Cultural transmission of grammatical structure: Using a web-based iterated learning paradigm to investigate the emergence of systematicity in syntax

Jacqueline Gutman

University of Rochester

A senior thesis for the Honors B.S. in Brain & Cognitive Sciences

Submitted in May 2011

Advisor: T. Florian Jaeger

Cultural transmission of grammatical structure: Using a web-based iterated learning paradigm to investigate the emergence of systematicity in syntax

Language is a culturally transmitted system, in that the data from which members of a community learn a language is provided by the linguistic behavior of other members of the same community. Each utterance contains information not only about the message intended to be conveyed, but also carries information about the properties of the system that generated that utterance (Christiansen & Chater, 2008; Kirby, Cornish, & Smith, 2008; Maynard Smith & Szathmáry, 1995). Language learners may be equipped with certain innate knowledge and biases, acquired through biological transmission, which constrain the ways they process the culturally transmitted data. Whatever the exact nature of this knowledge—whether in the form of specifically linguistic constraints, domain-general learning biases stemming from cognitive limitations, or some combination of the two —we know very little about what happens after these constraints have been established in the species. If cultural evolution is critical in shaping language change, then these biases differentially interact with pressures exerted by cultural transmission over historical time. If, however, cultural evolution does not play a role, then the effects produced by those biases have remained fundamentally static since the biological evolution of the language capacity. These claims cannot be evaluated by studying the behaviors of language learners in isolation. In order to evaluate the importance of the transmission process in language change, we need a model of cultural transmission.

The Iterated Learning Model (ILM; see Kirby & Hurford, 2002; Smith, Kirby, & Brighton, 2003) has recently emerged as a useful tool for isolating the effects of cultural transmission on this type of system in a controlled experimental setting. The ILM simulates the cultural transmission of language by taking the linguistic output produced by one language user, and feeding it as input to another language learner. This process can be iterated through multiple generations of learners, with the aim of investigating what types of trends begin to emerge. Weak inductive biases can create strong regularization effects over repeated iterations, even in cases where such an effect is not evident from the behavior of the individual learners (Reali & Griffiths, 2009; Kirby, Dowman, & Griffiths, 2007). Thus the ILM provides us with a powerful framework for studying the pressures shaping language over time, and can be used to uncover explanations of linguistic variation which cannot be fully accounted for under a synchronic analysis.

Studies conducted with the ILM have been used to provide support for the idea that cultural transmission can account for the emergence many of the most well-known typological universals and design features found in language. In particular, the ILM has been used to demonstrate the emergence of word order universals (Kirby, 1999), compositionality (Brighton & Kirby, 2001; Kirby, 2007; Kirby et al., 2008; Smith, Brighton, & Kirby, 2003), and recursion (Kirby, 2002). These studies have been used to demonstrate that pressures which derive from the nature of cultural evolution through a transmission bottleneck interact with the innate learning biases of language users to create a gradual shift towards patterns corresponding to observed typological universals (Kalish, Griffiths, & Lewandowsky, 2007; Kirby et. al, 2007 ; Smith, 2002, 2003). Because these are computational models, however, they require the experimenter to make explicit assumptions about the precise nature of the inductive biases that agents will be equipped with.

Motivation for this thesis

The difficulty of implementing a model that makes appropriate assumptions concerning learners' biases can be avoided by extending the iterated learning paradigm to human learners. Iterated artificial language learning (IALL) experiments expose human participants to a miniature artificial language, requiring the participants to produce some output which is then given to later participants in the experiment as their only input. Thus far, few IALL studies have been conducted. The limited work that has been done generally supports the results from computational models: language becomes more compositional in structure over time (Kirby et al., 2008; Cornish, Tamariz, & Kirby, 2009). That is, the words and phrases produced by participants in the later generations of these experiments exhibit a more systematic relationship between their form and the intended meaning than was present in their input.

At present, however, iterative artificial language learning experiments have focused exclusively the emergence of expressivity or semantic compositionality in the lexicon (e.g. Cornish, 2005, 2006; Kirby et al., 2008). Very limited investigation has been made into the role of cultural transmission in the origins of morpho-syntactic structure (but see Smith & Wonnacott, 2010, for an IALL study of plural marking). The aim of this thesis is to investigate whether cultural evolution can induce adaptive change in the structure of a language, replicating earlier findings with computational simulations and extending them to human participants. This is critical for purposes of responding to objections regarding the validity of the ILM and of computational models of language change more generally (e.g. Bickerton, 2007), as well as to investigate the role of cultural transmission where it interacts with the individual learning and domain-general learning biases of humans. Before we describe the specific scope of the experiments presented in this thesis, we provide a brief review of previous research.

Background

Cultural transmission, taken in combination with the innate learning biases of language users, may create functional pressures on the languages themselves to adapt in order to survive the transmission process by accommodating the needs and limitations of its users (Bates & MacWhinney, 1982; Brighton, Smith, & Kirby, 2005; Kirby & Hurford, 2002; Smith, Brighton, & Kirby, 2003). A language that adapts to its users' needs might be expected to change in the direction of increased learnability, expressivity, and structure (Cornish, et al., 2009; Kirby & Hurford, 2002).

Language as a biological adaptation

A strongly biological account of the development of structure in language places the explanatory burden on the initial emergence of the language faculty in humans. In this view, the complex structure of language developed via the only possible explanation for adaptive complexity in a biological system: natural selection (Pinker & Bloom, 1990). The shared ability to transmit some information to another individual, in order to engage in cooperative problem-solving with other members of the community, increases the chances of survival for individuals within that community. These individuals may use language to share the location of a food source, warn others of a potential danger, or make themselves more attractive to a potential mate, and they may accomplish these goals even when the referent itself is out of sight. Just as the impressive complexity of the eye suggests that it evolved by natural selection to fulfill the adaptive function of sight, the impressive complexity of grammar is taken to suggest that it too evolved by natural selection, to fulfill the adaptive function of information-sharing (Pinker & Bloom, 1990; Pinker & Jackendoff, 2005).

The biological adaptation by which humans evolved a capacity for language is typically conceived of here as encompassing a set of specifically linguistic constraints on the sorts of hypotheses that are possible—that is, Universal Grammar—brought to bear upon the task of language acquisition (Pinker & Bloom, 1990). This adaptation is assumed to have emerged in the species long ago and remained unchanged since that time, such that any and all grammatical structure seen in currently existing languages is a consequence of this static underlying system, and not from any pressure on language to be adaptive introduced later in the course of cultural evolution.

Language as a complex adaptive system

In contrast with the view described above, which focuses investigation on the adaptation in humans of a capacity for language, an alternative framework considers the perspective offered by viewing language itself as an adaptive system. In this view, the structure of language can emerge at any point in historical time as a result of pressures on the language to adapt to the needs of its users. Whereas humans without language abilities would be at a disadvantage compared to those with such abilities, languages without users are not merely at a disadvantage—they are extinct. The survival of a language is therefore entirely dependent on its learnability and communicative utility (Mufwene, 2001).

We can consider the goal of any viable communication system to be the faithful transmission of some message from one user of the system to another, without serious degradation of the informational content of that message. In order for this goal to be achieved, the conventions of the system must be shared by its users, and the system must be capable of differentially encoding messages relevant to the users of that system. In the case of human language, these two functions create competing pressures to develop learnability and expressivity in the system (Briscoe, 1998; Christiansen & Chater, 2008; Hoefler, 2006; Kirby et. al 2008).

These two pressures are directly in conflict with one another. A maximally expressive language would employ a unique, unambiguous signal for every possible meaning to be conveyed by the language. The large number of signal-meaning mappings contained in this type of language can be prohibitively costly, taxing the limitations on human systems of memory, processing, perception, and articulation. This system becomes increasingly difficult to learn as the number of distinct signals needed by the system increases. On the other hand, a maximally learnable language would convey meaning as simply as possible, by employing the same signal to encode all messages represented within that language. Learning this type of language would be a trivial task, but the system lacks all expressivity, since every signal is ambiguous—no meanings are differentiated. At best, this language is capable of expressing only that a speaker intends to convey some meaning, but not what that meaning might be. An adaptive language must optimize this trade-off between expressivity and learnability, so that it is learnable enough for each user to arrive at the same understanding of its conventions, and expressive enough to distinguish between the meanings of certain messages where that distinction is relevant (Briscoe, 1998; Christiansen & Chater, 2008; Cornish, 2005).

The role of the bottleneck in transmission

The learnability of a language is determined not only by the cognitive limitations of language users, but by the nature of the input available to learners. A language is learnable only when the data provided is sufficient for learners to acquire the signal for all possible meanings they might need to convey. When a language is needed to represent a greater number of messages than a learner will be exposed to, that language faces the problem of surviving the transmission bottleneck (Hoefler, 2006; Smith, Kirby, & Brighton, 2003). Though each learner is only exposed to a small subset of the signal-meaning mappings contained in the language, they must be able to acquire the signals for meanings they have not been exposed to. If the mappings between signal and meaning are completely random, this transmission bottleneck presents a serious challenge to the communicative viability of such a system; there would be no way to maintain learnability in the face of the bottleneck without greatly compromising the expressivity of the language. If, however, languages can adapt in ways that maximize their chances of survival, then the transmission bottleneck can actually provide a solution to the expressivity-learnability trade-off, by exerting a pressure on language to survive this transmission. Specifically, the bottleneck may force a language to evolve systematicity, such that the signal-meaning mappings are consistent and predictable, in order to allow learners to reproduce its entire internal structure from a subset of that language (Kirby & Christiansen, 2003; Kirby et al., 2008; Perfors, Tenenbaum, & Regier, 2006). With sufficient structure, learners of a language should be able to acquire the signal for meanings they were not exposed to by generalizing the structure seen in their input. When the relationship between form and meaning is systematic, it can be learned given only a limited amount of input. Systematicity, then, is one way that a language might optimize expressivity and learnability.

Biases against unpredictable variation

Previous work with artificial language learning (ALL) has demonstrated that under certain circumstances, when presented with a high level of inconsistencies in their input, both adult and child learners will regularize the grammar of an artificial language so as to reduce the level of unpredictable variation in the grammar—that is, variation which is not strongly conditional on the context in which it occurs (Hudson Kam & Newport, 2005, 2009; Smith & Wonnacott, 2010). When the input contains two or more distinct linguistic patterns, the ALL paradigm can be utilized to explore what patterns participants show a preference for regularizing (Culbertson, 2010; Fedzechkina, Jaeger, and Newport, 2011a, b, c). If two competing forms are equally functional for communication and favored by the learning mechanism, then different speakers may regularize by shifting the grammar in differing directions, with no clear preference across individuals for maintaining one pattern over another. However, if one of the inconsistently used patterns in the input is easier to learn than the other available pattern, we should see an interaction between the regularization bias and other learning biases to favor maintaining one variant over another in the input.

This reasoning has been used to test whether regularization proceeds in favor of typologically attested patterns of linguistic variation (Culbertson, 2010; Culbertson & Smolensky, 2010; Culbertson, Smolensky, & Legendre, 2011). Additionally, competition may occur between grammatical devices, such as case and word order, trading off to maximize functionality (Fedzechkina, Jaeger, and Newport, 2011a, b, c). Where variation persists over time, it may suggest that this stability is functionally motivated by some advantage in processing, acquisition, or suitability for communication.

Scope of the present study

The experiments in this thesis will explore whether unpredictable variation in the case-marking and word order of an artificial language is subject to regularization by adult learners, and whether this regularization proceeds asymmetrically, favoring one variant over another. The iterated learning paradigm we use here is particularly well-suited to examining this question, as a weak inductive bias that may not be evident from the changes introduced by a single learner can have a strong cumulative effect over the course of many generations. Furthermore, we can study these biases in the context of their interaction with functional pressures introduced by the cultural transmission process.

To the best of our knowledge, the studies described here constitute the first attempt to streamline the use of the iterated learning paradigm with human participants. By taking advantage of crowd-sourcing technologies that enable experimenters to access a wide pool of participants and run many chains in parallel, we can address a serious hurdle facing any investigation of the role of the transmission process in language learning. Because a large number of participants are needed to observe converging trends emerging across chains which tend to be highly variable, and because there is a certain slow-down associated with the difficulty of extracting the language produced by one generation in a form that can be passed down to a subsequent generation, running iterated learning studies in the laboratory has historically been prohibitively costly. These practical concerns have limited not only the number of studies that have been conducted within this paradigm, but the type of studies that can be feasibly run as well. Because of the difficulty implementing these types of experiments, there has been some tendency to keep the input languages as simple as possible, often by using only single-word utterances (i.e. Kirby, et al., 2008), and restricting the focus of investigation to emerging compositional structure. In exploring the feasibility of running iterated language learning studies over the web, we hope to address a major gap in the literature and introduce some exciting new possibilities for future study of the language transmission process.

This thesis describes three exploratory studies to examine in greater detail the nature of the changes introduced by learners. In the first two studies, we investigate language change within the course of a single generation to determine whether individual learners exhibit a bias towards increasing systematicity by reducing the amount of non-context-dependent (i.e. unpredictable) variation in the case-marking and word order of a language. In the third study, we explore how individual learning biases interact with transmission pressures by iterating the linguistic input through successive generations of language learners and what shifts in the grammar of an artificial language develop over time. If these languages demonstrate a shift towards increasing systematicity in their grammars, this suggests that cultural transmission can interact with learning biases to exert a strong pressure on languages to develop grammatical structure. More broadly, it would support the notion that language itself is an adaptive system that can change in ways that support its own survival. Increased systematicity may be an adaptive response by language to the problem of surviving the transmission bottleneck, allowing the language to be more easily reproduced from a subset. In the same vein, if the languages produced by the participants demonstrate a shift towards increasing transmissibility, such that a language is reproduced with increasing levels of fidelity by successive generations, it would further support the idea that languages are changing to promote their own survival by becoming increasingly learnable. This would extend previous findings by focusing on the emergence of morpho-syntactic structure with human learners.

To anticipate the big picture result, we encountered a variety of technical and design problems that resulted in high rates of lexical errors among the language learners. This in turn made it hard to interpret the data with regard to the variables of primary interest (word order and case-marker distribution). We do, however, find preliminary results consistent with the hypothesis that language learners exhibit a bias against unmotivated (free) variation, so that two freely varying forms in the input over time become associated with different distributions or that one of the forms dies out.

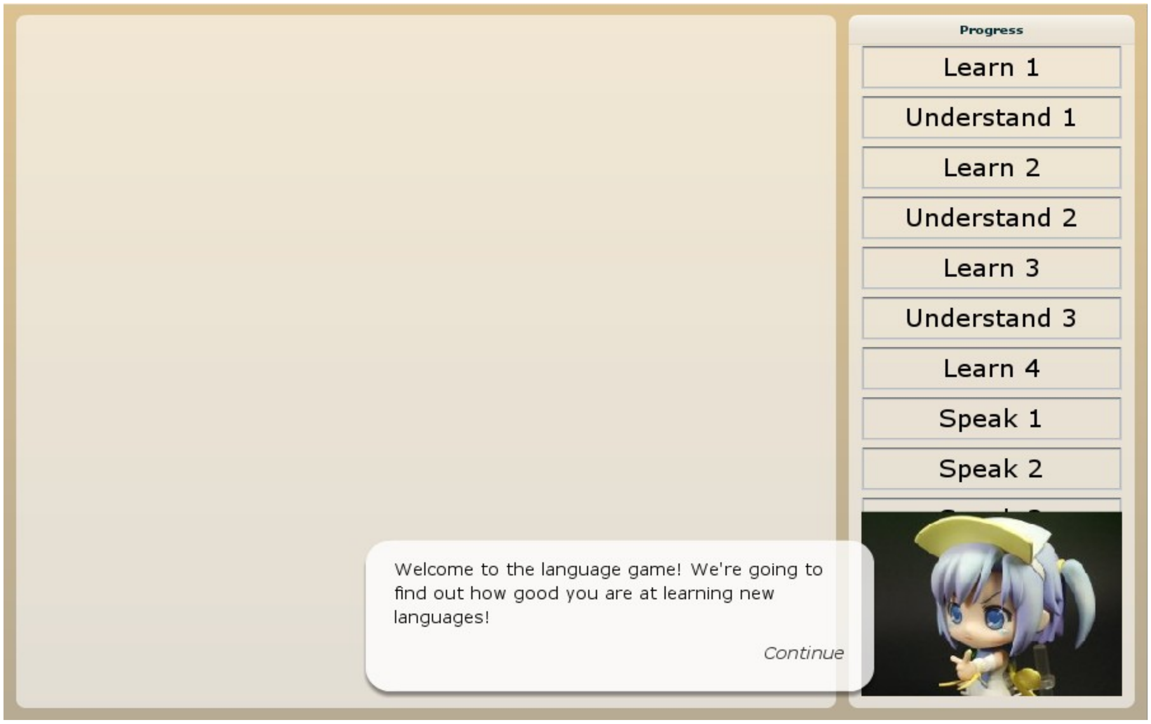
Experiment 1

The primary purpose of Experiment 1 was to explore the feasibility of the paradigm for future research. The critical questions here were as follows: 1) Were participants willing and able to complete the task?, 2) Can participants learn the language well enough in one exposure session to demonstrate a good grasp of the lexicon and some rudimentary processing/acquisition of the structures present in their input?, 3) Is the task sufficiently non-trivial that participants do not always perfectly reproduce their input (thus rendering the task uninformative for the study of language change)?, and 4) Are the changes introduced by participants subtle/gradual enough for the language to develop over time (i.e. is it worth studying learners for more than one generation?) The answers to these questions determine whether our current methodology is well-suited to studying the role of cultural transmission in language change.

Methods

Participants. Twenty participants drawn from Amazon's Mechanical Turk (MTurk) completed the full experiment online. (No information was collected on the number of participants who may have begun and subsequently abandoned the task.) Participants were certified to be at least 18 years old in compliance with the MTurk Participation Agreement. All participants were located in the United States and were self-reported native English speakers. Participants received $2.04 compensation for the experiment, deposited into their Amazon Payments Account within a week of completion of the study.

Procedure. Participants were told they would be participating in a language learning game in which they would receive exposure to, and subsequently be tested on, an alien language. The language game was presented in a Flash applet (developed by Harry J. Tily; see Tily, Frank, & Jaeger, 2011) that allowed participants to view their progress through the task in the sidebar (see Figure 1). Instructions were given to the participants prior to each experimental block by a cartoon character at the bottom of their screen. Participants were encouraged to turn up the sound on their computers and to repeat the sentences out loud to themselves during the experiment.



Welcome to the language game! We're going to find out how good you are at learning new languages!

Figure 1. Screenshot of the Flash applet used in the ex experiment. Participants saw their progress throughout the experiment in the right sidebar, and they received instructions from the cartoon character at the beginning of each block.

There were a total of 120 trials, spanning nine experimental blocks. The order of the blocks is summarized in Figure 2a. Participants saw a label for each block and received new task instructions at the beginning of each one. Blocks were labeled as either Learn (exposure trials), Understand (discrimination trials), or Speak (production trials). Because we were interested in studying the kinds of changes participants made to the language, perfect learning of the input was not necessarily desired. To avoid artificially discouraging drift and innovation in the grammar, participants received no explicit feedback on their performance at any point during or after the sentence discrimination and sentence production blocks. Each block is described in further detail below.

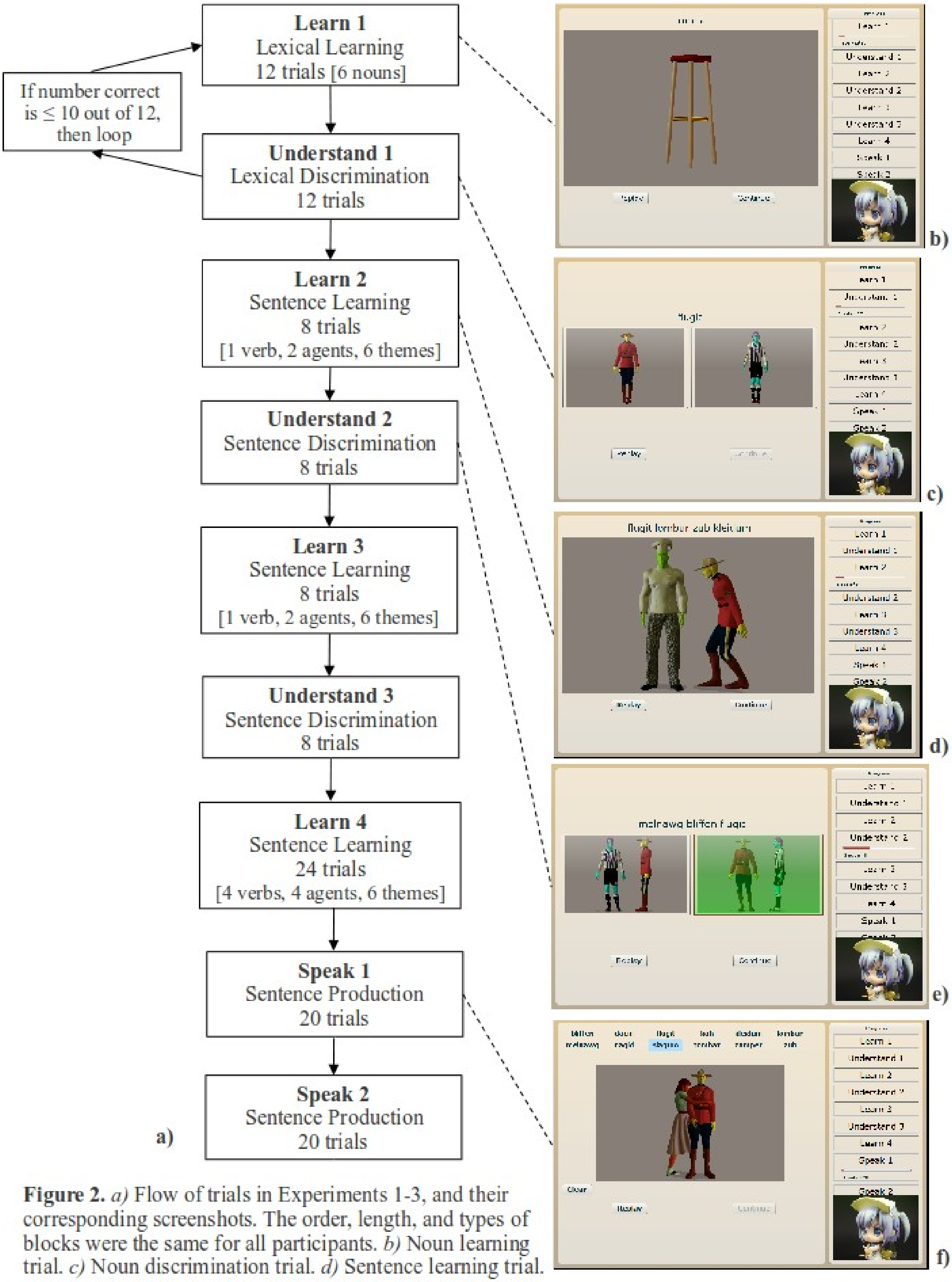


Figure 2. *a)* Flow of trials in Experiments 1-3, and their corresponding screenshots. The order, length, and types of blocks were the same for all participants. *b)* Noun learning trial. *c)* Noun discrimination trial. *d)* Sentence learning trial. *e)* Sentence discrimination trial. *f)* Sentence production trial.

Participants first saw 12 still pictures, accompanied by audio and text of the corresponding word, with each of the six nouns in the language being presented twice (**Learn 1**; see Figure 2b for an example). This learning block was followed by a forced-choice discrimination task testing their comprehension of the nouns they had just learned (**Understand 1**; Figure 2c). Participants heard the words while being shown the target picture and a distractor, and clicked on the picture corresponding to the word they heard. After completing all trials in the discrimination block, participants were given feedback on their performance, and if their performance fell below 11 out of 12 correct responses, they were required to repeat the task until target level of performance was reached. This was done to ensure good learning of the nouns so that variability in the language produced by participants later in the experiment could be attributed to grammatical differences rather than lexical error.

In the following learning block, participants heard 8 sentences in the artificial language while being shown videos of the corresponding action (**Learn 2**; Figure 2d). Participants were given the option of replaying the sentence, both audio and video, as many times as they liked. All 8 sentences in this block featured the same verb, with a single agent referent used in the each of the first four sentences, and a second agent referent used in each of the last four sentences. Thus the number of meaning features that differed from one learning trial to the next was kept to a minimum in order to better facilitate learning. Following the completion of this block, participants were tested on their comprehension of these 8 sentence items, displayed in a different order than during the learning block, using a discrimination task in which two videos were shown side-by-side (**Understand 2**; Figure 2e).

Pairs of target and distractor videos in the sentence discrimination trials always minimally differed in meaning. For all trials featuring an animate theme, the distractor video consisted of the same action as the target sentence, with the patient and agent roles reversed (e.g. *woman headbutting referee* contrasted with *referee headbutting woman*). For all trials featuring an inanimate theme, the distractor video consisted of the same action and agent as the target sentence, with a different noun functioning as the theme. Half of these used the alternate inanimate theme in the distractor video (e.g. *chef punching barstool* contrasted with *chef punching bookstand*), and half used an animate referent in the distractor (e.g. *chef punching barstool* contrasted with *chef punching mountie*). The verb was never relevant in discriminating between the target and distractor videos—the action depicted was identical for all trials.

Participants were then exposed to a new set of 8 sentences (**Learn 3**), presented in the same manner as in the previous sentence learning block, but featuring a novel verb. This was followed by another discrimination block with these 8 items (**Understand 3**). Following the comprehension task, participants were exposed to 24 novel sentences and their corresponding videos in random order (**Learn 4**). These 24 sentences included 16 items that contained two novel verbs, and 8 items that contained the two verbs that the participant had previously been exposed to.

Participants were then exposed to a new set of 8 sentences (**Learn 3**), presented in the same manner as in the previous sentence learning block, but featuring a novel verb. This was followed by another discrimination block with these 8 items (**Understand 3**). Following the comprehension task, participants were exposed to 24 novel sentences and their corresponding videos in random order (**Learn 4**). These 24 sentences included 16 items that contained two novel verbs, and 8 items that contained the two verbs that the participant had previously been exposed to.

In the final two blocks of the experiment, participants were shown videos without sound and instructed to produce sentences in the language by speaking a sentence describing the video, while clicking on these words they used one at a time from a list displayed at the top of their screen (**Speak 1** and **Speak 2**; see Figure 2f). Participants were able to play back the sentence they had just composed, along with the video they were describing, by clicking on the Replay button as many times as desired until they were satisfied with their sentence. Participants were asked to describe a total of 40 videos, in two sets of 20 sentences. About half (18 out of 40) of these videos were novel; that is, they contained a combination of agent, patient, and action to which the participant had not received prior exposure. Exactly half of these videos featured animate themes, and half featured inanimate themes. These 40 videos were chosen to correspond to the 40 videos that would comprise the exposure set provided to a subsequent generation in the iterated versions of this study. They were split into two separate production blocks in order to avoid participant fatigue and break up the monotony of a relatively long block, but there was no critical difference between the two.

As is customary for MTurk tasks, participants were given an opportunity to comment on the task before final submission of their data. The comment box was typically used by participants to convey their enjoyment of the experiment or to report technical glitches that had prevented completion of the task. All data from participants who reported a bug were excluded. Many participants expressed that they had found the task challenging but fun, that they were curious to receive feedback on their performance, and that they would be interested in participating in similar tasks again in the future.

Stimuli. A total of 80 videos were created using SmithMicro Poser animation software, with each video depicting an agent carrying out a transitive action towards another referent. In total the videos featured 4 human characters in the agent position (3 male and 1 female), 6 items in the theme position (the 4 previous characters as well as two inanimate objects), and 4 transitive actions. The 80 videos covered every plausible combination of the agent referents, theme referents, and actions (see *Description of the language* below for more details). Videos averaged approximately 2 seconds in duration. The position of the agent on the left or right side of the theme was counterbalanced across videos. Six still pictures, one for each character or object, were also created in order to introduce participants to each referent in isolation (see Figure 2b and the description of Procedure above).

Sentences were presented to participants aurally over their computer speakers, with the orthographic transcription of the sentence displayed on the screen simultaneously. The audio was prepared by synthesizing speech in a female voice to prepare individual words, that were then concatenated to form sentences. Thus the intonation of a word did not differ depending on its context. A total of 40 sentences were presented, corresponding to the subset of videos shown to each participant.

Description of the language. The language consisted of a set of 40 sentences. Every sentence consisted of a grammatical subject, object, a transitive verb, and a case-marker, although not in that order. Two word orders were permitted by the artificial grammar. Each sentence was presented in either subject-object-verb (SOV) word order, or object-subject-verb (OSV) word order, with 32 of the 40 sentences (80%) occurring as SOV, and 8 of the 40 sentences (20%) occurring as OSV. All sentences were verb-final. Two case-markers (*kah* and *zub*) appeared in the language in free variation, and both were used to mark the object of the action. The case-marker always immediately followed the noun it marked.

The lexicon consisted of 4 transitive verbs, 6 nouns, and the 2 case-markers. These nouns and verbs correspond to the characters, objects, and actions featured in the videos described above. All lexical items accorded with basic phonotactic constraints of English but were not words of English (e.g. *bliffen*). Of the 6 nouns, 3 referred to human male characters, 1 referred to a human female character, and 2 referred to inanimate objects.

Both case-markers occurred equally often in the language. This distribution was true not of the language as a whole, but with respect to every observable feature of the language. That is, both case-markers occurred equally often with each verb, with each noun, and with each word order. Both case-markers also occurred equally often with both animate and inanimate themes, and an equal number of times within each learning block. In short, variation between the two case-markers was maximally random and not contingent upon any feature of the context.

Each lexical item appeared in roughly the same contexts, with the same frequency, as other items in that class; there were no lexical biases in the input. Nouns with animate referents appeared in the both subject and object position, although they were twice as likely to occur as subjects than as objects. Each of these nouns appeared 10 times (out of 40 sentences) in the subject position, and 5 times in the object position. Nouns with inanimate referents appeared in the object position only. Each of these nouns appeared 10 times in the object position. Overall, sentences were equally likely to refer to an animate theme as an inanimate theme. Verbs did not occur with equal frequency. Two of the four verbs appeared in 12 sentences each, while the other two verbs appeared in 8 sentences each.

Coding

A total of 800 utterances produced by 20 participants were annotated by hand for word order, case-marking, and lexical error. A lexical error was defined as omitting the word referring to the action, agent, or theme depicted in the video, or substituting an incorrect word in place of the correct word. Lexical errors were then subdivided into errors in the verb and errors in one or both of the nouns.

We initially defined an instance of case-marking as any production of *zub* or *kah*. However, because this data primarily served as pilot data, we also attempted a more qualitative, impression-based analysis of participants' productions. Here we considered the set of utterances produced by a participant in its entirety, rather than annotating each sentence in isolation. Under this analysis, we occasionally revised annotation of case-marking when a repeated pattern of use suggested another interpretation. For example, a close look at one participant's output strongly indicated that they were using the word *zub* to refer to the action of *hug*, and so this was not counted as an instance of a case-marker. Case error was defined as the (apparent) use of a word other than *zub* or *kah* to mark case, or as the attachment of a case-marker to a word other than the object of the sentence (i.e. the subject or verb). We also included the omission of a case-marker as a third type of case error.

Word order was initially annotated only for sentences which contained no lexical errors (i.e. they contained all of the correct words to refer to the agent, theme, and action of the sentence, in some order), or which contained only a lexical error on the verb. However, we revised the annotation of word order in cases where a closer look at the data revealed a more intuitive description of the participant's output than what had been automatically annotated. For sentences containing lexical errors, the difficulty in determining what exactly the words that had been produced were intended to refer to made any annotation of word order unreliable.

Finally, a case-marker preference rating was calculated for each participant by subtracting the number of occurrences of *zub* from the number of occurrences of *kah*, and dividing by the overall number of case-markers present in that participant's output. Thus a magnitude of 1.0 of case-marker preference corresponds to exclusive use of either the *kah* or *zub* case-marker, and a rating of 0.0 corresponds to no difference in the frequency of occurrence of the two case-markers.

Results

We assess the feasibility of the paradigm by looking at results in four broad areas: lexical error, case error, regularization of word order, and regularization of case-marking. Table 1 provides a summary of the characteristics of the language produced by participants in comparison to their input.

Lexical error. Lexical errors in the verb were the most common error produced. 214 sentences (26.8% of all trials) contained an error in the verb used in the sentence. 84 sentences (10.5%) contained an error in one or both of the nouns used in the sentence. In total, 296 out of 800 sentences (37%) contained at least one lexical error, in naming one or more of the agent, theme, or action depicted in the video.

Case error. There were 75 errors (9.45% of all trials) involving the substitution of a word other than *zub* or *kah* in place of the case-marker. In all of these instances, participants used the word *daf* as a case-marker rather than a verb (*daf* was one of only three monosyllabic words used in the experiment aside from the two case-markers, so to avoid this type of lexical error in Experiments 2 and 3 we made all nouns and verbs equal length). An additional 59 sentences (7.4%) contained no case-marker at all. Out of the 738 sentences containing some type of case-marker, 89 (12.1%) marked case on the subject rather than the object. In total, 220 out of 800 sentences (27.5%) contained at least one case error.

Word order. There were 719 sentences, or 90% of the annotated trials, featuring a subject-object-verb word order. 77 sentences, or 10% of the annotated trials, featured an object-subject-verb word order. Twelve of the 20 participants used the less frequent word order (OSV) in fewer than 5% of sentences they produced. Five of the 20 participants came close to probability-matching their input, using the less frequent word order with similar frequency as they were exposed to (i.e. between 15 and 25% of sentences they produced). Two of the 20 participants used the less frequent word order in 30% of the sentences produced, slightly more often than seen in their input, but still much less frequently than they used the dominant SOV word order.

	% of productions	**% in input language**		**% of production**	**% in input language**
Kah case-marker	42.5	50.0	Verb error	26.7	
Zub case-marker	40.4	50.0	Noun error	10.5	
No case-marker	11.1	0.0	Any lexical error	31.8	
Case-marker on the object	81.1	100.0	Subject-before-object word order	89.9	80.0
Case-marker on the subject	11.1	0.0	Object-before-subject word order	9.6	20.0

Table 1. Summary of sentences produced in Exp. 1, compared to the input.

Case-marker distribution. Twelve of the 20 participants included either *kah* or *zub* in at least 90% of the sentences they produced, with another 5 participants including one of the two case-markers in 70%

to 80% of the sentences produced. Participants' preferences for *kah* versus *zub* varied greatly, with some participants producing one case-marker or the other exclusively, and other participants approximately matching the 50/50 distribution of *kah* and *zub* in their input. The distribution of case-marker preference for the 20 participants is shown in Figure 3. Overall usage for *kah* versus *zub* was roughly equal, with *kah* used marginally more often than *zub* (*kah*: 340 usages, *zub*: 323 usages).

Discussion

Experiment 1 provided an important test of the feasibility of the MTurk platform for conducting these types of studies. Rates of lexical error were slightly higher than typical for artificial language learning studies run in the laboratory, but were within an acceptable range given the limited amount of exposure to the language participants had received. Case error was also within an acceptable range, although the high rate of error involving the word *daf* motivated our decision to control for word length and stress pattern in subsequent experiments.

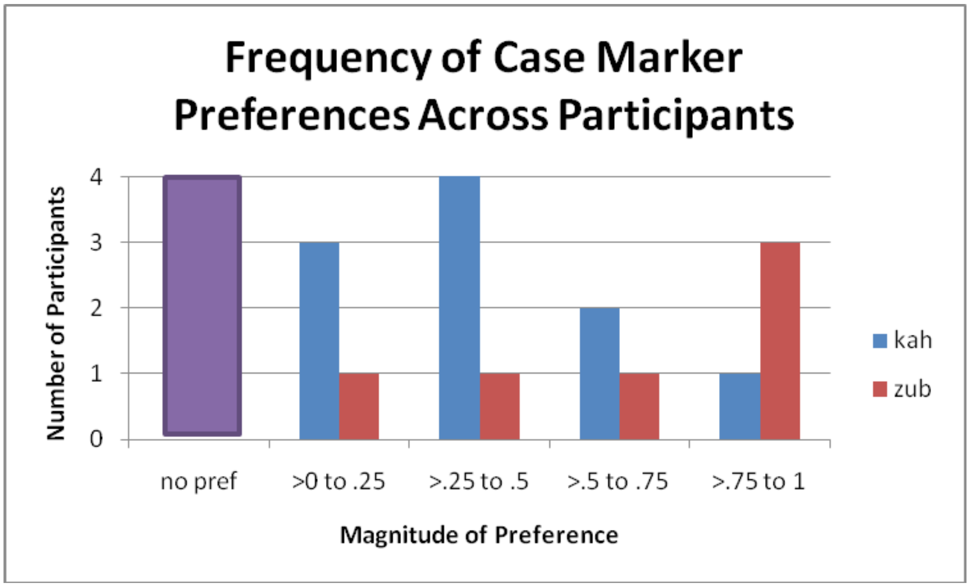


Figure 3. The number of participants displaying no preference, weak preference, moderate preference, and strong preference for either the *kah* or *zub* case-markers in Exp. 1. Overall, magnitude of case-marking preference was fairly evenly distributed, with a similar number of participants falling into each range. There were slightly more participants favoring *kah* (n=10) than participants favoring *zub* (n=6).

Participants were generally able to produce both case-markers in variation, with individual participants varying in the strength of their preference for one case-marker over another. Crucially, not all participants probability-matched the distribution of case-markers seen in their input. This suggests that it should at least be possible to use this platform to investigate how learners' inductive biases shape the production of case-markers in a language during cultural transmission.

The regularization bias for word order, however, proved more problematic. The majority of participants used the dominant word order nearly exclusively. The strength of this bias makes it difficult to address several questions of primary interest here, including whether variation can persist over time when it is functionally motivated, whether learners are able to reduce unpredictable variation in a language while maintaining or introducing variation that is conditioned upon some aspect of the context, and whether there is a complex trade-off between the regularization of case and word order in a language. In short, the overwhelming tendency to regularize the dominant word order when the input language features such a strong SOV bias obscures a priori any potentially interesting results we might otherwise find concerning the answers to these questions.

Experiment 2

Based on the pilot results from Experiment 1, Experiment 2 weakened the bias towards SOV sentences in order to allow for the possibility of probability-matching in learning the artificial grammar. Previous experiments with an 80% to 20% word order bias have demonstrated a nearly universal tendency to regularize the grammar, regardless of participant age, complexity of the grammar, or other factors known to affect regularization (Hudson Kam & Newport, 2005, 2009; Fedzechkina, Jaeger, & Newport, 2011a, b), and our results from Experiment 1 followed that trend. Because the strength of this regularization bias did not allow for much room to investigate how variation in word order develops over time, we conducted a slightly modified variation of the previous experiment this time using SOV as the dominant word order in 63% of all sentences, rather than 80%.

Methods

Participants. Twenty-two participants from MTurk completed the task. Two of these participants had previously participated in Experiment 1, and their data was excluded. No data was recorded for a third participant, either because of a technical error or early termination by the participant. In total, the data from 19 participants were included in the analysis. Participant eligibility and payment remained the same as in Experiment 1.

Description of the language. The language consisted of 40 sentences nearly identical to the sentences used in Experiment 1, with one major change. The ratio of SOV to OSV sentences in the language was altered so as to be less strongly biasing: the number of SOV sentences in the language was reduced from 32 to 25 (63%), while the number of OSV sentences was increased from 8 to 15 (37%). As in Experiment 1, this ratio was the same across the entire language, as well as for each lexical item, each case-marker, and for animate and inanimate referents. In all other respects, the language used in this experiment was the same as in Experiment 1. The details of this language are summarized in Table 2.

Distribution of sentences in the exposure set of a sample language

By action:		**By word order:**	
headbutt	12 (30 %)	SOV	25 (62.5 %)
hug	12 (30 %)	OSV	15 (37.5 %)
punch	8 (20 %)		
knockover	8 (20 %)	**By case-marker:**	
		kah	20 (50 %)
By agent:		zub	20 (50 %)
mountie	10 (25 %)	no case-marker	0 (0 %)
ref	10 (25 %)		

Distribution of sentences in the exposure set of a sample language

chef	10 (25 %)	**By theme animacy:**	
50swoman	10 (25 %)	male-animate	15 (37.5 %)
		female-animate	5 (12.5 %)
By theme:		inanimate	20 (50 %)
mountie	5 (12.5 %)		
ref	5 (12.5 %)	**By agent position:**	
chef	5 (12.5 %)	left of theme	~ 20 (50 %)
50swoman	5 (12.5 %)	right of theme	~ 20 (50 %)
barstool	10 (25 %)		
bookstand	10 (25 %)		

Table 2. Distribution of sentences in a sample language for Experiments 2-3. These distributions correspond to what a participant in Generation 1 might be exposed to. The distribution was the same for the initial generation of every chain, with the exception of which two verbs were the most frequent (this was counterbalanced across chains).

Stimuli. Videos and pictures were nearly the same as in Experiment 1, except that the position of the agent on the left or right side of the theme was now counterbalanced across videos, so that half the videos were mirror images of their counterpart in Experiment 1. Three of the 12 words were replaced so that all words (except for the two mono-syllabic case-markers, *zub* and *kah*) were as similar to each other as possible in length and stress pattern, i.e. disyllabic and trochaic, while still obeying basic English phonotactics. The three new words were generated by the same speech synthesizer used to create the audio for in Experiment 1. In all other respects, the audio and video used in this experiment were unchanged from the materials used in Experiment 1.

Procedure. The procedure was identical to the procedure described in Experiment 1.

Coding

A total of 760 utterances were annotated by hand for word order, case-marking, and lexical error. The same basic definitions and measures described in Experiment 1 were used.

Results

We look at results in four broad areas: lexical error, case error, regularization of word order, and regularization of case-marking. Table 3 provides a summary of the characteristics of the language produced by participants in comparison to their input.

Lexical errors. There were 158 sentences (20.8% of all trials) containing an error in the verb used in the sentence. There were 150 sentences (19.7%) containing an error in one or both of the nouns used in the sentence. In total, 250 out of 760 sentences (32.9%) contained at least one lexical error, in naming one or more of the agent, theme, or action depicted in the video.

	% of productions	**% in input language**		**% of production**	**% in input language**
Kah case-marker	50.1	50.0	Verb error	20.8	
Zub case-marker	45.2	50.0	Noun error	19.7	
No case-marker	4.5	0.0	Any lexical error	32.8	
Case-marker on the object	67.0	100.0	Subject-before-object word order	85.1	62.5
Case-marker on the subject	28.7	0.0	Object-before-subject word order	13.6	37.5

Table 3. Summary of sentences produced in Exp. 2, compared to the input.

Case error. There were no errors involving the substitution of a word other than *zub* or *kah* in place of the case-marker. 34 sentences (4.5%) contained no case-marker at all. Out of the 738 sentences containing some type of case-marker, 218 (29.5%) marked case on the subject rather than the object. In total, 254 out of 760 sentences (33.4%) contained at least one case error.

Word order. There were 647 sentences, or 86% of the annotated trials, featuring a subject-object-verb word order, while 103 sentences, or 14% of the annotated trials, featured an object-subject-verb word order. This ratio of SOV to OSV sentences in the participants' output was nearly the same as the ratio seen in the output of Experiment 1, despite the altered word order ratio of the input (Experiment 1: 80% SOV, 20% OSV; Experiment 2: 63% SOV, 37% OSV). Of the 19 participants, 13 regularized the word order, increasing the ratio of SOV to OSV sentences in the language, while 4 participants appeared to probability-match, maintaining the weak SOV bias present in the input language. Two of the 19 participants produced SOV and OSV sentences with roughly equal frequency, eliminating the weak SOV bias present in their input.

Case-marker distribution. Nearly all participants included explicit case-markers in every sentence produced, with only one of the 19 participants omitting both *kah* and *zub* more than once. Participants' preferences for *kah* versus *zub* varied greatly, with some participants producing one case-marker or the other exclusively, and other participants approximately probability-matching the 50/50 distribution of *kah* and *zub* present in their input. The distribution of case-marker preference for the 19 participants is shown in Figure 4. Overall usage for *kah* versus *zub* was roughly equal, with *kah* used marginally more often than *zub* (*kah*: 381 usages, *zub*: 344 usages).

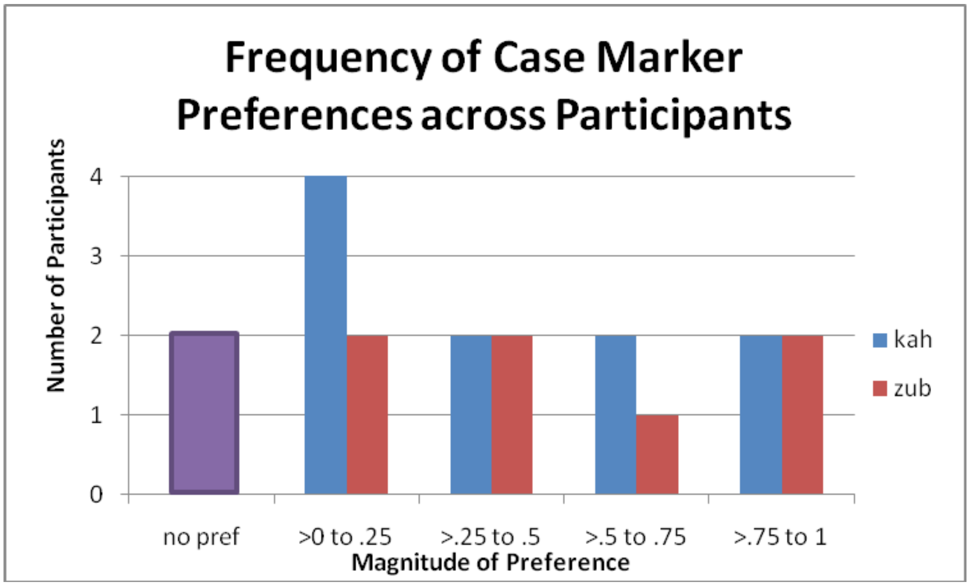


Figure 4. The number of participants displaying no preference, weak preference, moderate preference, and strong preference for either the *kah* or *zub* case-markers in Exp. 2. Overall, magnitude of case-marking preference was fairly evenly distributed, with a similar number of participants falling into each range. There were slightly more participants favoring *kah* (n=10) than participants favoring *zub* (n=7).

Across all participants, the probability of *kah* versus *zub* did not differ greatly depending on the animacy of the theme (*kah* with animate theme: 196 occurrences, *kah* with inanimate theme: 185 occurrences; *zub* with animate theme: 168 occurrences, *zub* with inanimate theme: 176 occurrences). However, a few participants did demonstrate some difference in the conditional probability of each case-marker depending on the animacy of the theme (see Figure 5). One participant in particular used *kah* exclusively to mark animate themes, and used *zub* exclusively to mark inanimate themes. The presence of some learners who—within one generation—strongly condition case-marker use on theme animacy suggests that such regularization could accumulate over generations, which we investigated in Experiment 3.

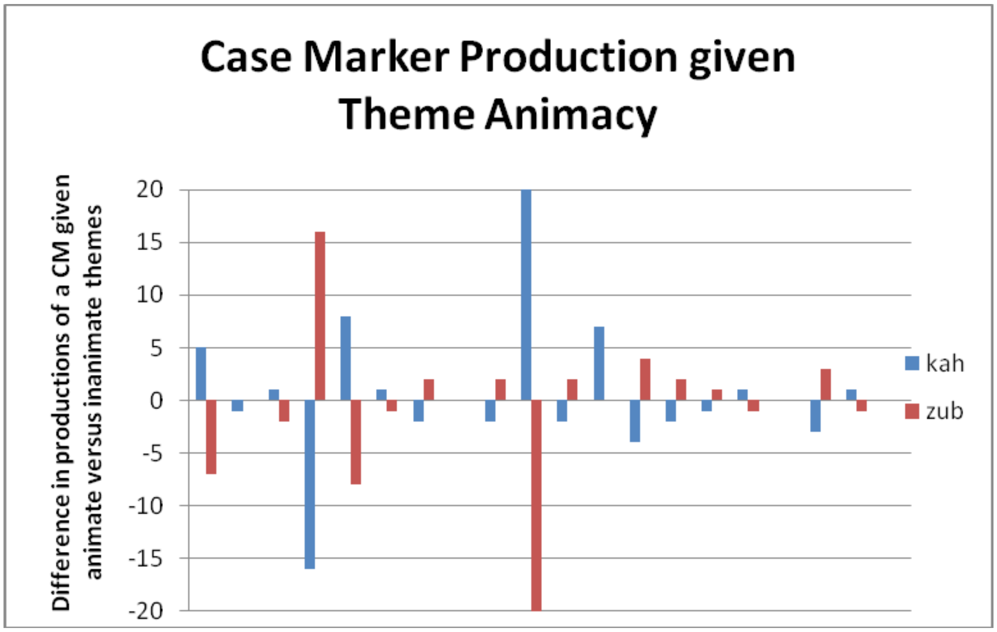


Figure 5. For each participant, we compared the frequency of producing a case-marker when the theme was animate to the frequency when the theme was inanimate. A positive difference indicates that the case-marker was produced more often with animate themes, and a negative difference indicates that the case-marker was produced more often with inanimate themes. Although most participants did not differentiate their case-marking depending on the animacy of the theme, a small number of participants demonstrated a strong or even perfect contingency between the case-marking and theme animacy.

Experiment 3

Experiment 3 expanded upon Experiments 1 and 2 by iterating the linguistic output from one generation of participants to the next providing our first test of the role of cultural transmission in effecting language change. This allows us to investigate the kinds of structure that may emerge in a language over time, and to evaluate the hypothesis that during the course of cultural transmission, there will be a general trend towards reducing the amount of random (unconditional) variation present in the word order and case-marking system of a language.

Methods

Participants. 41 participants from MTurk participated in the experiment, with 40 participants completing the task. Participant eligibility and payment were the same as in the first two experiments, and no worker who had participated in Experiment 1 or 2 was permitted to participate in Experiment 3.

Procedure. The procedure for each participant was absolutely identical to the procedure described in Experiment 2.

Stimuli. The videos and pictures used in this experiment were the same as those used in Experiment 2. The audio was created by concatenating the same synthesized speech clips used in Experiment 2.

Diffusion chain design. Participants were divided into four chains of ten generations. The language for the first generation of the first chain was exactly the same as the language used in Experiment 2. The language for the first generation of all other chains was generated by rotating the lexical items so that one verb replaced another verb, one animate replaced another animate, one inanimate replaced another inanimate, and one case-marker replaced the other case-marker. In the first generations of all chains, the language was balanced in the distribution of the two case-markers (i.e. their distribution was equal in frequency and not predictable from context). All chains were started with a 63% to 37% SOV to OSV word order bias. In all other respects, too, the input language for the first generation of each chain was the same as the input language used in Experiment 2 (see Table 2 for more details).

For all subsequent generations, the 40 sentences comprising their input language were taken from the 40 sentences output during the production blocks of the previous generation. We did not alter or filter this output in between generations in any way.

Coding

A total of 1600 utterances were automatically annotated for word order, case-marking, and lexical error by computer. Two measures of lexical error were used. The first measure defined a lexical error as a deviation from the word-meaning mappings presented in the initial instantiation of the language (i.e. in Generation 0). However, this includes changes to the word-meaning mappings introduced by later generations, and then reproduced correctly by their descendants. Therefore, this measure runs the risk of greatly overestimating the lexical error introduced by each generation. We calculated a second measure of lexical error which defined an error as a deviation from the words used in the corresponding sentence produced by the previous generation, excluding case-markers. (For example, producing *tombat bliffen lombur* when the previous generation produced *nagid bliffen lombur* would count as a lexical error, but producing *slagum zub flugit dacin* when the previous generation produced *flugit kah slagum dacin* would not be considered a lexical error under this measure.)

The number of case-markers in a sentence was calculated from the number of *kah*s produced plus the number of *zub*s produced. However, specific case-markers may also become lexicalized, so that they are more likely to co-occur with certain nouns. We therefore calculated the conditional probability of producing any case-marker, as well as the conditional probabilities of producing a particular case-marker, for each action, for each agent referent, and for each theme referent.

Explicit case-marking may be more critical in sentences containing animate themes, since two thematic roles, either agent or patient, are plausible for each character referred to in the description of the video. The conditional probabilities of producing any case-marker or a particular case-marker were therefore calculated given an inanimate theme, a male theme, and a female theme. These probabilities were then used to calculate the conditional entropies of case-marking over the action, over the theme referent, over the agent referent, and over theme animacy.

Word order was annotated only for sentences which contained both the correct name for the agent referent and the correct name for the theme referent. The word order was then defined by the relative position of those two words, and the third word, if present, was assumed to be the verb. Any occurrences of *kah* and *zub* were ignored in the annotation of word order.

Results and Discussion

First, we examine how well participants learned the language by analyzing their lexical errors in production. We then describe the changes in structures introduced by analyzing word order, presence of case-marking, and case-marker preference, and how these structures change over ten generations. Finally, we test whether these changes are predictable given any feature of the context by analyzing the conditional entropy in the case-marking systems and whether that entropy decreases with time.

Lexical errors. The total number of lexical errors contained in each sentence was recorded, and the results are shown in Figure 6. The percentage of trials containing at least one lexical error increased from 23% of sentence productions in Generation 1 (mean number of errors per sentence: 0.238, SD: 0.196), to 67% of sentence productions in Generation 10 when error was calculated by comparing to Generation 9 (mean number of errors per sentence: 0.954, SD: 0.751). In comparison to the word-meaning mappings present in the initial input language, 74% of sentence productions in generation 10 contained one or more of these type of errors (mean number of errors per sentence: 1.28, SD: 1.05), although by comparing Generation 10 to Generation 0, we overestimate the occurrence of error by conflating the error of one participant in a chain with the perfect reproduction of that participant's output by subsequent learners.

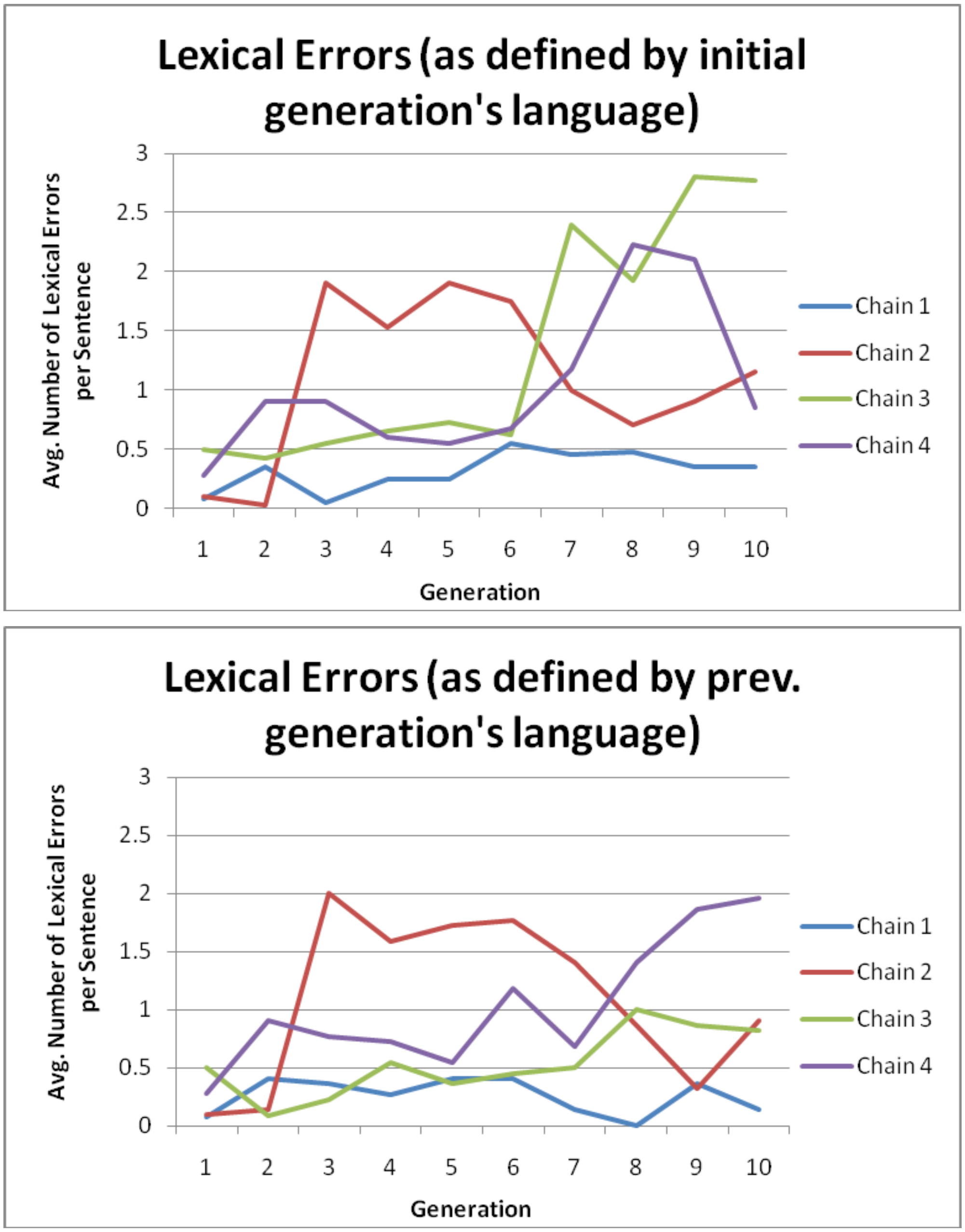


Figure 6. Average number of lexical errors by a produced in each sentence, by each speaker in a chain. a) Lexical errors where an error is defined as a departure from the initial word-meaning mappings (i.e. Generation 0). b) Lexical errors where an error is defined as a departure from the word-meaning mappings of the previous generation (i.e. Generation *k-1*). Rate of lexical error tends to increase over generational time, but error rates vary greatly between chains.

Word order. The number of instances a word order in each generation is shown in Figure 7. The total number of sentences that could be reliably annotated for word order information decreased with each generation, but even with the limited amount of data, we see a slight tendency to fix word order by eliminating the OSV order in favor of the dominant SOV word order. Most of this regularization of word order is done within the first generation. There is some indication that an English-like, SVO word order begins to emerge in the later generations, but this may actually be a mistaken annotation of word order due to the high degree of lexical error present in these later generations.

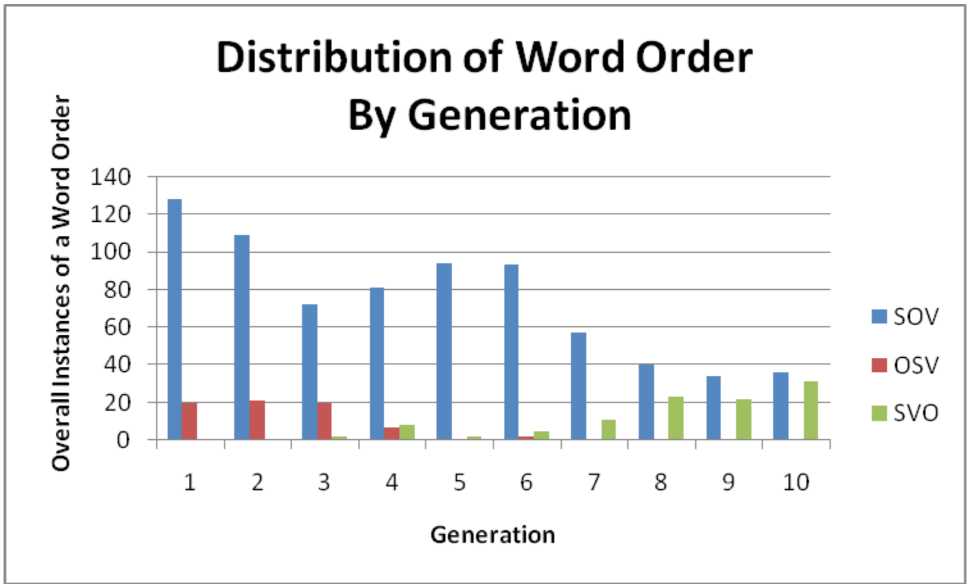


Figure 7. Total number of productions of a particular word order in each generation. The data with word order information available is extremely limited, especially in the later generations of Exp. 3, making it difficult to analyze trends in the word order over time.

Presence of case-marking. The average number of case-markers produced per utterance was recorded for each generation (see Figure 8). It is likely that some participants used *kah* and *zub* for reasons other than case-marking; for example, several participants in Chain 4 appear to have used *zub* as a kind of agreement marker on the verb. However, the total frequency of these two words is the best estimate of case-marking we have available. Total case-marking decreased greatly between the initial and final generations of three of the four chains observed (generation 1: M=1.0, SD=0.0; generation 10: M=0.38, SD=0.5). Only Chain 1 showed stability with respect to case-marking. This general trend of reducing overall case-marking with each generation was also true of each type of case-marker present in the input (*kah* generation 1: M=.68, SD=.47; generation 10: M=.22, SD=.47; *zub* generation 1: M=.32, SD=.47; generation 10: M=.16, SD=.41). The percentage of sentences containing either *kah* or *zub* for each generation is shown in Figure 9.

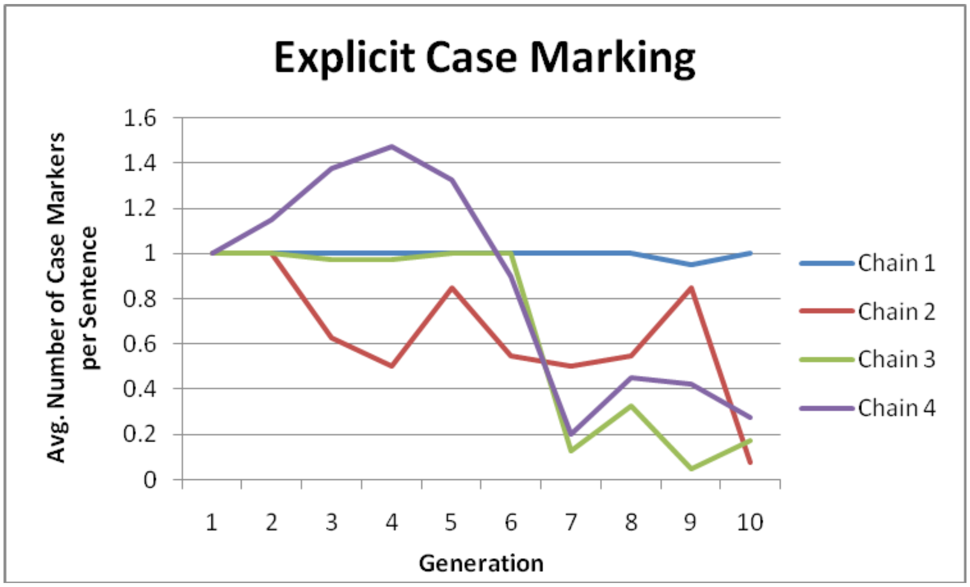


Figure 8. Average number of case-markers produced per utterance by generation. While Chain 1 remained completely stable, maintaining case-marking at a constant rate of 1 marker per sentence, Chains 2, 3, and 4 all demonstrated a partial decline or total abandonment of explicit case-marking by the tenth generation.

Case-marker preference. We next examined case-marker preferences, using the measure described for Experiment 1 above. It is important to keep in mind that a high preference for a particular case-marker tells us very little if the participants used case-markers rarely, as was the case in some of the later generations. Three of the four chains demonstrated an overall preference for *kah,* with the third chain demonstrating the reverse preference. The magnitude of these case-marker preferences for each generation is shown in Figure 10.

There was a great deal of variability in the direction and magnitude of these preferences across chains, and no clear trend to strengthen the magnitude of a preference for a *particular* case-marker is evident here: of the 40 participants observed, 23 increased (or maintained, if the preference was already at ceiling) the magnitude of a preference for a particular case-marker present in their input, while 17 participants reduced the magnitude of that preference. As can be seen in Figure 10, there is a general trend to have a dominant case-marker: except for Chain 1, all chains sooner or later exhibit a preference to mostly rely on one case-marker. This may point to a regularization of the case-marking system over generations (e.g. in terms of a reduction of the conditional entropy), which we investigated next.

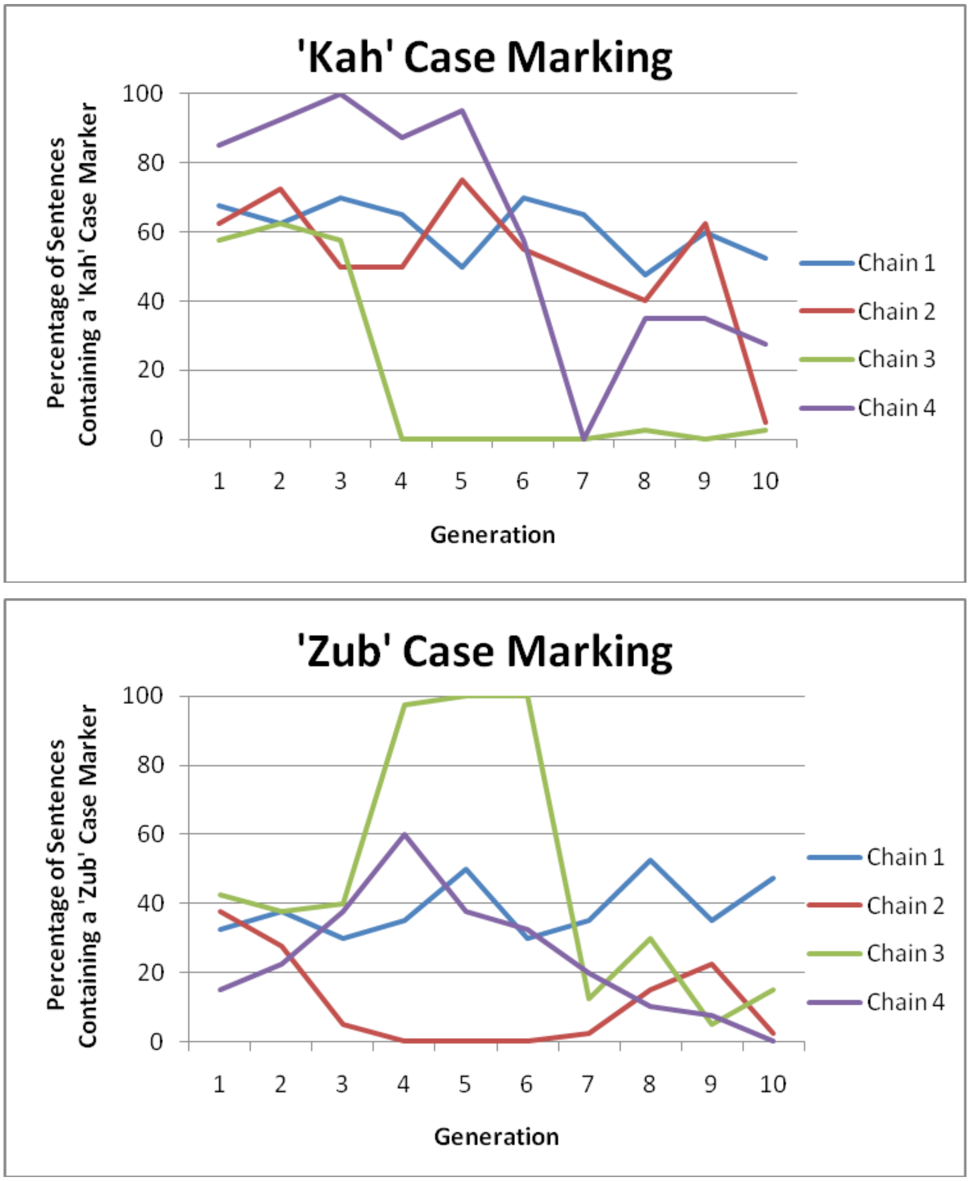


Figure 9. Percentage of sentences containing an instance of each type of case-marker by generation. Although some participants in each chain developed a strong preference for one case-marker over another, there was no clear trend to use one case-marker exclusively over another, and for most chains the overall frequency of both case-markers declined in the later generations. We note that in Chain 3, starting at generation 4 and lasting until generation 6, *zub* appears to have been used as a type of verb particle, following the verb in sentence final position in close to all or all sentences.

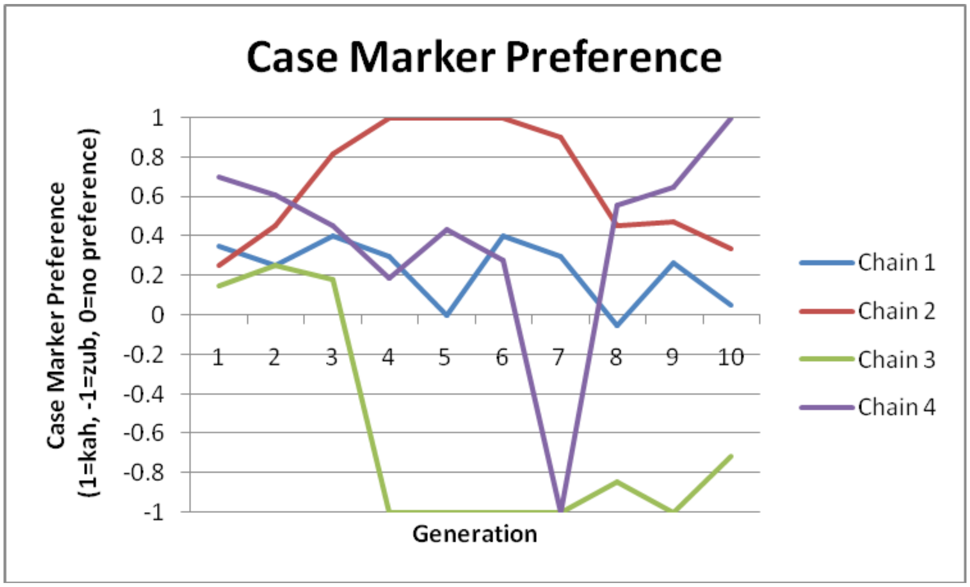


Figure 10. Strength of case-marker preference over generational time. A value of 1 represents a preference for *kah* 100% of the time, whereas a value of -1 represents a preference for *zub* 100% of the time. While we see some instances of chains strengthening a preference for a particular case-marker over time (represented by a shift away from zero in the later generations), several participants reverse the direction or greatly weaken the magnitude of case-marker preference.

Conditional entropy of case-marking. If the transmission process created a pressure to reduce variation in the language, regardless of context, we would expect to see each chain trending towards a case-marker preference rating of either 1.0 or -1.0 with each passing generation. In the four chains we observed, we did not see clear evidence for this kind of across-the-board reduction in variability. It may be the case, however, that participants maintained variability in their language while increasing the predictability of that variation in context.

To test whether this was the case, we examined the conditional entropy of case-marking given some feature of the context. If the variation in the production of a case-marker is random and not conditioned upon another variable in the input, then the degree of conditional entropy will be high. If this variation can be predicted when the value of the other variable is known, then the variation is not random and has lower conditional entropy. The increase in the predictability of variation of some feature of a language, and thus the systematicity of that language, can be quantified by the change in conditional entropy over time.

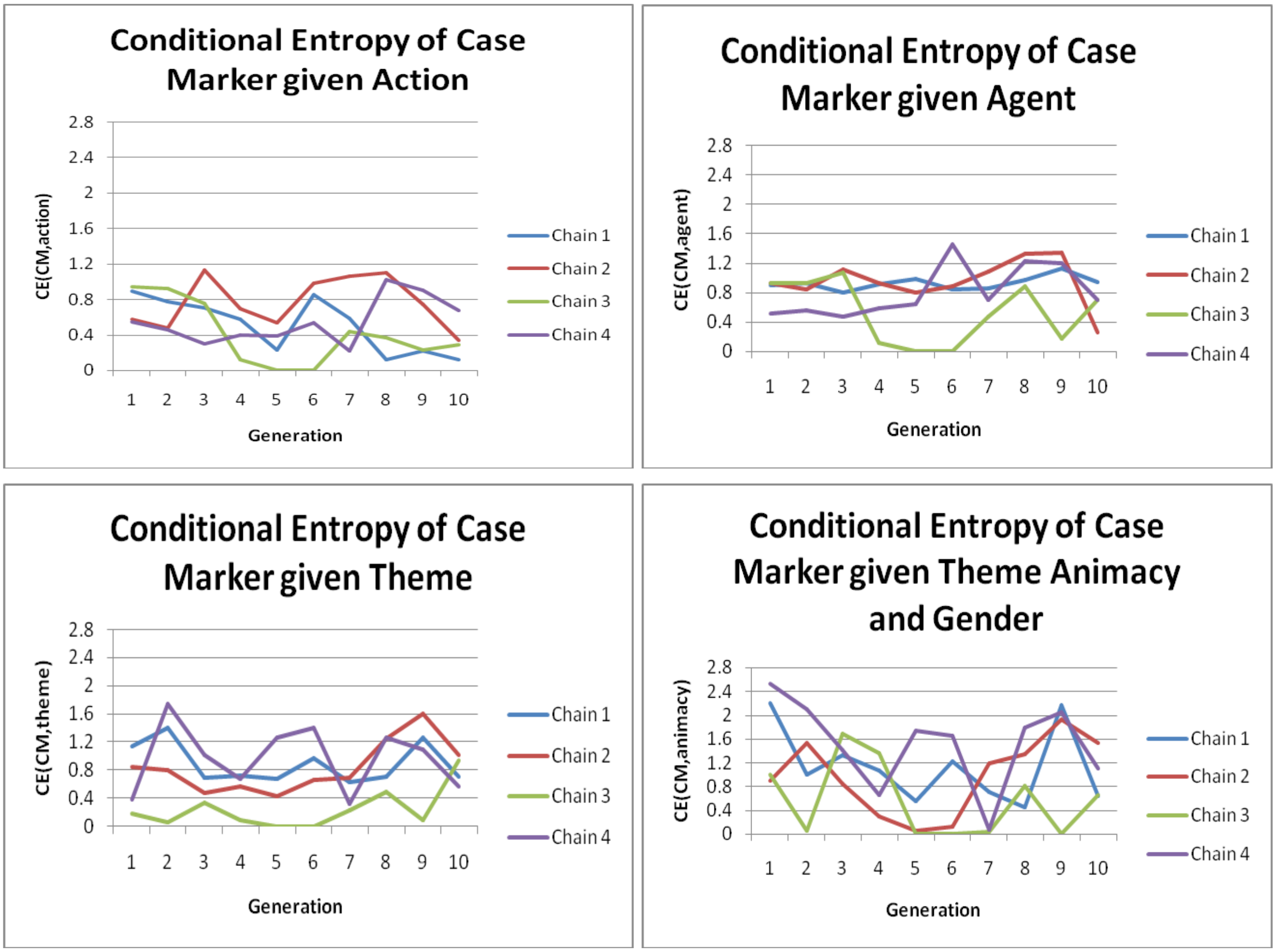


Figure 11. Conditional entropy of case-marking (*kah*, *zub*, or no case-marker) over time, given the action (top left), agent (top right), theme (bottom left), or animacy and gender (bottom right). If participants lexicalized the case-markers, we would expect to see a decrease in conditional entropy given one or more of these features. However, we do not see this trend in the data.

We evaluated the relationship between participant generation and the conditional entropy present in that generation's language given features of the input we identified as potentially relevant. We did not find clear evidence of a trend towards the reduction of conditional entropy in the case-marking over time (see Figure 11). For entropy of case-marking conditioned upon the action depicted, the average CE for all chains decreased from .57 in Generation 1 to .28 in Generation 10. For entropy of case-marking conditioned upon the agent referent, the average CE for all chains increased from .62 in Generation 1 to .83 in Generation 10. For entropy of case-marking conditioned upon the theme referent, the average CE for all chains increased from .43 in Generation 1 to .47 in Generation 10.

Finally, for entropy of case-marking conditioned upon the gender and animacy of the theme, the average CE for all chains increased from .39 in Generation 1 to .48 in Generation 10. Thus we do not find evidence that the amount of unconditioned variation present in the case-marking decreases with each iteration of the language.

Changes planned for future work

Lexical errors provided the greatest source of unwanted (i.e. theoretically uninformative) variability in our data, rendering potential trends in case-marking and word order difficult to detect and interpret. Unlike previous experiments which employed a lexical filtering process to prevent the proliferation of these errors from one generation to the next, we avoided any explicit manipulation of the language produced by participants in the experiment. Despite all participants' successful performance in the initial lexical learning block, lexical substitutions and inconsistencies introduced in the sentence learning and production portions of the experiment propagated throughout each chain and persisted in subsequent generations. These types of lexical substitutions greatly reduced the number of trials for which it was possible to extract reliable structural information about each sentence.

Our error rate was particularly high compared to the lexical error rates in the single generation studies (Experiments 1 and 2), and this was likely due, in part, to participants seeing the initial, Generation 0 word-meaning mappings during the lexical learning block, followed by the word-meaning mappings of the previous generation's language during the sentence learning blocks (this was not intended by design, but rather had escaped our attention). If these mappings had drifted from those present in the initial input, as was frequently the case, then the participant's input language was maximally conflicting. For these reasons, it will be paramount in future studies to either increase the amount of lexical training or to filter the output of each generation to be lexically consistent before it is given as input to the next generation. Furthermore, because most of the lexical errors occurred in the verbs, for which participants did not receive explicit instruction during lexical learning, it may also be necessary to either give the participants verb prompts during production, or to provide exposure for the verbs in isolation, as we did for the nouns.

Due to a technical mistake, all words of the Generation 0 lexicon, including both case-markers, were available for the participant's use during the production trials, whether that word had occurred in their input language or not. The result of this was that in some chains, case-markers which had completely died out were suddenly resurrected in later generations. Future studies will restrict the words appearing on the screen in the production trials to words that were actually used in the input language to avoid this problem.

General Discussion

One of the primary aims of this work was to evaluate the feasibility of conducting iterated learning studies through the Mechanical Turk platform. Our preliminary findings were very promising in this regard. Participants were able to learn the artificial language at a level comparable to the rate of learning demonstrated by typical participants in our laboratory, as measured by average performance in discrimination and naming tasks. Most participants were able to produce the full set of nouns in the lexicon after only a brief exposure period, but many had difficulty producing the verbs correctly without any explicit instruction. Feedback from the participants indicated that they found the task enjoyable and appropriately challenging, and would be willing to participate in similar studies again in the future. The use of MTurk facilitated a rate of participation in our experiments that iterated language learning studies have never previously approached: ten parallel chains of participants could typically be run in under an hour. The turnaround between generations was also a great improvement upon previous instantiations of the iterated learning paradigm. Without implementing any kind of lexical filtering mechanism in between generations, we were able to prepare the stimuli needed for subsequent generations of a chain in as little as ten minutes from the completion of the previous generation. In future work, we expect to implement server-side programming that will fully automate the process of preparing the next generation's language, so as to completely eliminate downtime between generations.

The data that can be collected within this paradigm is by nature both rich, in the sheer quantity of information that can be extracted, and noisy, in the inevitability of obtaining a certain amount of random, theoretically uninteresting error in the results. We can achieve robustness in the face of this noise by taking advantage of the reduced cost in both participant payments and time invested. The ease of running a larger number of participants makes new avenues of investigation possible within this paradigm. Future studies may increase the complexity of the language, the length of each participant's exposure to the language, or the number of chains and generations run. In addition, future studies increase the number of sources comprising a learner's input by sampling from the productions of multiple learners. This could provide a test of the simplifying assumptions made about the population dynamics of linguistic transmission by all studies done in the IALL paradigm to date (but see Kirby & Hurford, 2002; Livingstone & Fyfe, 1999; Winters, 2009, for an exploration of these dynamics in the ILM). All of these options create the opportunity for a more ecologically valid approximation of the natural language transmission process, and in addition to generally easing the cost of gathering data under this paradigm. Ideally, further work could help engender a shift away from studying the linguistic behaviors and biases of learners in isolation, and towards a more situated look at what learners do within the context of learning from the linguistic output of other language users.

Our findings at this stage are largely exploratory. We focus here on what kinds of innovation to the grammar appear to be possible within this paradigm, rather than what must necessarily be typical or expected. Participants received only brief exposure to the language and were then expected to produce full sentence descriptions for novel scenes not contained in their input. Because of the complexity of this task, and the relative difficulty of learning a vocabulary and a grammar without explicit instruction or feedback, there was considerable opportunity for learners to deviate drastically from their input even if they believed they had faithfully reproduced its structure. Many learners introduced innovations which were not suggested by their input, such as producing agreement-like markers on the verb, using subject-verb-object word order, and manipulating word order and word choice to maximize similarity in phonological onset of consecutive words. These may simply be failures to learn the language sufficiently given limited exposure, general inattention, or low motivation to perform the task well. On the other hand, these innovations could speak to something potentially more interesting and applicable to change in natural languages. The variability between participants in terms of how greatly they deviate from the input they receive likely reflects a complex interaction between differences in participants' general cognitive and motivational abilities and individual differences in propensity for linguistic innovation. With only one speaker representing each generation of a chain in our studies, these types of innovations may have appeared more detrimental to transmissibility than they might have in a situation where a learner receives input from a wide range of sources. In natural language situations where learners receive a greater quantity of input from a greater number of sources, these innovators of structure are likely to be hugely influential in terms of both the ways in which languages change and the rate at which they change.

More often than introducing completely novel features into a grammar, however, learners eliminated certain kinds of structure from their input. One trend that seems to emerge is that the learners in our studies generally avoid, to at least some degree, redundantly encoding information concerning thematic relations using both word order and explicit case-marking. Instead, most of these learners prefer to fix the word order and produce fewer case-markers than seen in their input. Although performance during the sentence discrimination task showed that learners were able to understand alternative (OSV) word orders, and that they could make use of case-markers to extract thematic role information from a sentence when other cues were not available or were not reliable, their sentence productions typically did not contain the same level of variability that participants dealt with as comprehenders, particularly with respect to word order. This regularization did not necessarily reflect a reduction in unpredictable variation, rather than an overall reduction in variation: the amount of random variation contained in each language, as measured by the conditional entropy of case-marking and word order, did not significantly decrease with each iteration – contrary to what we had hypothesized. However, several learners did produce a language that was in at least some ways more predictably structured than the language they had been exposed to. These individuals may exemplify a class of regularizers who, together with the linguistic innovators, help drive language change by increasing the predictability of using a structure in a specific context.

Advantages of the web-based paradigm

The web-based approach has a major advantage over previous lab-based iterative learning experiments: We were able to reach a large number of participants in a small fraction of the time required in previous studies (orders of magnitude smaller). This opens the door to a new type of iterative learning studies, with larger populations and more realistic transmission models, where the input to an individual is a weighted mixture of the outputs of individuals in the previous generation. Larger populations also allow for added robustness without artificially manipulating the input by filtering, as has been necessary in all previous IALL studies. The low cost of running additional participants also allows us to manipulate the input provided to learners, experiment with different types of languages in the initial input, and explore how these manipulations affect language change and how variation spreads in a population over a period of time. This paradigm also enables us to employ more complex artificial grammars in the input than would otherwise be possible. This makes the study of syntactic change using iterated learning feasible for the first time.

Ongoing studies

We are currently conducting a modification of the study described in Experiment 3 that replaces the qualitatively asymmetrical case-marking—*kah* versus *zub*—with the typologically more common pattern of optional case-marking—presence versus absence of a case-marker (Iggesen, 2008). In addition to the advantage of beginning with a system that more closely resembles what we typically see in currently existing languages, this provides us with the opportunity to attempt to replicate findings from single-generation artificial language learning studies and compare the changes introduced in the case-marking and word order systems of a language by individual learners with those that accumulate over the course of linguistic transmission. This study also implements several of the improvements to the methodology motivated by our analysis in Experiment 3.

Further work will examine to what extent and in what contexts we might see a trade-off in structure between case-marking and word order, and how the stability of those systems changes over time. If these types of studies can be used to demonstrate that languages change during the process of being transmitted in ways that increase the systematicity of each language's structure, such that the form a signal will take becomes increasingly predictable from the intended meaning, it would provide strong indication that there is something in the transmission process itself that pressures languages to adapt in order to ensure their own survival. This would suggest a new way of addressing the question of how, and for what purpose, languages change, and leaves us with the task of identifying what biases and functional pressures interact to shape the types of languages we see represented in the world today.

Conclusion

In this thesis we introduced a web-based paradigm for running iterated artificial language learning experiments. Because we are still in the exploratory stages of implementing this paradigm, we encountered a number of technical and methodological errors that rendered our data difficult to interpret. However, preliminary results presented here indicate that the web-based approach is a promising one for iterated learning studies. This paradigm drastically reduces the cost, in both time and money, of running these types of studies, making possible future avenues of investigation. While a high rate of lexical error in the present study introduced unwanted noise into in the data, we found results that suggested several interesting trends were emerging. In general, there were tendencies to fix the word order, to reduce overall presence of case-marking, and to develop a preference for one case-marker over another when that preference did not exist in the input. Ongoing work will explore whether we see these trends consistently emerging, and if they demonstrate a cumulative effect over time.

References

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 173–218). New York, NY: Cambridge University Press.

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua, 117*, 510-526.

Brighton, H., & Kirby, S. (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In J. Kelemen & P. Sosik (Eds.), *Advances in Artificial Life: Proceedings of the Sixth European Conference [ECAL01]* (pp. 592-601). Prague: Springer-Verlag.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews, 2*(3), 177-226.

Briscoe, E. J. (1998). Language as a complex adaptive system: Coevolution of language and of the language acquisition device. In P. A. Coppen, H. van Halteren, & L. Teunissen (Eds.), *Proceedings of the Eighth Meeting of Computational Linguistics in the Netherlands* (pp. 3-40). Amsterdam: Rodopi.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23*(2), 157-205.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*(5), 489-558.

Cornish, H. (2005). *The role of meaning within the iterated learning model of language evolution.* M.A. thesis, University of Edinburgh.

Cornish, H. (2006). *Iterated learning with human subjects: An empirical framework for the emergence and cultural transmission of language.* M.Sc. thesis, University of Edinburgh.

Cornish, H., Tamariz, M., & Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: What experiments can tell us. *Language Learning, 59*(s1), 187-205.

Culbertson, J. (2010). *Learning biases, regularization, and the emergence of typological universals in syntax.* Ph.D. thesis, Johns Hopkins University, Baltimore, MD.

Culbertson, J. & Smolensky, P. (2010). Testing Greenberg's Universal 18 using the Mixture Shift Paradigm for artificial language learning. To appear in *Proceedings of the 40th Annual Meeting of the North East Linguistic Society [NELS40].* Amherst, MA: Graduate Linguistics Student Association.

Culbertston, J., Smolensky, P., & Legendre, G. (2011). *Learning biases and constraints on syntactic typology: An artificial language learning approach.* Talk presented at the 85th Annual Meeting of the Linguistic Society of America, Pittsburgh, PA.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2011a). *Word order and case-marking in language acquisition and processing.* Poster presented at the 85th Annual Meeting of the Linguistic Society of America, Pittsburgh, PA.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2011b). *Functional pressures in (artificial) language learning.* Poster presented at the 24th Annual Conference on Human Sentence Processing, Stanford, CA.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2011c). Functional biases in language learning: Evidence from word order and case-marking interaction. To appear in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Boston, MA.

Hoefler, S. (2006). Why has ambiguous syntax emerged? In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *The Evolution of Language: Proceedings of the Sixth International Conference [EVOLANG6]* (pp. 123-130). Rome: World Scientific.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151-195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*(1), 30-66.

Iggesen, O. A. (2008). Asymmetrical case-marking. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. Retrieved from http://wals.info/feature/50

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review, 14*(2), 288-294.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. Oxford: Oxford University Press.

Kirby, S. (2002). Learning, bottlenecks, and the evolution of recursive syntax. In E. J. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge: Cambridge University Press.

Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In C. Lyon, C. Nehaniv, & A. Cangelosi (Eds.), *Emergence of communication and language* (pp. 253-267). London: Springer-Verlag.

Kirby, S., & Christiansen, M. H. (2003). From language learning to language evolution. In M. H. Christiansen & S. Kirby (Eds.), *Language evolution* (pp. 272-294). New York, NY: Oxford University Press.

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. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*(12), 5241-5245.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–148). London: Springer-Verlag.

Livingstone, D., & Fyfe, C. (1999). Modeling the evolution of linguistic diversity. In D. Floreano, J. Nicoud, & F. Mondada (Eds.), *Advances in Artificial Life: Proceedings of the Fifth European Conference [ECAL99]* (pp. 704-708). Berlin: Springer-Verlag.

Maynard Smith, J., & Szathmáry, E. (1995). *The major transitions in evolution*. New York, NY: Oxford University Press.

Mufwene, S. S. (2001). *The ecology of language evolution*. [Cambridge Approaches to Language Contact]. Cambridge: Cambridge University Press.

Perfors, A., Tenenbaum, J., Regier, T. (2006) Poverty of the Stimulus? A rational approach. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 663-668). Austin, TX: Cognitive Science Society.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707-784.

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition, 95*(2), 201-236.

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

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science, 14*(1), 65-84.

Smith, K. (2003). Learning biases for the evolution of linguistic structure: An associative network model. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, & J. Ziegler (Eds.), *Advances in Artificial Life: Proceedings of the Seventh European Conference [ECAL03]* (pp. 517-524). Berlin: Springer-Verlag.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: The cultural emergence of compositional structure. *Advances in Complex Systems, 6*(4), 537-558.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life, 9*(4), 371-386.

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

Tily, H. J., Frank, M. C., & Jaeger, T. F. (2011). The learnability of constructed languages reflects typological patterns. To appear in *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Boston, MA.

Winters, J. (2009). *Adaptive structure, cultural transmission, and language: Investigating a population dynamic in human iterated learning.* M.Sc. thesis, University of Edinburgh.