All-R language, while it is just over 0.60 for One-R. The All-R learner reaches 0.95 correct on final stress after 200 trials; the One-R learner does not reach that criterion until after 950 trials.

Figure 2: Proportion correct rightmost stress across the first 500 trials



Notably, the early trials show a dip in performance for One-R. At the beginning, the zero weights assign equal probability to the two stress patterns. Shortly thereafter, the majority of the words with initial stress lead to a relatively quick raising of the weight for the Stress-L constraint, and a corresponding lowering of the probability for correct final stress on the single word with that pattern. Regularization is being produced by the general constraint, combined with the effect of the weight update rule, much in the same way that learning on [b] generalized to [g] in the learning of the *bdg* language.

The second experiment with this constraint set examines what happens if two learners begin to interact after each having received 500 pieces of data from the All-R and One-R distributions. This experiment is inspired by earlier demonstrations that learning biases can be amplified and leave an imprint on linguistic systems if incomplete acquisition is transmitted from one learner to another (Hare and Elman 1995; see Zuraw 2003 and Wedel 2011 for overviews of agent-based learning in linguistics; see also Kirby and Hurford 2002 for a review of the related literature on iterated learning and language evolution and change). After the two learners are trained on the initial target distribution in the manner already described, a second phase begins in which each trial has one of the learners randomly chosen to be the 'teacher' from whom the other learns. As before, the trial begins by randomly choosing a Word, but now the pairing with the phonetic form is generated by sampling from the probability distribution defined by the current state of the teacher's grammar. This (Word, SR) pair then becomes the Winner for learning by the other 'agent'.

The full procedure consists of initial learning on the 500 pieces of data for each of the two learners (a sort of childhood), followed by 10,000 interaction trials (a kind of adolescence). Throughout, the learning parameters remained the same (learning rate = 0.1, zero minimum enforced). I ran this procedure 50 times for each of the initial One-R and All-R distributions described above. I then averaged over the probabilities assigned to each candidate by the two agents (which were always quite close). As a criterion for rightmost stress having been retained on Word 10, which had final stress in the original distribution in both cases, I required that its probability be greater than 0.8. The results are provided in (9); over twice as many of the runs for All-R produced retention of final stress.

(9) *Ratio of final stress retained on Word 10*
a. All-R: 41/50
b. One-R: 19/50

We have thus seen that the regularization observed in learning can have an impact on the final shape of a language when learners interact. To provide more information on the outcome of this interactive learning, (10) and (11) give the average probabilities that the two learners' final grammars assign to rightmost stress for each of the words for the first 12 of the 50 runs.

The column headed by 'T' indicates the probability in the original learning data, or Target. For the One-R experiments, shown in (10), rightmost stress had zero probability in all but the last word, labelled W10 in the table. Each of the columns R-1 through R-12 presents the outcome of a separate run, averaged over the two learners. All of the runs except R-3, R-9 and R-11 were taken as instances of loss of final stress on W10 for the figure in (9). Notably, in 7 of the 9 cases of loss of final stress on W10 (all but R-2 and R-10), the probability is quite a bit lower than the average outcome after 500 pieces of learning data shown in Figure 2, which is 0.64. As we will also see in the following experiments, the general tendency is for learners to come to agree on a single outcome, that is, to push the probabilities of a given candidate toward 1 or 0. Combined with the generalization of initial stress driven by the relatively high weight of the Stress-L constraint, we get the outcome that final stress on W10 tends toward 0 probability.

(10) *Average probabilities of rightmost stress, One-R*

	T	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9	R-10	R-11	R-12
W1	0	0.01	0.04	0.98	0.04	0.01	0.01	0.02	0.01	0.78	0.08	0.03	0.01
W2	0	0.01	0.03	0.99	0.03	0.01	0.01	0.03	0.01	0.51	0.06	0.02	0.01
W3	0	0.00	0.04	0.99	0.02	0.02	0.02	0.01	0.01	0.45	0.03	0.03	0.02
W4	0	0.01	0.04	0.99	0.08	0.02	0.05	0.01	0.01	0.55	0.04	0.02	0.01
W5	0	0.01	0.09	0.96	0.60	0.03	0.01	0.02	0.03	0.96	0.18	0.01	0.02
W6	0	0.01	0.11	0.99	0.12	0.01	0.02	0.01	0.01	0.36	0.22	0.03	0.01
W7	0	0.01	0.03	0.99	0.06	0.01	0.01	0.03	0.01	0.19	0.04	0.04	0.02
W8	0	0.01	0.03	0.95	0.02	0.01	0.01	0.01	0.01	0.33	0.02	0.01	0.03
W9	0	0.01	0.17	0.95	0.03	0.02	0.02	0.01	0.01	0.58	0.14	0.04	0.04
W10	1	0.13	0.50	1.00	0.05	0.05	0.09	0.15	0.15	0.99	0.60	0.83	0.04

The picture is of course quite different for the All-R learners, but some commonalities emerge. Most of the runs shown in (11) resulted in retention of final stress on Word 10. For the two that did not, R-9 and R-11, the loss of final stress is accompanied by a change to the rest of the system: all of the other words have probabilities tending toward zero. That is, even though final stress is lost, we still see a tendency toward systemic simplicity.

(11) *Average probabilities of rightmost stress, All-R*

	T	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9	R-10	R-11	R-12
W1	1	0.99	1.00	0.99	0.99	0.99	0.98	0.97	0.99	0.06	0.93	0.10	0.99
W2	1	0.99	0.99	0.97	1.00	1.00	0.99	0.96	0.99	0.20	0.92	0.03	0.99
W3	1	0.99	0.99	0.98	0.98	0.99	0.92	0.97	0.99	0.03	0.90	0.05	0.99
W4	1	1.00	0.99	0.99	0.99	0.99	0.97	0.97	0.99	0.05	0.96	0.12	1.00
W5	1	0.99	0.99	0.98	0.99	0.99	0.96	0.98	0.99	0.09	0.93	0.07	0.99
W6	1	0.99	0.99	0.99	0.99	1.00	0.96	0.97	0.99	0.09	0.93	0.09	0.99
W7	1	0.99	0.99	0.99	1.00	0.99	0.96	0.98	0.98	0.03	0.96	0.09	0.99
W8	1	0.99	0.99	0.99	0.99	0.99	0.96	0.97	0.99	0.03	0.97	0.10	0.99
W9	1	0.99	0.99	0.98	1.00	0.99	0.97	0.97	0.99	0.13	0.97	0.05	0.99
W10	1	0.99	0.99	0.99	1.00	0.99	0.99	0.96	0.99	0.01	0.92	0.12	0.99

I should emphasize that this experiment is intended only as a "proof of concept" that this grammar and learning model can produce regularization, and is not intended as a realistic model of any instance of linguistic change (whose development is an important direction for further research). I should also address a potential assessment of this demonstration of regularization as completely trivial, insofar as the postulation of general stress rules or constraints would seem to

lead inexorably to this sort of result. For a fully explicit account, however, one needs a learner with probabilistic success in the learning of exceptional stress (or some other explanation for why the exceptional case should be lost). There is no such proposal in the stress learning models of Dresher and Kaye (1990) or Tesar and Smolensky (2000), for instance. That is, regularization of exceptions is another instance of a tendency toward systemic simplicity that challenges many versions of current linguistic theory. It might also seem to be trivial insofar as it simply replicates results of Rumelhart and McClelland (1986b) (see *esp*. their rule of 78), Hare and Elman (1995), and other connectionist work. Here, the hope is that the broad applicability of constraint-based grammars to the analysis of regularities in natural language, which has been demonstrated in nearly two decades of sustained research on OT, will allow for the generalization of these sorts of results to a wider (or different) set of phenomena than would be possible with explicitly connectionist models, which tend to be more difficult to implement with highly structured linguistic representations (*cf.* Smolensky and Legendre 2006).

4 EMERGENCE OF CONSISTENT HEADEDNESS

In this section I present an experiment that is in many respects similar to the last one, though we move from phonology to a tentative exploration of some apparent parallels in syntax. In the last experiment, the general constraints were ones that applied across lexical items, and the specific constraints were those that applied to individual words. In this case, the general constraints apply across syntactic categories, while the specific constraints apply to individual categories. The constraints control the position of heads within their phrases, demanding that they be either leftmost or rightmost. As shown in (12), I use 8 specific constraints, and 2 general ones (there are left and right versions of each of the 5 in (12)). Pending a more serious investigation of the syntactic data, I have fairly arbitrarily assigned the heads and phrases with category labels.

(12) *Syntactic constraints*

Head-L/R	Assign a violation if the head is not left-/rightmost in its phrase
VP-L/R	Assign a violation if the head is not left-/rightmost in a verb phrase
PP-L/R	Assign a violation if the head is not left-/rightmost in a prepositional phrase
DP-L/R	Assign a violation if the head is not left-/rightmost in a determiner phrase
AP-L/R	Assign a violation if the head is not left-/rightmost in an adjective phrase

This experiment also diverges from the last one in that it eliminates the initial "childhood" phase. Instead, the weights start at zero and the learners begin to immediately interact. The zero weights produce equal probability for the two headedness possibilities for each type of phrase. The parameters were otherwise the same as in the last experiment, and I again ran 50 runs.

This time, the measure of interest is the tendency toward headedness being the same across categories. For each phrase type in each run, I simply took the candidate that got greatest probability as being the choice of headedness. Since there are 4 phrase types each with a binary choice, there are $2^4 = 16$ possible headedness combinations across them. Of these, only $2/16 = 0.125$ have consistent headedness. This thus provides the baseline probability of consistent headedness. In the outcome of the experiment, $40/50 = 0.80$ of the runs produced consistent headedness. I haven't submitted this difference from chance to a statistical test, but it seems reliable.

To again provide a more fine-grained view of a subset of the results, I show in (13) the outcome for the first 12 of 50 runs, each time averaged over the two learners. For each row, the column headed 'Phr.' indicates the phrase type, and 'H' the head location. After 10,000 trials, the learners have moved quite far from the uniform 0.50 probability of the initial state. It may well be of some independent significance that a tendency toward categorical outcomes is emerging from the interaction of learners operating with probabilistic grammar models. On top of this, we clearly see the tendency toward systemic simplicity: only R-9 and R-10 do not have the head in the same location across categories.

(13) *Average probabilities assigned to each head location*

Phr.	Head	R-1	R-2	R-3	R-4	R-5	R-6	R-7	R-8	R-9	R-10	R-11	R-12
VP	L	0.04	0.98	0.98	0.00	0.98	0.37	0.02	0.01	0.09	0.85	0.01	0.98
	R	0.96	0.02	0.02	1.00	0.02	0.63	0.98	0.99	0.91	0.15	0.99	0.02
PP	L	0.03	0.99	0.66	0.03	0.97	0.05	0.17	0.02	0.25	0.23	0.01	0.99
	R	0.97	0.01	0.34	0.97	0.03	0.95	0.83	0.98	0.75	0.77	0.99	0.01
DP	L	0.02	0.99	0.85	0.02	0.99	0.03	0.06	0.01	0.74	0.96	0.02	0.98
	R	0.98	0.01	0.15	0.98	0.01	0.97	0.94	0.99	0.26	0.04	0.98	0.02
AP	L	0.11	0.98	0.95	0.01	0.99	0.37	0.01	0.01	0.03	0.35	0.00	0.98
	R	0.89	0.02	0.05	0.99	0.01	0.63	0.99	0.99	0.97	0.65	1.00	0.02

The movement away from 0.50 probability is a sort of "rich get richer" effect. When the 'teacher' picks a form and the learner disagrees, the update will move probability toward the teachers' form. Over time, this process will tend to accumulate probability on one of the choices for headedness in each of the categories. The consistency across categories emerges from the activity of the general constraint. As an illustration, consider the update when the teacher supplies a left-headed VP, and the learner's grammar generates a right-headed one. The difference vector used in the update is shown as $W - L$ in (14).

(14) *A syntactic Winner-Loser pair*

		Head-R	Head-L	VP-R	VP-L
VP	Winner: Left	–1		–1	
	Loser: Right		–1		–1
	$W - L$	–1	+1	–1	+1

Based on this example of a left-headed VP, the learner will not only raise the weight of VP-L and lower the weight of VP-R, but it will also raise and lower the weights of the general Head-L and Head-R constraints. This increases the probability assigned to left headed phrases in general, and

hence increases the probability that when in the future this learner is the teacher it will produce a left-headed phrase as a learning datum, leading eventually to the "rich get richer" snowballing.

To verify that it is the inclusion of the general constraint that leads to the emergence of systemic simplicity, I ran the experiment again without it. This time, consistent headedness was only produced in 5 out of the 50 runs. This 0.10 probability comes close to the 0.125 predicted by chance, unlike the 0.80 probability of consistent headedness produced when the general constraint was included.

As mentioned in the introduction, consistent location of heads across phrases in a language is an observed cross-linguistic synchronic tendency (see Greenberg 1966 and Dryer 1992, amongst many others). Change toward consistent headedness is also an observed diachronic pattern (see e.g. Pintzuk 1999, Harris 2000). The source of this tendency is a matter of some controversy; the usual account seems to be a parsing based one (see Dryer 1992 for a review of the literature, as well as the modelling in Kirby 1999), though learning based accounts have also been proposed (on modelling, see Christiansen and Devlin 1997, Brady *et al.* 2008 and the papers cited there as well as Culbertson 2010). I am forced to leave a fuller exploration of the present approach and comparison with previous accounts to further work, but it's worth noting its potential to unify the parsing and learning explanations, insofar as the learning model here is interpreted as a model of the learning of parsing. Given the central role of prediction in parsing theories, it's likely an important feature of the learning algorithm that it incorporates the generation and testing of predictions.

5 EMERGENCE OF ECONOMIC INVENTORY STRUCTURE

We now return to the case we began with: phonological feature economy. Given the results of the intervening sections, one might imagine that if we extended the experiment in section 2 by adding a stage of learner interaction we would get greater loss of [g] when the learners start off by being exposed to the *bdg* distribution than when they initially get the *ptg* one. I in fact imagined just that, and ran a series of experiments that aimed to get this result, using a wide variety of permutations of the datasets, and the learning and grammar models. After a long series of failures, with initial *ptg* learners having just as strong a tendency for [g] preservation as initial *bdg* learners, I finally realized the flaw in this reasoning. Even though [g] is indeed reliably initially learned more readily if [b] and [d] are in the language, this does not mean that the full [p, t, k, b, d, g] system has an advantage over [p, t, k, g], and my failed experiments seem to indicate that it does not. A typical result was that both systems would tend to collapse to [p, t, k] or [b, d, g], both of which are ideal in terms of systemic simplicity. Furthermore, a weighting that produces generally voiceless consonants, with the exception of words with [g], does not seem to be any harder to reach than one that allows both voiced and voiceless consonants at all places of articulation.

One functional advantage of the *bdg* system is that it has more contrasts. In the successful experiment I will present in this section, the functional role of contrast is formalized by building on results in bidirectional OT (see Boersma and Hamann 2008 and papers cited there; see Wedel 2004 for a different approach to contrast maintenance). As in the interactive learning experiments presented in the last two sections, a randomly chosen teacher produces a form for learning by the other agent. This time, however, the learner is not supplied with the Word that the phonetic form is associated with, and there are multiple Words that may correspond to a phonetic form For example, [bi] is a possible phonetic form for both Word 1 and Word 2, as so a learner is given

just [bi], and must interpret it as the realization of Word 1 or Word 2. This is a case of hidden structure learning (Tesar and Smolensky 2000), and I adopt a probabilistic version of Tesar and Smolensky's (2000) Robust Interpretive Parsing to deal with it. The decision between underlying Words is made by sampling from the probability distribution defined by the grammar, in particular by the weighting of the Realization constraints that link phonetic form with meaning. The learner then proceeds as usual, generating its own phonetic form for the chosen Word, and updating the weights in case of mismatch.

Like the *bdg*/*ptg* experiment in section 2, there are 6 words at three places of articulation, but there are now three possible phonetic forms for each word: ones with voiced, voiceless or aspirated versions of the initial consonants. I have added another laryngeal feature so as to be able to see if the inventories that emerge from learner interaction tend to be economical, in Clements' (2003) sense. Will the learners tend to develop inventories that use the same laryngeal contrast across places of articulation? The candidate sets are shown in (15).

(15)

Word 1 [bi]/[pi]/[p^hi]	Word 4 [bi]/[pi]/[p^hi]
Word 2 [di]/[ti]/[t^hi]	Word 5 [di]/[ti]/[t^hi]
Word 3 [gi]/[ki]/[k^hi]	Word 6 [gi]/[ki]/[k^hi]

The constraint set included three Realization constraints for each Word, which demanded each of the three phonetic forms in (15). I assumed three monovalent laryngeal features [Voice], [Plain] and [Aspirated], along with the three monovalent place features [Labial], [Coronal] and [Dorsal]. There were general constraints assigning a positive reward for each of the laryngeal features, as well as more specific constraints rewarding the co-occurrence of each laryngeal feature with each place of articulation, along with the very specific Realization constraints that rewarded co-occurrence of a place-laryngeal pairing in a given word. Positive constraints, which assigned rewards of +1, were used rather than negative ones simply for convenience.

The constraints started out with zero weights, which resulted in a uniform probability of 0.33 being assigned to each of the three candidates for each Word (and also a uniform probability of 0.5 for each of the two possible Words for each phonetic form in Robust Interpretive Parsing). The learning parameters were as in all of the previous experiments, although each of the 50 runs had 20,000 rather than 10,000 trials.

The most basic result is that the learners tended towards a single phonetic form for each word. While the starting probability was uniformly 0.33, in 297 of the 300 Words (6 Words × 50 runs), the average probability assigned to one of the phonetic forms by the learners' final grammars was greater than 0.50, and usually much greater than that. This type of lexical convergence is well-known in agent-based learning, both in linguistics (e.g. Liberman 2002), and in robotics (e.g. Schultz *et al*. 2008). The convergence to a single headedness choice for each phrase demonstrated in the syntactic experiment in section 4 is another instance of this fundamental effect.

Perhaps more remarkably, the learners also seemed to tend to avoid homophony. That is, within each place of articulation, it was more often the case that the two Words (e.g. Word 1 and Word 4) had different choices for the laryngeal feature than would be expected by chance. The chance rate would be 6/9 = 0.66, since for each of the two Words there are 3 choices of laryngeal feature, and of the 9 combinations, 3 of them result in the same choice for the two Words. Of the

147 occasions when both Words each had one phonetic form with the majority of the probability, in 125 of these those two phonetic forms were different, yielding a rate of homophony avoidance of 0.85. This can perhaps be understood as an instance of error minimization: the more homophony there is, the more errors learners will make.

Finally, and perhaps most remarkably, the choice of laryngeal contrast across places of articulation tended towards uniformity. At each place of articulation, there are three possible patterns of contrast (aspirated *vs*. plain, plain *vs*. voiced, and voiced *vs*. aspirated). Combining these, there are $3^3 = 27$ possible patterns of contrast. Of these, only 3 use a single laryngeal contrast across places of articulation (0.11), while 18 use two laryngeal contrasts (0.67), and 6 use different laryngeal contrasts at each place of articulation (0.22). Of the 32 runs that had contrasts at each place of articulation, 13 (0.41) had the same pattern of contrast across places of articulation, 16 (0.50) had a distinct pattern of contrast at one place of articulation, and only 3 (0.09) had a distinct pattern at all three places of articulation. Thus, there is a skew towards uniformity of laryngeal contrast across places of articulation: the observed rate of uniform laryngeal contrast (0.41) is much higher than that expected by chance (0.11).

I have yet to subject these results to inferential statistical analysis, and I have also not run many similar experiments,[2] so it is hard to know how reliable these effects are. Given that the constituent pieces of the feature economy account (i.e. lexical convergence and error minimization in agent-based systems, systemic simplicity) have been observed in other experiments here and elsewhere, and have relatively clear sources, I feel reasonably confident at this point that the effect will be robust. Besides assessing the robustness of this account of feature economy, another important direction for further research is to better understand why the effect seems to be relativized to particular features (see Hall 2011 for relevant discussion), both in terms of the present model, and more generally.

6 CONCLUSIONS

It should be clear that this paper is largely programmatic. The aim has been to argue that the set of phenomena that I have characterized as instances of systemic simplicity pose fundamental problems for current linguistic theory, and to show that a unified account may be obtained by combining a standard assumption about the nature of linguistic constraints from linguistic theory, that they generalize across contexts, with some broadly applied assumptions from cognitive modelling and natural language processing: that linguistic models are probabilistic, and can be learned with a simple error-based learning procedure. Much remains to be done in terms of developing and further testing the proposed accounts of each of the instances of systemic simplicity discussed here, extending this approach to other instances of learning bias, historical change and typological skew, integrating it with explanations of "substantive" skews in typology (e.g. skews towards articulatory and perceptual fitness in phonology), and understanding the dynamics of these grammar and learning models. More generally, this is just one piece of a rapidly growing literature that shows the promise of approaches to linguistic explanation that combine explicit learning models with structured linguistic representations in probabilistic frameworks, and much will undoubtedly be gained by comparing the one explored here to alternatives.

[2] One further preliminary finding I can report is that when underlying representations (URs) are used and the Realization constraints are replaced by constraints demanding particular Word-UR mappings as well as faithfulness constraints, the homophony avoidance effect seems to weaken.

REFERENCES

Aristar, A.R. (1991). On diachronic sources and synchronic pattern: an investigation into the origin of linguistic universals. *Language* 67.1-33.

Aronoff, M. and Z. Xu. (2010). A Realization Optimality-Theoretic approach to affix order. *Morphology* 20. 381–411.

Bane, M. and J. Riggle (2008). Three correlates of the typological frequency of quantity-insensitive stress systems. In *Proceedings of the Tenth Workshop of the Association for Computational Linguistics' Special Interest Group in Morphology and Phonology*.

Boersma, P. (1997). Functional Optimality Theory. *Proceedings of the Institute of Phonetic Sciences of University of Amsterdam* 21, 37–42.

Boersma, P. and S. Hamann. (2008). The evolution of auditory dispersion in bidirectional constraint-based grammars. *Phonology*. 217–270

Boersma, P. and J. Pater. (to appear). Convergence properties of a gradual learning algorithm for Harmonic Grammar. In J. McCarthy and J. Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Christiansen, M. H. and J. T. Devlin (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates. 113-118.

Clark, B., M. Goldrick, and K. Konopka. (2008). Language Change as a Source of Word Order Generalizations. In R. Eckardt, G. Jäger, and T. Veenstra, *Variation, Selection, Development: Probing the evolutionary model of language change*. Berlin: Mouton de Gruyter. 75-102.

Clements, G. N. (2003). Feature economy in sound systems. *Phonology* 20. 287–333.

Coetzee, A. (2002). Between-language frequency effects in phonological theory. Ms., University of Massachusetts Amherst. [Available at http://www-personal.umich.edu/~coetzee/].

Coetzee, A. and J. Pater. 2011. The place of variation in phonological theory. In J. Goldsmith, J. Riggle, and A. Yu (eds.), *The Handbook of Phonological Theory* (2nd ed.). Blackwell, 401-413.

Culbertson, J. (2010). *Learning biases, regularization, and the emergence of typological universals in syntax*. Doctoral dissertation, Johns Hopkins University. [Available at http://web.jhu.edu/cogsci/people/alumni/CulbertsonDissertation2010.pdf].

Dresher, B. E. and J.D. Kaye. 1990. A Computational Learning Model for Metrical Phonology. *Cognition* 34: 137-195.

Dryer, M. (1992). The Greenbergian Word Order Correlations. *Language* 68. 81–138.

Goldwater, S. and M. Johnson. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. Eriksson, and O. Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. 111–120.

Greenberg, J. H. (1966). Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In J. H. Greenberg (ed.) *Universals of Language* (Second Edition). MIT Press. 73–113.

Hall, D.C. (2011). Labial place in phonology: Universal and variable. Paper presented at NELS 42.

Hare, M. and J. L. Elman (1995). Learning and morphological change. *Cognition* 56, 61–98.

Harris, A.C. (2000). Word Order Harmonies and Word Order Change in Georgian. In R. Sornicola, E. Poppe and A. Sisha-Halevy (eds.) *Stability, Variation and Change of Word Order Patterns Over Time*. Amsterdam, Netherlands: John Benjamins.133-163.

Hayes, B., K. Zuraw, P. Siptár and Z. Londe. (2009). Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85. 822-863.

Hochberg, J. (1988). Learning Spanish stress: Developmental and theoretical perspectives. *Language* 64. 683–706.

Jäger, G. (2007). Maximum Entropy models and Stochastic Optimality Theory. In J. Grimshaw, J. Maling, C. Manning, J. Simpson, and A. Zaenen (eds.) *Architectures, rules, and preferences: a festschrift for Joan Bresnan*. Stanford, California: CSLI Publications. 467–479.

Johnson, M. (2007). A gentle introduction to Maximum Entropy Models and their friends. Talk given at the Northeastern Computational Phonology Meeting, University of Massachusetts. [Available at www.cog.brown.edu/~mj/Talks.htm]

Jesney, K. and A.-M. Tessier. (2011). Biases in Harmonic Grammar: The road to restrictive learning. *Natural Language and Linguistic Theory* 29. 251-290.

Kirby, S. and Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (eds.) *Simulating the Evolution of Language*. Springer Verlag, London. 121-148.

Liberman, M. (2002). Simple Models for Emergence of a Shared Vocabulary. Paper presented at LabPhon 8, New Haven. [Slides available at http://languagelog.ldc.upenn.edu/myl/labphon.pdf]

Madiesson, I. and K. Precoda. (1992). The UPSID database. UCLA.

Martin, A. (to appear). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. In *Language*.

Moreton, E. and J. Pater. (2011). Learning artificial phonology: A review. Ms. University of North Carolina and University of Massachusetts Amherst.

Pater, J. 2008. Gradual learning and convergence. *Linguistic Inquiry* 39, 334-345.

Pater, J. 2009. Weighted constraints in generative linguistics. *Cognitive Science* 33, 999–1035.

Pater, J. (to appear). Universal Grammar with Weighted Constraints. In J. McCarthy and J. Pater (eds.), *Harmonic Grammar and Harmonic Serialism*. London: Equinox Press.

Pater J. and E. Moreton. (in prep.). Simplicity bias in phonological learning.

Phillips, B. (1984). Word frequency and the actuation of sound change. *Language* 60. 320–342.

Prince, A. and P. Smolensky (1993/2004). *Optimality Theory: Constraint interaction in generative grammar*. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version published by Blackwell, 2004.

Rumelhart, D., J. McClelland and the PDP Research Group (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. MIT Press.

Rumelhart, D. and J. McClelland. (1986). On learning the past tenses of English verbs. In J. McClelland and D. Rumelhart (eds.), *Parallel Distributed Processing, Volume II*, MIT Press. 216–271.

Schulz, R., G. Wyeth, and J. Wiles. (2010). Language change across generations for robots using cognitive maps. In H. Fellermann, M. Dörr, M. M. Hanczyc, L. L. Laursen, S. Maurer, D. Merkle, P.-A. Monnard, K. Stoy and S. Rasmussen (Eds.), *Artificial Life XII: Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems*. MIT Press. 581-588.

Smolensky, P. and G. Legendre. (2006). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*. Cambridge, Mass.: MIT Press.

Sonderegger M. and P. Niyogi. (in press) Variation and change in English noun/verb pair stress: Data, dynamical systems models, and their interaction. In A.C.L. Yu (ed.), *Origins of Sound Patterns: Approaches to Phonologization*. Oxford: OUP.

Tesar, B. and P. Smolensky. (2000). *Learnability in Optimality Theory*. MIT Press.

Travis, L. (2011). Of Micro- and Macroparameters: Ergativity, Austronesian and Bahasa Indonesia. Paper presented at Yale University.

Wedel, A. (2004). Category competition drives contrast maintenance within an exemplar-based production/perception loop. *Proceedings of the Workshop of the ACL Special Interest Group on Computational Phonology (SIGPHON)*. Association for Computational Linguistics.

Wedel, A. (2011). Self-organization in phonology. In E. A. H. Marc van Oostendorp, Colin J. Ewen and K. Rice (Eds.), *The Blackwell Companion to Phonology*. Blackwell. 130–147.

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

Zuraw, K. (2003). Probability in language change. In R. Bod, J. Hay, and S. Jannedy, eds. *Probabilistic Linguistics*. Cambridge, MA: MIT Press. Pp. 139-176.