Running head: LEARNABILITY AND CULTURAL UNIVERSALS

Greater learnability is not sufficient to produce cultural universals

Anna N. Rafferty

Computer Science Division

University of California, Berkeley, CA 94720 USA

Email: rafferty@cs.berkeley.edu

Phone: 650 450 3604

Thomas L. Griffiths

Department of Psychology

University of California, Berkeley, CA 94720 USA

Marc Ettlinger

Research Service

Veterans Affairs Northern California Health Care System, Martinez, CA 94553 USA

Abstract

Looking across human societies reveals regularities in the languages that people speak and the concepts that they use. One explanation that has been proposed for these “cultural universals” is differences in the ease with which people learn particular languages and concepts. A difference in learnability means that languages and concepts possessing a particular property are more likely to be accurately transmitted from one generation of learners to the next. Intuitively, this difference could allow languages and concepts that are more learnable to become more prevalent after multiple generations of cultural transmission. If this is the case, the prevalence of languages and concepts with particular properties can be explained simply by demonstrating empirically that they are more learnable. We evaluate this argument using mathematical analysis and behavioral experiments. Specifically, we provide two counter-examples that show how greater learnability need not result in a property becoming prevalent. First, more learnable languages and concepts can nonetheless be less likely to be produced spontaneously as a result of transmission failures. We simulated cultural transmission in the laboratory to show that this can occur for memory of distinctive items: these items are more likely to be remembered, but not generated spontaneously once they have been forgotten. Second, when there are many languages or concepts that lack the more learnable property, sheer numbers can swamp the benefit produced by greater learnability. We demonstrate this using a second series of experiments involving artificial language learning. Both of these counter-examples show that simply finding a learnability bias experimentally is not sufficient to explain why a particular property is prevalent in the languages or concepts used in human societies: explanations for cultural universals based on cultural transmission need to consider the full set of hypotheses a learner could entertain and all of the kinds of errors that can occur in transmission.

Keywords: cultural universals; iterated learning; learnability bias; cultural evolution; vowel harmony

Greater learnability is not sufficient to produce cultural universals

A comparison of how people speak and think across human societies reveals some surprising regularities. To give two examples, the syntax of human languages shows less variability than might be expected if languages were simply arbitrary communication schemes (Greenberg, 1963; Comrie, 1981; Croft, 2002), and religious concepts seem to follow a common schema (being "minimally counterintuitive") in a range of societies (Boyer, 1994). The existence of these cultural universals raises a natural question: Where do they come from? What makes particular languages or concepts more likely to appear in a society? Recent work has explored a possible answer to this question, based on differences in the ease with which languages and concepts are transmitted from person to person (e.g., Boyer, 1994, 2001; Boyer & Ramble, 2001; Culbertson, to appear; Finley & Badecker, 2007; Kirby, Cornish, & Smith, 2008; Moreton, 2008; Scott-Phillips & Kirby, 2010; Wilson, 2006). The basic idea behind this answer is that concepts and linguistic features that are easier to transmit are more likely to survive the process of transmission, and thus have the potential to become more prevalent: "[I]n order for linguistic forms to persist from one generation to the next, they must repeatedly survive the processes of expression and induction. That is, the output of one generation must be successfully learned by the next if these linguistic forms are to survive." (p. 303, Brighton, Kirby, & Smith, 2005). Since ease of transmission is presumably related to the compatibility of languages and concepts with human learning and memory, this provides a mechanism by which we should expect cultural objects to shape themselves to the structure of human minds.

The idea that cultural universals result from ease of transmission suggests an empirical strategy for explaining the prevalence of a particular property of languages or concepts, in which laboratory experiments are used to show that it is easier for people to learn or remember stimuli with that property than those without it (e.g., Boyer & Ramble, 2001; Culbertson, Smolensky, & Legendre, 2012; Finley, 2012; Finley & Badecker, 2007; Moreton, 2008; Tily, Frank, & Jaeger, 2011; Wilson, 2006). Moreton (2008) provides a description of how experimental evidence about learnability can shed light on biases: “In typological theories based on analytic bias, asymmetries between attested and unattested phonologies are attributed to cognitive predispositions which admit some phonological patterns and exclude others.” (p. 85). This empirical strategy simplifies the problem of investigating linguistic universals: “[E]xperimental techniques, such as artificial grammar learning paradigms, make it possible to uncover the psychological reality of claimed universal tendencies.” (p. 1, Finley, 2012). Similarly, Wilson (2006) notes that “[b]y demonstrating that participants generalize from a brief period of exposure in the way predicted by a formal, substantively biased learning model – not in the way predicted by an otherwise identical model that lacks substantive bias – the results reported here shift the debate from speculation over the source of typological distribution to experimental investigation of human learning.” (p. 968). In this description, it is clear that the goal is to draw conclusions about typological distributions based on what types of biased generalizations people make. While Boyer and Ramble (2001) are somewhat more circumspect about what conclusions can be drawn from better recall of one type of concept over another, they too use this evidence in support of what will become universal, saying, “[W]e can expect, all else being equal, concepts that are very easy to recall to spread in a cultural environment and concepts that are intrinsically difficult to recall to spread less.” (p. 538).

This experimental strategy relies on the premise that more accurate learning of a language or concept with a particular property is sufficient for that property to become widespread. We analyze whether this premise is sound using a combination of mathematical analysis and behavioral experiments. As a starting point, we use a simple linear transmission framework to model cultural evolution. A model of cultural evolution is linear if each agent observes data that were generated by a single other agent and forms a hypothesis about which language or concept generated these data. For example, in the case of language, each agent might hear a set of utterances and form a hypothesis about what language generated these utterances. After forming such a hypothesis, the agent then produces data that will be observed by another agent. After many such transmission events the distribution over the hypotheses that are learned converges to an equilibrium distribution, which indicates the relative prevalence of particular languages within the population.

Using a formal model of cultural transmission allows us to analyze how the ease with which particular languages and concepts are transmitted relates to their ultimate prevalence. We present two counter-examples showing that greater probability of being transmitted accurately is not sufficient for greater prevalence. These counter-examples correspond to cases that could plausibly arise for the transmission of languages and concepts. For simplicity we will refer to the transmission of hypotheses rather than differentiating the cases of languages and concepts. The first counter-example concerns a situation in which a hypothesis is transmitted with high probability, but once it disappears it is unlikely to reappear. This scenario could potentially arise with “minimally counterintuitive” religious concepts (Boyer, 1994, 2001), which are more memorable but less likely to be generated spontaneously. The second counter-example is a case in which a hypothesis has a sufficiently high probability of being transmitted successfully as to be more probable than any other single hypothesis, but there are many more other hypotheses. In this case, the other hypotheses may still dominate in the population. This situation can arise in transmission of languages, where the set of possible languages with a particular property may be far smaller than the set without.

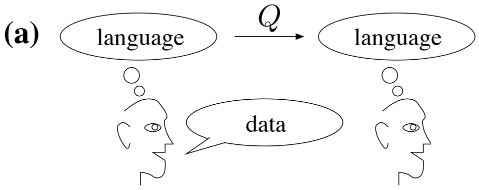
For each of our counter-examples we illustrate the theoretical possibility of greater learnability not resulting in a universal, and then provide an empirical demonstration of this phenomenon. For the first counter-example we conduct an experiment inspired by work on the transmission of religious concepts (Boyer & Ramble, 2001), showing that a distinctive item on a memorized list is transmitted with high probability, but nonetheless disappears from the list and does not return. For the second counter-example we use a paradigm similar to that of Finley and Badecker (2007) to explore learning and transmission of artificial languages. An initial experiment shows that an artificial language containing vowel harmony is transmitted more successfully than an arbitrary language. However, a second experiment demonstrates that vowel harmony quickly disappears when languages are transmitted across multiple generations.

Both of our counter-examples illustrate that it is not sufficient to show that some types of languages or concepts are more likely to be accurately transmitted in order to explain why these concepts become dominant across cultures. Determining the outcome of cultural transmission is a challenging problem, and accuracy of transmission is only one of the relevant factors. We argue that conclusions about cultural universals resulting from cultural evolution can only be obtained by making stronger assumptions about the transmission process, making more complete models that capture differences in learnability, characterizing the pattern of changes that result from information passing from one person to another, or simulating cultural transmission in the laboratory. We show that there are cases where learnability differences can be used to predict long-term outcomes, when certain assumptions about the transmission process are made. These predictions depend on the factors that were highlighted in the counter-examples, such as the number of hypotheses with a particular property, in addition to the strength of the learnability bias.

Formalizing Cultural Transmission

Languages and concepts change over time as they are transmitted from generation to generation (e.g., Bartlett, 1932; Labov, 2001). Our goal is to understand how the long-term consequences of this process of change are related to the factors that influence the success of a single transmission event. We begin by formalizing cultural transmission using a linear model, in which it is assumed that each person learns a concept or language from data produced by a single person in the previous generation. This model subsumes other popular models in the literature on cultural evolution, including simple versions of the iterated learning model (Griffiths & Kalish, 2007; Kirby, 2001) and the replicator dynamics (Komarova & Nowak, 2003; Schuster & Sigmund, 1983). We use this model to examine whether greater ease of transmission is sufficient to allow a language or concept to become prevalent.

We will use the term “hypothesis” to refer to any piece of information transmitted from one



(b)

$$\mathbf{Q} = \begin{bmatrix} 0.90 & 0.02 & 0.03 & 0.02 \\ 0.01 & 0.90 & 0.02 & 0.04 \\ 0.04 & 0.03 & 0.75 & 0.14 \\ 0.05 & 0.05 & 0.20 & 0.80 \end{bmatrix}$$

(c)

$$\boldsymbol{\pi} = \begin{bmatrix} 0.19 \\ 0.21 \\ 0.25 \\ 0.35 \end{bmatrix}$$

Figure 1. The linear model of cultural transmission. (a) A hypothesis is passed from one learner to another, and the transition matrix **Q** encodes the probability a learner will end up with hypothesis h_i when learning from data generated by somebody with hypothesis h_j. (b) An example transition matrix **Q** with four states. (c) The solution to the eigenvector equation $\mathbf{Q}\pi = \pi$ for this transition matrix. π gives the equilibrium probability that a learner will learn a particular hypothesis when hypotheses are transmitted via a process that has transition matrix **Q**.

person to another, such as a language or a concept. The first step in specifying our model is then to define the set of possible hypotheses, denoted H. Each element $h \in H$ is one possible hypothesis representing a specific concept or language. Transmission occurs when a new member of the population receives data from another member of the population and learns some $h \in H$. We assume that transmission occurs only from one person to another person, and that each person learns only one hypothesis. For example, one member of the population who knows language h_j might transmit that language to another member of the population, and that member might acquire language h_j. Alternatively, another language might be learned: The learner might not have heard enough data to fully specify h_j as the language or might have misheard something, and thus infers another language h_i that is consistent with the data she or he heard. More formally, we assume that for all $h_i, h_j \in H$, q_{ij} is the probability that someone will learn hypothesis h_i from someone who knows hypothesis h_j. These can be encoded in a *transition matrix* **Q** where the (i, j)th entry of the matrix corresponds to the *transition probability* q_{ij} (see Figure 1).

Using this framework, we can define learnability biases explicitly and determine whether a learnability bias for some property necessarily implies that this property will be present in the majority of hypotheses produced as a result of cultural transmission. As mentioned previously, a learnability bias means that one type of hypothesis is more likely to be transmitted accurately from one generation to the next than another hypothesis. This definition is similar to the notion of "cognitive bias" discussed in Wilson (2003). The learnability of a particular hypothesis is often found experimentally, as in experiments exploring how successfully languages or concepts with different properties are transmitted (e.g., Boyer & Ramble, 2001; Finley & Badecker, 2007; Wilson, 2003). Experiments that find that hypotheses with one property are more likely to be successfully learned than hypotheses without that property are establishing a learnability bias for that property. Formally, we define a learnability bias for some hypothesis h_i over some other hypothesis h_j as meaning that $q_{ii} > q_{jj}$. For example, one might expose one group of learners to language h_i and another group to language h_j. If more learners in the first group accurately learned the language they were exposed to, this would indicate a learnability bias for language h_i over language h_j. In a Bayesian model, a learnability bias might be expressed by having higher prior probability on one hypothesis than on another; however, our definition of a learnability bias simply expresses a tendency for one type of hypothesis to be more easily transmitted than another and does not rely on any particular model of learning.

We can extend the idea of a learnability bias to an abstract property of a hypothesis, rather than a specific hypothesis, by applying a similar definition to *sets* of hypotheses. Imagine there are two sets of hypotheses, H_1 and H_2. These sets might be defined by classifying all hypotheses with a particular property in H_1 and all concepts without the property in H_2. A learnability bias that favors a particular property means that each concept or language with that property is more likely to be transmitted successfully than each concept or language without that property. That is, for all possible pairs $h_i \in H_1$ and $h_j \in H_2$, $q_{ii} > q_{jj}$. This would indicate a general learnability bias for hypotheses in H_1 over hypotheses in H_2.

Using this definition of a learnability bias, we can now determine whether such a bias is sufficient to establish that the favored property will be present in the majority of hypotheses that learners learn. That is, if H_1 denotes the hypotheses with the property of interest, we want to determine whether a learnability bias for hypotheses in H_1 implies that the majority of the hypotheses in the population will be in H_1 and not in H_2 after hypotheses have been transmitted from person to person for some time. If this is the case, then demonstrating a learnability bias is sufficient to explain why a particular property might become universal. If not, we should be cautious in interpreting the evidence provided by learnability biases.

We can determine the consequences of cultural transmission by appealing to existing results on the equilibrium of this linear dynamical system. As mentioned above, this linear transmission model is related to two kinds of models that have been used to study language and cultural evolution: If we assume that learners are organized in a chain, this linear model is called iterated learning (Kirby, 2001); alternatively, if we assume that there exists an infinite number of learners in the population, each of whom learns from a single randomly selected learner, the model is called the replicator dynamics (Schuster & Sigmund, 1983). In either case, the probability that a learner will learn hypothesis h, assuming the population has reached equilibrium, is given by the solution to the eigenvector equation $\mathbf{Q}\pi = \pi$, normalized such that $\sum_{i=1}^{n} \pi_i = 1$ (for details, see Griffiths & Kalish, 2007). For hypotheses in H_1 to become dominant, it must be the case that these hypotheses occur the majority of the time. This condition will be met if $\sum_{h \in H_1} \pi_h > \sum_{h \in H_2} \pi_h$.

The analysis given in this section provides us with the tools needed to determine whether a particular property will become universal based on examination of the transition matrix. In the remainder of the paper, we use these tools to determine when a learnability bias will ensure that a property will appear in the majority of hypotheses. We primarily focus on two counter-examples in which this is not the case. The first counter-example focuses on the case where a hypothesis is transmitted accurately, but unlikely to be generated spontaneously. The second counter-example considers what happens when the favored set of hypotheses is much smaller than the alternative set. In each case, we suggest that the situation described by this counter-example could plausibly arise in the context of transmitting concepts or languages and present empirical results bearing out our predictions. We end by considering when a learnability bias does lead to a property becoming prevalent after many generations and show two special cases where this result holds.

Counter-example 1: Easy to transmit, hard to generate

Our first counter-example derives from a situation in which there exist two sets, one of which has hypotheses that all have high self-transition probabilities (H_1) and one of which has hypotheses with lower self-transition probabilities (H_2). However, the second set also has high inter-transition probabilities for hypotheses in the set: Learners who learn from someone with a hypothesis from H_2 tend to acquire a hypothesis from H_2 rather than a hypothesis from H_1. Thus, the self-transition probabilities for the sets of languages differs from the self-transition probabilities for individual languages within the sets. This pattern might occur in cases where learners rarely learn a particular hypothesis unless they receive data generated specifically from that hypothesis. For example, some hypotheses may be unlikely to be spontaneously generated by learners as a result of transmission errors. Additionally, this pattern might occur when hypotheses of one type are less likely to be accurately transmitted than those of another, but are likely to have transmission errors that result in the learner acquiring a hypothesis of the same type. In this case, hypotheses of the less learnable type might be more likely to be confused with one another, resulting in the whole set being more prevalent.

The transition matrix **Q** shown in Figure 1 (b) is an example of a matrix with this property. Let $H_1 = \{h_1, h_2\}$ and $H_2 = \{h_3, h_4\}$. We have that $q_{ii} > q_{jj}$ for all $i \in H_1$ and $j \in H_2$: Each hypothesis in H_2 has a lower self-transition probability than any hypothesis in H_1. Thus, we have a learnability bias for H_1 over H_2. However, the eigenvector π shown in Figure 1 (c) indicates that the equilibrium of this system, which will be reached after languages are transmitted from person to person many times, does not favor hypotheses in H_1. Instead, $\sum_{h \in H_1} \pi_h = 0.4$ while $\sum_{h \in H_2} \pi_h = 0.6$: A plurality of learners will acquire h_4, and most learners will adopt a hypothesis in H_2.

The discrepancy between what one might predict based on simply looking at the self-transition probabilities q_{ii} versus the actual equilibrium distribution comes from the fact that the q_{ii} do not take into account the low probabilities of transitioning from a hypothesis in H_2 to a hypothesis in H_1. For instance, while the probability h_3 is transmitted accurately is only 0.75, in almost all cases where the learner acquires a different hypothesis, that hypothesis will be h_4: only 5% of the time will a learner learn a language in H_1 when she hears data generated by someone who knows h_3. Rather than involving only the self-transition probabilities, the equilibrium probability that learners learn a concept in H_1 (and hence the prevalence of the property associated with H_1) is a function of the fidelity of transmission between all pairs of hypotheses.

In most cases, one cannot explain away problematic transition matrices by first collapsing the matrix into two states, one representing all hypotheses in H_1 and one representing all hypotheses in H_2, and then examining the resulting transition probabilities between these two sets. Such a transformation on a Markov chain only preserves the Markov property in cases where for all $i, j \in H_i$ $\sum_{n \in H_2} q_{ni} = \sum_{n \in H_2} q_{nj}$ (Burke & Rosenblatt, 1958; Kemeny & Snell, 1960): for all states h_i in H_1, the total probability of transitioning from that state to any state in H_2 must be the same. Given that one would expect, for example, some languages or concepts to be more similar to the hypotheses in the alternative set than others, and thus to have varying transition probabilities from one another, it seems relatively unlikely that this condition will hold in real cultural transmission situations. Verifying this condition also requires knowing the entire transition matrix, something that is rarely obtained in experiments; however, we discuss several special cases where this criterion holds later in the paper.

This counter-example implies that if the linear transmission model is an accurate model for understanding cultural evolution, then it is not sufficient to compare how accurately languages or concepts are maintained over a single generation in order to predict what trends will emerge after many generations. Instead, one must also look at the complete pattern of transition probabilities between hypotheses. The ways in which hypotheses change through transmission may be as important as the relative fidelities of transmission in determining long term trends. When one only looks for a learnability bias, the rate at which hypotheses change into other hypotheses is not accounted for, leaving open the possibility that predictions about long-term trends will be incorrect. To demonstrate that the situation in which this counter-example arises is plausible, we now turn to an empirical demonstration using a laboratory simulation of cultural transmission.

Experiment 1: Memory for Distinctive Items

We have shown mathematically that it is possible for a set of hypotheses to have a learnability bias but still not become universal after repeated cultural transmission due to having a low probability of being spontaneously generated. We now demonstrate this phenomenon in a behavioral experiment by simulating cultural transmission in the lab with human learners. In Experiment 1, participants completed a memory task in which they were exposed to a list of items and were then asked to reproduce this list from memory. Each new participant was exposed to the previous participant's list, creating a linear transmission structure. This procedure is an instance of the "serial reproduction" paradigm introduced by Bartlett (1932) for studying the effects of cultural transmission on items reproduced from memory, which has recently been analyzed as an instance of the linear transmission model (Xu & Griffiths, 2010).

The design of our experiment was motivated by previous work connecting a difference in memorability to cultural universals in religious concepts (Boyer, 1994, 2001; Boyer & Ramble, 2001). This work starts from the observation that across human societies, religious concepts tend to be "minimally counterintuitive," involving only a small number of changes from concepts that correspond to real objects or forces. For example, a statue that cries is more likely to appear as a religious concept than a statue that cries, levitates, and is invisible. One proposed explanation for this apparent universal is that minimally counterintuitive concepts are more memorable than mundane or extremely counterintuitive concepts, and thus come to dominate through cultural transmission. In support of this explanation, several experiments found a memory advantage for minimally counterintuitive concepts that appeared in stories (Boyer, 2001; Boyer & Ramble, 2001).

In this experiment, we explore whether a memory advantage is actually sufficient to cause a particular type of concept to dominate after many generations of cultural transmission. It seems plausible, for instance, that certain types of memorable stimuli are harder to spontaneously generate, and thus may never be regenerated if they are ever forgotten. To explore this possibility, we included a distinctive item in the initial list that people had to reproduce from memory. People were told that they were remembering a grocery list, and the first list included the unusual item *elephants*. We predicted that this distinctive item would be easier to remember (consistent with the classic Von Restorff effect, Von Restorff, 1933), but would be unlikely to be spontaneously generated, and would hence disappear as a result of repeated cultural transmission. If this is the case, it suggests that memorability evidence alone is not sufficient to explain why a concept is prevalent after many generations of cultural transmission. Such a finding would not invalidate the work of Boyer and colleagues, especially as we have not tested how counterintuitive our unusual item is and thus cannot directly compare to their conditions. However, this finding would suggest that a memory advantage is not a complete explanation for the prevalence of minimally counterintuitive religious concepts, since it would show that more memorable items are not necessarily highly prevalent after many iterations of cultural transmission.

Methods

Participants. Fifty members of a university community came into the laboratory and received $10/hour compensation for their participation in this and several other unrelated studies. An additional 150 participants were recruited via Amazon Mechanical Turk and completed the study online; they received a small amount of monetary compensation for their participation.

Stimuli. The stimuli for each participant consisted of ten words presented on a computer screen. Words were displayed sequentially, in the center of the screen. Each word was displayed for four seconds, followed by a half second blank screen.

Procedure. Participants completed the task on the computer. The program informed participants that they would be shown ten words to remember for a subsequent memory test. Participants were told not to write down or otherwise record the words. After exposure to the sequence of words, participants were distracted for 60 seconds with a reading and response task (either reading and completing a form, or reading a paragraph and answering questions). Participants were then asked to list the items that they had seen, either on the computer or on a sheet of paper. They were instructed to fill in all ten spaces and to give their best guess if they did not remember one of the ten items.

The lists to remember were generated as follows. The first list was composed of the following nine relatively common nouns, which are all items one would buy at a grocery store: *toilet paper*, *cheese*, *tomatoes*, *eggs*, *milk*, *lettuce*, *orange juice*, *bread*, and *bananas*. Additionally, there was one distinctive item (*elephants*) in this first list. Participants were organized into chains. All laboratory participants formed one chain, and the online participants were split into three separate chains, each consisting of 50 people. After the first participant in the chain, subsequent participants were given lists to remember composed of the ten items given by the previous participant in the chain; items remained in the same order as remembered by the previous participant. Participants' responses were left out if they failed to provide 10 distinct items.

Results and Discussion

The average number of items retained by each participant was similar for all chains, ranging from 8.82–9.18, so we analyze the results from laboratory and online participants together. This mirrors previous research showing that responses from online participants were similar to responses from lab participants (Sprouse, 2011). To ensure that the chains reached a stationary distribution, we calculated the expected time to convergence. As described in Appendix A, we bounded this time by relating our model to the coupon-collector problem (Feller, 1968), and found

Table 1

All items generated at least five times in Experiment 1.

Word	apple	juice	cheese	cereal	milk	bread	butter	egg	water	orange
Times Generated	9	8	7	7	6	6	6	6	5	5

that with 95% probability, the chains would have reached convergence within 45 iterations.

We observed that when *elephants* occurred in a list, it was remembered by participants 95% of the time, while other items were remembered by participants 87% of the time. To determine if this difference was significant, we used a logistic regression analysis. We compared a model with only a constant coefficient to one that had both a constant coefficient and a coefficient set to one for the item *elephants* and zero otherwise, and sought to predict the probability that each word would be remembered in a single generation. We found that the elephants coefficient was positive and that the model with this coefficient was a significantly better fit than the model without this coefficient ($\chi^2(1) = 5.42$, $p < 0.05$). This demonstrates a learnability bias for the distinctive item, as it was more likely to be remembered than other items. Across all chains, *elephants* was remembered for an average of 19.5 iterations.

However, in all four chains of participants, *elephants* eventually disappeared and was never re-generated. This is in contrast to more typical grocery list items that tended to be forgotten more frequently, but were also often spontaneously generated (see Table 1); for instance, *apple* was generated by nine different participants. To demonstrate the types of words that occurred frequently after many generations, Figure 2 shows the top words in the final ten iterations, aggregated across lists. *Elephants* does not appear among these words, since it never occurred in the last ten iterations of any of the chains.

Based on these results, the equilibrium distribution for grocery lists is unlikely to assign high probability to lists which include *elephants*, despite this item being highly memorable. This is

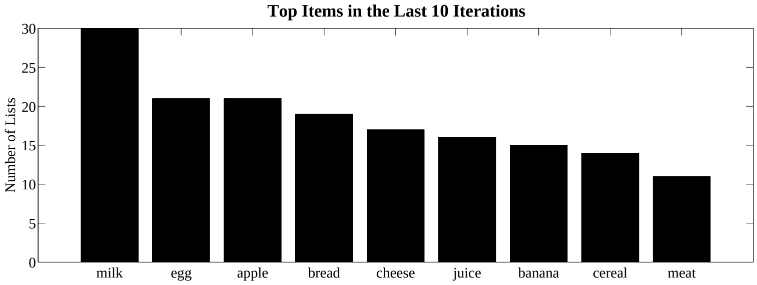


Figure 2. Results of Experiment 1, showing the most common words in the last ten iterations, aggregated across chains.

consistent with our predictions, and with the mathematical analysis presented in the previous section. Consequently, we should be cautious in using greater memorability as an explanation for why certain concepts seem to be universal – being easier to remember is not sufficient to allow a concept to dominate a population through cultural transmission, if that concept is not also reasonably likely to be generated spontaneously. An experiment that used a single generation of transmission would have found a memory bias for *elephants* (accurately transmitted 97% of the time) as compared to *apples* (accurately transmitted 89% of the time) and might lead one to erroneously conclude that *elephants* will be common in all future lists. However, to fully understand the way that memory biases interact with the process of cultural evolution, an experiment that simulates cultural transmission is required.

Counter-example 2: Differences in the Number of Hypotheses

Our first counter-example demonstrated that a learnability bias may not lead to a hypothesis having high probability in the equilibrium distribution resulting from cultural transmission. However, it is also possible that learnability biases might fail to translate to universals in the case where hypotheses with a particular property have higher equilibrium probabilities than hypotheses without. For example, consider the case where there are a limited number of hypotheses with the property of interest: even if each of these hypotheses is more probable in the equilibrium distribution than any hypothesis without the property, this property could still fail to become universal if there are many more hypotheses without that property. We next show how this situation can arise in language evolution, using a mathematical model and two artificial language learning experiments.

For this counter-example, we consider the property of vowel-harmony, although the counter-example is likely to apply in many other situations. Vowel harmony is an attested linguistic pattern wherein the vowels in words in a language must share some phonological feature. For example, in Turkish, the plural suffix is *-lar* in *bash-lar* 'heads', but *-ler* in *bebek-ler* 'babies' so as to adhere to the requirement that words are front-back harmonic. In the former, both vowels are back vowels, and in the latter, both vowels are front vowels. Vowel harmony is relatively common across the world's languages (van der Hulst & van de Weijer, 1995), but places a strong constraint on the structure of the lexicon. For instance, even if we imagine words being comprised of just two vowels, the requirement that vowels must share a phonological feature cuts down the space of possible words significantly.

Past work has shown that typologically attested vowel harmony patterns are generally more easily learned by English speakers than alternatives (Finley & Badecker, 2009; Moreton, 2008), although English does not possess this property. This demonstrates that there is a learnability bias for vowel harmony in human learners. In this section, we analyze language evolution using iterated learning with Bayesian agents in order to ascertain how large such a learnability bias must be in order to cause harmonic languages to dominate after many generations of cultural evolution. This allows us to explore how the necessary magnitude of the bias scales with the relative difference in number of hypotheses.

To analyze the prevalence of harmonic languages, we use iterated learning to explore the evolution of the lexicon of a language over time. In an iterated learning model, learners are organized in a chain, just as in Experiment 1. As in the more general transmission model, the dynamics of iterated learning depend on the transition matrix $\mathbf{Q}$. To define this matrix, we must specify the process by which learners select a language. We assume that learners are Bayesian, meaning that they infer a language h based on the data d that they receive according to Bayes' rule. The *posterior probability* assigned to h after observing d is $p(h|d) \propto p(d|h)p(h)$, where $p(d|h)$ (the *likelihood*) indicates the probability of d being generated from h, and $p(h)$ (the *prior*) indicates the extent to which the learner was biased towards h before observing d. If we assume learners select hypotheses with probability equal to their posterior probability, we obtain a transition matrix $\mathbf{Q}$ with entries

$$q_{ij} = p(h^{(t+1)} = i|h^{(t)} = j) = \sum_d p(h^{(t+1)} = i|d)p(d|h^{(t)} = j) \tag{1}$$

where $h^{(t)}$ and $h^{(t+1)}$ are the languages of learners at iterations t and $t+1$ respectively. Griffiths and Kalish (2007) have shown that in this model, iterated learning with Bayesian agents that sample from the posterior, the equilibrium distribution π is simply the learners' prior distribution $p(h)$. If a language h_i is more probable in the prior distribution and the likelihood function $p(d|h)$ takes the same form for all hypotheses, then h_i will have a self-transition probability q_{ii} that is greater than the self-transition probability of other languages, indicating a learnability bias. The equilibrium probability of this language will be $p(h_i)$, so if $p(h_i) > \sum_{j \neq i} p(h_j)$ then h_i will be more prevalent than all other languages at equilibrium. This implies that in order to dominate after many generations, it is insufficient for a particular language to have higher prior probability than any single other language. If there are many alternative languages, then the combined probability of these languages may overwhelm the bias towards the favored language.

To show that this situation can occur with the linguistic property of vowel harmony, we consider a particular lexicon. Assume that we have a finite number of words N, each of which has a vowel-harmonic variant and a variant that is not vowel harmonic. For example, words might be composed of a stem and one of two suffixes. With one suffix, the word is harmonic due to the suffix sharing a phonological feature of the vowel with the stem; with the other suffix, this feature is not shared. We define a language h as a binary vector of length N. If the ith position of the

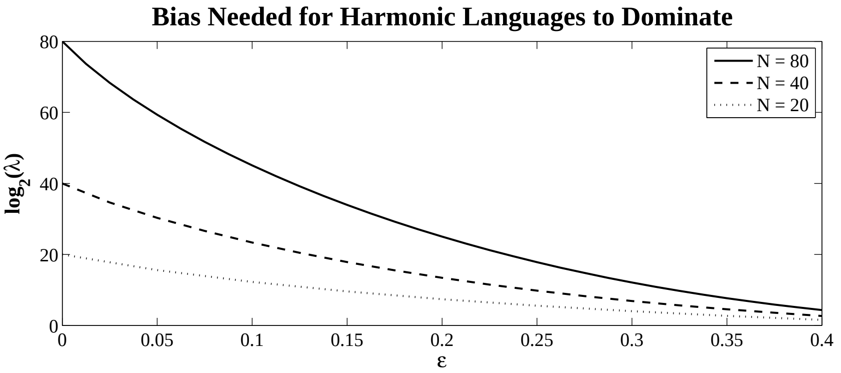


Figure 3. Bias λ necessary for harmonic languages to be more common than non-harmonic languages in equilibrium (λ is on a log scale). As the fraction ε of non-harmonic words allowed in a harmonic language increases, the bias λ that is necessary for harmonic languages to dominate decreases, since a larger ε leads to more harmonic languages relative to non-harmonic languages. As the total number of words *N* increases, the ratio of non-harmonic to harmonic languages also increases for each value of ε. This means that λ must be larger to compensate.

vector is a 1, the vowel-harmonic variant of the word is in the language; otherwise, the variant that is not vowel-harmonic is a part of the language. This results in 2^N possible languages. If we define a language as vowel-harmonic only if all words in the language are harmonic, then one of these 2^N languages is vowel-harmonic. More realistically, we might define a language as vowel-harmonic if the proportion of non-harmonic words is less than some ε. We can quantify the bias towards harmonic languages defining the prior probability on a hypothesis *h* as uniform within each type of language (harmonic and non-harmonic), and setting $p(h_{\text{harmonic}}) = \lambda p(h_{\text{non-harmonic}})$ for each hypothesis. Each harmonic language is thus λ times more likely in the equilibrium distribution than each non-harmonic language; for $\lambda > 1$, there is a learnability bias for harmonic languages.

Figure 3 shows the value of λ that is necessary for harmonic languages to be more prevalent than non-harmonic languages at equilibrium as a function of the proportion ε of non-harmonic words allowed: As ε decreases, the number of harmonic languages decreases, and thus the bias λ towards these languages must be larger (see Appendix B for mathematical details). If λ is greater than one, even much greater, but less than the value shown in the figure, there will be a learnability bias for harmonic languages over non-harmonic languages, but these languages will not be acquired by the majority of learners, and thus the learnability bias will not be sufficient to cause harmonic languages to be universal.

This mathematical analysis demonstrates that a relatively strong learnability bias will be required for languages with a particular property to dominate in cases where there are many more languages that lack the property. While past work has shown that some bias towards vowel harmony exists for English speakers (Finley & Badecker, 2009; Moreton, 2008), it is not clear whether this bias is sufficient to overcome the imbalance in the number of possible languages of each type. To explore this question, we conduct two experiments in which human learners learn an artificial language. In Experiment 2, we establish a learnability bias for a linguistic pattern based on vowel harmony over an arbitrary pattern, replicating results in prior work. In Experiment 3, we examine what happens when a language with the common pattern is transmitted multiple times among learners in the lab. Each learner learns a language and then produces data from this language to teach the next learner. By studying the languages that emerge after several generations of transmission in Experiment 3, we are able to determine whether the learnability bias found in Experiment 2 results in vowel harmony becoming widespread across the learned languages.

Experiment 2: Learnability of Vowel Harmony

Methods

Participants. A total of 40 members of the Berkeley community received either monetary compensation at $12/hr or course credit for their participation. All were native speakers of English.[1]

Stimuli. A trained linguist and native speaker of English was recorded saying 160 CVCVC words. Each word began with one of 80 CVC stems, twenty each with the vowels /i/, /e/, /u/ and /o/

and random consonants. Each stem was recorded with both variants, or allomorphs, of a suffix, [it] and [ut]. Thus, half the words were front-harmonic (e.g., pel-it, bis-it) and half were front-disharmonic (e.g., pel-ut, bis-ut). Alternatively, stimuli could be classified based on a dependency between the height of the first vowel and the frontness of the second value. Stimuli were height-front dependent if when there was a mid-vowel stem, there was a front vowel suffix (e.g., pel-it, bod-it), and when there was a high-vowel stems, there was a back-vowel suffix (e.g., bis-ut, tug-ut). Half of the words had this dependency, while the other half did not. The set of stimuli were constructed to be similar to those in Moreton (2008), but use a wider range of consonants to promote learning of the words and greater generalization (Gómez, 2002; Lively, Logan, & Pisoni, 1993).

Procedure. The procedure followed a modified artificial grammar paradigm. Participants were assigned to one of two conditions: the (attested) harmonic condition or the (unattested) height-front dependency condition. In both conditions, participants were exposed in training to 40 words from the language they were learning. In the harmonic condition, 40 harmonic words were selected. In the height-front dependency condition, 40 words were selected that met the height-frontness condition described above. This rule was chosen arbitrarily from the space of possible languages in order to test the hypothesis that vowel harmony would have a learnability bias over other unattested patterns.

Participants were familiarized with the words in the same way regardless of condition. They were given alternating blocks of passive listening and blocks in which for each trial, two words were played and they were required to choose which word they had previously heard. In the forced choice trials, the choice was between a word that had been played in the passive listening section and a word with the same stem and the alternate suffix. A total of five blocks of 40 trials each were included in training: three passive listening blocks with a forced choice block in between each. The forced choice trials were included as a method for consolidation of learning.

Following the training trials, participants completed one block of 80 test trials. On each test

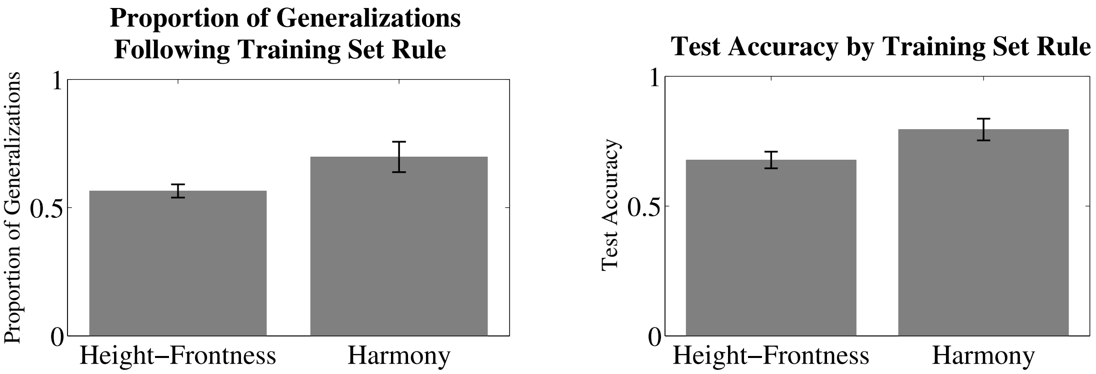


Figure 4. Results for harmonic versus height-frontness rule conditions in Experiment 2.

trial, participants were asked to choose which of two words they thought was from the language they had learned in the training trials. In each trial, the two words both had the same stem and differed in the suffix. 40 of the test trials included words from training, and 40 were generalization trials involving novel words.

Results and Discussion

As shown in Figure 4, we found a learnability bias for the harmonic language. Learners had significantly greater accuracy in test when they learned the vowel harmonic language than when they learned the height-frontness dependency language (80% correct for learners of the harmony rule versus 68% correct for the height-frontness rule, $t(38) = 2.23, p < .05$; Cohen's $d = 0.73$). Additionally, the proportion of generalizations that followed the learned rule was greater for learners in the harmony rule condition than learners in the height-frontness rule condition. 70% of generalizations made by learners of the harmonic language were harmonic in contrast to the 57% of generalizations by learners of the height-frontness dependency language that followed the height-frontness rule ($t(38) = 2.05, p < .05$; $d = 0.67$).[2] The result of these two phenomena was that the final languages produced by the learners in the harmony condition had a greater prevalence of harmonic words than the final languages of learners in the height-frontness dependency had of adhering words.

These results establish that the probability of transitioning from a harmonic language to another language with a high proportion of harmonic words is higher than the probability of transitioning from a height-frontness language to another language with a high proportion of adhering words. In terms of the transition matrix, this corresponds to $q_{\ell_{\text{harm}},\ell_{\text{harm}}} > q_{\ell_{\text{h-f}},\ell_{\text{h-f}}}$, where ℓ_{harm} is the set of languages with a high proportion of harmonic words and $\ell_{\text{h-f}}$ is the set of languages with a high proportion of words that follow the height-frontness rule. This satisfies our criterion for a learnability bias, and is the same quantity that is typically evaluated in arguments that relate learnability to typology in previous work (e.g., Finley & Badecker, 2007; Moreton, 2008; Tily et al., 2011; Wilson, 2006). This bias is of roughly the same magnitude as found in previous work.

Experiment 3: Transmission of Vowel Harmony

Methods

Participants. A total of 104 members of the Berkeley community received either monetary compensation of $12/hr or course credit for their participation. All were native speakers of English.[3]

Stimuli. The same stimuli were used as in Experiment 2.

Procedure. The procedure for this experiment was similar to the procedure in Experiment 2, but the way that words were chosen for training differed. For the first participant in each chain, a total of 40 stems were selected at random, and based on the starting condition of the chain, the suffix for each stem was selected. For example, for the 50% harmonic starting condition, 40 stems were chosen and of those stems, half were chosen to have the appropriate suffix to make the word harmonic and half were chosen to have the suffix to make the word non-harmonic. For subsequent participants in each chain, 40 words were chosen at random from those words which the previous

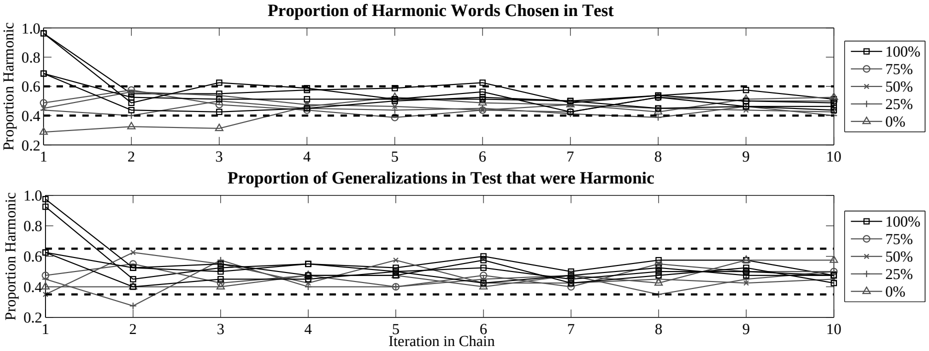


Figure 5. Iterated learning chain results for Experiment 3. Dotted lines show the two-tailed 95% confidence interval for chance responding; confidence intervals differ between the two graphs because there are 40 opportunities to generalize versus 80 opportunities to choose harmonic words.

participant had said was in the language. In order to exclude participants who had not actually learned the language in training, participants were not included in the chain if their accuracy in test on previously seen words was below 62.5%[4]; this is the lowest level of accuracy that is significantly different from chance guessing (binomial test, $p < 0.05$). Chains were started at 100%, 75%, 50%, 25%, and 0% harmonic. One chain with 10 participants was run for each starting point except for 100%. Four chains of 10 participants each were run at the 100% starting point as this is the point of most interest: given a learnability bias, does the percentage of harmonic words in a language remain consistently large?

Results and Discussion

While Experiment 2 showed a learnability bias for the harmonic language over an arbitrarily chosen language, the iterated learning chains in Experiment 3 did not favor the harmonic language. As shown in Figure 5, all chains tended toward languages with approximately 50% harmonic words. Grouping the chains into those starting with 100% harmonic words and those with a different starting point, we compared the number of harmonic words chosen at each generation via a t-test. While initial generations had significantly different numbers of harmonic words chosen

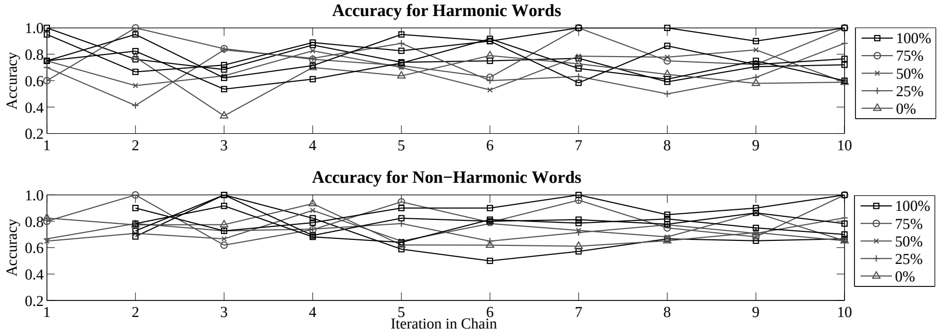


Figure 6. Accuracy on harmonic versus non-harmonic words by iteration in Experiment 3. Overall, there is no difference in accuracy.

based on the chains' starting points, they were not significantly different after 6 generations (all $p > 0.10$). There is also no difference in accuracy on the harmonic items based on generation and chain starting point (t-test with groups as above, all $p > 0.10$), as shown in Figure 6. This is empirical evidence that the pattern of behavior exhibited in the mathematical model reflects the behavior in human subjects: While one language is more accurately transmitted than others across one generation, the large number of possible languages prevent the biased language from predominating after multiple transmissions.

In contrast with the results of our experiment, harmony does exist in many languages of the world. Several factors might result in harmony being more common in these languages than in the final generations of our chains. Our experiments focus on cognitive learning biases, but it is likely that there are also sensorimotor biases that favor the articulation and perception of harmonic languages (Blevins, 2004). There may also be qualitative factors not included in our experiment that lead to the harmony bias being stronger in natural language than in the lab. For instance, children could have stronger harmony biases. Additionally, the quantitative bias towards harmony may be stronger than we found in Experiment 1. This could occur due to the existence of more words and a longer period of learning and use in naturalistic settings. Since all of our participants were adult speakers of English, their bias could also be weaker due to the fact that English is not a vowel-harmonic language. Another factor that could lead to a divergence between our results and natural language learning is the use of a linear transmission structure. In natural language learning, children may learn from people who are part of generations other than the prior generation. Transmission patterns are likely also influenced by other factors, such as language contact. This can result in speakers borrowing phenomena from other languages, resulting in the spread of properties that are unlikely to be generated spontaneously. Finally, there may be increased noise in transmission in lab experiments due to the fact that learning occurs over a relatively short period. This might mean that we would expect harmonic languages to eventually become less prevalent due to transmission errors, but that this process will be much slower in natural language than in the lab. In the experiment, we see that by the third generation the languages no longer contain more harmonic words than would be expected by random chance. This rapid shift may indicate that participants exhibit very little bias towards harmonic words when the input language is not 100% harmonic; in a naturalistic context, generalization is likely to be somewhat more robust due to the longer learning period and broader exposure to the language. Despite these differences, our experiment provides evidence for the fact that a bias need not lead to a universal tendency, something which is born out in the pattern of vowel harmony in existing languages: vowel harmony is relatively common, but it is not present in the majority of languages.

Cases where learnability leads to dominance

We have considered two counter-examples in which a mathematical analysis predicts that a learnability bias will not necessarily lead to a universal, and shown experimental evidence supporting these counter-examples. Yet, there also exist cases where learnability biases and the outcomes of cultural evolution are aligned (e.g., Griffiths, Christian, & Kalish, 2008; Kalish, Griffiths, & Lewandowsky, 2007; Reali & Griffiths, 2009). This leads to the question of how one can determine whether a learnability bias will lead to a universal through a process of cultural evolution. In this section, we use tools from Markov chain theory to identify two cases where learnability biases and long-term outcomes are aligned. We show that these cases require constraints on the quantities empasized in our counter-examples: the relative size of the two sets of hypotheses, and the likelihood of transmission errors from a hypothesis in one set to a hypothesis in the other.

To identify cases where learnability leads to universals, we use the strategy of collapsing the Markov chain into a related chain with one state for all hypotheses in one set and a second state with all hypotheses in the other set. For a transmission matrix with two hypotheses, Griffiths and Kalish (2007) show that if the learnability of h_1 is greater than that of h_2, then h_1 will be the majority hypothesis in the evolved population. Thus, if we could derive a 2×2 transition matrix corresponding to sets of hypotheses, then we could determine under what conditions a learnability bias would lead to a universal. However, as noted previously, we cannot in the general case transform a transition matrix over individual hypotheses into one over sets of hypotheses: the Markov property may not be preserved in the collapsed chain. This transformation can be made, though, when the matrix is *lumpable* with respect to the two sets of hypotheses (Burke & Rosenblatt, 1958; Kemeny & Snell, 1960). A matrix is lumpable with respect to a partition of the states into two sets H_1 and H_2 if the following criterion holds:

$$\forall i, j \in H_i : \qquad \sum_{k \in H_2} q_{ki} = \sum_{k \in H_2} q_{kj} \tag{2}$$

Intuitively, this criterion means that for all states h_i in H_1, the total probability of transitioning from that state to any state in H_2 must be the same. It is met if transmission errors in which the learner acquires a hypothesis in one set are equally likely to occur if the data came from any state in the other set.

Lumpability imposes relatively stringent constraints on the structure of the transition matrix, which we cannot expect to hold in all cases. However, there are several examples of transmission structures where this criterion does hold; we now turn to two special cases in which lumpability does hold and conditions for when learnability will lead to universals can be derived. First, consider a case where all hypotheses in set H_1 have self-transition probability equal to ℓ_1, and all hypotheses in set H_2 have self-transition probability equal to ℓ_2. Letting the probability of transmission errors from a given hypothesis be uniform, this results in a matrix with the following structure:

$$\begin{bmatrix} \ell_1 & & \frac{1-\ell_2}{n_1+n_2-1} \\ \vdots & \ddots & \vdots \\ \frac{1-\ell_1}{n_1+n_2-1} & & \ell_2 \end{bmatrix}, \tag{3}$$

where $n_1 = |H_1|$ and $n_2 = |H_2|$. It is easy to verify that the criterion in Equation 2 is met, so lumpability holds (see Appendix C for details). We can derive the transition matrix over the two sets and find conditions on how learnability relates to a set being dominant. For any single hypothesis in H_1 to be more prevalent than any single hypothesis in H_2, we need only the simple learnability condition that $\ell_1 > \ell_2$. However, hypotheses in H_1 will occur more often than hypotheses in H_2 if and only if

$$\frac{n_1}{n_2} > \frac{1-\ell_1}{1-\ell_2}. \tag{4}$$

If H_1 has more hypotheses than H_2 and hypotheses in H_1 are more learnable than hypotheses in H_2, then this condition will always hold and hypotheses in H_1 will be most prevalent. Otherwise, which set will dominate is dependent on the relative sizes of the sets of hypotheses and the strength of the learnability bias. This is suggestive of the issue with our second counter-example, which demonstrated that differences in the number of languages in each set could result in a learnability bias being overwhelmed. While one cannot create a general criterion for when such a size difference will negate a learnability bias, this special case shows the type of relationship that one should expect among these quantities.

The above example requires strong uniformity conditions. These conditions can be relaxed somewhat. Consider a system in which we again have $q_{ii} = \ell_1$ if $i \in H_1$ and $q_{ii} = \ell_2$ otherwise.

Now, we let the transmission error probabilities vary based on whether the error results in transmission to another hypothesis within the set or to a hypothesis in the other set. This is likely to occur if, for instance, hypotheses in the same set tend to be more similar to one another than hypotheses that are in different sets. Let o_1 be probability of a transmission error from a hypothesis in H_1 to any given hypothesis that is also in H_1. That is, for any $i, j \in H_1$, $q_{ji} = o_1$. Then we similarly define o_2 such that for any $i, j \in H_2$, $q_{ji} = o_2$. We also define the probability of transmission errors from a hypothesis i to any hypothesis j that is not in the same set to be uniform. Letting $n_1 = |H_1|$ and $n_2 = |H_2|$, this gives a transition matrix with the following structure:

$$\begin{bmatrix} \ell_1 & o_1 & & \frac{1-\ell_2-o_2(n_2-1)}{n_1} & \frac{1-\ell_2-o_2(n_2-1)}{n_1} \\ o_1 & \ell_1 & & \frac{1-\ell_2-o_2(n_2-1)}{n_1} & \frac{1-\ell_2-o_2(n_2-1)}{n_1} \\ & & \ddots & & \\ \frac{1-\ell_1-o_1(n_1-1)}{n_2} & \frac{1-\ell_1-o_1(n_1-1)}{n_2} & & \ell_2 & o_2 \\ \frac{1-\ell_1-o_1(n_1-1)}{n_2} & \frac{1-\ell_1-o_1(n_1-1)}{n_2} & & o_2 & \ell_2 \end{bmatrix}. \qquad (5)$$

This matrix is again lumpable, and we can derive conditions on the effects of a learnability bias: H_1 will occur more often than H_2 if and only if $\ell_1 + o_1(n_1 - 1) > \ell_2 + o_2(n_2 - 1)$ (see Appendix C for details, and for when a single hypothesis in H_1 will occur more often than a single hypothesis in H_2). This example is somewhat reminiscent of the first counterexample we presented, which highlighted the fact that one must consider not just how likely it is that a hypothesis will be transmitted accurately, but also how likely transitions are between hypotheses in the two sets.[5]

These two cases provide a positive account of how a researcher interested in linking learnability biases to universals might proceed (a topic we consider further in the General Discussion). If the researcher believes that the assumptions of the first model–consistent learnability within a set of hypotheses, uniform probabilities of moving to other hypotheses–are satisfied, then Equation 4 indicates the strength of a learnability bias that needs to hold in order for learnability to account for an observed universal. Alternatively, the observed strength of a learnability bias can be entered into this inequality to place bounds on the relative sizes of the sets of hypotheses that would have to hold for this explanation to be valid. If the more realistic assumptions of the second model are taken to hold, then the researcher needs to identify not just ℓ_1 and ℓ_2, as was done in most previous arguments from learnability to universals, but also o_1, o_2, and the size of each set of hypotheses. These estimates can then be used to determine whether the observed learnability bias is sufficient to overcome the rate of transitions between sets of hypotheses. Note that researchers creating models of cultural evolution can also check these conditions to determine whether a particular analysis applies to their models. While many cultural evolution models are not explicitly specified in terms of the transmission matrix, the transmission matrix can usually be calculated for a given mathematical model. Overall, the positive cases we have considered point to the fact that researchers must consider not only the strength of a learnability bias, but the relative sizes of the sets of hypotheses and probability of errors in transmission leading to the acquisition of a hypothesis in the same set versus the alternative set.

General Discussion

The learnability of languages and concepts clearly plays a role in their transmission and should be part of explanations of why languages and concepts with particular properties are more prevalent than others. However, greater learnability is not sufficient to explain how a property becomes a universal. Through mathematical analysis and behavioral experiments, we have demonstrated that a learnability bias does not always result in a property becoming prevalent across evolved languages or concepts. While the definition of a learnability bias that we use is relatively strict, looser versions of the criterion will generally result in learnability having less of an effect on the outcome, not more. The two counterexamples we considered both make clear that to determine if a property will become a universal, all of the transition probabilities must be considered.

The first counterexample highlighted the importance of considering the pattern of transition probabilities for what hypothesis is acquired when an error in transmission occurs. For a hypothesis to be widespread after the population has reached equilibrium, it is not sufficient for that hypothesis to be more likely to be transmitted accurately. The hypothesis must also have a non-negligible probability of being generated as the result of errors in transmission. Otherwise, if the hypothesis is ever not transmitted accurately, it will be unlikely to reappear in the population, as demonstrated in Experiment 1. This situation seems particularly relevant to arguments about the origin of religious concepts. Minimally counterintuitive concepts have been found to be more accurately transmitted than other religious concepts, but it might be the case that these concepts are unlikely to be regenerated in a population if they are ever forgotten. This counter-example is also relevant for the evolution and transmission of language. Clicks represent a relatively rare member of phonological inventories around the world, isolated to a handful of languages in sub-Saharan Africa (Miller, 2011). This is a curious distribution when one considers the fact that clicks are the most acoustically salient sounds; that is, languages with clicks have very high self-transition probabilities (Best, McRoberts, & Sithole, 1988; Traill & Vossen, 1997). One may explain this distribution of clicks, however, by noting that clicks are extremely unlikely to be spontaneously introduced into a language as they are not easily confusable with other sounds. Transitions from the set of languages without clicks to the set of languages with clicks are thus highly unlikely.

The second counterexample demonstrated that learnability must be evaluated relative to the number of hypotheses with and without the property of interest. If the set of languages or concepts that lack the property much larger than the set with the property, sheer numbers can dominate the effects of a learnability bias. This problem is likely to be particularly acute in the case of language, as shown in the Experiments 2 and 3: Harmonic languages are a small set of all possible languages, so even a relatively strong learnability bias was not sufficient to allow them to persist across multiple generations of cultural transmission.

While our main focus was to show that greater learnability need not lead to universals, there are cases where differences in learnability lead to one set of hypotheses being more prevalent than another. This has been shown experimentally in iterated learning experiments in which biases in the prior distribution are sufficient to lead to dominance in the stationary distribution (e.g., Griffiths, Christian, & Kalish, 2008; Kalish et al., 2007). We have also illustrated several special cases in which this result can be predicted from the structure of the transmission matrix.

In the remainder of the paper we consider the implications of our results for methods for connecting individual learning to cultural universals, identify some of the limitations of our analysis, and summarize our main conclusions.

Empirically Linking Learning and Universals

Our results demonstrate that showing that one hypothesis is more accurately transmitted than another is not sufficient to explain its prevalence. However, they do not rule out empirically investigating the relationship between individual learning and cultural universals. The problem with focusing on the accuracy of transmission is that it only captures one part of the transition matrix **Q** that characterizes cultural transmission – the diagonal of the matrix. There are at least three ways to get a more complete picture of the content of this matrix by studying individual learning.

The first method is simplest, but perhaps also most expensive in terms of time and the number of experimental participants required. This is to estimate the full **Q** matrix by recording not just whether people were accurate in acquiring each hypothesis, but also which hypotheses they selected when they were incorrect. Using this approach with each hypothesis that people could adopt would make it possible to estimate the set of conditional distributions that make up the matrix **Q**. However, estimating the conditional distributions in this way is likely to be feasible only when the set of hypotheses under consideration is relatively small; otherwise the amount of data that would be required to get an accurate estimate of the conditional distribution would be prohibitive.

A second method is to develop a computational model that characterizes human learning and use that model either to estimate **Q** or to directly link individual learning and the outcome of cultural transmission. In particular, the results of Griffiths and Kalish (2007) make establishing such a link straightforward if human learning is modeled as Bayesian inference, indicating that the equilibrium produced by the linear transmission model is just the prior distribution used in Bayesian inference. Under this approach, a Bayesian model could be used to capture the patterns seen in individual learning in the domain of interest, with the prior being estimated by examining which distribution best seems to capture people's behavior. Knowing the prior would then make it possible to determine which hypotheses would dominate after multiple generations of cultural transmission.

A third method is simply to simulate cultural transmission in the laboratory, and examine what emerges. This provides a way of directly estimating the equilibrium produced by cultural transmission, isolating the effects that individual learning is likely to have in producing universals. This approach has been used to examine how languages (Scott-Phillips & Kirby, 2010) and concepts (Griffiths, Kalish, & Lewandowsky, 2008) change through cultural transmission in the past, and is the most direct way to appeal to cultural transmission as a force that could result in cultural universals. Even simulating a small number of iterations of cultural transmission, as Barrett and Nyhof (2001) did for religious concepts, is potentially more informative than simply looking for differences in learnability.

Finally, as discussed above, the mathematical analyses of the two special cases where learnability biases do result in greater prevalence provide a more sensitive alternative to the existing empirical methods that have been used to try to link these two factors. The first case provides a clearer threshold on the strength that a learnability bias is required to have, and the second makes it clear how examining not just the accuracy with which a hypothesis is transmitted but also the rate at which hypotheses change form can lead to more accurate predictions about the outcome of cultural transmission. While these special cases are somewhat restrictive, Franceschinis and Muntz (1994) showed that the stationary distribution does not change dramatically in cases of "quasi-lumpable" transmission matrices, which violate the lumpability criterion only slightly. These mathematical analyses thus provide new criteria that can be used to make stronger arguments for a relationship between empirical measures of learnability and cultural universals.

Limitations of Analysis

The main limitation of our analysis is the use of the simple linear transmission model, in which each learner learns from one member of the previous generation. The experimental results that we show are dependent upon this assumption. It is easy to imagine variants on this model that make more realistic assumptions about cultural transmission. There is mixed evidence for whether a more complex transmission model would alter our findings. For instance, some research has found that allowing interaction between members of a generation can preserve concepts and properties within a cultural system (Fay, Garrod, & Roberts, 2008; Fay, Garrod, Roberts, & Swoboda, 2010). However, learning from multiple members of the previous generous generation is another potential change to the population structure that tends to dilute the effects of learnability on the languages produced by a population (Burkett & Griffiths, 2010; Smith, 2009).

Further research is necessary to clarify how combining such variations affects the outcome of cultural evolution, but to the extent that such modifications can be mathematically specified, our work provides guidance for drawing conclusions about long-term trends from the results of transmission in a single generation. For instance, in the case of multiple new members of the current generation learning from multiple members of the previous generation, the transition matrix $\mathbf{Q}$ can be specified over the possible combinations of languages in generation n and the possible combinations in the next generation $n+1$. Explicitly enumerating these sets will result in a large increase in the number of possible states, but the resulting matrix can be analyzed to determine if the counter-examples provided here, or the lumpability criterion, apply to the model. Conversely, if behavioral experiments show that learnability tends to lead to long-term prevalence in a particular context, it demonstrates that cultural evolution must be modeled such that the transmission matrix does not fall into one of the counterexamples we have presented.

While many of our analyses considered only the transmission matrix, and thus are applicable to any model where the transmission matrix meets the constraints in the analysis, we also included a Bayesian model in which we assumed that learners acquire a hypothesis by sampling from the posterior distribution over hypotheses given the observed data. Under this sampling assumption, the equilibrium distribution corresponds to learners' prior distribution (Griffiths & Kalish, 2007). However, different models of acquisition, such as maximizing over the posterior distribution, would result in different patterns of transmission. For example, previous work has shown that maximization tends to magnify biases in the prior (Kirby, Dowman, & Griffiths, 2007; Smith & Kirby, 2008), and that the extent to which the language with maximal prior probability is favored is dependent on the amount of noise in transmission (Griffiths & Kalish, 2007). If language acquisition is more accurately characterized as maximizing over the posterior, rather than sampling, then it suggests that the counter-examples we have described will be relatively infrequent in cultural evolution. Whether this is the case has not yet been established. While Smith and Kirby (2008) make a strong argument that a maximization model would be favored evolutionarily, more behavioral experiments should examine the question of whether and in what circumstances language acquisition follows the maximization pattern. Smith and Wonnacott (2010) show some evidence that language evolution can magnify individual learner biases, but because this work only considers a single starting language for establishing the effects of language learning within a generation, the magnification does not necessarily support the hypothesis that learners are maximizing over the posterior distribution; as the authors note, their work demonstrates the difficulty of predicting long term trends based on limited data from a single generation of language evolution. By testing the assumptions of the Bayesian model experimentally and relaxing the constraints of this model in theoretical work, we hope that future research will help to develop a more complete picture of how individual learning occurs and the implications of this process for the outcome of cultural transmission.

In considering limitations, it is also important to note that while we think the situations identified in our counter-examples arise sufficiently often with languages and concepts that we should assert caution in interpreting explanations of universals in terms of learnability, we still expect learnability to play an important role in shaping the languages and concepts that appear across human societies. The results of Griffiths and Kalish (2007) make it clear that factors that influence how easy it will be for people to learn or remember a hypothesis will directly affect whether cultural transmission will favor that hypothesis. Our goal is simply to point out that greater learnability is not in itself sufficient to produce a universal, and consequently not necessarily a complete explanation for why certain properties of concepts and languages are prevalent.

Gaining a Deeper Understanding of Cultural Evolution

The approach we have taken suggests a strategy for gaining a deeper understanding of existing proposals about cultural evolution. For a given model of the cultural evolution process, it is usually possible to calculate the patterns of transmission that occur and specify these patterns as a Q-matrix. For example, many models of cultural evolution involve a two-step process: first, possible variants of a language or utterance are generated, and then, a variant is selected from these possibilities (e.g., Blythe & Croft, 2012; Niyogi, 2006). This mirrors genetic evolution, which involves both random variation (reproduction and mutation) and selection. While the specification of these models is on the surface very different from the Bayesian language learning model that we have discussed, the Q-matrix for these models can still be calculated. Each entry in the matrix represents the probability that the language was generated given the observed data and that once it was generated, it was selected by the learner.

Explicitly calculating the Q-matrix for a given model provides the opportunity to analyze what predictions the model makes about cultural evolution after many generations and to differentiate models from one another. For instance, by checking whether a model meets the lumpability conditions that we have described, one can verify whether that model predicts that learnability biases will result in dominance after many generations. If these conditions are not satisfied, one can turn to establishing whether either of the counterexamples we have described is relevant. Via the *Q*-matrix, one can also explore finer-grained predictions such as whether a larger learnability bias will result in a hypothesis being more prevalent than a smaller learnability bias. In the case of disagreement between models, further experimentation and verification of which outcome occurs in the real-world transmission of concepts and languages can provide support for one model over another. These analyses thus provide a tool for better understanding individual models of cultural evolution and for comparing models to one another based not on immediate processes of evolution but on predictions about the outcome of evolution after many generations.

Conclusion

Our results suggest that the relationship between learnability and cultural universals is more complex than assumed in previous work. This complexity is congruent with the evidence that all languages and cultures do not exhibit all properties for which learnability biases have been found (e.g., as discussed in Evans & Levinson, 2009). Indeed, in historical linguistics, the general principle is one of language divergence, rather than convergence on some universal language (e.g., Greenberg, 1971). Given this relationship, one must rethink using empirical evidence for particular learnability biases to explain why particular cultural or linguistic tendencies occur. Instead, one must either gain a more complete picture of cultural transmission by understanding how hypotheses change when transmitted, or actually simulate multiple transmissions in the lab to establish whether a particular property is actually maintained over many generations. Ultimately, we hope that by identifying a stronger criterion for connecting individual learning to cultural universals we have provided a tool that can be used to definitively understand how human societies relate to the structure of human minds.

References

Barrett, J., & Nyhof, M. (2001). Spreading nonnatural concepts. *Journal of Cognition and Culture*, *1*, 69-100.

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology*. Cambridge: Cambridge University Press.

Best, C., McRoberts, G., & Sithole, N. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 45–60.

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

Blythe, R. A., & Croft, W. (2012). S-curves and the mechanisms of propagation in language change. *Language*, *88*(2), 269–304.

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion*. Berkeley, CA: University of California Press.

Boyer, P. (2001). *Religion explained: The evolutionary origins of religious thought*. New York: Basic Books.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*, 535-564.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), (pp. 291–309). Oxford University Press.

Burke, C. J., & Rosenblatt, M. (1958). A Markovian function of a Markov chain. *The Annals of Mathematical Statistics*, *29*(4), 1112–1122.

Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8).*

Comrie, B. (1981). *Language universals and linguistic typology*. Chicago: University of Chicago Press.

Croft, W. (2002). *Typology and universals*. Cambridge University Press.

Culbertson, J. (to appear). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Linguistics and Language Compass*.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329.

Evans, N., & Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–492.

Fay, N., Garrod, S., & Roberts, L. (2008). The fitness and functionality of culturally evolved communication systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3553–3561.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, *34*(3), 351–386.

Feller, W. (1968). *An introduction to probability theory and its applications*. New York: Wiley.

Finley, S. (2012). Typological asymmetries in round vowel harmony: Support from artificial grammar learning. *Language and Cognitive Processes*.

Finley, S., & Badecker, W. (2007). Towards a substantively biased theory of learning. *Berkeley Linguistics Society*, *33*.

Finley, S., & Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of Memory and Language*, *61*, 423–437.

Franceschinis, G., & Muntz, R. (1994). Bounds for quasi-lumpable Markov chains. *Performance Evaluation*, *20*(1), 223–243.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431–436.

Greenberg, J. (Ed.). (1963). *Universals of language*. Cambridge, MA: MIT Press.

Greenberg, J. (1971). *Language, culture, and communication*. Stanford: Stanford University Press.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science*, *32*, 68-107.

Griffiths, T. L., Kalish, M., & Lewandowsky, S. (2008). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3503.

Griffiths, T. L., & Kalish, M. L. (2007). A Bayesian view of language evolution by iterated learning. *Cognitive Science*, *31*, 441-480.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, *14*, 288-294.

Kemeny, J., & Snell, J. (1960). *Finite markov chains*. Princeton, NJ: van Nostrand.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Dowman, M., & Griffiths, T. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Komarova, N. L., & Nowak, M. A. (2003). Language dynamics in finite populations. *Journal of Theoretical Biology*, *221*, 445-457.

Labov, W. (2001). *Principles of linguistic change. Volume II: Social Factors*. Blackwell.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242.

Miller, A. (2011). The representation of clicks. In E. H. Marc van Oostendorp Colin J. Ewen & K. Rice (Eds.), *The blackwell companion to phonology* (pp. 416–439). Wiley-Blackwell.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology*, *25*(1), 83–127.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328.

Restorff, H. von. (1933). Über die wirkung von bereichsbildungen im spurenfeld. *Psychological Research*, *18*, 299-342.

Schuster, P., & Sigmund, K. (1983). Replicator dynamics. *Journal of Theoretical Biology*, *100*(3), 533 - 538.

Scott-Phillips, T., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

Smith, K., & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1509), 3591–3603.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 1–13.

Tily, H., Frank, M., & Jaeger, T. (2011). The learnability of constructed languages reflects typological patterns. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Traill, A., & Vossen, R. (1997). Sound change in the Khoisan languages: new data on click loss and click replacement. *Journal of African Languages and Linguistics*, *18*, 21–56.

van der Hulst, H., & van de Weijer, J. (1995). Vowel harmony. In J. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 495–534). Blackwell.

Wilson, C. (2003). Experimental investigation of phonological naturalness. *Proceedings of the 22nd West Coast Conference on Formal Linguistics*.

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

Xu, J., & Griffiths, T. L. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, *60*(2), 107–126.

Appendix A

Convergence of Experiment 1 to Stationary Distribution

To bound the probability that the chain in Experiment 1 reached the stationary distribution, we can relate it to a stochastic process with known behavior. We assume that when an item in the list is forgotten, the replacement item is sampled from the prior distribution over words. It is thus a sufficient condition for convergence to the stationary distribution to have forgotten all of the original items in the list. Since each one is resampled from the prior, its new value is independent of the starting state of the chain. This condition is equivalent to the *coupon-collector* problem (Feller, 1968), which bounds the number of cereal boxes necessary to collect all n coupons given that each box contains one coupon and coupons are distributed uniformly across boxes. When the first box is sampled, a new coupon is collected with probability $\frac{n}{n}$. With the next box, the probability of a new coupon falls to $\frac{n-1}{n}$, and similarly, after collecting i coupons, the probability of a new coupon is $\frac{n-i}{n}$. The expected number of boxes needed to find all coupons is thus

$$n\sum_{i=1}^{n}\frac{1}{i} = nH_n, \tag{6}$$

where H_n is the nth harmonic number. To bound the time to reach the stationary distribution in Experiment 1, we first seek to bound the number of replacements necessary to have resampled all 10 items; let the number of replacements necessary be a random variable r. By Equation (6), $E[r] = 10H_{10} = 29.3$. We then need to convert this into an expected number of iterations to convergence by calculating the expected number of items that will be resampled at each iteration, $n\hat{p}$. $\hat{p}$ is the empirical probability of any given item will be resampled; in our data, averaged across chains, this is 0.107. Thus, the expected number of iterations to reach convergence is $\lceil \frac{10H_{10}}{10\hat{p}} \rceil = 28$.

The variance of the coupon collector problem is also known and is equal to $n^2\sum_{i=1}^{n}\frac{1}{i^2} - nH_n$. Substituting $n = 10$, we have standard deviation $\sigma = 11.2$ replacements. The one-tailed 95% confidence interval for reaching the stationary distribution corresponds to $E[r] + 1.65\sigma = 47.8$

replacements, which is equivalent to $\frac{47.8}{n\hat{p}} = 45$ iterations.

Appendix B

Modeling the Evolution of Vowel Harmony Using Iterated Learning

In this appendix, we give further mathematical detail concerning how we calculated the results in Figure 3. This figure shows the bias λ that is necessary to compensate for the fact that harmonic languages are outnumbered by non-harmonic languages as a function of ε. To calculate this value, we need to specify the form of the prior for this model and calculate the number of harmonic languages. We consider hypotheses that are binary vectors of length N, with a one in the ith position if the language includes the vowel harmonic variant of word i and a zero if the language includes the non-harmonic variant. We define a language as harmonic if the proportion of non-harmonic words in the language is less than or equal to some ε. The number of harmonic languages n_H can then be calculated as:

$$n_H = \sum_{k=0}^{\lfloor \varepsilon N \rfloor} \binom{80}{k}, \tag{7}$$

as each harmonic language is a vector of length N with no more than $\lfloor \varepsilon N \rfloor$ zeros.

In the model, we define the prior as uniform within each type: All harmonic languages have the same prior probability as one another, as do all non-harmonic languages. To incorporate the bias, we specify that if the prior on each non-harmonic language $p(h_{\text{non-harmonic}})$ is equal to some constant p, then the prior on each harmonic language $p(h_{\text{harmonic}})$ is equal to λp. We can calculate the value of p using the constraint that the prior probability of all hypotheses must sum to one:

$$p(2^N - n_H) + \lambda p n_H = 1$$

$$p = \frac{1}{2^N - n_H + \lambda n_H}, \tag{8}$$

where there are $2^N - n_H$ non-harmonic languages. Harmonic languages will be more common after many generations if they have more than half of the mass in the equilibrium distribution; in this case, that means they must have more than half of the mass in the prior distribution:

$$\begin{aligned}
\lambda p n_H &> \frac{1}{2} \\
\frac{\lambda n_H}{2^N - n_H + \lambda n_H} &> \frac{1}{2} \\
\lambda n_H &> 2^{N-1} - \frac{n_H}{2} + \frac{\lambda}{2} n_H \\
\lambda &> \frac{2^N}{n_H} - 1. \qquad (9)
\end{aligned}$$

Thus, for Figure 3, we vary ε and calculate the minimum λ necessary for Equation 9 to hold.

Appendix C

Proof of Conditions for When Learnability Leads to Universals

In this section, we expand upon the technical details in the section on cases where learnability leads to dominance. For the 2×2 transition matrix, Griffiths and Kalish (2007) give the equilibrium probability π_1 for hypothesis h_1 as:

$$\pi_1 = \frac{q_{12}}{q_{21} + q_{12}} = \frac{1 - q_{22}}{2 - q_{11} - q_{22}}. \tag{10}$$

From this, we can calculate that π_1 will be greater than or equal to π_2 if and only if $q_{11} \geq q_{22}$:

$$\begin{aligned}
\pi_1 \geq \pi_2 &\Leftrightarrow \frac{1 - q_{22}}{2 - q_{11} - q_{22}} \geq 1 - \frac{1 - q_{22}}{2 - q_{11} - q_{22}} \\
&\Leftrightarrow \frac{1 - q_{22}}{2 - q_{11} - q_{22}} \geq \frac{1 - q_{11}}{2 - q_{11} - q_{22}} \\
&\Leftrightarrow 1 - q_{22} \geq 1 - q_{11} \\
&\Leftrightarrow q_{11} \geq q_{22}, \qquad (11)
\end{aligned}$$

with equality only holding in the case that $q_{11} = q_{22}$. To prove the conditions for the matrices which have many hypotheses, each in one of two sets, we derive the transition matrix over the sets H_1 and H_2 and then determine when the above condition will hold. Kemeny and Snell (1960) show that for a matrix that is lumpable on a partition over two sets H_1 and H_2, the transformed transition matrix has entries $q_{12} = \sum_{k \in H_1} q_{ki}$ where $h_i \in H_2$ and $q_{21} = \sum_{k \in H_2} q_{kj}$ where $h_j \in H_1$. q_{11} and q_{22} are then set such that the columns of the matrix sum to 1.

In the first case we describe, all hypotheses h_i in set H_1 have self-transition probability $q_{ii} = \ell_1$, and all hypotheses h_j in set H_2 have $q_{jj} = \ell_2$. For any given hypothesis h_s, $q_{ts} = q_{rs}$ when $t \neq s$ and $r \neq s$; this encodes the condition that transmission errors are uniform. Let $n_1 = |H_1|$ and $n_2 = |H_2|$. This results in a transmission matrix where each column k has some ℓ_g in the diagonal, corresponding to the set H_g to which h_k belongs, and the remaining entries in the column are uniform.

This matrix is lumpable: For each $h_j \in H_2$, $\sum_{i \in H_1} q_{ij} = \frac{n_1(1-\ell_2)}{n_1+n_2-1}$, and for each $h_i \in H_1$, $\sum_{j \in H_2} q_{ji} = \frac{n_2(1-\ell_1)}{n_1+n_2-1}$. Thus, we can transform the process into a 2×2 transition matrix on H_1 and H_2 with the above sums forming the non-diagonal entries:

$$\begin{bmatrix} \frac{n_1-1+n_2\ell_1}{n_1+n_2-1} & \frac{n_1(1-\ell_2)}{n_1+n_2-1} \\ \frac{n_2(1-\ell_1)}{n_1+n_2-1} & \frac{n_2-1+n_1\ell_2}{n_1+n_2-1} \end{bmatrix}. \tag{12}$$

By Equation 11, we know that H_1 will occur more often than H_2 if and only if $\frac{n_1-1+n_2\ell_1}{n_1+n_2-1} > \frac{n_2-1+m_1\ell_2}{n_1+n_2-1}$. Simplifying, we find:

$$\begin{aligned} \frac{n_1-1+n_2\ell_1}{n_1+n_2-1} > \frac{n_2-1+n_1\ell_2}{n_1+n_2-1} &\Leftrightarrow n_1-1+n_2\ell_1 > n_2-1+n_1\ell_2 \\ &\Leftrightarrow n_1+n_2\ell_1 > n_2+n_1\ell_2 \\ &\Leftrightarrow n_1-n_1\ell_2 > n_2-n_2\ell_1 \\ &\Leftrightarrow n_1(1-\ell_2) > n_2(1-\ell_1) \\ &\Leftrightarrow \frac{n_1}{n_2} > \frac{1-\ell_1}{1-\ell_2} \end{aligned} \tag{13}$$

where the above transformations preserve the direction of the inequality since $n_1+n_2-1>0$ as there must be at least two hypotheses and $1-\ell_2>0$ since by the conditions for stationarity to exist, $\ell_2<1$.

We can also consider when a single hypothesis in H_1 will be more prevalent than a single hypothesis in H_1. We can use Equation 10 to derive the stationary probability of the sets H_1 and H_2:

$$\pi_{H_1} = \frac{n_1(1-\ell_2)}{n_1(1-\ell_2)+n_2(1-\ell_1)} \tag{14}$$

$$\pi_{H_2} = \frac{n_2(1-\ell_1)}{n_1(1-\ell_2)+n_2(1-\ell_1)} \tag{15}$$

By symmetry, we know that all hypotheses within a set will occur equally often at stationarity, so the stationary probability of any $h_i \in H_1$ is $\frac{\pi_{H_1}}{n_1} = \frac{(1-\ell_2)}{n_1(1-\ell_2)+n_2(1-\ell_1)}$. Thus, $h_i \in H_1$ will occur more often than $h_j \in H_2$ if and only if $\ell_1 > \ell_2$; this is the same condition as for the original 2×2 matrix.

For the second example of a matrix where conditions on learnability can be derived, we consider a system in which again $q_{ii} = \ell_1$ if $i \in H_1$ and $q_{ii} = \ell_2$ otherwise. We let o_1 be the probability of a transmission error from a hypothesis in H_1 to any given hypothesis that is also in H_1. That is, for any $i, j \in H_1$, $q_{ji} = o_1$. We similarly define o_2 such that for any $i, j \in H_2$, $q_{ji} = o_2$. We also define the probability of transmission errors from a hypothesis i to any hypothesis j that is not in the same set to be uniform. We define $n_1 = |H_1|$ and $n_2 = |H_2|$. See main text for a graphical depiction of the structure of this matrix.

Again, the lumpability criterion is met, so we can compress this to a 2×2 matrix:

$$\begin{bmatrix} \ell_1 + o_1(n_1 - 1) & 1 - \ell_2 - o_2(n_2 - 1) \\ 1 - \ell_1 - o_1(n_1 - 1) & \ell_2 + o_2(n_1 - 1) \end{bmatrix}. \tag{16}$$

By Equation 11, H_1 will occur more often than H_2 if and only if $\ell_1 + o_1(n_1 - 1) > \ell_2 + o_2(n_2 - 1)$. In the main text, we note that the strong uniformity conditions in the structure of original matrix can be relaxed: the total mass $o_i(n_i - 1)$ need not be spread uniformly across other hypotheses in the same set, and the transmission error probability $1 - \ell_i - o_i(n_i - 1)$ may similarly be placed non-uniformly. This does not change our results since the sum over these quantities remains the same, and to form the compressed transition matrix, we need only the sum.

We can also calculate the condition for when a single hypothesis in H_1 will be more common than a single hypothesis H_2, returning to the assumption that probability $o_i(n_i - 1)$ and $1 - \ell_i - o_i(n_i - 1)$ are uniformly distributed across hypotheses in the set and not in the set, respectively. By Equation 10, the stationary probabilities of H_1 and H_2 are as follows:

$$\pi_{H_1} = \frac{1 - \ell_2 - o_2(n_2 - 1)}{2 - \ell_1 - o_1(n_1 - 1) - \ell_2 - o_2(n_2 - 1)} \tag{17}$$

$$\pi_{H_2} = \frac{1 - \ell_1 - o_1(n_1 - 1)}{2 - \ell_1 - o_1(n_1 - 1) - \ell_2 - o_2(n_2 - 1)}. \tag{18}$$

Then symmetry again gives us that all $h_i \in H_i$ will have the same stationary probability. Thus, the stationary probability of $h_i \in H_1$ will be greater than $h_j \in H_2$ if and only if:

$$\begin{aligned} \frac{\pi_{H_1}}{n_1} > \frac{\pi_{H_2}}{n_2} &\Leftrightarrow \frac{1 - \ell_2 - o_2(n_2 - 1)}{n_1(2 - \ell_1 - o_1(n_1 - 1) - \ell_2 - o_2(n_2 - 1))} > \frac{1 - \ell_1 - o_1(n_1 - 1)}{n_2(2 - \ell_1 - o_1(n_1 - 1) - \ell_2 - o_2(n_2 - 1))} \\ &\Leftrightarrow \frac{1 - \ell_2 - o_2(n_2 - 1)}{n_1} > \frac{1 - \ell_1 - o_1(n_1 - 1)}{n_2}. \end{aligned} \tag{19}$$

This quantity is dependent on the probability of an error in transmission leading to hypothesis in the same set as well as the relative sizes of the two sets of hypotheses.

Author Note

Preliminary results from this work were presented at the 2nd Workshop on Cognitive Modeling and Computational Linguistics (part of *ACL-HLT* 2011). This work was supported by a National Science Foundation Graduate Fellowship and a National Defense Science & Engineering Graduate Fellowship to Anna N. Rafferty, by grant numbers IIS-1018733 and BCS-0704034 from the National Science Foundation to Thomas L. Griffiths, and grant number T32 NS047987 from the National Institutes of Health to Marc Ettlinger.

Footnotes

[1] 57% of participants were monolingual, and 43% were bilingual. None of the languages spoken by bilingual participants (e.g., Mandarin or Spanish) had vowel harmony.

[2] For Experiment 3, participants who had low accuracy ($< 62.5\%$ of previously heard words chosen in test as "from the language") were excluded. Performing the same exclusion in this experiment preserves the results: Mean accuracy of 87% for the harmonic condition versus 73% for the height-frontness rule condition ($t(28) = 2.74, p < .05$; $d = 1.04$), and 77% mean proportion of generalizations following the rule for the harmony condition versus 58% for the height-frontness rule condition ($t(28) = 2.43, p < .05$; $d = 0.92$). This exclusion criterion resulted in removing five participants from each condition.

[3] 53% were bilingual, and 47% were monolingual. As in Experiment 2, none of the languages spoken by bilingual participants exhibited vowel harmony.

[4] 25% of participants in the first experiment and 24.5% of participants in the second experiment failed to meet the accuracy criterion. Failure to meet this criterion was not reliably associated with condition or generation.

[5] Note that while we have stated the uniformity conditions here strongly in order to illustrate the example and because we believe this is likely to be the most relevant to behavioral researchers, the same criterion holds if the non-diagonal entries are not uniform but only total some value o_i for transmission errors from a hypothesis in set H_i to all hypotheses in the same set and $1 - \ell_i - o_i$ for transmission errors from a hypothesis in set H_i to all hypotheses in the other set.