A radically emergentist approach to phonological features: implications for grammars

Aleksei Nazarov
University of Massachusetts Amherst

Abstract

Phonological features are often assumed to be innate (Chomsky & Halle 1968) or learned as a prerequisite for learning grammar (Dresher 2013). In this paper, I show an alternative approach: features are learned in parallel with grammar. This allows for addressing an interesting question: is it really optimal that the phonological grammar only use phonological features to refer to segmental material (Chomsky & Halle 1968), or could it be more advantageous for the grammar to refer to segmental material on more than one level of representation? The learner considered here finds that it is only optimal for the grammar to use phonological features to refer to multiple segments in the same pattern (e.g., the class of nasals), but when a pattern refers to a single segment, it may be at least equally good for the grammar to refer to this single segment as a bare segment label (for instance, [m] instead of [labial, nasal]). In this way, the grammar uses different kinds of representational units (features and non-features) for the same sound – which mimics models with multiple layers of representation (such as Goldrick 2001, Boersma 2007).

1. Introduction[*]

In this paper, I will introduce a novel way of looking at models of emergent phonological structure (in particular, models of emergent phonological features). I will argue that such models can make predictions not only about typology (Blevins 2004, Mielke 2004), but also about the shape of grammars of individual languages. I will propose a radically emergentist model of learning segmental representations, and I will show through computational simulations that this model predicts grammars which diverge from what is standardly assumed: the emergentist approach described here predicts that constraints in phonological grammars will not always refer to phonological features, in contrast to the standard approach.

In the last decade, the idea that various kinds of phonological structure may be "emergent" (see, for instance, Blevins 2004; Wedel 2003, 2011; Mielke 2004) has gained traction in the literature. These ideas depart from the canonical hypothesis that most or all elements of phonological structure are innate and universal (see, for instance, Chomsky & Halle 1968 for universal phonological features). When some aspect of phonological structure is emergent, I take this to mean that it is not necessary to stipulate the particulars of that aspect (such as individual phonological features) in Universal Grammar – which is similar to the approach taken by Dresher (2013, 2014).

The debate on whether certain aspects of phonology are emergent or innate (see, for instance, Blevins 2006 and Kiparsky 2006 for a debate on the emergence of final devoicing and the rarity of final

[*] Many thanks to Adam Albright, Paul Boersma, Lisa Davidson, Brian Dillon, David Erschler, Martin Krämer, John Kingston, John McCarthy, Marc van Oostendorp, Joe Pater, Presley Pizzo, Clàudia Pons-Moll, Robert Staubs, Peter Svenonius, and Kristine Yu for very helpful discussion, and many thanks to audiences at UMass Amherst, the University of Tromsø, Leiden University, and the University of Manchester. All mistakes are my own.

© 2014 Aleksei Nazarov. *Nordlyd* 41.1: 20-57, special issue on Features edited by Martin Krämer, Sandra-Iulia Ronai and Peter Svenonius. University of Tromsø – The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd

voicing) has mostly centered on typological facts. In the realm of phonological features – which will be the focus of this paper – Mielke (2004) motivates his Emergent Feature Theory by appealing to the typology of phonological patterns: no universal feature theory has a good account of the range of phonologically active classes in his cross-linguistic database (Mielke 2007), while his Emergent Feature Theory has a better explanation for the patterns that are found in the data.

Mielke's (2004) Emergent Feature Theory states that phonological features are entities induced by the language-acquiring infant, rather than predefined entities (as, for instance, the list of features in SPE). This view is the inspiration for the model that will be proposed here. However, Mielke does not explicitly specify what motivates the induction of phonological features. I propose here that the motivation for learning features is the desire of the language-learning infant to induce grammars which have maximally general constraints for every phonological pattern.

As pointed out by Halle (1978), the formulation of a grammatical statement (rule or constraint) in terms of phonological features means that this statement will be applicable to all segments with that feature description, including novel segments. Halle gives the example of the name Bach [bɑx], which, despite ending in a non-English consonant [x], still triggers devoicing of the suffix [-z]: /bɑx+z/ → [bɑxs], *[bɑxz] "Bach's/Bachs". This provides evidence for the idea that the rule of devoicing [-z] refer to the feature [(±)voice], and that [x] is classified with respect to that feature. This methodology is called "Bach-testing" (Halle 1978 credits the idea to Lise Menn). Other tests (such as wug-testing; Berko 1958) have also shown that language users generalize rules of mental grammar to new tokens.

Based on this evidence, I follow Chomsky & Halle's (1968) and Albright & Hayes' (2002, 2003) ideas about generalization as a driving force in the acquisition of grammar. In fact, I will assume that finding the most general formulation of a phonological pattern is the main driving force behind learning phonological grammars.

If a grammar does not have access to classificatory phonological features, it will miss many generalizations – every pattern which applies to more than one segment will have to be triggered by a series of constraints. For instance, a process of final devoicing (as, for instance, in Dutch – see Booij 1995) will have to be triggered by constraints that refer to each of the individual segments undergoing final devoicing: *b#, *d#, *v#, ...

Since features help the learner reach generality in grammar (as was explained above), it can be said that the presence of features is motivated by learning the grammar. In other words, the learner could have lived without phonological features, but it is the desire to state phonological patterns as broadly as possible that makes the learner induce these features. In this sense, phonological categories such as classificatory phonological features can be hypothesized to be a side effect of learning entire grammars.

Because my hypothesis regarding the induction of features also refers to the construction of grammars, I will explore a model which jointly induces phonological grammar and phonological features – the initial state of the learner is an empty grammar and no phonological features, and its final state is a grammar which accounts for the phonological patterns of the language and refers to phonological features (when necessary). Induction of phonological and morphological grammars has been simulated computationally before (Albright & Hayes 2002, 2003; Hayes & Wilson 2008), and machine learning of phonological features has also been explored (Niyogi 2004, Jansen & Niyogi 2008, Lin 2005, Lin & Mielke 2008).

Joint induction of grammar and features has been explored at least in one instance: Archangeli et al. (2012) explore a model which induces grammatical constraints and allows them to refer to sets of segments. An example of this is the (positively formulated) vowel harmony constraint {i, u}→ ¬{e}: if the first vowel is one of {i, u}, then the second vowel is not a member of the set {e}.

However, Archangeli et al. do not restrict the sets of segments that may occur in constraints, whereas a model with feature labels limits the sets of segments that may be referred to (for instance, the feature set from SPE (Chomsky & Halle 1968) cannot define a set of segments like {i, q}). The model I present here, to my knowledge, is the first in which a distinction is made between segments and phonological feature labels within the grammar, and these labels are induced based on improvement of the grammar.

Throughout this paper, the term "phonological feature" will only refer to classificatory phonological features, which are entities that classify the segment (allophone or phoneme) categories of a language. Chomsky & Halle (1968) are careful to distinguish such classificatory features from phonetic features, which are real-valued articulatory or acoustic dimensions. I will assume a framework in which segments are learned from acoustics, and classificatory features are learned from segments. This is in line with work such as Pierrehumbert (2003a) and Peperkamp et al. (2006), which exemplifies a two-stage approach to learning in which low-level abstract representations (e.g., allophones) are learned in a bottom-up fashion, and higher-level abstract representations (phonemes) are learned based on these low-level abstract units. If the grammar is given access to a feature like [voice] (for voiced obstruents), the constraint that triggers final devoicing can simply be stated as *[voice]#. This expresses the pattern of final devoicing with maximal generality.

This may be contrasted with the traditional idea, which is even implicitly present in the recent literature on learning phonological features inductively (see, for instance, Lin 2005, Lin & Mielke 2008), that phonological categories such as features are learned because there is an *a priori* (innate or otherwise) requirement to have such categories (the grammar requires the presence of phonological features, for instance). Under this view, the learning of features is completely independent from the learning of grammar, and features are only emergent as far as their content is not specified innately. The model proposed here, however, maintains that the presence of features is motivated by an external factor – namely, the goal of having maximally general constraints in the grammar.

The type of feature induction reported here is also different from the previous work cited above because I will not induce phonological features from acoustic representations, like Lin (2005) and Lin & Mielke (2008) did. Instead, I chose to start from representations in terms of segments (allophones), and find classificatory features which unite some of these segments.

There has been a good amount of work on learning segments from acoustics (see, for instance, Niyogi 2004, Vallabha et al. 2007, Elsner et al. 2013, and Boersma & Chladková 2013). However, the learning of classificatory features from the phonological behavior of segments, as outlined in Mielke's (2004) Emergent Feature Theory, has not been explored as much (barring Archangeli et al. 2012). For this reason, I chose to focus on this aspect of learning. Of course, it would be desirable for a more extended implementation of the model described in section 2 of this paper to take the learning of segments from acoustics into account as well.

Since I assume that the building blocks of phonological representation are not innately specified and thus subject to between-language variation, the constraints that are part of grammars cannot be universal (as is standardly assumed in Optimality Theory (OT) – see, for instance, Prince & Smolensky 1993/2004). If phonological constraints cannot refer to non-phonological representations, and if the building blocks of phonological representations are not fixed before the reception of language input, then neither can phonological constraints be fixed before language input is received. For work discussing and implementing the induction of phonological constraints, see, for instance, Hayes (1999), Hayes & Wilson (2008) and Wilson (2010).

In the model proposed in this paper, induction of phonological constraints and induction of phonological features interact in a cyclic way: the induction of phonological constraints prompts the induction of features, and the induction of features prompts the induction of new constraints. I will explain the details of this interaction in section 2.2. It is this cyclic interaction that leads to a gradual increase in generality by gradually recruiting phonological features into the grammars. This interaction also leads to the general effect that phonological features are only recruited when necessary – which is precisely the property highlighted by the case study taken up in this paper, which I will now describe.

The model of feature learning just discussed, which may be called radically emergentist, is implemented computationally, and applied to a toy language which has the crucial properties of existing patterns in natural languages. The toy language has the following three phonological patterns:

(1) *phonological patterns in the toy language*
    a. no [m] word-finally
       baman, bamab; *bamam
    b. no nasals word-initially
       baman; *maman, *naman
    c. no labials in between high vowels
       baman, binin; *bimin, *bibin

The relevant property of this data set is that pattern (1a) can be expressed either in terms of the allophone [m] alone – as the constraint *m# – or in terms of the intersection of the two classes used in (1b,c): *[labial, nasal]#. The standard, innate feature approach predicts that the constraint *[labial, nasal]# will be chosen to represent this pattern. However, I will show that my model predicts that the constraint *m# will be chosen.

This constraint, *m#, which appeals to linguistic sound without the mediation of classificatory features, co-exists with other constraints that do appeal to classificatory features in the grammars that are learned by the radically emergentist model. In other words, my model leads to grammars in which some constraints refer to sound through features, but other constraints refer to sound through a lower-level type of representation: segments (allophones). This means that the grammars generated by the emergentist model refer to multiple levels of phonological abstraction.

This latter finding is at odds with the usual assumption that all constraints in phonological grammars are encoded in terms of features. The idea that grammars only refer to features is a natural consequence of the assumption that phonological features are the innate units of classifying linguistic sound for the purposes of grammar (Chomsky & Halle 1968). Even models that do acknowledge that the processing of linguistic sound may take place at various levels of abstraction (acoustics, segments, features, ...) usually maintain that the phonological grammar itself refers to phonological features only (see Pierrehumbert 2003b).

The model proposed here, however, generates situations in which the grammar itself has access to multiple levels of abstraction for the same sound. In the toy language, for instance, a sound fragment which sounds like [m] may be referred to in one constraint by the feature combination [labial, nasal], and in another constraint by a featurally unanalyzed label “m”.

I am building on the assumption that segment categories (whether these are allophone-like or rather phoneme-like) are learned from complex acoustic cues (see, for instance, Lin's 2005 work for various models of this process). This learning step is not modeled here, but it is assumed that “m” is a symbol which has been learned by the language-acquiring infant as corresponding to certain acoustic cues in certain contexts.

Of course, it is not obvious that the learning of phonological, classificatory features takes place after segment categories have been acquired. It may be the case that classificatory features are learned simultaneously with segment categories – see, for instance, Dresher (2014) for discussion of this scenario.

However, the crucial concept is that classificatory features are based on some categorical cutoff (in Chomsky & Halle's 1968 system, this is the cutoff between the + and – value of a feature), and segment categories such as “m” and “n” provide a basis for establishing such cutoff values: even though a feature like [labial] may be realized variably, one can be certain that the segments [p], [b], and [m] will be assigned that feature, regardless of the precise acoustic and articulatory values. It is merely because of this that I have set up the model to have this shape.

The emergent feature learning scenario proposed here makes predictions that are different from standard assumptions about grammar. Instead of only appealing to features, the grammars predicted by this model sometimes refer to atomic segment units, and sometimes to features. Some consequences of this will be seen in sections 2.4 and 5, but the crucial difference between this state of affairs and canonical assumption about grammars is that atomic segment units (e.g., [m]) are not shorthand for a feature bundle, as in SPE (Chomsky & Halle 1968) for example.

It is in this sense that the model of emergent features which will be presented here makes predictions about within-speaker grammatical structure. Some (non-prosodic) constraints in the grammar are predicted not to be encoded in terms of feature bundles, but in terms of atomic, featurally unanalyzed labels denoting segment categories. This has important conceptual consequences, but testable empirical predictions can also be derived from this: section 5 will sketch these predictions. However, the main result obtained in this paper is that grammars appealing to a mix of features and atomic segments are a natural consequence of the model described here.

The rest of this paper is organized as follows. Section 2 will present my theory of phonological category emergence in more detail, and will also describe in more detail the toy language sketched above and the intuitions as to the consequences of the model as applied to the toy language. After this, section 3 will describe the computational implementation of the model applied to the toy language, and section 4 will show the results of this simulation. Finally, section 5 will offer discussion and concluding remarks.

2. A radically emergentist model of grammar and feature learning

2.1 Classificatory phonological features as byproducts of grammar learning

The model that will be proposed here is one which jointly induces phonological grammars and phonological features. This is inspired by inductive perspectives on the learning of both phonological features (Mielke 2004) and phonological grammars (Hayes 1999, Hayes & Wilson 2008, Wilson 2010). However, the synthesis made here is novel: I will assume a minimum of prior knowledge of representations – since the learner is not given any phonological features in advance, and the way towards finding phonological features is dictated by the grammar.

In my model, the presence of phonological features is not motivated by an explicit requirement to have phonological features per se. Instead, features are induced by the learner because they help state constraints in a way that generalizes over more forms.

For instance, a feature such as [voice], if assigned to voiced obstruents only, helps state the generalization that Dutch does not allow word-final voiced obstruents (Booij 1995) in one single constraint (for instance, *C[voice]#). If there is no feature [voice], then the same generalization should be expressed by banning each single voiced allophone in Dutch separately (*b#, *d#, *v#, *z#, etc.), which lacks a general acknowledgment of the pattern of final devoicing (in terms of traditional phonological analysis, this would be a "missed generalization"). It is in this sense that phonological features enable a more general statement of phonological patterns.

The type of phonological feature appealed to here is the classificatory phonological feature – a unit which generalizes over individual segment (allophone or phoneme) categories. As pointed out as early as SPE (Chomsky & Halle 1968), classificatory phonological features are distinct from phonetic features, which can be seen as real-valued dimensions of articulatory and acoustic space. Even in the SPE model, there were rules converting discrete-valued classificatory features (e.g., [+voice]) to real-valued phonetic features (e.g., [4.1 voice]).

Of course, SPE does not have an explicit level of segments. Classificatory features are arrived at by finding correspondences between + or – and a continuous scale. These correspondences are expressed in contextual rules (e.g., [+ voice] → [3 voice] / __ [+ vocalic]), which, in a sense, can be thought of as taking on the function of segments.

However, I will take an approach in which classificatory features are not homonymous with phonetic features. Rather, classificatory features are induced from groupings of segments that are active in a certain context. I assume that segments categories themselves are extracted from phonetic features. This yields the following picture:

(2) *The relation between phonetic and classificatory features*
Classificatory features (categorical)
↑

Segments
↑
Phonetic features (gradient)

The classificatory features induced in my model will be based solely on evidence from phonological processes, not on acoustic or articulatory dimensions. However, this choice was made for the sake of simplicity of implementation, so that classificatory features could be induced based on phonological behavior alone, but it is possible and desirable to build phonetics into the feature induction mechanism by adding acoustic and articulatory dimensions to the feature induction mechanism (which, itself, is described in sections 3.2.3-4).

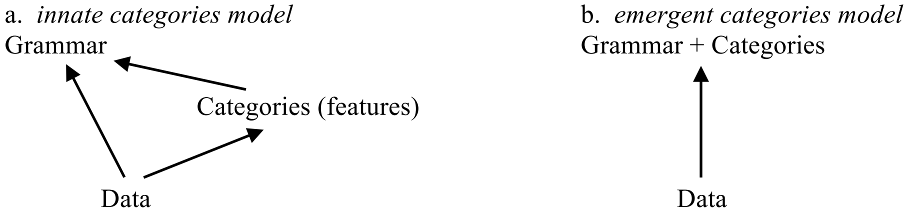
It has been demonstrated experimentally (see, for instance, Cristia et al 2013 for results from perception experiments with infants; see also the evidence reviewed by Moreton & Pater 2012) that phonetic factors do play an important role in the mental grouping of segments. An ideal version of the current model should take this into account: evidence for classificatory features should be taken from both the acoustic and the structural realm. However, in order to focus the model on the idea of feature induction for the sake of summarizing and grouping segments, I concentrated on the structural factors – since inducing phonological features without motivation from the grammar can be done on the basis of phonetics alone (see, for instance, Lin & Mielke 2008), but inducing phonological features as motivated by grammar learning can only be done with structural factors present.

Therefore, I would like to emphasize that the absence of phonetic factors in the current model is purely an implementational choice. Even though the exclusion of substantive or phonetic factors in inducing features is reminiscent of Substance-free Phonology (Morén 2006), it is not a tenet of the theory behind this model. Including phonetic information in this type of model would most definitely be an important avenue for future research.

In the current emergentist model, phonological categories are learned jointly with the grammar. This can be contrasted with a traditional, innate model of phonological representation: the categories of segmental representation are innate classificatory features. The most logical learning scenario for this model would be one in which the mapping from acoustics to features is learned, and separately from that, grammatical statements are learned from the encoding of acoustic tokens into features.

The figures below contrast the learning scenarios implied by the standard approach with innate features, and the one followed by the current emergent model, respectively.

(3) *Comparison between models*


a. *innate categories model*
b. *emergent categories model*

In this paper, I will only consider a part of the path drawn in (3b) above: I will look at the learning of classificatory phonological features from segment (allophone) units. I will assume that segment units have already been learned from the acoustic signal, and my simulations only apply to the part of the learning path after segments have been learned.

The precise way in which segments are learned from acoustics in this radically emergent model remains to be crystallized in future work (see, for instance, Elsner et al. 2013, among others, for proposals of how segment units are learned from acoustics). However, there is evidence that allophone or phoneme units (the distinction between allophones and phonemes will not be important for the data investigated here) are active in sound processing and phonological grammar. Experiments in speech perception and production (see, e.g., Jesse et al. 2007, Nielsen 2011) have demonstrated that the processing of speech makes crucial reference to segment units. For instance, Nielsen (2011) shows in an imitation study that segment categories play a role separate from both phonological features and exemplars in accounting for how listeners generalize the presence of an eccentric speech attribute (exaggerated aspiration) to new words.

Furthermore, there is evidence from phonological processes in which segment identity plays a separate role from feature identity. For instance, consonant OCP processes (see Coetzee & Pater 2008 for an overview) tend to avoid featurally similar consonants within a certain domain, but tend to allow phonemically identical consonants in the same domain. For instance, Cochabamba Quechua (Gallagher 2014) does not allow two non-identical consonants in a disyllabic word to both be ejective, e.g., [tʃʼaka] ‘bone’, but *[tʃʼak’a] (Gallagher 2014:2 (1a-b)). At the same time, if two consonants in a disyllabic word are identical, they may both be ejective, e.g., [tʃʼatʃʼa]‘to soak’ (Gallagher 2014:2 (1c)).

These opposite tendencies (avoidance of similarity in non-identical pairs, non-avoidance of similarity in identical pairs) may, in principle, be explained by appealing to features only. However, this leads to a rather convoluted statement of constraints: a constraint such as OCP-[+constricted glottis] will be violated once for every two consecutive consonants which both have the feature [+constricted glottis], but only if there is at least one other feature (e.g., a place feature, as in *[tʃʼak’a]) which the two consonants do not share. This means that a constraint specific to one feature of a consonant needs to search through all other features associated with that consonant before finding whether it is violated. A much more elegant formulation would be one where segment identity is a separate piece of information, separate from featural identity. In that case, one could say that OCP-[+constricted glottis] is violated once for every sequence of consonants which are identical in their [+constricted glottis] specification, but non-identical in their segment specification. This creates a much less computationally intensive definition of the OCP-constraint[1].

These findings show that the phonological grammar can appeal to segments as a unit. For this reason, it is not unreasonable to assume that segment units are somehow induced by language learning infants, and can be the basis for learning the classificatory features that are important in the grammar of a language. See also Dresher (2009, 2013, 2014) for other work that assumes that segments are the basis for learning classificatory features.

The following section will explain how the two main components of the model – constraint induction and feature induction through clustering – interact to yield maximally generalizing grammars and also abstract phonological units such as classificatory phonological features.

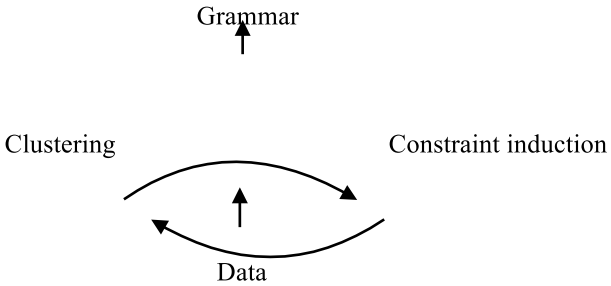
2.2 The components of the model

The radically emergent model proposed here has two main components: induction of phonological constraints, and induction of classificatory phonological features through clustering. As I will show, it is the interaction of these two components that leads to finding a grammar with maximally generalizing constraints. The induction of features opens the way for the induction of more general constraints – see, for instance, the example of [voice] in section 2.1. In this sense, feature induction is motivated by an external factor: the goal of maximally generalizing constraints (which will be fleshed out below).

The two components of the model interact in a cyclic way: induction of a group of constraints is followed by an attempt to find features through clustering, and the representational units found in the process of clustering are then allowed to be used in another instance of inducing constraints.

(4) *Cyclic model of grammar learning*

[1] Many thanks to Joe Pater for pointing this out to me.



Inducing a group of highly relevant constraints makes it possible to state the patterns observed in the data. For instance, if the observed tokens have no word-final [m], but there is no constraint which states that pattern, then adding a constraint against word-final [m] to the grammar will make the grammar more accurate, but adding a constraint which would require word-final [m] would not make the grammar more accurate.

Of course, it is not trivial that the non-occurrence of a certain segment in a certain position in the lexicon will lead to the induction of a constraint against it – any lexicon may contain accidental gaps. However, there is empirical evidence that at least some positional gaps in the lexicon may lead to the induction of phonotactic constraints. For instance, [ŋ] never occurs word-initially in the English lexicon, and all 8 native speakers of American English that I informally surveyed agreed that [ŋ]-initial non-words such as [ŋæpi] are ungrammatical.

Furthermore, the statistical learning model considered here is built so that it can distinguish between constraints that represent random artifacts in the data, and constraints that truly bring the grammar's predictions closer to the learning data – which makes a constraint relevant to the data at hand. The relevance of a constraint can be measured in various ways – see Hayes & Wilson (2008), Wilson (2010). Information gain (as used by Wilson 2010) was chosen here (see section 3.2.1.2).

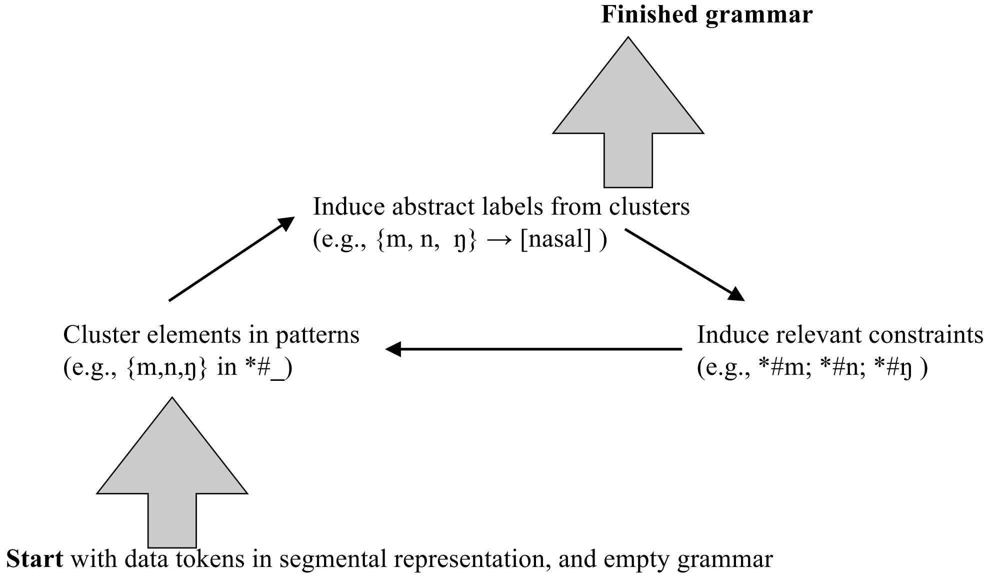
The model is set up so as to find a grammar in which every distinct phonological pattern is stated with the fewest number of constraints. This is what is meant by “maximally generalizing constraints” – instead of stating a phonological pattern with a large number of specific constraints (for instance, *#m, *#n, *#ŋ for a pattern which prohibits word-initial nasals), aiming for a single constraint which triggers the pattern wherever possible (for instance, *#[nasal]).

Clustering makes it possible for the learner to move toward this goal. If a language bans word-initial nasals, the constraints *#m, *#n and *#ŋ will be found to be highly relevant. These constraints all share the context *#_ (i.e., the word-initial context), and cluster analysis will find that {m, n, ŋ} form a cluster of high relevance out of all segments of the language, when inserted in the context *#_ .

Clusters of representational elements thus found may be interpreted as classes of segments which undergo one and the same process. By the logic employed in Emergent Feature Theory (Mielke 2004), a previously non-existent feature label can be created and assigned to this class. In this fashion, the cluster {m, n, ŋ} will be assigned a feature label – let us name it [nasal]. Once abstract labels have been induced from clustering, these labels can be used to induce new constraints.

The figure below summarizes how the model just described operates to induce grammars:

(5) *Summary of the model*


Finished grammar
Induce abstract labels from clusters
(e.g., {m, n, ŋ} → [nasal])
Cluster elements in patterns
(e.g., {m,n,ŋ} in *#_)
Induce relevant constraints
(e.g., *#m; *#n; *#ŋ)
Start with data tokens in segmental representation, and empty grammar

2.3 Data

The data to which this model will be applied is a toy language, rooted in properties found in attested natural languages. These crucial properties have to do with the size of segment classes to which phonological patterns refer. I will show in 2.4 that it is to be expected in an emergent feature model like the one described here that the size of segment classes influences their representation in the grammar. In this subsection, I will introduce single-segment and multi-segment phonological patterns.

Phonological patterns may apply either to groups of segments, or to a single segment. For instance, English epenthesizes a vowel between a stem-final sibilant [s, z, ʃ, ʒ, ʧ, ʤ] and a following [z] (belonging to a plural or possessive suffix) – which is a pattern that applies to groups of segments (at least in part of the context of the rule; Jensen 1993). This is exemplified in (6) below.

(6) *Inter-strident epenthesis*
ø → [ɨ] / [+strident] (=[s, z, ʃ, ʒ, ʧ, ʤ]) __ +/z/
examples:

/roz/ + /z/ → rozɨz	"roses"
/bʊʃ/ + /z/ → bʊʃɨz	"bushes"
/bæʧ/ + /z/ → bæʧɨz	"batches"
but: /mov/ + /z/ → movz, *movɨz	"mauves"

At the same time, English allows word-initial three-consonant consonant only if the first consonant [s], and this first consonant is followed by a stop and a liquid (Jensen 1993, Mielke 2007). This constraint can be interpreted as one that assigns rewards instead of penalties, such that all triconsonantal initial clusters are banned, but a constraint that rewards [s] followed by a stop and a liquid outweighs this constraint (Presley Pizzo, p.c.)[2].

[2] Constraints that reward candidates instead of penalizing may be problematic in Standard OT (Prince 2007), but they are certainly possible in other frameworks. See, for instance, Kimper (2011) on positive constraints in Harmonic Serialism. Positive constraints are also possible in Harmonic Grammar (Pater 2009:1006). The framework of Maximum Entropy (Berger et al. 1996, Della Pietra et al. 1997, Goldwater & Johnson 2003, Hayes & Wilson 2008, Wilson 2010), used in

(7) *Three-consonant clusters consist of* [s] *followed by a stop and a liquid*
[s][-son,-cont][+son,+cont] ([s]= [+strident, -voice, +anterior])
examples (Jensen 1993:67):

[stre] "stray"	*[ftre], *[ntre]
[splæʃ] "splash"	*[fplæʃ], *[mplæʃ]

This pattern is by no means the only one which appeals to just one segment. An automated search of P-base (Mielke 2007) yields 13 patterns in geographically and genetically disparate languages which are encoded as applying to one single segment. A manual search of a subset of P-base – namely, all languages whose names begin with A – yielded 11 additional patterns within that sample alone which apply to a single segment, implying that a much greater number of one-segment patterns could be found by further manual inspection[3]. From these findings, it can be established that the type of pattern in (9) is not an artifact of English, but exists across different languages.

One further fact about these English data is that [s] is the intersection of two segment classes appealed to by other phonological patterns in English. One of these segment classes is the class of sibilants (as in (6)), and the other is the class of voiceless anterior coronals [t, θ, s], which are the only segments that may occur at the end of a word-final three-consonant cluster (Jensen 1993, Mielke 2007):

(8) *Three-consonant clusters must end in* [t, θ, s]
C1 C2 C3 # → C3 = [+coronal, +anterior, -voice] (= [t, θ, s])
examples (Jensen 1993:69):

[nɛkst] "next"	*[nɛksp]
[sɪksθ] "sixth"	*[sɪksf]
[mʌmps] "mumps"	*[mʌmpʃ]

The intersection of [s, z, ʃ, ʒ, ʧ, ʤ] and [t, θ, s] is exactly [s] – as can be seen in the diagram in (9) below.

(9) *English consonants: voiceless anterior coronals in gray boxes, and sibilants in clear box*

p		t			k
b		d			g
f	θ	s	ʃ	ʧ	
v	ð	z	ʒ	ʤ	
m		n			ŋ
w		ɹ l	j		

The emergent feature model presented in 2.1-2 above induces feature labels from segment classes active in phonological patterns. This means that, if the model were to be applied to the facts summarized in (9) above, sibilants and [t, θ, s] would be assigned a feature label each (sibilants = [X], [t, θ, s] = [Y]). This, in turn, would allow the model to encode the single segment [s] in terms of these two labels ([s] = [X,Y]). The model has no other way of encoding single segments in terms of features. Since phonological features are not available *a priori*, but induced from segments' behaving as a group, one cannot invoke any features "inherent" to a single segment to describe it in terms of features (since there are no "inherent" features, but only features induced from group behavior).

the current simulations, uses the same mechanism of constraint interaction as Harmonic Grammar, and therefore may also use rewarding rather than penalizing constraints (Pizzo 2013).

[3] A summary of the patterns found can be inspected under "One-segment phonological patterns" at http://blogs.umass.edu/anazarov/feature-learning/.

Features as induced by this procedure are not a direct basis for explanation of typological patterns, in the same sense as feature-geometrical systems such as Avery & Rice's (1989) or Clements & Hume's (1995) feature systems are. Evidently, a theory or even just a model of features must do more than simply describe the facts of a single language. However, as pointed out by Kirby & Hurford (2002) and others, typology can and should be explained by more than just innate biases.

It has been argued by, for instance, Heinz (2009) and Staubs (2014) that learning of grammars shapes typology. In this sense, the current model is not unlikely to make typological predictions: if an iterated learning schema is followed (Kirby & Hurford 2002), the biases in the model may be amplified over generations of speakers to yield over- and underattestation of certain typological possibilities.

Furthermore, it has been proposed by Bybee (2001) and Blevins (2006) that recurrent patterns of language use shape phonological typology. For the domain of phonological features, specifically, Mielke (2004) argues that a language use-based approach makes better predictions with the respect to the cross-linguistic typology of natural classes employed in phonological patterns. If this line of reasoning is followed, then the current model can make additional quantitative typological predictions when combined with knowledge of phonetics and language use.

These properties lay the empirical framework for answering the question asked in the introduction, namely: will the emergent feature model sketched here lead to grammars that encode even single-segment phonological patterns in terms of phonological classificatory features? Since cases like the one just sketched allow the emergent feature model to encode a single-segment pattern (only [s] word-initially in CCC) in terms of features, it is interesting to see whether the model will do this, or opt for some other type of representation. Section 2.4 will elaborate on this question, but first I will describe the data which were actually offered to the model.

In order to reduce the English case just described to the bare basics – which is desirable for modeling learning, since the hypothesis space for learning both constraints and features simultaneously is quite large – I used a toy language which shared the crucial properties of the English case. The toy language that was used for the simulations had the following properties. All words in the toy language had the shape CVCVC, and the segment inventory is as in (10) below:

(10) *Segment inventory of the toy language*

a. consonants

p	t	k
b	d	g
m	n	ŋ

b. vowels

i		u
	a	

The toy language had exactly three phonotactic restrictions, one of which is stated over one single segment (like the three-consonant onset generalization with respect to its first segment) and two of which are stated over groups of segments (like the other two generalizations in English). These phonotactic restrictions are shown below:

(11) *Phonotactic restrictions in toy language*

a. single segment restriction: no word-final *m
✔panab, ✔panan, ✔panaŋ *panam

b. multiple segment restriction: no word-initial nasals [m, n, ŋ]
✔tadig, ✔badig, ✔kadig *nadig, *madig, *ŋadig

c. multiple segment restriction: no labials [p, b, m] between two high vowels [u, i]
✔daban, ✔duban, ✔dabun *dubun, *dupun, *dumun, *dubin, *dibun

The single segment targeted by the restriction in (11a) can be defined by the intersection of the two groups of segments employed in the other two restrictions – labials and nasals – since [m] is the only labial nasal in the language. This is the same situation as in the English case sketched above. I will now turn to discussing how these properties of the English case and the toy language bear on the question of whether a radically emergentist model of feature learning is likely to lead to grammars which only refer to classificatory features.

2.4 Intuitions about predictions

In the introduction, it was stated that the standard innate model of phonological features and the emergentist model presented in sections 2.1-2.2 makes different predictions about grammars with regard to the representations they use. The standard model generates grammars which only have constraints referring to features by definition (as I will show below), while it turns out that the emergent model has constraints referring to a range of levels of abstraction, including features and featurally unanalyzed segments (as will be shown in section 4). In this subsection, I will explain the intuitions behind these predictions.

The canonical view of phonological representation is that (non-prosodic) phonology refers only to bundles of (classificatory) phonological features (see, for instance, Chomsky & Halle 1968). Whether these bundles be organized in autosegmental representations (Goldsmith 1976) or not, it remains a common assumption that whenever a notation like [b] or [m] occurs, it is shorthand for an intersection of features. This view is in line with the assumption that phonological computation operates on an alphabet of universal (innate) representational elements.

If all phonological units are innate, then the best hypothesis seems to be that these innate units are phonological features (Chomsky & Halle 1968). To my knowledge, there have been no claims that non-prosodic phonological units of any other level of abstraction (for instance, atomic segment/allophone units such as [b] or [i]) could be innate. Evidence in favor of the innateness of allophone units seems to be absent, and evidence against it seems abundant – languages differ widely in the size and nature of their segment inventory (UPSID; Maddieson & Precoda 1990, Reetz 1999), and languages differ equally widely in the particular segment types that they employ (UPSID finds 919 unique segment types across 451 languages, and none of these segment types occurs in each language examined); finally, the level at which cross-linguistic generalizations can be made over segment inventories appears to be the feature rather than the individual segment (see, for instance, Clements 2003).

Thus, a view in which all phonological units are innate implies that phonological features are the only available units for phonological computation. This means that all constraints of the grammar (whether these constraints be innate or induced) should refer to these phonological features – since the translation from raw acoustic data to phonological structure proceeds by mapping acoustics onto bundles of phonological features. For this paper, I will not run a simulation to show this – but the (strong) view that features are the only permissible building blocks of non-prosodic phonological structure logically entails that all constraints in the grammar will refer to phonological features when dealing with non-prosodic structure.

The radically emergentist view proposed here, however, does not set requirements on the units employed in the constraints in the grammars: the only goal is to build a grammar which states patterns in the most general and concise form possible. At first sight, it appears that the goal of having maximally general constraints, and that of having maximally abstract representations (i.e., features) in every position of every constraint, converge – since features help to state constraints in a more general format.

However, there are some situations in which these two goals diverge. Some patterns can be stated with maximal generality without appealing to features: this is true of patterns which refer to no more than one segment. For instance, this is true of the one-segment pattern in the toy language described in section 2.3, which prohibits only the single segment [m] from occurring word-finally:

(12) *Single segment restriction: no word-final *m (= (11a))*
✔panab, ✔panan, ✔panaŋ *panam

This pattern may be stated in terms of features only:

(13) *Featural formulation of pattern*
*[labial, nasal]# : One violation for every labial nasal at the end of a word.

However, the pattern can also be stated with maximal succinctness if [m] is not decomposed into phonological features:

(14) *Segmental formulation of pattern*
*m# : One violation for every [m] at the end of a word.

In cases of this type, then, the emergent category approach does not specify a reason why the featural formulation of the pattern (as in (13)) should be preferred over the segmental one (as in (14)). Since the featural formulation only becomes available after all the features necessary to define the one segment [m] have been induced, and the segmental formulation (*m#) is available before the induction of any of these features, it seems likely that the segmental formulation will be added to the grammar first, and the featural formulation (*[labial,nasal]#) would not add any generality to the grammar and would therefore be dispreferred.

This implies that one-segment patterns such as "no word-final [m]" (or [s]-initial English clusters – see section 2.3) will not necessarily be stated in the grammar in terms of features, but there will be a bias toward stating such patterns in terms of individual segment categories. This is distinct from the canonical, innate feature model, in which all constraints must always refer to features.

The difference between *[labial,nasal]# and *m# is not merely conceptual. As mentioned in the introduction, "Bach-testing" reveals the generality of a phonological pattern. Multi-segment patterns such as voicing assimilation readily spread to novel segments (as in the example given in the introduction: the loan segment [x] is recognized as voiceless, and voicing assimilation applies to the sequence /x + z/ to yield /bɑx + z/ → [bɑxs] "Bach's/Bachs"). However, it is not clear how one-segment patterns behave on this test. The constraints *[labial,nasal]# and *m# make different predictions with respect to this.

The formulation *[labial,nasal]# implies that whatever other segments come to be classified as [labial,nasal] (for instance, a novel segment [w̃] or [ɱ] or [m̥]) will also be banned word-finally. In other words, the one-segment pattern which bans word-final [m] should behave the same as multi-segment patterns such as voicing assimilation.

On the other hand, the formulation *m# does not make such predictions: it predicts that only [m], but not novel [labial,nasal] segments, will be banned word-finally – since *m# applies only to sounds labeled as the segment unit [m]. In section 5, I will return to this distinction and suggest ways in which this could be turned into empirically testable behavioral predictions.

I have reasoned above that the innate model and the emergent model vary in their predictions with respect to the representation of constraints in the grammar: the innate model leads to a grammar with features only, while the emergent model should lead to a grammar which refers to sounds in a variety of ways. However, implementing the emergent model computationally is necessary to ensure that the model does indeed make these predictions.

In the following section, I will describe a computational implementation of the emergent model described above. This implementation will perform learning of feature-like units from data expressed in terms of segments alone. As will be seen in section 4 afterwards, it turns out that the predictions of the emergent category model are borne out.

3. Computational implementation

3.1 General structure of the algorithm

The computational implementation of the preceding model will make use of two established machine learning techniques: modeling in a Maximum Entropy framework (Berger et al. 1996, Della Pietra et al. 1997, Goldwater & Johnson 2003, Hayes & Wilson 2008, Wilson 2010), and cluster analysis with Gaussian mixture models (Everitt 2011). The novelty of the current implementation lies in the way in which these two are combined.

Maximum Entropy is a general-purpose machine-learning framework (Berger et al. 1996, Della Pietra et al. 1997) which has been applied successfully by Hayes & Wilson (2008) as well as by Wilson (2010) to the learning of phonotactic constraints from distributional patterns in a set of data tokens. See section 3.2.1 for a brief exposition of the mechanism of this type of learning model, and its application to phonological learning.

However, Hayes & Wilson's model assumes that phonological features are available to the learner from the outset, while the model described in the preceding section states that features are a byproduct of building a grammar. To incorporate this latter aspect, I let the mechanism of inducing constraints interact with cluster analysis. The details of this interaction will be described in section 3.2.3 below. Cluster analysis will perform the function described for it in section 2.1: clustering of patterns in the constraint set allows for discovery of feature categories from segment categories.

The general schema of the algorithm is as follows. The procedure starts with a set of phonotactic data drawn from the toy language as sketched in section 2.2. The data are presented as strings of segments in the shape of indivisible symbols such as <p>, <t>, <i>:

(15) *Segment inventory of the toy language*
Consonants: p t k b d g m n ŋ
Vowels: i a u

The data offered to the learner are all the CVCVC forms that can be made out of the segment inventory above and which also obey the phonotactic restrictions of the toy language:

(16) *Some forms offered to learner*
a. offered to the learner:
panab, panan, panaŋ, ...
tadig, badig, kadig, ...
daban, duban, dabun, ...
b. not offered to the learner:
*panam (violates “no final [m]”), ...
*nadig, *madig, *ŋadig (violate “no initial nasals”), ...
*dubun, *dumun, *dibun (violate “no labials between high vowels”), …

At this initial stage, the Maximum Entropy model has no constraints, meaning that every possible representation has equal likelihood (see section 3.2.1.1). The set of possible representations considered by this learner consists of all CVCVC combinations that can be made out of the segment inventory – so that this includes both the phonotactically legal forms, as exemplified in (16a), and the phonotactically illegal forms, as exemplified in (16b).

The reason why only CVCVC forms are considered is a purely practical one: these forms suffice to observe the activity of all three phonotactic constraints in the toy language; adding other forms would only increase the workload for the learning algorithm. A fuller and more realistic simulation would include forms of other phonotactic shapes (such as VC, CVC, CVCV).

However, since all data offered to the learner are of the shape CVCVC, it is necessary to also restrict the hypothesis space of the grammar to CVCVC forms – otherwise, the grammar would to account for the absence of data of the shape VC, CVC, CVCV, CVCC, etc. This is because of the nature of the Maximum Entropy learner used here: the grammar must fit the statistical distribution in the data, including gaps, with maximal accuracy (see Berger et al. 1996, Della Pietra et al. 1997, Goldwater & Johnson 2003, Hayes & Wilson 2008, Wilson 2010 for more on Maximum Entropy learners).

To get from this initial stage to the goal of having a grammar model with maximally general constraints, a loop in which constraint induction and clustering interacted was repeated until the grammar was judged as being maximally general. The loop contained five steps:

(17) *Steps inside the looped part of the algorithm*

1. constraint selection
2. selection of contexts for clustering
3. clustering itself
4. creation of feature labels from clusters
5. weighting of constraints selected at step 1 in the grammar model

A grammar was judged as being maximally general when the constraints currently in the grammar, and their weights, created such a probability distribution that the grammatical CVCVC forms had at least 95% of all the likelihood. The presence of regularization (see section 3.2.5) in the model makes it impossible for the model to assign a large amount of likelihood to the grammatical (observed) forms if the constraints in the model are too specific. This means that a high likelihood for the grammatical forms entails that the constraints have a certain level of generality.

Section 3.2 will now give descriptions of each of the five steps in the loop summarized in (17). The full code of the algorithm can be found in the "Full code of the implementation (in R)" section at http://blogs.umass.edu/anazarov/feature-learning/.

3.2 The steps of the algorithm

3.2.1 Selection of constraints

The first step of the iterated part of the algorithm (see (17) above) is the selection of a small group of constraints. The intuition behind the selection procedure is to find one constraint or a small group of closely related constraints which correspond to a pattern in the data. In practice, the best choice was a hill-climbing algorithm that made use of a rudimentary form of evolutionary algorithms (Ashlock 2006) – to find a local peak of information gain relative to the current grammar.

In order to explain this, I will first briefly introduce Maximum Entropy grammars, and then introduce information gain. After this, I will explain the mechanism of constraint selection in more detail.

3.2.1.1 Maximum Entropy learning for phonotactics

MaxEnt models (Berger et al. 1996, Della Pietra et al. 1997, Goldwater & Johnson 2003, Hayes & Wilson 2008, Wilson 2010) make use of constraints (for instance, OT-style phonological constraints)[4] to generate a probability distribution over objects/events (for instance, phonological input/output mappings or phonological output forms). The distinctive characteristic of the MaxEnt model is that every constraint is assigned a weight such that overall information-theoretic entropy is maximized (i.e., the model makes minimal assumptions about unknown objects/events to determine the weights of constraints).

Because this type of model weights constraints only based on the training data given to the model, MaxEnt models are very useful for modeling the learning of phonological grammars from positive data.

[4] In the field of Natural Language Processing, the statements which phonologists call "constraints" are instead called "features" (see, e.g., Della Pietra et al. 1997). I will follow the phonologists' usage.

Since language-acquiring infants learn their grammars from positive data (Brown & Hanlon 1970, Marcus 1993), these models can give us insight into the acquisition of phonological grammar (Hayes & Wilson 2008).

MaxEnt models for phonology are similar to Harmonic Grammar (Pater 2009, Potts et al. 2010): both analytic frameworks share the property of having gradually violable constraints which have weights instead of ranks. In contrast to (non-probabilistic versions of) Harmonic Grammar, however, MaxEnt models do not appoint a winning candidate out of a list of candidate outputs, but, instead, they defines a probability distribution over output candidates.

Since learning in this case will be purely phonotactic (following Hayes & Wilson 2008), the grammar defines a distribution over all possible output forms, as if they all come from the same input. As had already been mentioned in section 3.1, the toy language only has "words" of the shape CVCVC, so that the set of possible output forms was also built on that pattern (see (10)). For modeling a more realistic language that has a variety of word shapes, a more varied set of output forms would have to be considered.

Each of the possible output forms is assigned some probability by the grammar. Even though each form will have a very small probability (because the number of possible forms is very large), phonotactically better candidates will still have a much higher share of probability than phonotactically worse candidates.

As had been said above, the goal of a MaxEnt model is to maximize entropy in the system. Entropy is maximal when the probability distribution assigned to candidates by the grammar equals the distribution of data points input to the learner (Manning & Schütze 1999). For this reason, the weights of the constraints in the model are adjusted so as to minimize the discrepancy between predicted probabilities (*q*) and observed probabilities (*p*) of data points, which are defined as follows:

(18) *observed probability of a candidate: count of observations for candidate x divided by total number of observed tokens of any candidate (Ω stands for the set of all candidates)*

$$p(x) = |x| \,/\, \Sigma_{y\in\Omega} |y|$$

(19) *predicted probability of a candidate given a MaxEnt model*

$$q(x) = e^{H(x)} \,/\, \Sigma_{y\in\Omega}\, e^{H(y)}$$

where $H(x) = \Sigma_i\ [w_i \text{ x } C_i(x)]$

$C_i(x)$ is the penalty constraint C_i assigns to candidate *x*, and w_i is the weight given to C_i

The discrepancy between these two distributions is obtained by finding the Kullback-Leibler (K-L) divergence (Kullback & Leibler 1951) of *q* from *p*. K-L divergence is defined as follows:

(20) *Kullback-Leibler divergence of model distribution w from sample distribution t*

$$D_{KL}\ (t\ ||\ w) = \ \Sigma\ (\ t(x) * \log\ [\ t(x) \,/\, w(x)\ ]\ )$$

To maximally approximate the point of maximum entropy, the weights of constraints are adjusted so as to minimize the K-L divergence of the model distribution, *q*, from the sample distribution, *p*[5]:

[5] This function was minimized by the L-BFGS-B method built in into R (Bates et al. 1997-2014). See http://stat.ethz.ch/R-manual/R-patched/library/stats/html/optim.html for documentation and references.

(21) *Objective function of MaxEnt grammar*

$$\text{Obj} = \min_w [\ D_{KL}(p \parallel q)\ ] = \min_w [\ \Sigma\ (\ p(x) * \log [\ p(x) / q(x)\ ]\ )\ ]$$

As had been mentioned in section 3.1, the model also had a component in it that encourages generalization: this was an L2 regularization prior. This is a term which penalizes the grammar for high constraint weights. Specifically, the difference between every constraint weight in the model and a model mean is squared, and the results of this are added together and divided by a constant (twice the variance). The mean was set to 0, and the variance was set to 1,000.

(22) *Objective function of MaxEnt grammar with L2 prior*

$$\text{Obj} = \min_w [\ D_{KL}(p \parallel q) + \Sigma_i\ (\ (\ w_i - \mu\ )^2 / 2\sigma\ )\ ]$$

This has the effect that, when a pattern is represented by many small constraints (e.g., *#m, *#n, *#ŋ), these constraints will not be allowed to have high enough weight to truly distinguish between grammatical and ungrammatical candidates, while assigning weight only to a single constraint for the same pattern (e.g., *#[nasal]) be punished much less, so that this single constraint will be allowed to have a higher weight. This is because the objective function must be minimized, so that the optimum of the function strikes a compromise between maximum entropy and minimum squared summed constraint weights. The constraints *#m*#n, *#ŋ each need the same weight as *#[nasal] to account for the data, but if *#m, *#n, *#ŋ are actually given these weights, their penalty is three times the penalty for *#[nasal]; to mitigate this penalty, the model is forced to give each of *#m, *#n, *#ŋ a lower weight than they need to fully account for the pattern.

This is the reason why the criterion of 95% total model-assigned likelihood (q) on grammatical candidates, mentioned in section 3.1, makes sure that the grammar has enough generalization in it – because a lack of generalization in the constraints bars the grammar model from giving constraints high enough weight to assign a high enough portion of likelihood to the grammatical forms. I will return to the effects of the L2 prior in section 3.2.5, where I will discuss its effects on the way in which constraint weights are updated in the course of the simulation.

3.2.1.2 Information gain

The concept that will be crucial to constraint selection is information gain (see Della Pietra et al. 1997, Wilson 2010). This is an information-theoretic measure which estimates how much a new constraint could maximally improve the grammar model (in terms of its objective function).

In more technical terms, information gain measures the maximal drop in the divergence of the predicted distribution (*q*) from the observed distribution (*p*) that could happen if the new constraint *C** were added to the grammar with weight *w**. The weight of the new constraint may be varied for the purpose of maximization (but not the weights of the other constraints in the grammar). This is expressed in the following formula (where q_{w*C*} stands for "the distribution over candidates defined by the current grammar q to which constraint *C** is added with weight *w**"):

(23) *Information gain*

$$G(w^*,C^*) = \arg\max_{w^*} [\ D_{KL}(p \parallel q) - D_{KL}(p \parallel q_{w^*C^*})\ ]$$

I will use this measure in the procedure of constraint selection, which will be described in the next subsection.

3.2.1.3 Constraint selection

The constraints used in this implementation were negative (penalty-assigning) constraints which penalized sequences of two or three elements (i.e., bigrams or trigrams). These elements could be either a word boundary, or members of a set consisting of all the segments in the segment inventory of the toy language, and all the phonological features currently in the model[6]. I will call this latter set σ′. The following are examples of possible constraints:

(24) *Possible constraints*

*t#
*a[labial]
*amu

The penalty (violations) assigned to a candidate by a constraint was computed by encoding the sequence of elements in the constraint definition as a regular expression, and counting the number of occurrences of the expression in the candidate.

Constraints were induced through a procedure driven by the concept of information gain as laid out above. The process of inducing a constraint (or small group of constraints) consisted of two steps: a random search step, and an optimization step, each of which will be detailed below.

At the random search step, the algorithm generated a random bigram or trigram built out of the elements of σ′ (plus #, the word boundary) and computed that constraint's information gain value. This was repeated until a constraint with an information gain of at least 0.01 nats was found. For instance, this could have been the constraint *#ba.

This constraint with an information gain of at least 0.01 nats was then used as the seed for the optimization step. The optimization step continuously modified its input constraint until a peak of information gain was reached. Three types of modification were used:

(25) *Types of modification*

1. deletion of a random element (for trigrams)
   or insertion of a random element in a random position (for bigrams)
2. change a random segment to another segment (not applicable to features)
3. change a random feature or segment to another feature[7]

Whenever a modification resulted in a higher information gain, this modification was kept, and the result of this modification was used as the basis for a new series of modifications. (This is a technique loosely based on evolutionary algorithms (Ashlock 2006).) The modifications were halted as soon as 10 modifications in a row yielded no improvement in information gain, it was assumed that a peak in information gain value had been found. One possible path of improvement is the following:

(26) *Example of improvement path*

*#ba → *#ma → *#m

The output of this optimization step was the constraint from which no improving modifications could be made, plus whichever other constraints that were found by modifying that constraint which had the same information gain (within some range). For instance, for *#m, these neighboring constraints with the same information gain were *#n and *#ŋ. The output of the optimization step for this example, therefore, is the set {*#m, *#n, *#ŋ}.

If there were no neighboring constraints with the same information gain value, only the constraint which represented the peak in information gain was output by the procedure (for instance, *m# had no neighboring constraints with similar information gain values). The one or several constraints selected by

[6] As will be seen in section 3.4, features were actually implemented as sets of segments – which were inserted in constraint definitions as alternative sets (“m OR n OR ŋ”, encoded as [mnŋ] in regular expressions).

[7] If only one feature is available at the time when this modification is performed, then either a segment is replaced by that single feature, or that feature is replaced by a random segment.

this two-step algorithm were then passed on to the second step – which was the creation of contexts in which feature induction was to take place, based on the constraints just induced.

3.2.2 Context creation

Once a group of constraints had been selected at the first step – for instance, the set {*#m, *#n, *#ŋ} – these constraints became the basic material for feature induction. Since feature induction was done by clustering constraints per context, the first step in this process was to find all possible contexts inherent in the constraints just selected.

A context was defined as a phonological configuration with one time slot missing. As a consequence of this definition, the set of contexts that could be created from a single constraint consisted of every way of removing one time slot from that constraint. For instance, the single constraint *ubi would have yielded the following set of contexts:

(27) *Set of contexts for* *ubi
*_bi
*u_i
*ub_

This procedure of finding contexts was repeated for every constraint that was in the set selected at step 1 – and then all unique (non-repeating) contexts thus found were retained and passed on to the next step. For the set of constraints {*#m, *#n, *#ŋ}, which had been selected at step 1, this yields the following contexts:

(28) *Set of contexts for* {*#m, *#n, *#ŋ}
*_m (found for *#m)
*_n (found for *#n)
*_ ŋ (found for *#ŋ)
*#_ (found for all three constraints)

The contexts thus found were passed to the next step of the loop, which was clustering – a crucial step toward finding features.

3.2.3 Clustering

Once contexts were found, as in the preceding step, these contexts were tabulated against the segment inventory of the toy language. The table constructed for the contexts obtained at the previous step (*_m, *_n, *_ŋ , *#_) is shown below:

(29) *Context table for* {*#m, *#n, *#ŋ}

	a	i	u	p	t	k	b	d	g	m	n	ŋ
*_m												
*_n												
*_ŋ												
*#_												

This table was populated by the information gain values (see 3.2.1.2 above) for the constraints constructed by inserting the segment corresponding to the column in the context corresponding to the row. For instance, the cell in the "u" column and the "*_m" row stands for the constraint *um.

The table below shows information gain values for the *#_ row of the previous table:

(30) *Context table values for *#_*

	a	i	u	p	t	k	b	d	g	m	n	ŋ
*#_	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.015	0.015	0.015

Every row of the table, filled out according to this procedure, was then passed on to the clustering procedure. Clustering was done by a Mixture of Gaussians model (which is a special case of the Finite Mixture model – see Everitt 2011).

For every context (as represented by a row in the table), a 2-component, equal-variance Mixture of Gaussians model was fit to the vector of information gain values corresponding to that context[8]. This was done to separate the segments that were active in a context from the segments inactive in a context.

The information gain values in every cell of a row of the table represent the degree to which inserting a segment in a constraint context and adding it to the grammar improves the grammar's fit to the data. For a given constraint context (e.g., *#_) and a given segment (e.g., [b]), this measure provides an estimate of the likelihood of whether that segment participates in the pattern denoted by that context ("How likely is it that [b] participates in the pattern [no word-initial _] ?").

The participation of a segment in a phonological pattern reveals its phonological function. It is in this sense that a clustering model over information gain values in a context can reveal something about the phonological function of groups of segments.

As can be seen above visually, the values for [m], [n], and [ŋ] are much higher than for the other segments within the context *#_ (which is a reflection of the fact that the data lack precisely [m n ŋ] word-initially). The 2-component Mixture of Gaussians model found exactly that division:

[8] Because the vector of values to cluster over was so short, and the information gain values within a cluster tended to be identical, the model was unable to fit combinations of statistical distributions to these data. This was remedied by adding random noise to the data (an amount of 0.0000001 was added to random table cells).

(31) *Division into higher-mean component (bolded) and lower-mean component (not bolded)*

	a	i	u	p	t	k	b	d	g	**m**	**n**	**ŋ**
*#_	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	0.002	**0.015**	**0.015**	**0.015**

After the Mixture of Gaussians model had been fit, the component with the highest mean (corresponding to the "active" segments) was used as the basis for a new feature, when appropriate (see the next section on what was seen as "appropriate"). The clustering model was run on every context in the table, so that every context discovered at the preceding step (in our example, this would be *_m, *_n, *_ŋ, *#_) could potentially give rise to a new feature.

3.2.4 Feature induction

Once a Mixture of Gaussians model was fit to the data for one of the contexts, the Gaussian component with the higher mean was taken (since this Gaussian component stands for the segments active with that context), and the likelihood of every segment under that Gaussian component was computed. An approximation of the likelihood vector for the context *#_ is displayed below:

(32) *Likelihood vector for higher mean Gaussian in context *#_*

	a	i	u	p	t	k	b	d	g	m	n	ŋ
*#_	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	5.2×10^{-9}	1	0.99	0.99

This likelihood vector was used for deciding whether a certain segment was part of the class of segments active in a context. As is standard practice in Finite Mixture models, a segment was classified as part of the active class whenever the likelihood of that segment was at or above 0.5. In the example above, this yields [m, n, ŋ] as being part of the class active in the context *#_ .

There were exactly two situations in which such a likelihood vector was not stored. One situation was when a Mixture of Gaussians model of the desired kind could simply not be fit to the information gain values of a context, for instance because there was too little variation between information gain values in that context[9].

The other situation is when the model gave only one segment ≥ 0.5 likelihood under the higher mean Gaussian – in other words, when the model threatened to induce a feature that is assigned to one single segment only[10].

[9] I used the Mclust implementation of Mixture of Gaussian clustering from the mclust package, version 4.2, in R (Fraley et al. 2011). When this implementation failed to fit a model with 2 equal variance components to the data for a context, no likelihood vector could be recorded.

[10] While it is advantageous in hierarchical feature assignment models such as the one proposed by Dresher (2009) to assign a feature to a single segment (because the function of features is to contrast segments), it is not advantageous to do so in the current model, because the function of features in this model is to summarize phonological behavior and make phonological patterns more easily representable. Despite this, it of course remains essential to account for the notion of phonological contrast in any model, and in future work the model may be extended to account for contrast as well.

Such was the case for the context *_#, for instance: only [m] was banned word-finally, so that only the segment [m] would be predicted by the model to be one of the active segments in that context. Since intersections of features were computed and allowed in constraints, and the intersection of [labial] and [nasal] yielded exactly the segment [m] in the toy language, the grammar still has a way of appealing to the single segment [m] through features alone.

The reason why I did not allow feature labels that stand for one segment is to let feature labels always be more general than their corresponding segments. If the generality of a representational unit is measured in terms of how broad a part of the articulatory and acoustic spectrum it covers, then this means that feature labels must cover more articulatory or acoustic ground. Since the classificatory features in my model have no intrinsic definition (as opposed to Chomsky & Halle's (1968) features, which are always grounded in some phonetic dimension), a feature which classifies only one segment of a language denotes only that segment, and nothing more. For example, a feature [X] which stands for the segment [m] and no other segments would not be more general than [m], since the set of acoustic realizations of [m] would be the same as the set of acoustic realizations of [X].

Crucially, however, the absence of features that refer to one segment does not bar the system from referring to a single segment by feature labels. As I will explain toward the end of this subsection, features were allowed to combine to define a set of segments. For instance, the combination of [labial] (={p,b,m}) and [nasal] (={m,n,ŋ}) yields [labial, nasal], which stands for [m]. It is in this way that the model was able to refer to single-segment classes through features – and, as will be seen in the results section, 2 of the 32 runs of the simulation actually did have a constraint *[labial,nasal]#. This means that the result that grammars obtained from my learning model generally did not refer to single-segment classes through features is not pre-encoded in the model itself, but truly is an emergent fact.

The assignment of feature labels to segments proceeded through the likelihood vectors that were stored for each Mixture of Gaussians model (and, thus, for each context). A feature label was assigned to the segments that had at least 0.5 likelihood under that model.

To prevent the same segment class from being assigned several feature labels (since the same class can be active across multiple contexts; for instance, the set [m,n,ŋ] could be active in several contexts, but it should not be assigned 4 different labels), the following safeguard was used. Whenever a new likelihood vector had been found through Mixture of Gaussians analysis of some context, the similarity of that new likelihood vector to every previously stored likelihood vector (if any were present at that point) was computed. The similarity between two likelihood vectors was computed based on the "profile" of the two vectors: are relatively larger values in the same place in both vectors[11]? If the similarity to some existing likelihood vector exceeded a certain threshold (0.9 on a scale of 1 in this case), then the new likelihood vector was stored under the label of that existing likelihood vector. Otherwise, the new likelihood vector was stored under a new, randomly generated label.

Thus, to summarize, a new label was generated every time a fit of the Mixture of Gaussians model found a likelihood vector over the set of segments which was unlike any previously stored vector. For each labels which was found at a given iteration of the cycle, its likelihood vector was converted into a set of segments by taking all the segments that had at least 0.5 likelihood under that label (the same criterion

[11] This was done by normalizing both vectors with an L2 normalizer so that they both summed to 1, and then taking the dot product of these vectors. The large numbers in these vectors will be closer to 1, and the smaller numbers will be closer to 0. The only pairings which will be able to bring the dot product close to 1 are pairings of two large numbers.

$$\text{Similarity}(v,w) = \left([v_1, ..., v_n] / \sqrt{(v_1^2 +... + v_n^2)} \right) * \left([w_1, ..., w_n] / \sqrt{(w_1^2 +... + w_n^2)} \right)$$

as described above)[12]. For instance, if the likelihood vector in (32) above was assigned the random label "XM", then the procedure just described yields a statement like "XM = {m,n,ŋ}".

Even though labels assigned to clusters of segments were arbitrary, I will use the names [labial] and [nasal] for {p,b,m} and {m,n,ŋ}, respectively, when presenting the results of the simulations in section 4. However, the use of the labels [labial] and [nasal] does not mean that the classes that they denote have any meaningful phonetic interpretation.

Since single segments can only be referred to by a combination of classificatory features and not by single features (given the definition of features given above; for instance, [m] can only be referred to by [labial, nasal], not by any single feature), it was necessary to also define segment sets which correspond to a combination of features.

After the segment sets corresponding to the new labels induced at an iteration had been determined (for instance, "XM = {m,n,ŋ}"), the following procedure was followed: for each newly induced label, the intersection of that label's segments and each other label's segments was computed. For instance, if the set of previously induced labels was as in (33) below, the intersection between the newly induced label "XM = {m,n,ŋ}" and these labels would be as in (34).

(33) *Examples of labels induced for sound classes*
LO = {p,b,m}
KV = {a,i,u}
TR = {i,u}

(34) *Intersections between XM ={m,n,ŋ} and the features in (32)*
{m}
{}
{}

The sets of segments corresponding to new single labels and new combinations of labels were subsequently added to σ′, the repertoire of representational units out of which constraints could be built (see section 3.2.1.3). In this fashion, the new features were allowed into constraints that were to be induced at the next iteration of the cycle.

3.2.5 Constraint weighting

Once the constraints selected at the first step had been exploited for feature induction at steps 2 to 4, these originally selected constraints were added to the grammar. This was done by adding the newly induced constraints to the previously induced constraints (if any), and adjusting the weights to minimize the objective function (see (22) in section 3.2.1.1).

The newly induced constraints started out with a weight of zero, while the old constraints started out at their previous weight; these initial weights were then optimized to minimize the objective function.

For instance, if the constraints *#m, *#n, *#ŋ were added to an empty grammar, the constraint vector and the weight vector before and after weight optimization would look as displayed in (35) below. The weight of 6 happens to be ideal for each of these constraints to express the absence of word-initial [m], [n], and [ŋ]. The constraints end up with equal weights since they punish the same number of candidates, because the number of candidates starting in [m] is the same as that of candidates starting in [n], or [ŋ].

[12] For labels which had more than one likelihood vector stored under them, all these vectors were averaged and the segments which had at least 0.5 likelihood under that averaged likelihood vector were said to belong to that label.

(35) *Reset of specific constraints to zero*

Before optimization	
Constraint	Weight
*#m	6
*#n	6
*#ŋ	6
*#[nasal]	0

After optimization	
Constraint	Weight
*#m	0
*#n	0
*#ŋ	0
*#[nasal]	8

This zero-reset effect follows from the fact that regularization was very strong: the variance in the prior was set very low[13] (variance = 1,000), so that there was a strong bias against assigning any amount of weight to constraints that were not general enough. For this reason, the optimization algorithm found that the best set of weights was such that none of the specific constraints *#m, *#n, *#ŋ received any weight, and the general constraint receives all the weight.

If there were no zero-reset effect, less specific constraints such as *#m, *#n, *#ŋ this would actually be more advantageous, since this paper seeks to find evidence that segments and features each have separate roles in grammar.

However, the zero-reset effect that appears here is perfectly in line with the standard hypothesis that phonological grammar refers to features only: since the constraints that refer to the individual segments [m], [n], and [ŋ] have zero weight in the right-hand part of (35), the grammar no longer refers to the individual segments [m], [n], [ŋ] as far as banning them in word-initial position is concerned.

In this manner, the zero-reset effect gives the standard hypothesis of an all-feature grammar a head start: if the zero-reset effect happened for all instances where a feature-based constraint replaces a segment-based constraint, the learner would end up with an all-feature grammar.

However, the zero-reset effect, as described above, only happens when a new constraint is more general and covers a strict superset of the forms that one or more old constraints account for. When a new constraint is homonymous with an old constraint, both constraints receive non-zero weight after optimization. For instance, when a grammar has the constraint *m#, and the constraint *[labial,nasal]# is added to that grammar, both constraints retain non-zero weight after optimization, so that the grammar refers to both features and segments.

This difference between the two situations just described (feature-based constraint punishes a strict superset of forms punished by segment-based constraints; feature-based constraint is homonymous with segment-based constraint) follows from the nature of the L2 prior. This is because a new constraint like *[labial,nasal] does not take over the function of more than one old constraint – so that assigning all weight to that constraint cannot lessen the penalty imposed on high weights for individual constraints (since adding weight to *[labial,nasal] does not take away weight from more than one constraint).

After constraints were weighted, the total likelihood assigned by the resulting model to the grammatical candidates was assessed. If this likelihood was less than 95%, another iteration of the cycle

[13] Simulations run with a weaker bias (for instance, with variance = 5,000) do not exhibit a zero-reset effect, but these simulations also do not find more general constraints which supersede previously induced specific constraints (as is the case for *#[nasal] versus *#m, *#n, *#ŋ).

described in this subsection (3.2) was initiated. Otherwise, the constraints and weights as fixed at the latest iteration were output as the final grammar.

4. Results

A run of the computational simulation as described in the previous section yields a grammar in the form of a set of constraints, and a weight for each constraint. As explained in section 3.2.1.1, these constraints and their weights generate a probability distribution over the space of possible CVCVC distributions. An example of a grammar generated by a run of the simulation is given here:

(36) *Results of example run*

Constraint	Weight
*m#	2.45
*ubi	0
*upi	0
*umi	0
*umu	0
*imu	0
*imi	0
*ipi	0
*ipu	0
*upu	0
*ubu	0
*ibu	0
*ibi	0
*#n	0
*[high][labial]	0.08
*m[high]	0
*#[nasal]	3.87
*[labial][high]	0.05
*[high]b[high]	1.57
*i[labial]i	0
*u[labial]i	0
*i[labial]i	0
*[high][{p,m}][high]	2.26

32 independent runs of the simulation were performed, and it is the 32 grammars resulting from these runs that will be analyzed.

For the purposes of analysis, I will be interested in only one aspect of these grammars: what type of units do the constraints in the grammar refer to – segmental units or feature labels? To isolate this aspect, I extracted, for every grammar, all constraints that had non-zero weight in that grammar, and I inspected their definitions.

Only constraints with non-zero weight were included, because constraints with zero weight have no influence whatsoever on the outcome of the grammar, and the grammar would have had the same effect if these constraints did not exist. Since grammars were continuously updated with new constraints, and new constraints could reset old constraints to zero (as explained in section 3.2.5), but constraints were never removed, some constraints in the final grammar had zero weight[14]. The table of constraints in (36) above yields the following constraints with non-zero weight:

(37) *Non-zero constraints from (36)*

Constraint	Weight
*m#	2.45
*[high][labial]	0.08
*#[nasal]	3.87
*[labial][high]	0.05
*[high]b[high]	1.57
*[high][{p,m}][high]	2.26

This grammar (which is the grammar generated at the first run of the simulation) was represented as the following set of constraints for the purpose of analysis:

(38) *(37) as a set of constraints*
{*m#, *[high][labial], *#[nasal], *[labial][high], *[high]b[high], *[high][{p,m}][high]}

The non-zero-weight constraints for each grammar were subsequently sorted according to which of the three phonotactic patterns of the toy language (see section 2.3) they encoded:

(39) *The three phonotactic patterns*
a. no final [m]
b. no initial nasals
c. no labials between high vowels

Most constraints clearly represented one of these three patterns. However, 8 out of 32 grammars also included one of the following two constraints[15]:

[14] The optimization procedure used (**optim** in R; see http://stat.ethz.ch/R-manual/R-patched/library/stats/html/optim.html for documentation) distinguishes between true zero and very small numbers in its output.

[15] 4 grammars had *Vm; 4 grammars had *mV.

(40) [m] *plus vowel constraints*
a. *Vm (no vowel followed by [m])
b. *mV (no [m] followed by a vowel)

These constraints do not represent a unique pattern among the three listed in (40) above. Rather, (41a) partially represents patterns (40a) and (40c). Word-final [m] always occurs after a vowel (since only CVCVC word shapes were considered), but not every postvocalic [m] is word-final. [m] is one of the labial sounds that may not occur in between high vowels, and V_ is a partial description of that context.

In a similar way, *mV (=(41b)) is a partial description of patterns (40b) and (40c). Every word-initial [m] (which is one of the nasals, that are prohibited word-initially) is also prevocalic, but not every prevocalic [m] is word-initial. Every [m] in between two high vowels is prevocalic, but not every prevocalic [m] is in between two high vowels.

These 2 constraints will be excluded from the discussion below, as the discussion will focus on the shape of constraints per phonotactic pattern. Specifically, I will look at how often features and unanalyzed segments were used to encode these patterns. I will first examine the word-final pattern ("no word-final [m]"), then the word-initial pattern ("no word-initial nasals"), and, finally, the word-medial pattern ("no labials in between two high vowels").

A document containing the complete results of the 32 runs of the simulation can be found under "Full simulation results" at http://blogs.umass.edu/anazarov/feature-learning/.

4.1 No word-final [m]

The pattern which prohibits word-final [m] was always represented in the grammars with one or both of the following two constraints:

(41) *Constraints for "no word-final [m]"*
a. *m# (penalty of 1 for every word-final segment which is [m])
b. *[labial,nasal]# (penalty of 1 for every word-final segment which is a labial nasal)

However, there was a very strong preference for using *m#. 30 out of 32 grammars only had the segment-based constraint *m#, not the feature-based *[labial,nasal]#. One grammar had both *m# and *[labial,nasal]#, and it was just the one remaining grammar that had *[labial,nasal]# without *m#.

Thus, we can say that 30 grammars encoded this pattern with reference only to segments, 1 grammar encoded it with reference both to features and segments, and 1 grammar encoded it with reference only to features. This reveals a strong bias to represent the one-segment pattern "no word-final [m]" with segment-based constraints.

Comparison between this pattern and the other two patterns will reveal if this tendency towards segmental representation is unique to the "no final [m]" pattern, as was predicted by the emergent feature model (section 2.4).

4.2 No word-initial nasals

The pattern which bans word-initial nasals [m, n, ŋ] in the toy language was represented by the constraint *#[nasal] in 29 out of 32 grammars. In 3 of these 29 grammars, this constraint was accompanied by *#[nasal]V.

(42) *Most frequent representations for "no word-initial nasals"*
a. *#[nasal] (penalty of 1 for every word-initial nasal segment)
b. *#[nasal]V (penalty of 1 for every word-initial prevocalic nasal segment)

These two constraints are homonymous for the forms examined here, since all words have the shape CVCVC.

The remaining 3 (out of 32) grammars use feature labels which do not correspond to traditional features: [{m,ŋ}] (a feature assigned to [m], [ŋ], but not [n]) and [{n,ŋ}] (a feature assigned to [n], [ŋ], but not [m])[16]. These grammars are shown below.

(43) *Grammars that appeal to strange "features"*

Run #	Constraints representing word-initial pattern
11	*#m, *#[{n,ŋ}], *#[{n,ŋ}]V
16	*#[{m,ŋ}], *#[{n,ŋ}]
17	*#m, *#[{n,ŋ}]

Of these three deviant grammars, only the grammars at runs 11 and 17 employ a segment-based constraint (*#m) – and even then it is in conjunction with at least one feature-based constraint.

From these data, we may conclude that there is a strong bias toward representing the pattern which prohibits word-initial nasals with features. None of the 32 grammars represented the pattern exclusively with segment-based constraints, and only 2 grammars represented the pattern with a combination of a segment-based constraint and one or more feature-based constraints (see runs 11 and 17). The other 30 grammars refer only to features with respect to the word-initial restriction.

4.3 No labials between high vowels

Finally, the word-medial pattern (which penalized labial consonants in between high vowels) found a much more variant grammatical representation, owing to its complexity.

17 out of 32 grammars represented this pattern with the constraint *[high][labial][high], as was expected. In 3 of these 17 grammars, *[high][labial][high] co-occurred with the constraint *[high][labial], and in 1 of the 17, *[high][labial][high] co-occurred with *[labial][high].

(44) *Most frequent representation for "no labials between high vowels"*

a. *[high][labial][high] (penalty of 1 for every sequence of a high vowel, a labial consonant and another high vowel)
b. *[high][labial] (penalty of 1 for every sequence of a high vowel followed by a labial)
c. *[labial][high] (penalty of 1 for every sequence of a labial followed by a high vowel)

In any event, these 17 grammars all refer to features only. The remaining 15 grammars had constraints of various shapes and sizes: both two-position and three-position constraints ((45a-b)), and constraints which referred to segments only, features only, or combinations of features and segments ((45c-e)).

[16] I would like to thank an anonymous reviewer for pointing out that these do at least correspond to intersections of phonological features proposed in the literature. [m, ŋ] can be described as [+nasal, -coronal] in Chomsky & Halle's (1968) system, or as [nasal, peripheral] according to the systems proposed by Dogil (1988) and Avery & Rice (1989). [m,n] can be described as [nasal, lingual] in Clements & Hume's (1995) system.

(45) *Other constraints for "no labials between high vowels"*

a. two-position constraint: *[high]m
b. three-position constraint: *[high]m[high]
c. segments-only constraint: *mu
d. features-only constraint: *[high][{p,b}][high]
e. features-and-segments constraint: *u[labial][high]

Although there was considerable variation in the constraints representing the word-medial pattern in these 15 grammars, there was one clear generalization: none of these grammars represented the pattern only in terms of segment-based constraints. The table below provides some examples of how the word-medial pattern was represented in these grammars:

(46) *Other constraints for "no labials between high vowels"*

Run #	Constraints representing word-initial pattern
8	*[high]m; *[high][{p,b}][high]; *[high][labial]i; *[high][labial]u
14	*[high]m[high]; *[high][{p,b}][high]
27	*mi; *mu; *[high]m; *u[{p,b}][high]; *[high][{p,m}]{high]

Summarizing, none of the grammars represented the word-medial pattern in terms of segment-based constraints only; 15 out of 32 grammars represented the pattern while appealing to a mixture of segments and features, and the 17 remaining grammars represented the pattern with appeal to features only.

4.4 Summary of results

The three subsections above showed clear differences in grammatical representation between the three phonotactic patterns in the toy language. The table below summarizes these differences – with type of units appealed to (segments, segment/features, features only) tabulated against phonotactic pattern:

(47) *Type of unit appealed to for each phonotactic pattern*

Phonotactic restriction	Constraints only appeal to segments: # of grammars	Constraints appeal to mixture of segments and features: # of grammars	Constraints only appeal to features: # of grammars
no word-final [m]	30	1	1
no word-initial nasals	0	1	31
no labials between high Vs	0	15	17

This table shows that there was an overwhelming preference for the word-final pattern, which appeals to the single segment [m], to be represented in terms of the segment [m] only. At the same time, neither of the two other patterns, both of which are based exclusively on multi-segment classes, are represented with segments only in any of the grammars.

While the word-initial pattern has an almost absolute preference for being represented with features only, the word-medial pattern is represented with features only in about 1 in 2 runs of the simulation. This latter fact can be attributed to the length and complexity of the constraint *[high][labial][high], in which all three slots are occupied by features.

In any event, there is a clear asymmetry between the one-segment pattern (“no final [m]”) and the two multi-segment patterns (“no initial nasals”, “no labials in between high vowels”): the one-segment pattern is almost always represented with just a segment in the constraint, whereas this does not happen for the multi-segment patterns. This matches the intuition given in section 2: a truly emergentist model of feature learning leads to a grammar which appeals to a spectrum of levels of abstraction (such as segments/allophones and features), instead of a grammar which always appeals to the highest available level of abstraction (classificatory phonological features).

5. Discussion and conclusion

This paper introduced a radically emergentist model of phonological feature induction. This model has induction of phonological features alongside induction of OT-style constraints – a combination that had not been explored before. The model had as its goal to create grammars with maximally general constraints – following applications of the same principle in Albright & Hayes (2002, 2003) and Hayes & Wilson (2008). This created a model in which the presence of phonological features is motivated by the learning of grammar itself: phonological features are one of the instruments by which constraints can be made more general. This is why features in this model are truly emergent.

One reason why this model is interesting is because it sees features as truly emergent – in the sense that their presence is motivated by external factors (namely, generality in the grammar's constraints). However, even more interestingly, I showed that the model derives predictions that can be tested within an individual language, as opposed to typological predictions which can be tested across languages (an approach explored by Blevins 2004, 2006, Mielke 2004, and others).

The canonical view, in which the vocabulary for segmental phonology (in the form of phonological features) is predefined and innate (see, for instance, Chomsky & Halle 1968), allows only for grammars which refer to features in constraints which represent segmental patterns. However, I showed through computational implementation of the radically emergentist model proposed here that this model predicts grammars which refer to a mixture of various levels of abstraction. In the case investigated here, these levels of abstraction were unanalyzed segment (allophone/phoneme) units and phonological features. Such grammars emerged from this learning model because this model's goal is to learn grammars with maximally general and concise constraints, independently of the level of representational abstraction (segment, feature, ...) to which these constraints refer.

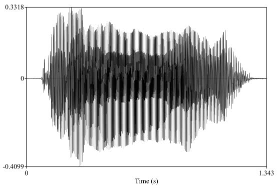
Grammars referring to different levels of abstraction in phonological representation have interesting properties in and of themselves. For instance, the segment and the feature in the current model can be seen as distinct levels of representation. One and the same sound event could be described as either [m], or as [labial, nasal], and a constraint can refer to either mode of description, as shown in (48) below:

(48) *Different levels of abstraction at which a sound transcribed as [m] can be represented*

featural representation: [labial,
nasal]

segmental representation: [m]

sound:



The current model does not allow mismatches between segments and features (i.e., a segment [m] is always attached to [labial] and [nasal], and *vice versa*). However, segment units, in this model, can exist without features (as in the initial state of the learner, for instance). Moreover, the feature representations for each segment are learned in the process of learning the grammar. These ideas together suggest that, conceptually speaking, the grammar should represent the mappings between features and segments – which was left out of the current implementation for the sake of simplicity.

A more extended model in which the mappings between segments and features are represented in the grammar would be forced to implement these mappings as violable constraints of the sort displayed in (49). This is because in the general framework of Optimality Theory/Harmonic Grammar in which the current model is set, every grammatical statement is a violable constraint.

(49) *Constraints for regulating segment/feature matching in a more extended model*

a. x ∈ {[p], [b], [m]} → labial(x): One violation mark for every segment which is one of {p, b, m} and does not have the feature [labial].

b. x ∈ {[m], [n], [ŋ]} → nasal(x): One violation mark for every segment which is one of {m, n, ŋ} and does not have the feature [nasal].

In a more extended model which contains this type of constraints, the GEN component of the grammar will have to consider candidates in which the intended segment/feature pairings are not respected. Some examples are given in (50):

(50) *Examples of mismatching representations in a more extended model*

a. [nasal] ([m] is not assigned the feature [labial])
[m]

b. [labial] ([b] is assigned [nasal])
[nasal]
[b]

Even though the representations in (50) violate the intended segment/feature mappings, there are certain constraint rankings under which such representations might receive high probability. For instance, (49a) might receive higher probability than a candidate in which [m] is classified as both [labial] and [nasal] in a grammar in which a constraint against the feature [nasal] in a certain context far outranks constraint (49a).

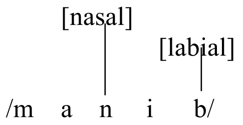
The conceptual possibility of mismatches between independent levels of representation is reminiscent of such models as Turbidity Theory (Goldrick 2001), Colored Containment (Oostendorp 2008), Abstract Declarative Phonology (Bye 2006) and Bidirectional OT (Boersma 2007, 2011). Models of this kind allow for accounts of phonological opacity (Kiparsky 1968, 1973) without the use of derivational mechanisms (see, for instance, Goldrick 2001, Bye 2006, and Boersma 2007 for examples of such accounts).

Conceptually speaking, there is no reason why grammars in which different levels of abstraction are allowed to mismatch (along the lines sketched above) could not be used to the same effect. This is an interesting and important line of research that is worth pursuing in the future.

One final note is due with regard to the multi-leveledness of representations that arise in these simulations. Since both segments and features are independent parts of the surface representations that emerge, Richness of the Base (Smolensky 1996) requires that the grammar pick a grammatical surface representation for underlying forms that have /m/ without a corresponding feature specification, as well as underlying representations that only have a feature specification [nasal, labial] but no segment /m/ to go with it. Such underlying representations are given in (51).

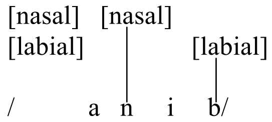
(51) “*Mismatching” underlying representations*

a.



(first segment is /m/ but does not have a [labial, nasal] specification)

b.



(first segment has a feature specification but no segment specification)

Since the current grammar model has the mapping between [nasal, labial] and /m/ specified outside the grammar, every input in which [labial, nasal] occurs without /m/, or *vice versa*, would only be mapped onto outputs in which [m] and [labial, nasal] go hand in hand.

In the more extended model sketched in the paragraphs above (where the mapping between segments and features is regulated by violable constraints), inputs such as the ones in (51) could potentially surface as fully faithful outputs, if the constraints regulating the link between [labial, nasal] and [m] are sufficiently low-ranked. However, such a low ranking would only be justified by some kind of opaque phenomenon in which outputs with [m] but no [labial, nasal] specification (or [labial, nasal] but no [m] specifications) are somehow needed. Otherwise, the constraint that regulates the co-occurrence of [labial, nasal] and [m] would be high-ranked and would force every winning output for the inputs in (51) to have [labial, nasal] wherever [m] occurs, and *vice versa*.

Another interesting property of grammars like the ones induced in this model is that they call to mind the existence of evidence from work in speech production and perception (e.g., McQueen et al. 2006, Jesse 2007, Nielsen 2011) that speech is processed at multiple distinct levels of abstraction (exemplars, segments, features). The techniques used in this work (which all have to do with asking subjects to expand a pattern presented to them in training data) could also be applied to test the level of abstraction referred to by various constraints in the grammar.

The latter means that the language-internal predictions of the current emergentist model are testable through behavioral experiments. This is an innovation for models with emergent explicit grammatical structure – which can be contrasted with exemplar models, in which grammatical structure is emergent but implicit (see, for instance, Wedel 2003, 2011; Blevins' 2004, 2006 model does have implicit grammatical structure, on the other hand). Models such as Mielke’s (2004) have mainly relied on typological patterns, but this model allows for potential experimental evidence in favor of the emergent feature scenario.

One example of such a behavioral test of the within-speaker predictions of the model would be to combine the methodology of the studies mentioned above with Halle's (1978) “Bach-testing”, briefly discussed in section 2.4. For instance, speakers of English may be confronted with a novel segment which conforms to the feature specification of [s], which is the only segment allowed as the first consonant in #CCC clusters – see section 2.2.

Since one might expect that the pattern should be formulated in its most concise form, the features that match [s] in the constraints are expected to be [-voice,+anterior,+strident]. An example of a novel (non-English) segment that matches that description is [s̪] (dental [s]). The question is whether speakers that are taught this new segment unit [s̪] will automatically extend the pattern to this segment (i.e., forms such as [s̪plɪt] will be as acceptable as [splɪt]).

It is not important that [s̪] may have features that do not match those of [s] – for instance, [±distributed]. All that matters is that [s] and [s̪] have the same values for [±voice], [±anterior], and [±strident] – since these are the three features which are sufficient to distinguish /s/ from the other phonemes in English. The reason why only this matters is that I assume some economy mechanism which has the same effect as Chomsky & Halle's (1968) Evaluation Metric: every constraint only refers to features that are necessary to distinguish the sounds that the constraint applies to from those that it does not apply to. In this manner, the (reward-assigning) constraint ✔#[-voice,+anterior,+strident]CC does not care about the value of [±distributed] or [±dorsal] of its first segmental position.

The canonical innate feature model predicts that the grammatical statement of the pattern of only-[s]-starting-#CCC must appeal to features only, and thus, speakers will automatically extend the pattern to the novel segment (e.g., [s̪]). On the other hand, the emergent feature model predicts that the single-segment domain ("only [s]") makes it very likely that the pattern will be represented with [s] as a segment only, without featural specification. This means that there will not be a strong impetus to generalize the pattern, so that the novel segment ([s̪]) will not be allowed at the beginning of #CCC clusters.

This and similar experimental methodologies, when applied to this problem, will shed light on the predictions of both models. Consequently, behavioral tests of this sort constitute an important direction for future work.

Another direction for future research is to extend the model sketched and implemented here to a larger part of the path between raw acoustic data and grammar – specifically, the model should be applied to learning segment categories from acoustic input. This and other projects are of high importance for the results presented here. However, the current results in themselves provide a novel perspective on the problem of phonological abstraction and the dialogue between innate and emergent approaches to phonological structure.

References

Albright, A. & B. Hayes. 2002. Modeling English Past Tense Intuitions with Minimal Generalization. In: *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology*, ACL, pp. 58-69.

Albright, A. & B. Hayes. 2003. Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study. *Cognition* 90: 119-161.

Archangeli, D., J. Mielke & D. Pulleyblank. 2012. From Sequence Frequencies to Conditions in Bantu Vowel Harmony: Building a grammar from the ground up. In: B. Botma & R. Noske (eds.), *Phonological Explorations: Empirical, Theoretical and Diachronic Issues*, Mouton de Gruyter, Berlin, pp. 191-222.

Ashlock, D. 2006. *Evolutionary Computation for Modeling and Optimization*. Springer, New York, NY.

Avery, P. & K. Rice. 1989. Constraining underspecification. In: *Proceedings of the Northeast Linguistic Society 19*, Graduate Linguistics Students Association, Amherst, MA, pp. 1-15.

Bates, D., J. Chambers, P. Dalgaard, S. Falcon, R. Gentleman, K. Hornik, S. Iacus, R. Ihaka, F. Leisch, U. Ligges, T. Lumley, M. Maechler, D. Murdoch, P. Murrell, M. Plummer, B. Ripley, D. Sarkar, D. Temple Lang, L. Tierney, & S. Urbanek. 1997-2014. The R project.

Berger, A.L., S.A. Della Pietra, & V.J. Della Pietra. 1996. A Maximum Entropy approach to Natural Language Processing. *Computational Linguistics* 22: 39-71.

Berko, J. 1958. The Child's Learning of English Morphology. *Word* 14: 150-177.

Blevins, J. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Blevins, J. 2006. A Theoretical Synopsis of Evolutionary Phonology. *Theoretical Linguistics* 32 2: 117-166.

Booij, G. 1995. *The Phonology of Dutch*. Clarendon Press, Oxford.

Boersma, P. 2007. Some listener-oriented accounts of h-aspiré in French. *Lingua* 117: 1989-2054.

Boersma, P. 2011. A programme for bidirectional phonology and phonetics and their acquisition and evolution. In: A. Benz & J. Mattausch (eds.), *Bidirectional Optimality Theory*, John Benjamins, Amsterdam, pp. 33-72.

Boersma, P., & K. Chládková. 2013. Detecting categorical perception in continuous discrimination data. *Speech Communication* 55: 33-39.

Brown, R., & C. Hanlon. 1970. Derivational complexity and order of acquisition on child speech. In:
J. Hayes (Ed.), *Cognition and the developmenf of language*, Wiley, New York, NY, pp. 11-53.

Bybee, J. 2001. *Phonology and language use*. Cambridge University Press, Cambridge.

Bye, P. 2006. *Grade alternation in Inari Saami and Abstract Declarative Phonology*. Ms., Universitetet i Tromsø.

Chomsky, N. & M. Halle. 1968. *The sound pattern of English*. Harper and Row, New York, NY.

Clements, G. N. 2003. Feature economy in sound systems. *Phonology* 20 3: 287-333.

Clements, G. N. & E. Hume. 1995. The Internal Organization of Speech Sounds. In: J. Goldsmith (ed.), *Handbook of Phonological Theory*, Basil Blackwell, Oxford, pp. 245-306.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. 2013. Constrained generalization of implicitly learned sound patterns. *Journal of Laboratory Phonology* 4 2: 259-285.

Della Pietra, S., V.J. Della Pietra & J.D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 380-393.

Dogil, G. 1988. Phonological configurations: natural classes, sonority and syllabicity. In: H. van der Hulst & N. Smith (eds.), *Features, segmental structure and harmony processes*, Part 1, Foris, Dordrecht, pp. 79-103.

Dresher, B.E. 2009. *The Contrastive Hierarchy in Phonology*. Cambridge University Press, Cambridge.

Dresher, B.E. 2013. The arch not the stones: Universal feature theory without universal features. Talk given at the Conference on Features in Phonology, Morphology, Syntax and Semantics: What are they?, Center for Advanced Study in Theoretical Linguistics (CASTL), University of Tromsø, October 31–November 1, 2013.

Dresher, B.E. 2014. The arch not the stones: Universal feature theory without universal features. Nordlyd 41.2: 165–181.

Elsner, M., S. Goldwater, N.H. Feldman, & F. Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 42-54.

Everitt, B. 2011. *Cluster analysis*. 5th edition. Wiley, Chichester, West Sussex.

Fraley, C., A.E. Raftery, T.B. Murphy & L. Scrucca. 2012. *mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation*. Technical report N. 597. Department of Statistics, University of Washington.

Gallagher, G. 2014. *An identity bias in phonotactics: Evidence from Cochabamba Quechua*. Ms., New York University.

Goldrick, M. 2001. Turbid output representations and the unity of opacity. In: M. Hirotani, A. Coetzee, N. Hall & J.-Y. Kim (eds.), *Proceedings of the Northeast Linguistic Society 30, Rutgers University*, Graduate Linguistics Student Association, Amherst, MA, pp. 231-245.

Goldsmith, J. 1976. Autosegmental phonology. Doctoral dissertation, MIT.

Goldwater, S., & M. Johnson. 2003. Learning OT constraint rankings using a Maximum Entropy model. In: *Proceedings of the workshop of variation within Optimality Theory*, pp. 113-122.

Halle, M. 1978. Knowledge unlearned and untaught: what speakers know about the sounds of their language. In: M. Halle, J. Bresnan & G.A. Miller (eds.), *Linguistic theory and psychological reality*, MIT Press, Cambridge, MA, pp. 294-303.

Hayes, B. 1999. Phonetically-Driven Phonology: The Role of Optimality Theory and Inductive Grounding. In: M. Darnell, E. Moravscik, M. Noonan, F. Newmeyer, and K. Wheatly (eds.), *Functionalism and Formalism in Linguistics*, Volume I: General Papers, John Benjamins, Amsterdam, pp. 243-285.

Hayes, B. & C. Wilson. 2008. A maximum entropy model of phonotactics and phontactic learning. *Linguistic Inquiry* 39: 379-440.

Heinz, J. 2009. On the role of locality in learning stress patterns. *Phonology* 26: 303-351.

Jansen, A., & P. Niyogi. 2008. Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition. *Journal of the Acoustic Society of America* 124 3: 1739-1758.

Jensen, J. 1993. *English phonology*. John Benjamins, Amsterdam.

Jesse, A., J.M. Page & M. Page. 2007. The locus of talker-specific effects in spoken-word recognition. In: *Proceedings of ICPhS XVI*, pp. 1921-1924.

Kimper, W. 2011. Positive constraints and Finite Goodness in Harmonic Serialism. Ms., University of Massachusetts Amherst.

Kiparsky, P. 1968. Linguistic universals and language change. In: E. Bach and R. Harms (eds.), *Universals in Linguistic Theory*, Holt, Rinehart and Winston, New York, NY, pp. 170-202.

Kiparsky, P. 1973. Abstractness, opacity and global rules. (Part 2 of “Phonological representations”). In: O. Fujimura (ed.), *Three Dimensions of Linguistic Theory*, TEC, Tokyo, pp. 57-86.

Kiparsky, P. 2006. The Amphichronic Program vs. Evolutionary Phonology. *Theoretical Linguistics* 32 2: 217-236.

Kirby, S. & J.R. Hurford. 2002. The emergence of linguistic structure: an overview of the iterated learning model. In: A. Cangelosi & D. Parisi (eds.), *Simulating the evolution of language*, Springer, London, pp. 121-147.

Kullback, S. & R.A. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics* 22 1: 79-86.

Lin, Y. 2005. Learning features and segments from waveforms: A statistical model of early phonological acquisition. Doctoral dissertation, UCLA.

Lin, Y. & J. Mielke. 2008. Discovering place and manner features: what can be learned from acoustic and articulatory data? In: J. Tauberer, A. Eilam & L. MacKenzie (eds.), *Penn Working Papers in Linguistics* 14 1: 241-254.

Maddieson, I., & Precoda, K. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104-111. Department of Linguistics, UCLA.

Manning, C. & H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, G.F. 2003. Negative evidence in language acquisition. *Cognition* 46: 53-85.

McQueen, J.M., A. Cutler & D. Norris. 2006. Phonological abstraction in the mental lexicon. *Cognitive Science* 30: 1113-1126.

Mielke, J. 2004. The emergence of distinctive features. Doctoral dissertation, Ohio State University.

Mielke, J. 2007. P-base, version 1.92. Software, University of Ottawa. http://137.122.133.199/~Jeff/pbase/index.html

Morén, B. 2006. Consonant-vowel interactions in Serbian: features, representations and constraint interactions. *Lingua* 116 8: 1198-1244.

Moreton, E. & J. Pater. 2012. Structure and substance in artificial-phonology learning. Part II: Substance. *Language and Linguistics Compass* 6 11: 702-718.

Nielsen, K. 2011. Specificity and abstractness in VOT imitation. *Journal of Phonetics* 39: 132-142.

Niyogi, P. 2004. Towards a computational model of human speech perception. In: *Proceedings of the Conference on Sound to Sense, MIT (In Honor of Ken Stevens' 80th birthday)*, pp. 208-222.

Oostendorp, M. van. 2008. Incomplete Devoicing in Formal Phonology. *Lingua* 118: 1362-1374.

Pater, J. 2009. Weighted Constraints in Generative Linguistics. *Cognitive Science* 33: 999-1035.

Peperkamp, S., R. Le Calvez, J.-P. Nadal, & E. Dupoux. 2006. The acquisition of allophonic rules: statistical learning with linguistic constraints. *Cognition* 101: B31-B41.

Pierrehumbert, J. 2003a. Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech* 46 2-3: 115-154.

Pierrehumbert, J. 2003b. Probabilistic Phonology: Discrimination and Robustness. In: R. Bod, J. Hay, & S. Jannedy (eds.), *Probability Theory in Linguistics*, The MIT Press, Cambridge, MA, pp. 177-228.

Pizzo, P. 2013. Learning phonological alternations with online constraint induction. Talk given at the Old Word Conference in Phonology, Boğaziçi University, İstanbul, January 16–19, 2013.

Potts, C., J. Pater, K. Jesney, R. Bhatt and M. Becker. 2010. Harmonic Grammar with Linear Programming: From linear systems to linguistic typology. *Phonology* 27: 77-117.

Prince, A. 2007. The pursuit of theory. In: P. de Lacy (ed.), *Cambridge handbook of phonology*, Cambridge University Press, Cambridge, pp. 22-46.

Prince, A. & P. Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Technical report, Rutgers University and University of Colorado at Boulder, 1993. ROA 537, 2002. Revised version published by Blackwell, 2004.

Reetz, H. 1999. Web interface to UPSID. http://web.phonetik.uni-frankfurt.de/upsid_info.html

Staubs, R.D. 2014. Computational modeling of learning biases in stress typology. Doctoral dissertation, University of Massachusetts Amherst.

Vallabha, G.K., J.L. McClelland, F. Pons, J.F. Werker, & S. Amano. 2007. Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences* 104 33: 13273-13278.

Wedel, A. 2003. Self-Organization and Categorical Behavior in Phonology. *Proceedings of the Berkeley Linguistics Society* 29: 611-622.

Wedel, A. 2011. Self-Organization in Phonology. In: M. van Oostendorp, C. Ewan, E. Hume and K. Rice (eds.), *The Blackwell Companion to Phonology*, Vol. 1, Blackwell Press, Oxford, pp. 130-147.

Wilson, C. 2010. Searching for phonological generalizations. Talk given at the Cornell Workshop on Grammar Induction, Cornell University, Ithaca, NY, May 15 2010.