Journal of Language Evolution, 2017, 94–112
doi: 10.1093/jole/lzx013
Research article

Structure and abstraction in phonetic computation: Learning to generalise among concurrent acquisition problems

Bill Thompson* and Bart de Boer

Artificial Intelligence Lab Vrije Universiteit Brussel Pleinlaan 2, Brussels, B-1050, Belgium

*Corresponding author: Bill@ai.vub.ac.be; Bart@ai.vub.ac.be

Abstract

Sound systems vary dramatically in their lower-level details as a result of cultural evolution, but the presence of systematic organisation is universal. Why does variation pattern differently at these two levels of abstraction, and what can this tell us about the cognitive mechanisms that underpin human acquisition of speech? We explore an evolutionary rationale for the proposal that human learning extends to, and is perhaps even specialised for, making inferences at the higher-order level of abstraction. The ability to infer systematicity from distributional cues, by identifying signatures of structural homogeneity and anticipating subtle exceptions, can bootstrap lower-level learning, and is not subject to the moving target problem, a major evolutionary objection to specialisation in speech cognition. We examine this idea from a statistical perspective, by studying the representational assumptions that underpin generalisation among concurrent phonetic category induction problems. We present a probabilistic model for jointly inferring individual sound classes *and* a system-wide blueprint for the balance of shared and idiosyncratic structure among these classes. These models lead us to an evolutionary conjecture: culture pushes cognitive adaptation up the hierarchy of abstraction in learning

Key words: generalisation; statistical inference; learning to learn; cultural evolution; phonetic computation.

1. Introduction: concurrent learning problems in phonetic systems

What computational abilities allow human infants to so readily extract complex conventional structures from ambient acoustic data, in the face of dramatic cross-linguistic variety in sound systems? Like many aspects of language, speech sounds are not acquired in isolation, but concurrently as a system of related behaviours. This basic property of cultural transmission sets up a layer of learning problems that is often overlooked in discussions of speech cognition: over and above individual classes of speech sounds, there exist regularities in the *relationships between* individual classes that determine the systematicity of a sound inventory. Are these higher-order regularities learned or assumed?

In this article, we examine this aspect of the learning problem from an evolutionary perspective. Lower-order properties of individual speech sounds vary rapidly via cultural evolution, and may therefore present a *moving target* (Christiansen and Chater 2008; Chater et al. 2009) for biological evolution of specialised learning. In contrast, higher-order system-level regularities can be

© The Author 2017. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

understood as stable features of the learning problem across languages and populations (Dunbar and Dupoux 2016): this may defuse the evolutionary objection to cognitive specialisation in speech cognition, by providing a consistent pressure for individuals to be good at recognizing the signatures of systematic organisation, and anticipating subtle exceptions. Moreover, the ability to learn system-level generalisations can accelerate learning of lower-level properties of speech sounds by providing top-down constraints that can be acquired early (Smith et al. 2002; Kemp et al. 2007). This view offers a computational account for the apparent flexibility of human phonetic abilities: *specialise in the abstract to learn the concrete*.

We explore this evolutionary hypothesis from a computational perspective, by studying the representational assumptions underpinning one of the core mechanisms thought to facilitate speech acquisition: generalisation among classes of speech sounds (Kleinschmidt and Jaeger 2015). Faced with a system of related behaviours, any kind of learner is confronted with a statistical dilemma that must be addressed: what is the appropriate degree of structural generalisation among classes of sounds? In other words: to what extent does learning about one class of sounds also inform learning about another? What is the balance of idiosyncrasy and homogeneity among an inventory of related speech sounds? Recent experimental (e.g. Maye et al. 2008) and modelling (Pajak et al. 2013) work has begun to shed light on human performance in problems with this kind of structure; the present analysis asks how an evolutionary perspective can contribute to this understanding.

Our hypothesis, which is in line with the general movement towards understanding language and speech acquisition as probabilistic inference (see e.g. Kleinschmidt and Jaeger 2015; Pajak et al. 2016, for overviews of recent work) but motivated by an evolutionary perspective, is that language learners *learn* the answers to these questions, and perhaps impose domain-specific expectations at this level. They learn the structure of the learning problem by jointly making inferences about abstract properties of a *system* of sounds, which in turn bootstrap learning about individual sounds. We illustrate these arguments using unsupervised probabilistic models of phonetic category induction, drawing on ideas from the literature around non-parametric Bayesian statistics. We analyse three models: each model represents a different set of assumptions about latent sources of common and idiosyncratic structure among a system of speech sounds. By exploring the performance of these models at learning miniature artificial phonetic systems, we show how and when these assumptions translate into benefits that evolution could capitalise on.

1.1 Generalisation among sound classes

Human speech systems are universally composed from inventories of re-usable building blocks (de Boer et al. 2012; Ladd 2014): related phonetic behaviours (phonemes) that must be learned from acoustic data, with little supervision, concurrently. Individual languages don't tend to employ a random subset of possible speech sounds, but rather are composed of an inventory that exhibits structural regularity along multiple dimensions (see e.g. Clements 2003; Dunbar and Dupoux 2016). This structure means that phonetic systems can be carved up into coherent classes of related sounds based on shared acoustic or articulatory features. To give a simple example, the voiceless stops /p/, /t/, and /k/ form a natural class of sounds in English, related by a common manner of articulation and voicing: within this class the three sound categories /p/, /t/, and /k/ —*phonemes* in traditional terminology—are distinguished by variation in place of articulation. This example illustrates how sound systems can be understood to possess group-like structure at multiple levels: phonemic categories of related acoustic tokens, and higher level classes of phonemes (such as stops). Throughout, we will cast this structure generally in terms of classes that contain distinct categories (phonemes) as described, and ask how this structure could be exploited by an ideal learner.

Here is a simple question that evades a simple answer: *does the fact that infants must learn multiple phonemes make it harder or easier to learn individual phonemes?* A key basic insight into the nature of our phonetic capacities is that during learning, existing phonetic representations guide the acquisition of new patterns of behaviour: the sounds we already know influence how we learn new sounds (Maye et al. 2008). It is common to cast this ability in statistical terms, which allows us to understand and model the transfer of learned information as *generalisation* of phonetic features to new groups of sounds (see e.g. Maye et al. 2008).

Generalisation is a well-studied human ability: understanding the conditions and computational principles that lead learners to go beyond their experience is a fundamental goal for cognitive science in general (Shepard 1987; Tenenbaum and Griffiths 2001), and is the foundation for entire theories of linguistic knowledge (see e.g. Goldberg 2009). Specifically in phonetics and phonology, generalisation over categories of speech sounds has been widely studied using exemplar theory

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

(see e.g. Kirchner et al. 2010; Soskuthy 2013, for recent perspectives) and connectionist models (see e.g. Hare et al. 1995), and is beginning to be studied from a statistical perspective (see e.g. de Boer and Kuhl 2003; Lin and Mielke 2008; Feldman et al. 2013; Kleinschmidt and Jaeger 2015). Experimental evidence suggests that both infants and adults are able and inclined to generalise phonetic categories or distinctions to new learning problems (see e.g. Maye et al. 2008; Finley and Badecker 2009; Perfors et al. 2010; Pajak and Levy 2011), indicating that abstraction and generalisation are important components of human phonetic abilities. But why?

1.2 The utility of abstraction and generalisation

The utility of abstraction and generalisation are well-understood theoretically: for example, a recent analysis by Kemp et al. (2007) showed how the ability to learn what they call *overhypotheses*—higher-order generalisations concerning abstract features of a learning problem—can lead to a number of benefits in learning. For example, drawing generalisations allows learners to make remarkably accurate predictions from extremely limited data, sometimes referred to as *one-shot learning*. Kemp et al., citing earlier sources (Goodman 1983), provide a simple example that is worth paraphrasing. Imagine observing an experimenter empty out three bags of marbles, in turn: two of the three contain all black marbles, and the third contains all white marbles. Afterwards, observing a single black marble drawn from a new bag leads people to predict that future draws from the new bag will also be black: a strong conclusion that is not licensed by the singleton observation alone, but dependent upon having inferred the abstract generalisation that colour is uniform within bags, and extending this to the new bag.

Somewhat counter intuitively, higher-order generalisations like this can sometimes even be justified *before* lower-level details have been learned well (Kemp et al. 2007), since the evidence base for higher order properties can be larger than the evidence base relevant to specifics (*all bags of marbles contribute to learning the abstract generalisation, but only one bag informs learning specific colors*). Crucially, Kemp et al. also show, in line with comparable research in other domains (e.g. Tenenbaum et al. 2011), that regularities at this higher level of abstraction can in theory be learned via the same principles of statistical inference from distributional evidence that are thought to underpin learning about lower level phonetic variability (though see e.g. Calamaro and Jarosz 2015, for an alternative perspective on whether distributional statistics can account for generalisation in phonology).

These principles have also been shown to apply to generalisation in phonetic computation (e.g. Pajak and Levy 2011, 2014; Feldman et al. 2013). To give a concrete example, Pajak and Levy (2011) show that adult participants instinctively generalise a two-category segmental length distinction from sonorant to obstruent classes of sounds in an artificial language, indicating that adult learners have a system-level expectation that phonetic structures will be re-used in new classes. This finding mirrors the well-established idea that learners can generalise to new categories in language acquisition in general (See e.g. Arnon and Clark (2011) for a range of contemporary of perspectives) and second language acquisition in particular (see e.g. Pajak et al. (2016) for a review). Pajak et al. (2013), building on work by Feldman et al. (2013), also provide a statistical model for this kind of learning, based on the idea that learners believe phonetic categories can be shared among sound classes: the model even accounts for absolute feature differences (e.g. absolute length in long versus short categories) between classes, showing how generalisation through inference of abstract systematic properties of a language can be divorced from the implementation of those features in specific classes of sounds.

1.3 Generalisation by assumption?

The simple example above shows how an ability to learn at higher levels of abstraction, and to generalise across related inference problems, can lead to considerable improvements in learning, speeding it up or extending it to new domains, particularly in cases of noisy, lossy, or sparsely observed data (see e.g. Kemp et al. (2007), for more detail). However, these benefits depend upon a crucial assumption: the assumption that generalisation is actually *licensed*; that there are system-level regularities to exploit. In simple terms, extending already-encountered structures to new domains is only a good idea when that new domain does indeed possess similar structure. Human sound systems are well known for their re-use of structural features—their *feature economy* (Clements 2003)—suggesting that generalisation is in general a good strategy.

However, sound systems often include exceptions to generalisations (sometimes called *gaps*), and the particular features that are re-used change from language to language. Attaining a quantitative estimate for the degree of structural re-use in real sound systems is challenging, but a recent analysis of many languages (Dunbar and Dupoux 2016) suggests that sound systems

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

are rarely more than half as systematic as they theoretically could be, and most of the time are even less systematic than this (while still being significantly systematic). These estimates are made with respect to a particular representation of sound systems and specific measures of structural reuse, but are broadly in accordance with previous analyses (see e.g. Mackie and Mielke 2011), and with the general idea that sound systems appear to be meaningfully but not *maximally* systematically organised.

More anecdotally, it is easy to find examples of this kind of structure in phonological databases of the world's languages (see e.g. Maddieson and Disner 1984; Moran et al. 2014). To give a simple example, many varieties of Dutch employ a systematic two-category voicing distinction in bilabial (/p/ and /b/) and dental (/t/ and /d/) plosives. Extending this generalisation to velar plosives (i.e., assuming /k/ and /ɡ/) however would be a mistake: /k/ is in the inventory, but /ɡ/ only occurs in loans. Similarly, Yoruba (UPSID No. 4137) has a two-category voicing distinction that contrasts /d/ with /t/, /ɡ/ with /k/, and /ɡb/ with /kp/, but does not extend to bilabial plosives (/b/ exists, but /p/ does not phonemically). A learner inclined towards blanket generalisation would make different mistakes in each of these languages; a learner who did not generalise at all would miss out on a substantial regularity. How can a learner know when to generalise?

More generally, we should be cautious of a circularity in reasoning about the value of generalisation: generalisation is thought to be a shortcut to knowledge in new or under-observed domains, yet by definition new domains are those in which the learner should have least confidence that the generalisation actually holds. The problem is clear when expressed in statistical terms: when trying to make concurrent inferences in a set of related inductive problems, any kind of learner must start from some assumption about the degree of shared structure, or statistical dependence that exists between those problems. The degree of generalisation, and its utility for the learner, will always depend on this assumption and its validity.

Languages in general, and sound systems in particular (Sóskuthy 2015; Vaux and Samuels 2015), adapt culturally to the biases of individuals who learn and use them (see e.g. Griffiths and Kalish 2007; Christiansen and Chater 2008; Kirby et al. 2014). This dynamic suggests that, in the long term, cultural evolution will lead towards a good fit between learners' expectations about generalisation and the true structure of the language. However, sound systems may also be shaped by other factors, such as speech production mechanics (Martinet 1939) and communicative functions (Liljencrants and Lindblom 1972; Vaux and Samuels 2015), meaning that fixed prior expectations about generalisation may not always correspond to the underlying structure of a language, potentially leading to a mismatch that must be solved either through learning or through biological evolution of appropriate domain-specific expectations at this level. We will revisit this question of directionality later in the article (in Section 4).

2. Methods: probabilistic inference of speech sound systems

A range of problems in the acquisition of speech sounds can be understood as examples of *category induction* (see e.g. de Boer and Kuhl 2003; McMurray et al. 2009; Feldman et al. 2013; Pajak et al. 2013). Speech sounds can be divided into classes based on their properties (e.g. by place of articulation, or vowels versus consonants, etc). Each class of sounds has an associated set of tendencies concerning individual speech features (e.g. voicing, duration, spectral features), and these tendencies can be represented as categories within a class (e.g. voiced versus voiceless inter-dental fricatives). Casting speech acquisition in these terms allows us to study the computational structure of the problem faced by learners, and in turn by evolution, which shapes learners. Our approach is to lay out an ideal observer model for the basic inferential task, then to study the behaviour of the model in scenarios that expose the model's strengths and weaknesses.

2.1 Problem specification and notation

The learner is learning about n classes of speech sounds, $\mathcal{C} = \{C_1, C_2, \ldots, C_j, \ldots, C_n\}$, where C_j denotes the jth out of n classes ($j = 1, \ldots, n$). Each class can contain multiple phonemes, which we model as categories. For reasons of space and simplicity, we will focus on learning the behaviour of a single speech feature f which distinguishes phonemes throughout these classes of sounds, but note that our analysis generalises naturally to the case of multiple features, and our arguments would be strengthened if those features are also statistically non-independent. Each class of sounds C_j is associated with a distribution F_j over the variation in the feature f. As a concrete example, we could say the partition into classes $\mathcal{C}$ reflects place of articulation among stop consonants, so that for example class C_1 is understood as bilabial closures, C_2 alveolars, and C_3 velars. If the feature f is understood as voice-onset-time (VOT), for instance, then F_1 would be the distribution over VOT among

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

bilabial stops. In a language like English, F_1 would be a bimodal distribution encoding two phonemic categories which correspond roughly to voiced and voiceless stops /b/ and /p/, as would F_2 (/t/ and /d/) and F_3 (/k/ and /g/). In a language that has a three way voicing distinction in bilabial stops, F_1 would be a trimodal distribution[1].

2.2 An ideal observer model

The learning problem is to make informed inferences about the kind of distribution F_j that underpins the use of feature f in sound class C_j: when cast as probabilistic inference by an ideal observer, the solution is to compute a posterior distribution $p(F_j|Y_j)$, where Y_j is a body of observed evidence (observed phonetic tokens) relevant to the task. For any hypothesised distribution F_j, an ideal learner's degree of belief in that hypothesis is proportional to the likelihood of the observed evidence under the hypothesis, $p(Y_j|F_j)$, and any *a priori* beliefs the learner already holds, $\pi(F_j)$, independent of the data:

$$p(F_j|Y_j) \propto p(Y_j|F_j)\,\pi(F_j)\,. \tag{1}$$

Natural distributions of speech features within a language are often complex (e.g. bimodal), meaning that their structure cannot be well captured by standard parametric models (for instance, a Gaussian distribution cannot have two peaks). A common solution is to capture this structure using large or infinite *mixtures* of parametric distributions (de Boer and Kuhl 2003; Feldman et al. 2013; Doyle et al. 2014; Eaves et al. 2016), which can engender arbitrary expressive power for complex distributions, and can be analysed efficiently using stochastic sampling or approximation methods.

The structure of each F_j is therefore captured by a mixture (such as a bimodal category distinction) of univariate Gaussian distributions (see e.g. Feldman et al. 2013): the learner's task is to infer this mixture based on the acoustic data via Bayesian inference. Estimating an unknown mean and variance for a single Gaussian via Bayesian inference is a standard statistical procedure (Gelman et al. 2003, chapter two). Estimating a *mixture* of Gaussians involves hypothesising a configuration of multiple Gaussians in the feature space, each with its own estimated mean and variance, and computing the likelihood of the acoustic data under this hypothesis. Bayesian inference for this mixture involves specifying a prior probability distribution over this space of possible mixtures, and computing the posterior distribution given the acoustic data. The space of possible infinite mixtures is naturally infinite itself. Nevertheless, standard techniques specify an appropriate prior over this space of functions, and provide sampling routines for computing the posterior under this prior (Ferguson 1973).

The specific algorithm we adopt cycles through the datapoints in a sound class, assigning each to a Gaussian category: each assignment step of the algorithm decides whether to assign the datapoint to an existing category (to which other datapoints have already been assigned) or to hypothesise a new category. The algorithm repeats this sweep through all datapoints over and over. Samples collected using this technique (known as *Gibbs sampling*) converge to the true posterior distribution as their number grows. Full technical details can be found in Appendix A. The reader is free to treat this learning model as a (rational) black box that implements Bayesian inference for arbitrary F_j given Y_j in accordance with equation (1).

2.3 Miniature artificial languages

We define a small number of simplified miniature artificial languages (partial phonetic systems), hand designed to vary along two dimensions of interest: the specific structures in the language, and the manner in which those structures are re-used throughout the system. Each artificial language has $n = 3$ sounds classes $\mathcal{C} = \{C_1, C_2, C_3\}$ whose internal structure must be learned from data $Y = \{Y_1, Y_2, Y_3\}$. These data represent noisy realisations of the true underlying structures: for example, the dataset Y_3 contains values for f for all acoustic tokens known to belong to class C_3. Within a class, variation in the feature f can distinguish multiple phonemes: the learner's task is to infer the phonemic categories distinguished by feature f in a particular sound class. All three languages are shown in Fig. 1. The languages are designed to have the minimal structure necessary to illustrate our arguments concerning the principles of re-use and exception, rather than to be representative of specific natural languages.

2.3.1 L1: Systematic unimodal structure

Language L1 (Fig. 1, top row) encodes a single category that is reused throughout the system: it exhibits the systematic structural generalisation that the distribution of f in all three classes is unimodal Gaussian, with matched means and variances. With respect to the voice-onset time example above, the phonemes in this language

[1] Note that this example is not meant as a contentful claim about the hierarchy place >> VOT: the model could equally be defined over classes determined by voicing with place of articulation as the varying feature, or some other division of interest.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

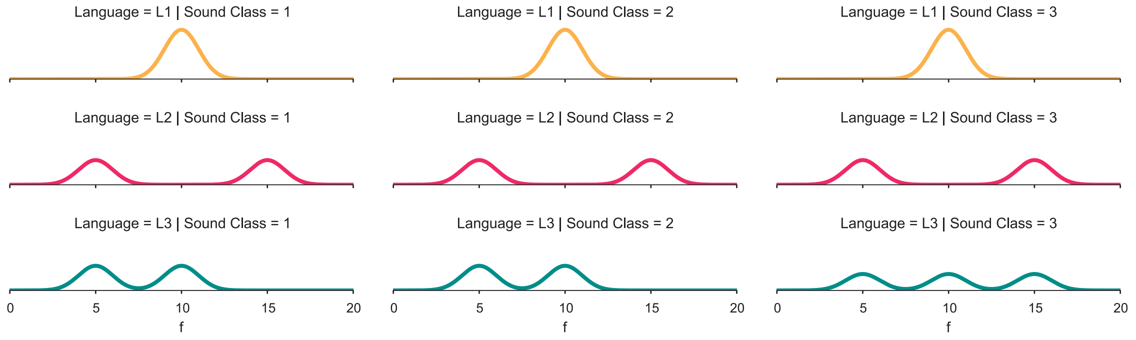


Figure 1. Miniature artificial phonetic systems. Languages along the rows, sound classes along the columns. In Language L1 (top), a unimodal structure is shared uniformly across the sound classes. Language L2 (middle) also shares structure across sound classes uniformly, but this structure is bimodal. Language *L*3 violates the assumption of uniformly re-used structure: classes C_1 and C_2 share a bimodal common structure, but C_3 has a trimodal structure. Lines show the probability density function over *f*: the y-axis is left blank to highlight that only the shape of the distribution is important; scales are arbitrary but fixed across classes and languages.

could be understood to represent /g/, /d/, and /b/, for example: each sound class has only one phonemic category in this language. Here and throughout, absolute scales for *f* are arbitrary but fixed and shared across classes.

2.3.2 L2: Systematic bimodal structure

Language L2 (Fig. 1, middle row) also represents the assumption that structure is re-used across classes, but the shared structure is of a different form: all sound classes are characterised by the same *bimodal* distribution of sounds. This case illustrates that the degree of structural re-use among classes can itself be understood as an important property of a language (Clements 2003), orthogonal to the specific structures re-used (or not). Each class has two phonemic categories, and these categories sit at analogous locations in feature space across classes. Such a system could be understood to represent a simple two-way voicing distinction, for example: class C_1=/k/, /g/; class C_2=/t/, /d/; and class C_3=/p/, /b/. In this respect, this structure could represent the voicing distinction in English plosives.

2.3.3 L3: Bimodal system, trimodal exception

Language *L*3 (Fig. 1, bottom row) violates the systematic structural generalisation that structures are faithfully re-used across all sound classes. Here, classes C_1 and C_2 share a bimodal structure similar to language L2, but class C_3 represents an exception. This example captures the balance of re-use and exception often found in real sound systems (Dunbar and Dupoux 2016), which rarely exhibit total structural isomorphism and often contain exceptions to generalisations. The trimodal structure in class C_3 extends the bimodal structure common to the other two classes, by allowing an additional phonemic category higher in the feature range[2]. This language could represent a system like that found in Aizi (UPSID No. 4131), which would correspond to: C_1=/k/, /g/; C_2=/t/, /d/; C_3=/p/, /b/, /ɓ/. We stress again that these are simply example correspondences with real languages: the crucial aspect is the structural principle of re-use and exception embodied by this miniature system. *L*3 introduces both a new structural form (trimodal) and a new degree of re-use: something between systematic and idiosyncratic class structure. The goal of the next section is to ultimately provide a statistical model of this *something inbetween*.

3. Analysis and results: probabilistic inference for inventories of behaviours with non-parametric structure

The general learning model (equation 1) defined in subsection 2.2 provides a normative solution to the inference problem underpinning acquisition of an

[2] An alternative way to construct this language would be to assume the bimodal structure encodes phonemes that are maximally separated, sitting at the extremes of the feature range, with the exceptional phoneme in C_3 falling in between these two extremes. Dispersion of phonemic categories to take advantage of the available space in this way might naturally be expected from considerations of expressiveness or self-organisation (de Boer 2000). Our model does not deal with these aspects of sound systems, and the feature range in these miniature systems is largely arbitrary, so the *bunched up* bimodal structure is adequate for our purposes. We thank an anonymous reviewer for raising this point.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

individual class of sounds C_j., but does not specify a model of the structural relationships *among* these individual learning problems. Here we consider three statistical models for this structure, each of which represents a different overarching assumption the learner could make about the degree of shared structure in the phonetic system as a whole, $\mathcal{C}$.

It is our aim to show that certain assumptions can lead to improvements in learning that are independent of the specific structures being learned, a key desideratum for cognitive specialisation in our phonetic capacities on an evolutionary timescale. We begin with extreme cases, in which sound classes are assumed to be independently structured (M1), or underpinned by a single common structure (M2), before considering a probabilistic generalisation which allows a mixture of shared and idiosyncratic structures (M3). Readers who wish to pursue the statistical foundations of these models are directed to Muller et al. (2004), on whose models our analysis is based.

Each model will be trained twice on each phonetic system: once under a dataset that is sufficient to learn each component class well (the *full data* condition: $N_j = 60$, where N_j gives the number of datapoints observed in class C_j), and once under a dataset half this size (the *half data* condition: $N_j = 30$). Datapoints are randomly generated from the underlying Gaussians. The *half data* condition is a simple proxy for adverse conditions during learning: in an evolutionary context, benefits can be gained through an ability to acquire a phonetic systems faster (from fewer data) or more reliably (from noisier data). Unless stated otherwise, we assume a relatively uninformative prior over the specific kinds of structures the learner expects to encounter in a language: a weak expectation that individual categories within a class C_j won't be too diffuse, with neutral expectations regarding the number of categories in a class, and their locations in the feature space (see Appendix A for further details).

3.1 Evaluation criteria

Each of the models was trained separately on each of the artificial languages. In an evolutionary context, the relevant quantity is future behaviour: after being exposed to a sound system and learning its structure, the learner goes on to produce those sounds in the future, and will be rewarded in proportion to how consistent its own usage is with usage in the community at large. An appropriate measure of future behaviour is the *posterior predictive* distribution for data under the model, $p(Y^{new}|Y)$, which describes how we expect the learner to behave when producing data (Y^{new}) after learning from data (Y), taking into account the possible outcomes of learning (see Appendix A for details). We report the Kullback–Leibler divergence between this distribution and the ground truth as an error statistic for model performance.[3]

3.2 M1: Independent individual structures

Perhaps the most basic assumption is to treat each individual class of sounds C_j as an isolated learning problem, independent of any other sound classes also under consideration. This assumption can be written:

$$p(y_j) = F_j(y_j), \text{ for } j = 1, \ldots, n, \tag{2}$$

where y_j is an acoustic token, known to belong to class C_j, exhibiting a value for f. This equation states that the data distribution $p(y_j)$ for feature f in any class C_j follows the distribution F_j. No other source of structure is relevant to the behaviour of this class of sounds. Sound classes in this model are linked only through the hyperparameters of a shared prior, $\pi(F) = \pi(F_j)$, for all classes $j = 1, \ldots, n$. And this is only through stipulation: the model assumes one shared prior underpins individual learning in all classes, but there is no logical reason that this should always be true.

Figure 2 shows how well the model has learned in both learning conditions. The model is able to learn languages L_1 and L_2 reasonably well (see Fig. 3 for error rates), especially in the *full data* condition, but is worse at learning language L_3. In particular, the model does not learn the trimodal class C_3 in language L_3 well at all. This class is inherently more difficult to learn than the others, due to a more complex internal structure. Model performance declines across the board in the *half data* condition, to a degree that is proportional to the decrease in data. This decrease in learning accuracy is most pronounced in language L_3.

[3] An alternative approach to measuring model performance would be to ask whether the model has learned the correct number of categories. We chose to measure the deviance of the posterior predictive distribution instead for two reasons, conceptual and practical. Conceptually, given our evolutionary focus, it is natural to directly measure behaviour rather than the internal representation underpinning behaviour. Practically, measuring how many categories the model has learned is potentially complicated in model M3, where shared categories can contribute to classes probabilistically. Our measure avoids this complication and is uniform across models and languages. We thank an anonymous reviewer for raising this point.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

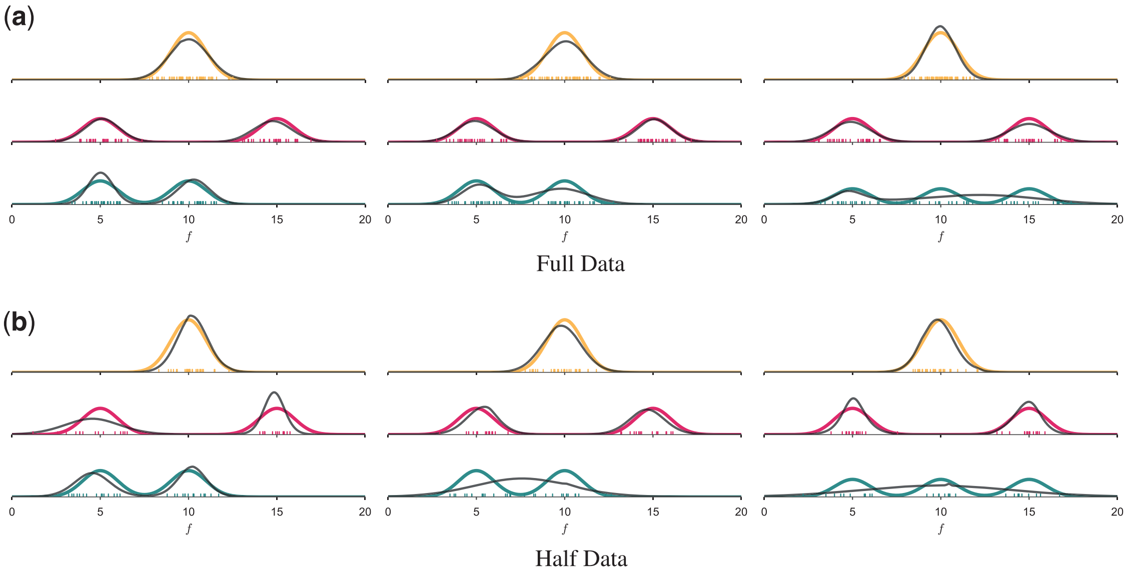


Figure 2. Results for model M1 after being trained on the full dataset (a) and a dataset of half the size (b), for languages L_1 (top row), L_2 (middle row), and L_3 (bottom row). The coloured lines give the ground truth artificial phonetic system to be learned; black lines show the model's predictions after being trained on the data. The carpet plot (upright ticks on the x-axis) shows the data used to train the model (sampled randomly from the true underlying Gaussians). These results represent one run of the model, but are representative of typical model inferences: other runs result in different errors, but the errors are of similar number and kind. See figure 3 for error statistics computed over many runs of the model.

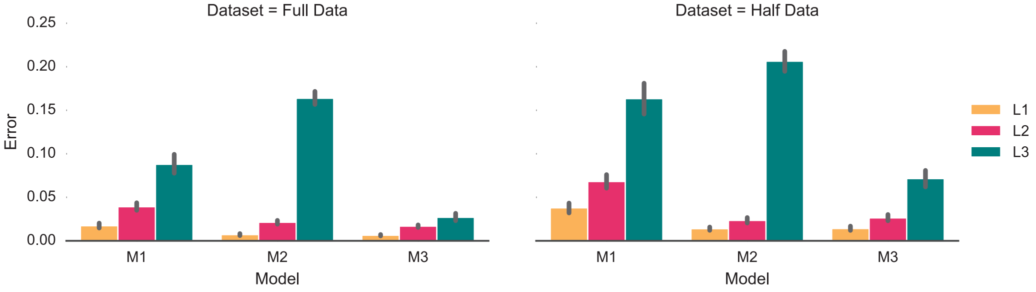


Figure 3. Error statistics (Mean KL divergence between the ground truth of the artificial phonetic system and the posterior predictions of the model after training) by language and data condition (left: *full data*; right: *half data*). Lower errors (smaller bars) correspond to better model performance. Bars represent the mean error averaged over 50 simulation of the model, based on 100 MCMC samples from the posterior for each simulation (after burning 1500 initial samples and thinning by 2).

Even in favourable cases where structure is repeated across classes (like L_1 and L_2), the model has no way to exploit this inventory-wide regularity during learning. All else being equal, learning in this model can only be improved through refinements to the prior $\pi(F)$. In traditional cultural evolution literature, this kind of adaptation would be called a *content bias* (Boyd and Richerson 1985), since it reflects preferences for specific structural forms.

If the prior is assumed to be shared across learning problems, then consequent improvements in learning are dependent on the idea that structural forms are shared across sound classes, since a prior preference in one direction would *hinder* learning in any exceptional classes whose structure is at odds with this bias. For example, a prior favouring a bimodal structure improves learning about language $L2$, but less so in language $L3$, where this prior is at odds with the trimodal class C_3. If we assume prior beliefs are *not* shared across learning problems, but are independent facets of the learning system, then this problem is avoided, but at a cost: in evolutionary terms, the cost associated with building these

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

improvements increases multiplicatively with the number of learning problems, since each prior must be built individually. This solution to the computational problem is richly structured in advance: given the plasticity of human learning, we should strive to avoid this hypothesis where possible.

Analogously, but more generally, refinements to the prior are subject to the problem of diversity *across languages* discussed earlier: prior preferences that improve learning in one language may be unsuited to the structures in another language. In the face of rapid cultural evolution, so the argument goes (Christiansen and Chater 2008), evolved biases for specific structures quickly become out-dated. Only weak biases are thought to be evolvable at this level (Thompson et al. 2016).

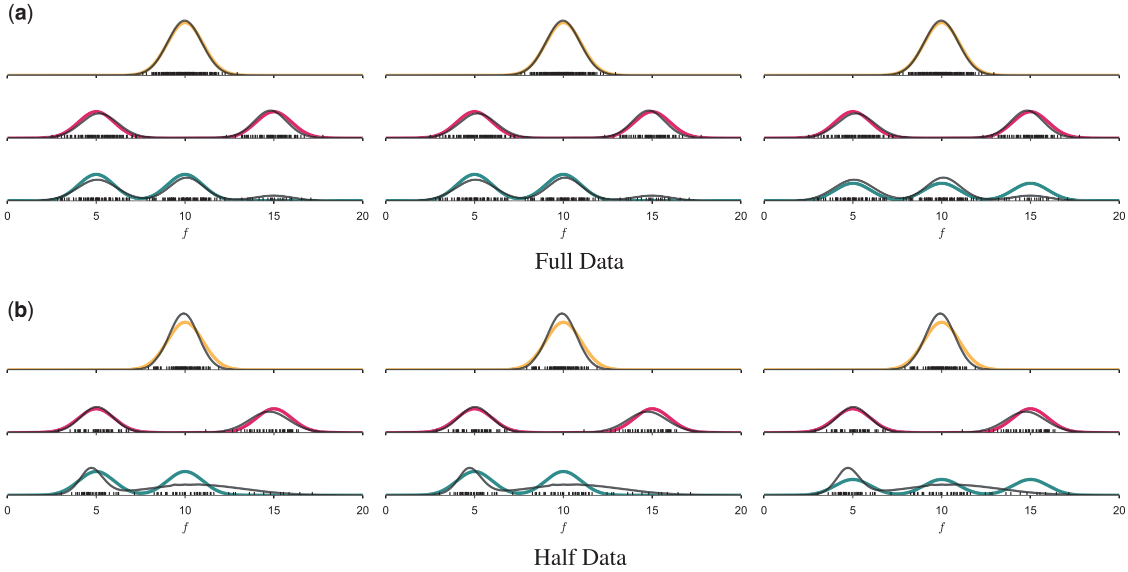
3.2.1 M2: Group structure

In cases where structure is repeated across classes, the model above (M1) is unable to exploit this regularity. Improvements to learning can be achieved by the assumption that the learning problems are *not* independent, but instead form a group bearing on a common structural question. Treating sounds classes as a group of closely related phenomena allows sharing of statistical power during learning: this is the key assumption that underpins the inferential advantages outlined by Kemp et al. (2007), discussed in the introduction. Observations in this model are *exchangeable* across sound classes. This means that anything learned about a class of sounds C_j transfers directly to all other classes $C_{i \neq j}$. This model can be written:

$$p(y_j) = F_*(y) \quad \text{for} \;\; i, j = 1, \ldots, n \qquad (3)$$

The core assumption is that a single common distribution, F_*, underpins the behaviour of f in all classes of sounds (hence why the subscript j is dropped in the final term). Figure 4 shows the results of learning under this assumption. Note that the model's predictions are identical across classes within a language. Unlike M1, model M2 treats all the data as representative of a single underlying shared distribution, so is able to learn that distribution accurately even in relatively adverse learning conditions: compared to model M1, model M2 has three times the data that bear on learning F_*, since $N_* = N_1 + N_2 + N_3$. For example, even in the *half data* condition, model M2 learns languages L_1 and L_2 roughly as well (Fig 4(b), top two rows) as model M1 learns these languages in the *full data* condition (Fig 2(a), top two rows).

Following the same logic, the generalisation that results from assuming shared structure could also allow this model (M2) to make accurate predictions about a completely unseen class. Crucially, if the model received no data from the third class in languages L_1 and L_2, it could still make fairly accurate predictions, because the



Full Data

Half Data

Figure 4. Results for model M2 after being trained on the full dataset (a) and a dataset of half the size (b), for languages L_1 (top row), L_2 (middle row), and L_3 (bottom row). The coloured lines give the ground truth artificial phonetic system to be learned; black lines show the model's predictions after being trained on the data. In this plot, the carpet ticks are coloured black and repeated across all classes within a language to indicate that all datapoints contributed to inference for all classes (within one language).

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

structure in the unseen class C_3 is shared with C_1 and C_2, which have been observed. This model turns small datasets into relatively large datasets: it is robust to adverse learning conditions, allowing rapid generalisation and *one-shot* learning through transfer of structure from one class to another. Unlike model M1, the accuracy of learning about a class C_j is not always at the mercy of the quality of the data available for that class: instead, system-wide regularities can bootstrap learning.

However, these benefits are only accrued when the assumption underpinning this model—that all classes share a common structure—is in fact true of the language to be learned. Language L_3 demonstrates how this assumption can have negative consequences when it conflicts with the true phonetic system. Even in the *full data* condition, this model cannot learn an accurate representation of language L_3: the model has no way to capture the trimodal structure of the exceptional class C_3 without also imposing that structure on classes C_1 and C_2 (as it has in the *full data* condition—see Fig. 4(a), bottom row). The assumption of group-structure among classes can hinder learning about exceptional classes *and* about the more homogeneous classes, since the model is always forced to compromise. This is the key weakness of the model: its assumption of uniform structure throughout sound classes is powerful when correct, but is not always correct, and cannot be overturned.

Both of these models (M1 and M2), and the structural assumptions they represent, posses features that are beneficial in the right conditions. Ideally, we would be able to define a model that can combine these advantages by allowing a broader space of structural blueprints that interpolate between the assumptions of idiosyncratic and group structure among sound classes. In the next section, we present a model that allows this broader space of structural assumptions, and show how these structures can be learned from data.

3.2.2 M3: Partial structure sharing in learned degree

Real languages tend to sit somewhere between the extremes of completely systematic shared structure and completely idiosyncratic structure. In these cases, a learner who makes either of the assumptions captured in the models above would face difficulties: model M1 allows too little sharing of structure between classes (none), while model M2 assumes too much (uniform structural regularity). *Partial* pooling of structure is a classic problem in statistics, and we can lean on results from this literature to formulate a model of phonetic computation. We require a model that allows a group of acoustic behaviours to be represented by a balance of shared and idiosyncratic structures, and a learning mechanism that is capable of identifying these sources of structure in data, and their relative importance. In simple terms, to capture our basic proposal, we must be able to specify a model of how the learner can *learn when to generalise, and to what degree*.

Capturing this balance in a *phonetic* model is complicated by the non-parametric nature of the distributions that frequently arise in speech (e.g. multi-modal distributions). Representing dependencies among non-parametric distributions has received attention recently in Bayesian statistics, and powerful contemporary models such as the Hierarchical Dirichlet Process (Teh et al. 2006) have already begun to be applied to sound system acquisition data (see e.g. Pajak et al. 2013), and other categorisation problems humans solve (see e.g. Griffiths et al. 2007; Canini et al. 2010). We choose the Dependent Dirichlet Process model put forward by Muller et al. (2004) as an appropriate structure, since this model allows an explicit parametrization of the balance between shared and idiosyncratic structures among a group.

Under this model, the structure of any individual class of sounds is represented by a weighted mixture between a core structural component common to all classes, F_*, and class-specific idiosyncratic structure F_j. In this way the learner can identify commonalities between classes, and learn about this common component from all contributing classes as in model M2, buying the inferential advantages discussed above, without the restriction that any class-specific idiosyncratic features must also be captured in this component. Formally, this model can be written:

$$p(y_j) = \epsilon_j F_*(y) + (1-\epsilon_j)F_j(y_j), \text{ for } j = 1, \ldots, n \quad (4)$$

where $0 \leq \epsilon_j \leq 1$ specifies the relative contributions of the shared and idiosyncratic structural components to the behaviour of feature f in sound class C_j. We will refer to the balancing parameter ϵ_j as the *borrowing strength* of class C_j. Collecting these into a single parameter vector $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, we can treat $\vec{\epsilon}$ as fully determining a system-level measure of behavioural homogeneity among sound classes: we will call this quality of a language the *common degree*.

Models M1 and M2 are in fact special cases of this model. Model M1 results from setting $\epsilon_j = 0$. Model M2 is achieved by setting $\epsilon_j = 1$. We could imagine an intermediate model, M2.5, which allows this broader range of structures, but assumes that the balance of regularity and idiosyncrasy is a fixed assumption made by the learner: this could be a fixed system-wide property

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

(which respects the constraint $\epsilon_j = \epsilon_*$), or specific to each class such that each ϵ_j has its own fixed value. We focus on the superclass of these models, in which the *common degree* $\vec{\epsilon}$ is learned. Figure 5 shows the statistical structure of these models, set in a model-schema that highlights how all these cases can be seen as specific instances of a general family of models defined by equation (4).

Appendix B provides details of a model for how $\vec{\epsilon}$ can be learned from the same distributional statistics that allow F to be inferred. Intuitively, given some knowledge of F_j and F_*, any hypothesised borrowing strength ϵ_j implies a likelihood distribution for the observed data. By specifying a prior $\pi(\vec{\epsilon})$, we can characterise the posterior $p(\vec{\epsilon} \mid Y)$ and make inferences about the underlying balance. The learning task becomes a problem of joint inference for $\vec{\epsilon}$ and $\mathcal{F} = (F_*, F_1, \ldots, F_n)$, where inferences about one variable can bootstrap learning about the other, allowing both to be learned simultaneously. Formally, we assume a shared, uniform prior distribution $\pi(\epsilon_j) = \text{Beta}(\epsilon_j; 1) \sim 1$ for the borrowing strengths, and Binomial likelihood for a binary indicator that specifies whether each datapoint was generated by the common or the class-specific component. An estimate for ϵ_j is straightforward to obtain in each step of the posterior sampling algorithm by resampling the indicator variables based on the current estimate of ϵ_j, updating the posterior for ϵ_j and sampling a value (see Muller et al. (2004), for details—our formulation differs from theirs only in the assumption that borrowing weights can vary between classes).

Figure 6 shows example results in this model. These results make most sense in the context of what the model has learned about $\vec{\epsilon}$. Samples from the posterior for ϵ_j, by language and sound class, are shown in Fig. 7. Focus first on the results in the Full Data condition, shown in Fig. 6(a). The model should be able to learn that structure is re-used throughout sound classes in languages L_1 and L_2. Figure 7 shows that the posterior samples for ϵ_j in these languages are indeed concentrated near $\epsilon_j \approx 1$. Model M3 has learned to behave like model M2: as such, it should have a low error rate on these two languages in both the *full data* and *half data* conditions, since in both cases the model has many datapoints that bear on the shared component F_*: the model has learned that a single common structure dominates the phonetic system. Figure 3 confirms this predicted pattern of errors.

However, the ability to learn $\vec{\epsilon}$ also allows the model to behave more like M1 where the data dictate. Consider language L_3: as a human observer, it is easy to make approximate predictions about the balance of shared and idiosyncratic structures in this language. A common bimodal structure in the lower range for f is shared throughout the system. However, this shared structure describes the full behaviour of classes C_1 and C_2, but not of class C_3, which includes an additional category higher in the range for f that is not present elsewhere in the sound class inventory. Since the three categories within the trimodal class C_3 are equiprobable, a sensible guess would be to suppose the common component accounts for roughly two-thirds of structure in this sound class, while one-third is special to this class. This is exactly what the model learns: Fig. 7 shows that

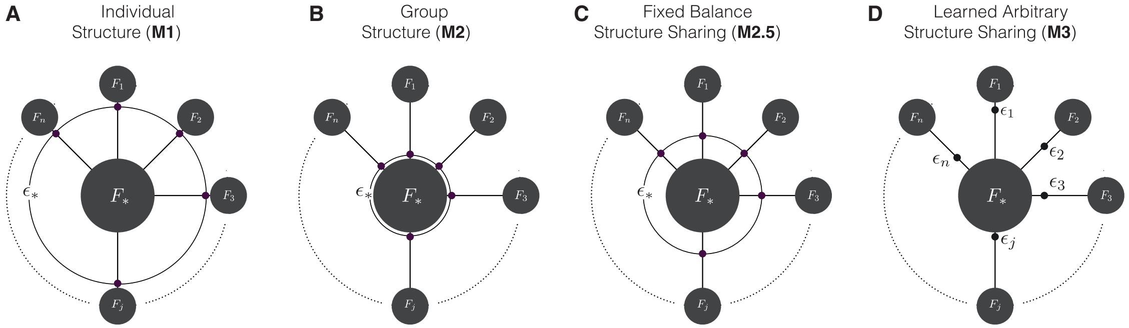


Figure 5. A schema for the family of possible models defined by equation (4), in four examples. Outer shaded circles represent F_j, the idiosyncratic structural component of a sound class C_j. Dotted lines between these circles just denote continuation, since the number of classes in the model is unbounded. The central shaded circle represents F_*, the shared structural component common to all classes. For each class, a radius connects the idiosyncratic component and the shared component: the position of the shaded point on this radius indicates the balance of the two structural components for that class (ϵ_j). The closer the point to F_0, the heavier is F_0 in the mixture, and vice versa. For example, panel A shows model M1, where all classes are statistically independent structures (radius point at F_j indicates $\epsilon_j = 0$). Panel B shows the opposite extreme (model M2), where all sound classes share a common structural form. In panels A–C, the radius points are connected to highlight the assumption that $\epsilon_i = \epsilon_j = \epsilon_*$ for all $i, j = 1, \ldots, n$. Panel D shows model M3, which relaxes this assumption.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

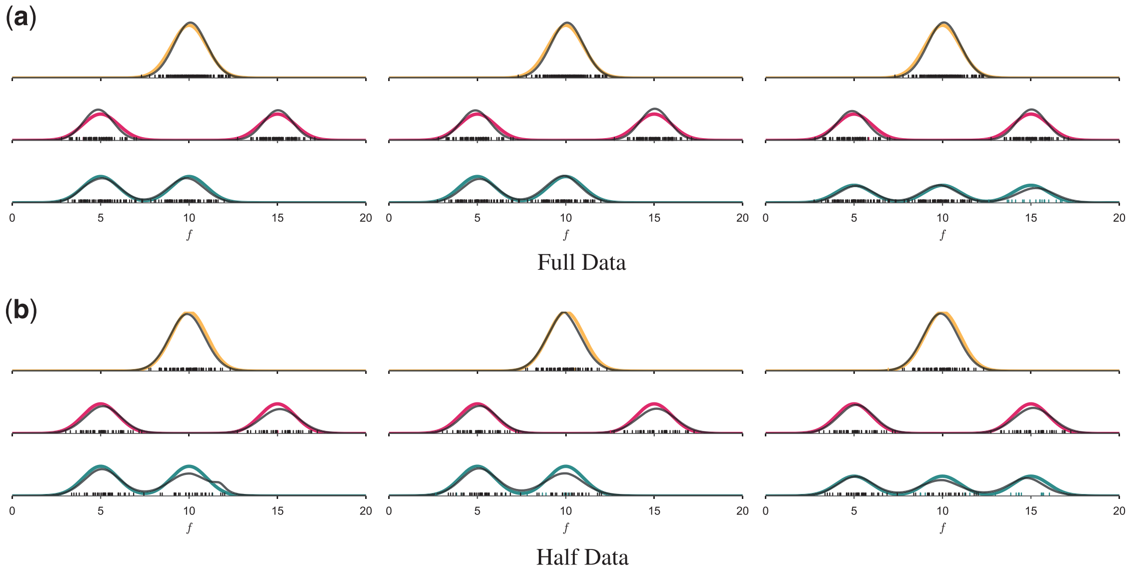


Figure 6. Results for model M3 after being trained on the full dataset (a) and a dataset of half the size (b), for languages L_1 (top row), L_2 (middle row), and L_3 (bottom row). The coloured lines give the ground truth artificial phonetic system to be learned; black lines show the model's predictions after being trained on the dataIn this plot, carpet ticks are plotted for a given class is the data-point represented has contributed to inference for that class. Black ticks have been assigned to the common component, coloured ticks have been assigned to the class-specific component.

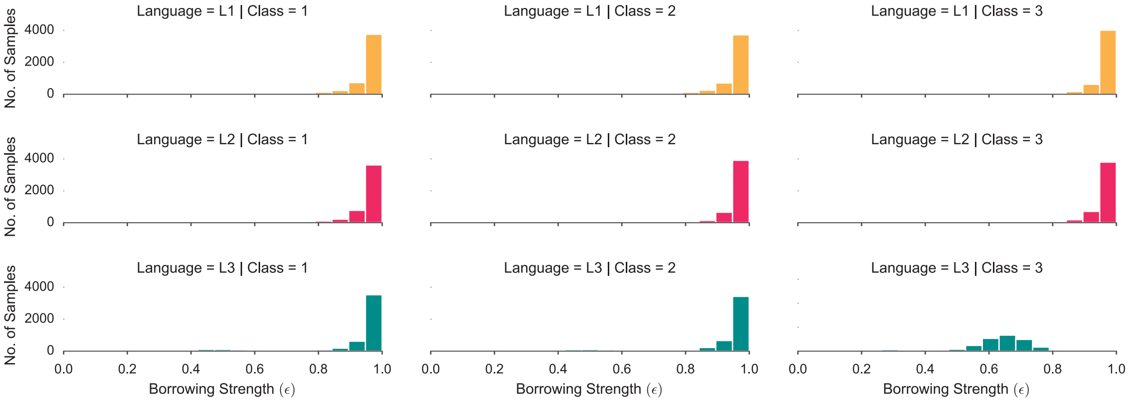


Figure 7. Samples from the posterior distribution for the borrowing strength, ϵ_j, under model M3, by language and class. Samples are combined from fifty simulations of the model: for each of the fifty simulations, we collected 100 Markov chain Monte Carlo samples from the posterior using Gibbs sampling. Samples were collected after burning 1500 earlier samples to counter the influence of initial conditions, and thinned by a factor of 2 to counter sample autocorrelation.

posterior samples for the borrowing strength in classes C_1 and C_2 of L_3 concentrate near $\epsilon_1 \approx \epsilon_2 \approx 1.$; equivalent samples for class C_3 concentrate around $\epsilon_3 \approx .66$. Model M3 is the only model able to learn the structure of language L_3 accurately, as a result of the fact that it can infer the common degree: it identifies a bimodal common structure, and identifies an exception to the generalisation in C_3, allowing it to make accurate inferences in both learning conditions.

One final feature of this model worth noting is that inferences about the common degree are gathered under a prior distribution $\pi(\vec{\epsilon})$. We analysed the model under an independent unbiased prior over borrowing strength, which allows the data to determine the common degree the model infers, but this need not be the case. It is straightforward to model a broad class of biased expectations, for or against homogeneous sound inventories, simply by changing the parameters of the prior.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

Moreover, it would be simple to extend hierarchical structure to the model for ϵ_j too: in this way, inferences about the borrowing strength in one class could inform expectations for the borrowing strength in another, allowing generalisation even at this level. We leave this extension to future research.

4. Discussion

We presented a computational analysis of a basic statistical dilemma that any learner of sound systems must in theory address: how independent are a group of related behaviours? We showed how the ability to learn this abstract quality of a language can bootstrap learning of lower-level details, potentially providing improvements to learning. Our analyses were motivated by the idea that cognitive specialisation at this more abstract level may not be subject to the moving target problem, and may therefore provide a plausible mechanistic explanation of the human proclivity for systematic sound inventories, rooted in abstraction and generalisation. Before discussing this idea further (subsection 4.2), we outline how our analyses relate to previous work.

4.1 Relation to previous work

Our methods are closely related to previous models of generalisation in speech acquisition. Much previous computational work has concerned algorithmic- or implementation-level analyses, whereas ours is a computational-level analysis. In general, implementation-level neural-network based models of generalisation (see e.g. Hare and Elman 1995) concentrate on whether or not specific representational architectures are able to learn and transmit generalisations. Analyses at the algorithmic level of analysis, such as exemplar-based models (see e.g. Soskuthy 2013), tend to concern how computationally frugal, data-driven inference algorithms can lead to phonetic generalisation. Computational-level analyses from a statistical perspective have been less prominent historically, but are becoming more common (de Boer and Kuhl 2003; Vallabha et al. 2007; McMurray et al. 2009; Toscano and McMurray 2010; Feldman et al. 2013; Pajak et al. 2013). This level of analysis allows us to study the nature of the computational problem being solved, and to isolate explicit components of the problem: in our case, the nature and consequences of assumptions a learner can make about latent sources of common structure among classes of sounds.

The present model is most closely related to the model of Pajak et al. (2013), which also provides a statistical formulation of generalisation among sound classes. Where we use the Dependent Dirichlet Process model to induce dependencies among nonparametric distributions, Pajak et al. use the Hierarchical Dirichlet Process. While we judged the Dependent Dirichlet Process model to be the more suitable structure for demonstrating our particular evolutionary arguments, the model developed by Pajak et al. is a more complete formulation of the acquisition problem, in at least two respects. Firstly, it can handle certain cases that ours cannot: for example, in our model (version M3) any sound class inheriting categories from the common distribution must inherit all such common categories, whereas the Hierarchical Dirichlet Process used by Pajak et al. allows sound classes to pick and choose which categories to share with other classes. This corresponds to a richer space of representational structures for capturing dependencies among sounds: this extra structure is useful for capturing human performance, but is not required for our theoretical analysis. Understanding exactly how rich these structural blueprints must be to capture human performance well is an exciting opportunity for future research.

Secondly, our model does not accommodate purely structural isomorphism among classes, such that the absolute acoustic values of a shared category cannot be realised differently in different classes. Pajak et al.'s model does allow this feature. Their model is designed to shed light on the mechanisms underpinning participants' generalisations in an experimental length-distinction task, and these features of the model improve its ability to achieve this. Sharing statistical power among concurrent non-parametric inference problems is an active area of research in Bayesian statistics (see e.g. Lin et al. 2010; Lin and Fisher 2012; Paisley et al. 2015), and we are enthusiastic that these advances can be brought to bear on human cognition. Given our evolutionary focus, we chose the Dependent Dirichlet Process model since it balances expressivity and simplicity: it allowed us to interpolate cleanly between the two extremes of group and individual structure, and explicitly parametrise this balance as a mixture model.

4.2 Evolutionary considerations

An influential idea in the language sciences (Christiansen and Chater 2008) is that biological evolution could not have shaped human cognition to anticipate features of language that vary rapidly as a result of cultural evolution, such as sound systems. This idea relies on the principal that hypothetical biological changes that result in improvements to language cognition are beneficial only for learning specific languages.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

Our analysis illustrates that there are cases where this principal may not hold: in particular, we showed that the ability to learn when to generalise (or, equivalently, to learn the degree of systematicity in a sound system) can provide improvements to learning that are not tied to specific languages, and may in fact increase the range of learnable languages. This analysis represents a concrete computational example of the often underspecified idea that language-related pressures could theoretically result in improvements to 'general' learning capacities. While the analysis strictly concerns individual learning, it raises a number of evolutionary questions which are discussed in more detail below.

4.2.1 What has selection shaped?

The present analysis is consistent with at least two interpretations of how selection could shape human learning. Firstly, our arguments could be understood to imply that the pressure to acquire language may have resulted in the ability to learn at this level of abstraction per se. A natural objection to this argument is that fine-grained command of generalisation among related problems does not result from language-related changes to human cognition, but is instead an antecedent capacity, possibly shared with other species. A large literature on animal cognition suggests that the capacity for generalisation and abstraction is present in a number of non-human animal species, from non-human primates (Vonk and MacDonald 2004; Ravignani and Sonnweber 2017) to birds (Smirnova et al. 2015; Spierings and Ten Cate 2016), and that there are interesting species-level differences in the inclination to generalise (Spierings and Ten Cate 2016). This may imply an ancient foundation for these abilities.

Secondly, our analysis is consistent with the idea that language-related pressures may have shaped specific expectations about the *level* of licensed generalisation in sound systems, a specialisation built on top of the antecedent domain-independent capacity to learn at this level. This question strikes us as productive. We hope that the model can contribute towards testing this empirical question: are learners more or less prone to assume structural reuse in speech systems than in other behavioural systems?; do people's generalisations among sound classes vary with evidence for more or less system-wide structural re-use, or are they stronger biases that cannot be easily overruled?; can people be trained to generalise strongly in one part of the sound system and not in another?; By representing these inclinations parametrically, the model provides a quantitative inference scheme for answering these kinds of questions experimentally.

Evolutionary modelling suggests that inductive biases which evolve to sub-serve learned conventional systems are likely to be defeasible pre-dispositions (Thompson et al. 2016); experimental evidence indicates that learners initially over-generalise then roll back these expectations when exceptions are encountered, in many linguistic domains (e.g. Ambridge et al. 2013). In this respect, evolutionary theory and experimental results align to make a strong prediction about the kinds of priors we should expect to infer from target experiments: they should favour generalisation, but be weak enough so that data can overrule that expectation.

More generally, we conjecture that large-scale cultural transmission of multiple behavioural systems provides an environment that favours learners who are well equipped to learn at this level. This would imply that language is one of many under-constrained conventional domains in which it is beneficial to identify or impose group-like structure on concurrent learning problems, and therefore that there is an abundance of motivation for refinement of these abilities: the reasoning we have outlined for sound systems may be a specific instance of a more general dynamic. This idea remains to be studied formally in evolutionary models.

4.2.2 Languages adapting to priors

A foundational insight from the language evolution literature suggests that languages adapt culturally to reflect the biases of the individuals that learn them (Griffiths and Kalish 2007), either faithfully or by exaggerating these preferences. Specific instances of this general dynamic have been demonstrated in speech-related experiments. For example, Verhoef et al. (2014) showed that, when people are asked to observe and recreate a repertoire of continuous acoustic signals, and the process is repeated over transmission chains to simulate cultural evolution, those repertoires of signals develop an internal structure that is built around re-use of structural primitives. This suggests that an inclination towards generalisation not only benefits individual learning, but can also have long-term structure-forming benefits over the course of cultural transmission. One interpretation of results like these is that the sound system of any mature language will already have evolved culturally to posses approximately the degree of shared structure expected by a learner, resolving any mis-match, and diffusing the need for this quality of a language to be learned beyond the learner's initial assumptions. Our arguments concern a learner whose task is to infer the structure underpinning an established linguistic system, but must ultimately be understood as part of this cultural process.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

If sound systems have converged on prior expectations in this respect, the degree of shared structure observable in the sound systems of existing languages should be a good indication of people's priors over the common degree. Any variety in the systematicity of sound inventories across languages would imply that this quality is learned under a relatively weak prior, rather than a strong, fixed assumption made by the learner. If, as is likely, factors other than inductive biases also shape the structure of sound systems (Martinet 1939; Vaux and Samuels 2015; Wedel and Fatkulin 2017), then we should expect that this structure reflects an unpredictable mixture of learner-internal and learner-external constraints and must therefore be learned. Either way, we hope our analysis can provide tools to help distinguish between these possibilities, and we are enthusiastic about the prospect of combining cognitive models of the kind studied here with evolutionary simulations of cultural transmission.

4.3 Future directions

Several natural extensions of this work strike us as exciting. In computational terms, it would be straightforward to induce dependencies among the borrowing strength variables. Doing so would model a learner who makes generalisations about the borrowing strength: a generalisation about the degree of generalisation. Similarly, existing techniques for inferring, rather than fixing, more of the model's variables (see e.g. Escobar and West 1995) would be valuable additions to the model, particularly with respect to the parameters of the prior over individual structures (e.g. bimodal, trimodal categories). In experimental terms, we hope to apply this model to experimental instantiations of cultural transmission in which initial generations learn unsystematic phonetic systems and transmit their inferences to new participants: the aim will be to estimate the common degree at each generation, and make inferences about the participants' priors at this level. Cross linguistic studies of the inclination to assume systematic generalisations are another exciting avenue for exploring the influence of language experience on expectations at this level.

More generally, we aim to draw attention to the idea that solving inference problems at higher-order levels of abstraction is a distinctive feature of human cognition that may underpin our proclivity for systematic organisation: language, speech, and other systems of shared, structured, conventional behaviours can be maintained in human populations thanks to higher-order principles of organisation that bind these behavioural inventories into learnable packets.

5. Conclusion

Speech sounds, like many aspects of language, come in systems. Perhaps it is the organisation of the system, rather than the details of its elements, that is ingrained in speech cognition. The individual elements in sound systems may vary dramatically over time and between populations as a result of cultural evolution, but the presence of system-level organisational principles is universal. The ability to learn these higher-order qualities of sound systems can bootstrap learning of lower-level features. We showed how system-level conclusions can be learned via statistical inference in concurrent phonetic category induction problems. This leads to an evolutionary hypothesis: culture forces cognitive specialisation further up the hierarchy of abstraction—away from lower level details, and towards the ability to extract and extend abstract generalisations that make future learning faster and more accurate.

References

Ambridge, B. et al. (2013) 'The Retreat from Overgeneralization in Child Language Acquisition: Word Learning, Morphology, and Verb Argument Structure', *Wiley Interdisciplinary Reviews: Cognitive Science*, 4/1: 47–62.

Arnon, I. and Clark, E. V. (eds) (2011) *Experience, Variation and Generalization, volume 7 of Trends in Language Acquisition Research*. Amsterdam: John Benjamins Publishing Company.

Boyd, R., and Richerson, P. J. (1985) *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

Calamaro, S., and Jarosz, G. (2015) 'Learning General Phonological Rules From Distributional Information: A Computational Model', *Cognitive Science*, 39/3: 647–66.

Canini, K. R., Shaskov, M. M., and Griffiths, T. L. (2010) 'Modeling Transfer Learning in Human Categorization with the Hierarchical Dirichlet Process'. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*.

Chater, N., Reali, F., and Christiansen, M. H. (2009) 'Restrictions on Biological Adaptation in Language Evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 106/4: 1015–20.

Christiansen, M. H., and Chater, N. (2008) 'Language as Shaped by the Brain', *Behavioral and Brain Sciences*, 31/05: 489–509.

Clements, G. N. (2003) 'Feature Economy in Sound Systems', *Phonology*, 20/3: 287–333.

de Boer, B. (2000) 'Self Organization in Vowel Systems', *Journal of Phonetics*, 28: 441–65.

—— and Kuhl, P. K. (2003) 'Investigating the Role of Infant-Directed Speech with a Computer Model', *Acoustics Research Letters Online*, 4/4: 129.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

——, Sandler, W., and Kirby, S. (2012) 'New Perspectives on Duality of Patterning: Introduction to the Special Issue', *Language and Cognition*, 4/4: 251–9.

Doyle, G., Bicknell, K., and Levy, R. (2014) 'Nonparametric Learning of Phonological Constraints in Optimality Theory'. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1094–103.

Dunbar, E., and Dupoux, E. (2016) 'Geometric Constraints on Human Speech Sound Inventories', *Frontiers in Psychology*, 7: 1061.

Eaves, B. S. et al. (2016) 'Infant-Directed Speech Is Consistent With Teaching', *Psychological Review*, 123/6: 758–71.

Escobar, M. D., and West, M. (1995) 'Bayesian Density Estimation and Inference Using Mixtures', *Journal of the American Statistical Association*, 90/430: 577.

Feldman, N. H. et al. (2013) 'A Role for the Developing Lexicon in Phonetic Category Acquisition', *Psychological Review*, 120/4: 751–78.

Ferguson, T. S. (1973) 'A Bayesian Analysis of Some Nonparametric Problems', *The Annals of Statistics*, 1/2: 209–30.

Finley, S., and Badecker, W. (2009) 'Artificial Language Learning and Feature-Based Generalization', *Journal of Memory and Language*, 61/3: 423–37.

Gelman, A. et al. (2003). *Bayesian Data Analysis*. Fl, USA: CRC Press.

Goldberg, A. E. (2009) 'The Nature of Generalization in Language', *Cognitive Linguistics*, 20/1: 93–127.

Goodman, N. (1983) *Fact, Fiction, and Forecast*. Cambridge, MA, US: Harvard University Press.

Griffiths, T. L. et al. (2007) 'Unifying Rational Models of Categorization via the Hierarchical Dirichlet Process'. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, p. 323–328.

——, and Kalish, M. L. (2007) 'Language Evolution by Iterated Learning with Bayesian Agents', *Cognitive Science*, 31: 441–80.

Hare, M., and Elman, J. L. (1995) 'Learning and Morphological Change', *Cognition*, 56/1: 61–98.

——, ——, and Daugherty, K. G. (1995) 'Default Generalisation in Connectionist Networks', *Language and Cognitive Processes*, 10/6: 601–30.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007) 'Learning Over hypotheses with Hierarchical Bayesian Models', *Developmental Science*, 10/3: 307–21.

Kirby, S., Griffiths, T., and Smith, K. (2014) 'Iterated Learning and the Evolution of Language', *Current Opinion in Neurobiology*, 28C: 108–14.

Kirchner, R., Moore, R. K., and Chen, T.-Y. (2010) 'Computing Phonological Generalization Over Real Speech Exemplars', *Journal of Phonetics*, 38/4: 540–7.

Kleinschmidt, D. F., and Jaeger, T. F. (2015) 'Robust Speech Perception: Recognize the Familiar, Generalize to the Similar, and Adapt to the Novel', *Psychological Review*, 122/2: 148–203.

Ladd, D. R. (2014). *Simultaneous Structure in Phonology*. Oxford, UK: Oxford University Press.

Liljencrants, J., and Lindblom, B. (1972) 'Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast', *Language*, 48/4: 839–62.

Lin, D., and Fisher, J. (2012) 'Coupled Dirichlet Processes: Beyond HDP'. In *NIPS 2012 Workshop on Bayesian Nonparametric Models* (BNMP) for Reliable Planning and Decision-Making Under Uncertainty.

——, Grimson, E., and Fisher, J. W. (2010) 'Construction of Dependent Dirichlet Processes based on Poisson Processes'. In: Lafferty, J. D. et al. (eds), *Advances in Neural Information Processing Systems 23*, pp. 1396–404. Barcelona, ES: Curran Associates, Inc.

Lin, Y., and Mielke, J. (2008) Discovering place and manner features: What can be learned from acoustic and articulatory data. University of Pennsylvania Working Papers in Linguistics, 14(1).

Mackie, S., and —— (2011) 'Feature Economy in Natural, Random, and Synthetic Inventories'. In: Clements, G. N. and Ridouane, R., (eds), Camridge, UK: *Where Do Phonological Features Come From?: Cognitive, Physical and Developmental Bases of Distinctive Speech Categories*, pp. 43–64.

Maddieson, I., and Disner, S. F. (1984) 'Patterns of Sounds'. *Cambridge studies in speech science and communication*, pp. 1–4.

Martinet, A. (1939). 'Rôle de la corrélation dans la phonologie diachronique'. In *Travaux du Cercle linguistique de Prague VIII*, pp. 273–88. Prague: Cercle Linguistique de Prague.

Maye, J., Weiss, D. J., and Aslin, R. N. (2008) 'Statistical Phonetic Learning in Infants: Facilitation and Feature Generalization', *Developmental Science*, 11/1: 122–34.

McMurray, B., Aslin, R. N., and Toscano, J. C. (2009) 'Statistical Learning of Phonetic Categories: Insights from a Computational Approach', *Developmental Science*, 12/3: 369–78.

Moran, S., McCloy, D., and Wright, R. (eds) (2014) *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Muller, P., Quintana, F., and Rosner, G. (2004) 'A Method for Combining Inference Across Related Nonparametric Bayesian Models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66/3: 735–49.

Paisley, J. et al. (2015) 'Nested Hierarchical Dirichlet Processes', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37/2: 256–70.

Pajak, B., Bicknell, K., and Levy, R. (2013) 'A Model of Generalization in Distributional Learning of Phonetic Categories'. In *Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics*, pp. 11–20.

——, and Levy, R. (2011) 'Phonological Generalization from Distributional Evidence'. In Carlson, L., Holscher, C., and Shipley, T. (eds), *Proceedings of the 33rd Annual Con-*

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

ference of the Cognitive Science Society, pp. 2673–8, Austin, Texas.

——, and —— (2014) 'The Role of Abstraction in Non-Native Speech Perception', *Journal of Phonetics*, 46: 147–60.

—— et al. (2016) 'Learning Additional Languages as Hierarchical Probabilistic Inference: Insights From First Language Processing', *Language Learning*, 66/4: 900–44.

Perfors, A., Tenenbaum, J. B., and Wonnacott, E. (2010) 'Variability, Negative Evidence, and the Acquisition of Verb Argument Constructions', *Journal of Child Language*, 37/03: 607–42.

Ravignani, A., and Sonnweber, R. (2017) 'Chimpanzees Process Structural Isomorphisms Across Sensory Modalities', *Cognition*, 161: 74–9.

Shepard, R. (1987) 'Toward a Universal Law of Generalization for Psychological Science', *Science*, 237/4820.

Smirnova, A. et al. (2015) 'Crows Spontaneously Exhibit Analogical Reasoning', *Current Biology*, 25/2: 256–60.

Smith, L. B. et al. (2002) 'Object Name Learning Provides on-the-Job Training for Attention', *Psychological Science*, 13/1: 13–9.

Soskuthy, M. (2013) 'Phonetic Biases and Systemic Effects in the Actuation of Sound Change', PhD thesis, Edinburgh: The University of Edinburgh.

Sóskuthy, M. (2015) 'Understanding Change through Stability: A Computational Study of Sound Change Actuation', *Lingua*, 163: 40–60.

Spierings, M. J., and Ten Cate, C. (2016) 'Budgerigars and Zebra Finches Differ in How they Generalize in an Artificial Grammar Learning Experiment', *Proceedings of the National Academy of Sciences of the United States of America*, 113/27: E3977–84.

Teh, Y. W. et al. (2006) 'Hierarchical Dirichlet Processes', *Journal of the American Statistical Association*, 101/476: 1566–81.

Tenenbaum, J. B., and Griffiths, T. L. (2001) 'Generalization, Similarity, and Bayesian Inference', *Behavioral and Brain Sciences*, 24/04: 629–40.

—— et al. (2011) 'How to Grow a Mind: Statistics, Structure, and Abstraction', *Science (New York, N.Y.)*, 331/6022: 1279–85.

Thompson, B., Kirby, S., and Smith, K. (2016) 'Culture Shapes the Evolution of Cognition', *Proceedings of the National Academy of Sciences of the United States of America*, 113/16: 4530–5.

Toscano, J. C., and McMurray, B. (2010) 'Cue Integration With Categories: Weighting Acoustic Cues in Speech Using Unsupervised Learning and Distributional Statistics', *Cognitive Science*, 34/3: 434–64.

Vallabha, G. K. et al. (2007) 'Unsupervised Learning of Vowel Categories from Infant-Directed Speech', *Proceedings of the National Academy of Sciences of the United States of America*, 104/33: 13273–8.

Vaux, B., and Samuels, B. (2015) 'Explaining Vowel Systems: Dispersion Theory vs Natural Selection', *The Linguistic Review*, 32/3: 1–24.

Verhoef, T., Kirby, S., and de Boer, B. (2014) 'Emergence of Combinatorial Structure and Economy through Iterated Learning with Continuous Acoustic Signals', *Journal of Phonetics*, 43: 57–68.

Vonk, J., and MacDonald, S. E. (2004) 'Levels of Abstraction in Orangutan (Pongo abelii) Categorization', *Journal of Comparative Psychology*, 118/1: 3–13.

Wedel, A., and Fatkulin, I. (2017) 'Category Competition as a Driver of Category Contrast', *Journal of Language Evolution*.

Appendix

Appendix A. Unsupervised Bayesian inference for non-parametric mixtures of univariate Gaussians

In line with common practice, we model each sound class as an unbounded mixture of univariate Gaussians with unknown means and variances. The prior over this space of mixtures is given by the Dirichlet Process (Ferguson 1973) with base measure G and concentration/total mass parameter α. For G, we choose the Normal-Inverse-Chi-Square distribution (Gelman et al. 2003). We set $\alpha = 1$ in all analyses, as a relatively neutral expectation for the number of components represented in each mixture. We use a collapsed Gibbs sampler for posterior inference, marginalising over unknown means and variances for individual components/clusters, using the posterior predictive distribution as the likelihood for new data y_k^{new} given existing data y_k assigned to the cluster indexed by k:

$$\begin{aligned} p(y_k^{new} \mid y_k, \Phi) &= \int\int p(y_k^{new} \mid \mu, \sigma^2) p(\mu, \sigma^2 \mid y_k)\ d\mu\, d\sigma^2 \\ &= t_{\nu n}\left(\mu_n,\ \sigma_n^2 \cdot \frac{1+\kappa_n}{\kappa_n}\right)\ , \end{aligned} \tag{5}$$

where $t_\nu(\cdot)$ is the non-central Student t density with ν degrees of freedom, $\Phi = (\mu_0, \kappa_0, \sigma_0^2, \nu_0)$ are parameters of the prior $p(\mu, \sigma^2)$, and:

$$\kappa_n = \kappa_0 + n_k \tag{6}$$

$$\mu_n = \frac{1}{\kappa_n} \cdot (\kappa_0\, \mu_0 + n_k\, \bar{y}_k) \tag{7}$$

$$\nu_n = \nu_0 + n_k \tag{8}$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0 \sigma_0^2 + s^2 + n_k \kappa_0 \cdot \frac{(\mu_0 - \bar{y}_k)^2}{\kappa_n}\right) \tag{9}$$

$$s^2 = \sum_{i=1}^{n} (y_k^i - \bar{y}_k)^2 \tag{10}$$

$$\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_k^i \tag{11}$$

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

$$n_k = |y_k| \tag{12}$$

$$y_k = \{y_i \in Y_j : z_i = k, i = 1, \ldots, N_j\}\,. \tag{13}$$

See Gelman et al. (2003) for more details of the model set out in equations (5)–(13), which represent standard techniques and results for Bayesian analysis of the Gaussian distribution. We set $\mu_0 = 10$ and $\kappa_0 = 10^{-2}$ to ensure effectively uninformative expectations about the location of the categories in feature space, $\sigma_0^2 = 1.$ and $\nu_0 = 10^{-2}$ to ensure a very weak expectation for categories that aren't too diffuse in feature space. With this acoustic model for the individual categories, the sequential representation of the Dirichlet process can be used to sample category assignments for datapoints within a class C_j. If $\mathbf{z_j} = (z_1, \ldots, z_{N_j})$ are assignment variables for the internal structure of class C_j, such that $z_i = k$ indicates that datapoint $y_i \in Y_j$ has been assigned to category k in class C_j, and $\mathbf{z_{j-i}}$ denotes the full set of assignments in class C_j *excluding* the assignment z_i, then the posterior distribution for category assignment is proportional to:

$$p(z_i = k \mid Y_j \, \mathbf{z_{j-i}}) \propto p(y_i \mid z_i = k, \mathbf{z_{j-i}}) p(z_i = k \mid \mathbf{z_{j-i}}). \tag{14}$$

The likelihood term is given by equation (5). The prior $p(z_i = k \mid \mathbf{z_{j-i}})$ has a simple form under the Dirichlet process:

$$p(z_i = k \mid \mathbf{z_{j-i}}) = \begin{cases} n_k/n_* & \text{if } n_k > 0 \\ \alpha/n_* & \text{if } n_k = 0 \\ & \text{(i.e. category } \kappa \text{ is a} \\ & \text{new category),} \end{cases} \tag{15}$$

where $n_* = N_j + \alpha$. Note that N_j is defined in the main text as the number of datapoints observed in class C_j. In model M3, N_j has a slightly more subtle interpretation: it counts the total number of datapoints assigned to the parent DP (either F_j or F_*) of the category under consideration. See below. Equation (14) allows us to sample the full model posterior using a Gibbs sampler: iteratively cycling through the datapoints and resampling a category assignment based on all other assignments, over and over. Marginalising over the posteriors for the mean and variance for each category with equation (5) means there is no need to resample point estimates for these values.

The results we report in the main figures show the posterior predictive distribution for new data, computed over the final posterior sample after running the sampler for 1500 iterations. One sampling iteration corresponds to re-sampling category assignments for all datapoints.

Appendix B. Inducing dependencies between Dirichlet Processes (DPs)

Inducing dependence between the DPs is in most respects just like any other mixture model. Full details of the statistical model can be found in Muller et al. (2004): here, following Muller et al. but simplifying, we provide an algorithmic guide to implementing sampling for ϵ_j. Within a class C_j, any datapoint can be owned either by the class-specific component F_j or by the shared component F_*. Allow a second indicator variable for each datapoint: $r_i = 0$ if datapoint y_i (with class index j implicit) was generated by any of the categories in the local component F_j, and $r_i = 1$ if it was generated by any of the categories in the shared component F_*.

When re-sampling category assignments for a datapoint, compute the posterior probabilities (using equation 14) for each of the local categories (categories belonging to F_j) and weight each by $1 - \epsilon_j$. Compute posterior probabilities for all categories in the shared component F_*, and weight these by ϵ_j. Concatenate these probability vectors and re-normalise. Sample a category assignment from this mixture posterior and set r_i accordingly. After a full sweep through all datapoints, sampling category assignments and indicator variables, sample a new value for ϵ_j at each iteration of the sampler as follows. Let $\mathbf{r_j} = \{r_1, \ldots, r_{N_j}\}$ be the set of all ownership indicator variables for class C_j. Then, under the assumption that each r_i is an i.i.d. Bernoulli trial with success probability ϵ_j, the number of datapoints allocated to the common component $q_j = \sum_{i=1}^{N_j} r_i$ is a binomial random variable from which we can reverse-engineer ϵ_j (through standard Bayesian analysis of the Beta-Binomial distribution. See Gelman et al., 2003). Muller et al. (2004) assume a mixture prior over ϵ that includes components which reserve positive probability mass for $\epsilon = 1$ and $\epsilon = 0$. At the expense of these extreme cases, we simplify and assume a conjugate Beta prior density $\epsilon_j \sim \text{Beta}(1, 1)$. Sampling ϵ_j then reduces to sampling from $\text{Beta}(1 + q_j, N_j - q_j + 1)$. This procedure is given in pseudo-code below

```
Initialization. assign all datapoints in a class to one category owned by $F_j$; sample $\epsilon_j$ from the prior, for all $j$.;
for $t = 1, \ldots, \infty$ do
    for each sound class $j = 1, \ldots, n$ do
        for each datapoint $i = 1, \ldots, N_j$ do
            i) compute assignment posteriors $p(z_i = k)$ for all categories $k$ owned by $F_j$, using equation (14);
            ii) weight this posterior vector by $1 - \epsilon_j$;
            iii) compute assignment posteriors $p(z_i = k)$ for all categories $k$ owned by $F_*$ using equation (14);
```

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017

```
            iv) weight this posterior vector by $1 - \epsilon_j$;
            v) concatenate posterior vectors and renormalise to obtain full category assignment posterior;
            vi) sample a category assignment $z_i$ from the full posterior;
            vii) update $z_i$, update $r_i$;
        end
        compute $q_j$ from $\mathbf{r}_\mathrm{j}$;
        sample $\epsilon_j$ from Beta$(1 + q_j, N_j - q_j + 1)$
    end
end
```

Algorithm 1: Psuedo-code for the Gibbs sampler implementing posterior inference in model **M3**.

Downloaded from https://academic.oup.com/jole/article-abstract/2/1/94/3893599
by guest
on 04 December 2017