It takes a village: The role of community size in linguistic regularization

Annemarie Kocab, Jayden Ziegler, & Jesse Snedeker

Department of Psychology, Harvard University

Annemarie Kocab (corresponding)
Dept. of Psychology
Harvard University
33 Kirkland St.
Cambridge, MA 02138
kocab@fas.harvard.edu

Abstract

Studies of artificial language learning provide insight into how learning biases and iterated learning may shape natural languages. Prior work has looked at how learners deal with unpredictable variation and how a language changes across multiple generations of learners. The present study combines these features, exploring how word order variation is preserved or regularized over generations. We investigate how these processes are affected by (1) learning biases, (2) the size of the language community, and (3) the amount of input provided. Our results show that when the input comes from a single speaker, adult learners frequency match, reproducing the variability in the input across three generations. However, when the same amount of input is distributed across multiple speakers, frequency matching breaks down. When regularization occurs, there is a strong bias for SOV word order (relative to OSV and VSO). Finally, when the amount of input provided by multiple speakers is increased, learners are able to frequency match. These results demonstrate that both population size and the amount of input per speaker each play a role in language convergence.

Keywords: artificial language learning; word order; iterated learning; learning biases

1. Introduction

Human languages are systematic in two ways. First, each individual language has its own set of systematic grammatical rules (e.g., Baker, 2001; Givón, 1985). For example, in English, we reliably use subject-verb-object (or SVO) word order to mark the relationship between the participants in an event. Second, while these regularities show cross-linguistic variation (e.g., Japanese uses SOV order), there is nevertheless great systematicity *across* languages in the patterns of variation, leading to typological and implicational universals, and strong statistical associations (e.g., Comrie, 1989; Croft, 2003; Greenberg, 1963; Maslova, 2003). For example, of the six possible word orders, the two subject-initial orders account for nearly 90% of the languages reported to have a dominant order (i.e., 48% SOV and 41% SVO; see Dryer, 2005; 2011; 2013; Greenberg, 1963).

How can we account for this systematicity both within and across languages? There are two broad proposals. The first concerns the innate capacities of the learner, which guide what we produce. For example, the structure of language has been proposed by some to stem from innate syntax (e.g., Gleitman, 1990; Grimshaw, 1981; Pinker, 1984), or from our prelinguistic conceptualization of the world (e.g., Göksun, Hirsh-Pasek, & Golinkoff, 2017; Hartshorne, O'Donnell, Sudo, Uruwashi, & Snedeker, 2016; Lakusta, Wagner, O'Hearn, & Landau, 2007; Strickland, 2017). Such accounts offer an explanation for certain recurring regularities across different languages, such as the existence of grammatical subjects and objects. A second account is that the structure of languages is shaped over long periods of time through historical processes (e.g., Christiansen & Chater, 2008; Kirby, 2001; Kirby, Cornish, & Smith, 2008; Tomasello, 2008). On this view, language is a product of cumulative cultural evolution, where each generation builds on the work of the preceding generation, resulting in increasing complexity (e.g., Tomasello, Kruger, & Ratner, 1993; Tomasello, 1999). On these proposals, the prevalence of certain word orders over others can be attributed to the cognitive biases of individual learners, which are amplified over time through intergenerational transmission and iterated learning, leading to these typological patterns (e.g., Griffiths & Kalish, 2007; Reali & Griffiths, 2009).

Recently, findings from two different empirical approaches, the study of artificial grammar learning in the laboratory (e.g., Culbertson & Newport, 2015; Culbertson et al., 2012; Fedzechkina et al., 2012; Hudson Kam & Newport, 2005; Kirby, Cornish, & Smith, 2008; Smith et al., 2016; Smith & Wonnacott, 2010; Wonnacott, 2011; Wonnacott, Brown, & Nation, 2017) and the study of naturally emerging languages (e.g., Kocab, Senghas, & Snedeker, 2016; Pyers et al., 2010; Senghas, 2003; Senghas & Coppola, 2001) offer evidence for a third possibility. The structure of language may arise rapidly through learning and transmission, where learners change the input as it is filtered through their cognitive systems (e.g., Bickerton, 1981; Gleitman & Newport, 1995; Pinker, 1994; Senghas, Kita, & Özyürek, 2004). The study of emerging languages, such as Nicaraguan Sign Language (NSL), demonstrates how a language may quickly take on structure within one or two generations (Senghas & Coppola, 2001). Further, findings from this research program have revealed different patterns of development depending on the domain under consideration. Some aspects of language seem to emerge quickly, such as a stable lexicon and words for cardinal numbers (Richie, Yang, & Coppola, 2013; Senghas, 1995; Flaherty & Senghas, 2011) while others, such as the use of spatial morphological marking to indicate the roles of agents and patients in an event, require additional generations (Senghas, 2003). These findings suggest that the answer to what drives the development of linguistic structure may vary depending on the domain under consideration.

In this paper, we leverage the laboratory to ask how intergenerational transmission shapes emergent structure when learners are exposed to variable word orders, and how this structure may be affected by population size. In the remainder of this introduction, we do four things. First, we review the artificial grammar learning studies, focusing on what they reveal about the role of individual learning biases and regularization (section 1.1). Second, we review findings from iterated learning studies which employ diffusion chains, where an experimenter models a behavior for a participant who learns and reproduces that behavior for another participant, and so on, allowing researchers to test the effect of multiple generations on the structure of language (section 1.2). Next, drawing on historical studies and computational models of language change, we discuss how population dynamics could interact with language transmission and convergence (section 1.3). These three sections motivate the current study, which addresses how the structure of the language community—the size of the population and the amount of input provided from each person—affects regularization across multiple generations of learners. The fourth and final section presents our approach to addressing these questions, in a domain that has not yet been investigated using diffusion chains: word order (section 1.4).

1.1. Learning biases, regularization, and frequency matching in single-generation studies

As noted above, languages are characterized by regularity, as seen in the prevalence of subject-initial word orders. At the same time, there exists variability: As an example, while the canonical word order in Turkish is SOV, all six possible word orders are grammatical. Much of this variability appears to be systematic and meaningful: The word order used in Turkish varies depending on discourse structure (Erguvanli, 1984). Any theory of language must capture all of these facts: how regularities arise, how variability is preserved, and what predicts which processes will prevail. Studies of artificial grammar learning have begun to make progress on this topic by exploring how learners reorganize languages with grammatical variation. Below, we discuss findings from studies looking at three different linguistic domains: determiner-noun pairs, harmonic word orders, and differential case-marking.

Hudson Kam and Newport (2009), using a single generation of adult and child learners, experimentally varied noise in the input provided to participants. Participants learned an artificial language where the use and frequency of multiple determiners with nouns occurred probabilistically. They found that adults frequency matched, reproducing the inconsistencies, except in cases of extreme noise. In contrast, children regularized the inconsistencies.

In a study that varied word order, adults learned a language with a mix of harmonic (where heads are consistently initial or final, such as Adjective-Noun and Numeral-Noun, or Noun-Adjective and Noun-Numeral) and non-harmonic orders (where the heads are not consistent, like Adjective-Noun and Noun-Numeral; Culbertson, Smolensky, & Legendre, 2012). Rather than reproducing the input, adult learners showed greater use of harmonic orders compared to non-harmonic orders. In another study, child learners showed an even stronger bias for harmonic word orders, producing harmonic patterns at a high frequency, even when they were less frequent in the input (Culbertson & Newport, 2015). Instead of reproducing the variability, both child and adult learners reorganized their input to decrease this noise, and the emergent structure aligned with what is seen typologically across languages.

In another study looking at word order, adult learners were taught an artificial language with two constituent orders used in different frequencies and with differential case-marking (Fedzechkina, Jaeger, & Newport, 2012). In differential case-marking systems, certain types of subjects and objects, and not others, are marked. Referents that are less typical for their grammatical function are overtly marked. This marking often occurs in a principled manner—for example, following animacy. As subjects are frequently animate and objects inanimate, in natural languages with differential case systems, inanimate subjects and animate objects are case-marked. In Fedzechkina et al. (2012), however, participants were taught a language where the case-marking was *not* conditioned on animacy (that is, not adhering to the pattern typically seen in natural languages). Instead of reproducing this less predictable pattern, adult learners reorganized the input so that case-marking became conditioned on animacy, following the pattern typically seen in languages. Once again, learners seem to have a bias against variability, and in the course of a single experimental session, regularized inconsistent input and preferentially learned forms mirroring structures that appear cross-linguistically.

Together, these findings demonstrate three things. First, rather than reproducing the variation in the input, learners will regularize, reorganizing the language to become more systematic.

Second, while variable input is dispreferred by both child and adult learners, child learners regularize much more readily than adults. The input must be highly complex for adults to regularize. For example, in Hudson Kam and Newport (2009), regularization was observed in adults when the input language contained 16 noise determiners each occurring 2.5% of the time and one main determiner occurring 60% of the time. One explanation that has been proposed for this difference in regularization between children and adults is that adults may be better able to reproduce a greater degree of variability due to better memory or processing (e.g., Hudson Kam & Chang 2009; Hudson Kam & Newport, 2009; see also Newport, 1988; cf. Perfors, 2016; Smith et al., 2017; West & Stanovich 2003). An alternative, though not mutually exclusive, explanation that has been proposed for why adults frequency match more than children is pragmatic-- adults expect the underlying variation to be by design and assume they should attempt to correctly learn that variation (Perfors, 2016). In addition, in a recent study where adult learners were taught an artificial language with unpredictable variation in word order, regularization was strongest when the learners believed they were communicating with a human interlocutor compared to a computer (Fehér, Wonnacott, & Smith, 2016). There are also differences in how adults and children regularize variable input. Adults generally regularize by selecting the most frequent construction in the input. In contrast, children show individual variability in which form is regularized. Some children regularized the most frequent determiner, but others regularized by omitting all determiners, or by using determiners with nouns in transitive but not in intransitive sentences (Hudson Kam & Newport, 2009).

Third, there is a tendency for both children and adults to regularize towards typologically preferred patterns, such as word orders that are harmonic (where there is consistency in head directionality) and case-marking systems that are conditioned on animacy (Baker, 2001; Dryer, 1992, 2013a; 2013b; Greenberg, 1963). These findings suggest that recurrent linguistic patterns may stem from learning biases against irregularity, and forms that mirror those seen in the world's languages are preferred and regularized by learners of a novel artificial language in the course of a single experimental session.

1.2. Iterated learning and transmission in multiple-generation studies

In the single-generation studies reviewed above, participants are trained on an artificial language constructed by the experimenters, and then are asked to produce descriptions in this language (e.g., Culbertson & Newport, 2015; Culbertson et al., 2012; Fedzechkina et al., 2012; Hudson Kam & Newport, 2005; Wonnacott, 2011; Wonnacott, Brown, & Nation, 2017). In multiple-generation studies, chains are created such that the first participant in each chain learns the language provided by the experimenters in a single session. They then produce descriptions and this output is recorded. The output of the first learner is given as input to a second participant, whose output in turn serves as input for the third participant, and so on. This continues until the desired number of iterations, or "generations," is achieved (e.g., Kirby, Cornish, & Smith, 2008; Smith et al., 2016; Smith & Wonnacott, 2010). In these experiments, multiple chains are often constructed, but each link in each chain consists of a single person as described above. Thus, every learner receives input from a single language model.

Diffusion chains allow experimenters to better understand factors underlying language evolution. In these paradigms a language is learned and relearned by multiple minds, a process parallel to the filtering of natural languages through generations of learners. The utility of this approach is bolstered by computational work suggesting that through iterated learning, even weak biases, given enough time, may shape linguistic structure (Griffiths & Kalish, 2007; Reali & Griffiths, 2009). Indeed, findings from diffusion chain studies demonstrate that artificial languages take on structure and become more learnable across participants (e.g., Kirby et al., 2008; Smith & Wonnacott, 2010).

In one such study, adult participants learned an artificial language composed of written labels for colored objects in motion, such as a blue circle bouncing (Kirby et al., 2008). The labels in the initial language devised by the experimenters were random and generated by concatenating between 2 and 4 syllables (e.g., *hopa*, *manehowu*, *lemipo*). With each round of iterated learning, the set of lexical items took shape, becoming more structured. For example, a spiraling motion was described with one syllable string (e.g., *pilu*), the color grey with another syllable (e.g., *ne*), and a triangle shape with yet another syllable string (e.g., *aho*). The number of changes needed to convey related meanings, such as objects sharing the same shape, decreased over generations, and the language became more learnable, where the degree of change between the output of one generation and that of the next decreased. While not all diffusion chains converged on a structured language, these experiments show how structure *can* emerge through the process of learning and transmitting a language from one individual to another.

Further evidence for the effect of intergenerational transmission comes from a study by Smith and Wonnacott (2010) which compared single-generation learners with participants in diffusion chains. Adult participants were taught a semi-artificial language for describing scenes involving one or two animals performing a movement. The descriptions consisted of an English noun, a nonsense verb, and for scenes containing two animals, one of two possible marker of plurality (e.g., *fip* or *tay*). One marker was used 75% of the time, and the other 25% of the time. In the initial input language, there was no consistent relationship between the plurality marker and the marked noun. The majority of the single-generation participants frequency matched, maintaining the variability and producing output that was very similar to the input. For participants in the diffusion chains, however, use of the plural marking became more predictable over generations, such that by the final generation, a conditioned system of variance had emerged (9 of the 10 chains). Six of the nine chains that converged on a system of conditioned variance maintained the two forms of the plural markers, lexicalizing them such that each noun eventually became associated with one plural marker.

1.3. Size and structure of a language community

Diffusion chains represent a critical step towards increasing the ecological validity and scope of artificial language learning studies. The early work on diffusion chains, however, abstracts away from one critical feature of natural language learning and language creation: Real linguistic communities consist of multiple learners in each generation who receive input from multiple members of the prior generation (as well as each other).

Intuitively, the presence of multiple speakers could either slow down the process of convergence, preserving variability, or it could accelerate regularization, resulting in more rapid convergence. On the one hand, when there is just one learner in the community (one speaker in a generation), that learner is free to latch onto any generalization that they perceive. Whatever regularization (or convergence) occurs in their head can be passed down, with no additional interference, to the next generation. When multiple learners are present, each might latch onto a somewhat different generalization, resulting in more variable input and slowing down the process of convergence (Smith et al., 2017). On the other hand, introducing multiple sources of input as each person is learning logically increases the complexity of the learning problem by adding another factor for the learner to track (speakers) and allowing for more complex data patterns, such as speaker-specific sources of variability. When complexity is increased, or cognitive resources are taxed, previous work suggests adults are less likely to frequency match and more likely to regularize (e.g., Hudson Kam & Newport, 2009). Thus, we might expect that increasing the size of the community will accelerate regularization.

Some evidence for a dampening effect of multiple sources of input on regularization comes from a recent study comparing diffusions chains where a single participant in each generation learns from a single source with chains where there are two participants in each generation who learn from two sources (Smith et al., 2017). Using the same semi-artificial language as in Smith and Wonnacott (2010), Smith et al. (2017) looked at participants' learning and use of plural marking in three conditions. In the *one-person* condition, each generation of the chain consists of a single participant who learns from a single speaker. In the *two-person identity-unknown* condition, each generation consists of two participants who learn from two speakers. The descriptions are presented in a dislocated speech bubble with no information about the speaker; thus, participants do not know that there are multiple speakers producing different descriptions and are consequently unaware of the speaker's identity. In the *two-person identity-known* condition, each generation also consists of two participants who learn from two speakers. Participants are provided with a picture of one of two aliens alongside a speech bubble during the training. Half of the descriptions are presented with alien 1 and the other half with alien 2. One participant is asked to provide the label produced by alien 1, and the other participant is asked to provide the label produced by alien 2. Smith et al. (2017) observe fast convergence on a system of conditioned variance (where there is a one-to-one mapping of nouns to markers) in the one-person condition and slower convergence on a system of conditioned variance when there are multiple speakers (two-person identity-unknown and two-person identity-known conditions). We return to this result in the general discussion.

The present study builds on the use of diffusion chains by adding a layer of complexity that is present in natural language transmission: multiple learners at each link in the chain (see also Samara et al., 2017, for another study comparing learning from two speakers). This work, and others like it, allows us to begin examining how differences in population size can affect learning and convergence, building connections between the artificial language learning literature and work in linguistics suggesting that population size is a relevant factor for predicting typological variation and how languages change over time. A well-known observation is that languages used by small and isolated communities tend to have richer morphological and agreement systems, while languages with larger communities of speakers, such as English, tend to have simpler systems (Lupyan & Dale, 2010; Nettle, 2013; Trudgill, 2011; Wray & Grace, 2007). This inverse relationship between population size and morphological complexity has also been observed in language change. Historically, as a language community grows, the language tends to exhibit less complex morphological and agreement systems, the classic case being that of English (MacWhorter, 2002). In contrast, in the lexical domain, larger communities are reported to have larger vocabularies than smaller communities, likely in part due to the effects of trade, literacy, and technology (Bromham, Hua, Fitzpatrick, & Greenhill, 2015; Goulden, Nation, & Read, 1990; Pawley, 2006).

1.5. Our approach

The present study brings together several lines of research on regularization and convergence in artificial language learning. In these experiments, we ask how learning biases interact with features of the input: number of speakers and tokens per speaker. We probe adults' learning of two competing word orders, one dominant and one less frequent, to understand how variability in word order is preserved or eliminated as a language is passed down across multiple generations of learners. We expand on the diffusion chain design, which has traditionally used only one or two speakers per generation, by including multiple learners in each generation who create the input for the next generation.

To date, there have been no studies of word order variation employing diffusion chains. As the findings from the single-generation studies discussed in section 1.1 illustrate, as well as typological observations, testing the effects of iterated learning on word order is informative because learners seem to have strong cognitive biases for certain word order patterns (see also Bever, 1970; Dowty, 1991; Halliday, 1967; MacWhinney, 1977; Lambrecht, 1996; Osgood, 1980). In a separate vein, typological work has established that subject-initial word orders are highly prevalent in the world's languages (Dryer, 2005, 2011; 2013; Greenberg, 1963). Interestingly, this preference for subject-initial constituent orders has also been observed in laboratory studies using gestural language creation paradigms, where participants are asked to create novel communication systems (Gibson et al., 2013; Goldin-Meadow, Ozyurek, Sancar, & Mylander, 2008; Hall, Mayberry, & Ferreira, 2013; Kocab, Lam, & Snedeker, 2017; Langus & Nespor, 2010). Together, these findings suggest that word order is a domain where we can begin to probe the effects of cognitive bias and intergenerational transmission.

Across seven experiments, we vary (1) the word orders that are present in the initial input provided by the experimenter, (2) whether the input to each participant stems from a single source (one speaker) or whether it is divided among several sources (four speakers), and (3) the total amount of input that each learner is given.

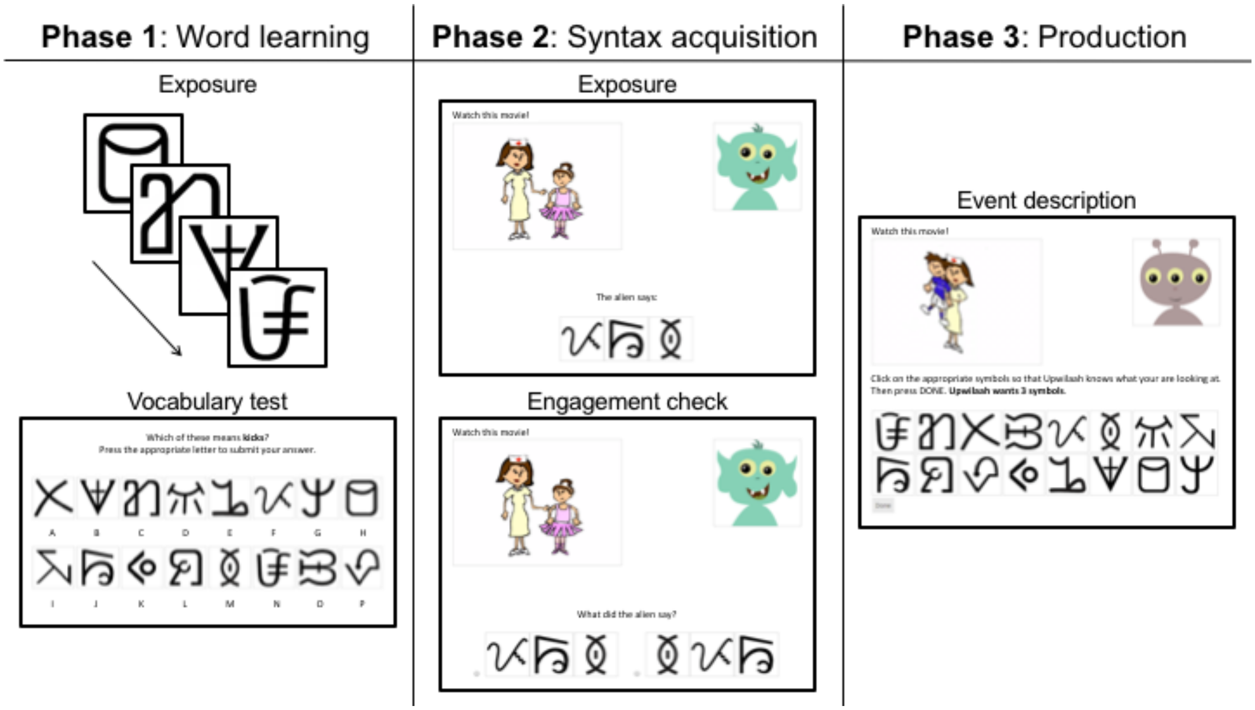
In Experiments 1 and 2, adult learners are exposed to an artificial language that uses two word orders, SOV and OSV, with the input produced by a single source (one alien speaker). In Experiment 1, SOV is the dominant word order, occurring 71% of the time, with OSV occurring the other 29% of the time. In Experiment 2, the proportions are the same, but OSV is the dominant order. We ask whether, when learning from a single speaker, adults will frequency match, reproducing the proportions of the two word orders, or converge on a single word order.

In Experiments 3-5, we ask whether the presence of multiple speakers will accelerate regularization. Experiments 3 and 4 have the same structure as Experiments 1 and 2, respectively, but with input provided by four different sources (four alien speakers) instead of one. If the presence of multiple speakers accelerates regularization, then we might expect learners to converge on a single order when learning from multiple sources. We also ask whether convergence will vary by word order: Will participants always converge on the dominant order (SOV in Experiment 3, OSV in Experiment 4), or will they instead converge on the more typologically common word order (SOV)? Experiment 5 extends this question to a new dominant word order: VSO (with SOV non-dominant).

Finally, in Experiments 6 and 7, learners again receive input from four sources, but we increase the amount of input each source produces four-fold. If the number of speakers interferes with frequency matching, then we might expect regularization in this case. However, given dense enough data from each speaker, learners may be able to overcome the complexity of balancing multiple sources and once again frequency match.

2. Methods Overview

To administer these experiments, we created an artificial language-learning platform using JavaScript and jsPsych (de Leeuw, 2015) that interfaces with Amazon Mechanical Turk via psiTurk (Gureckis et al., 2016). Participants were told that they would be learning an alien language. The first generation of learners was given a set of sentences (and scenes) that were created by the experimenters. We call this input Gen0. The output of the first group of learners (Gen1) was used to construct the input for the next generation of learners, and this process was repeated with each generation. We had initially planned to run five generations per experiment, but in cases of regularization, participants produced near-categorical word order preferences by the third generation (Gen3). We therefore collected only three generations of participants per experiment. Each experiment included three phases: (1) word learning, (2) syntax acquisition, and (3) production (Fig. 1). The critical differences across experiments were in the nature of the input in the syntax acquisition phrase.



Which of these means **kicks**?
Press the appropriate letter to submit your answer.

Watch this movie!

The alien says:

Watch this movie!

What did the alien say?

Watch this movie!

Click on the appropriate symbols so that Upwilaah knows what your are looking at. Then press DONE. **Upwilaah wants 3 symbols.**

Figure 1. Example stimuli and procedure.

2.1. Word learning phase

In the word learning phase, participants learned pairings of individual alien words (nouns and verbs) and meanings (animate creatures, inanimate objects, and actions). The vocabulary consisted of 16 hieroglyph symbols that were paired with 7 nouns (boy, girl, man, woman, ball, pizza, table) and 9 verbs (chase, drop, eat, hug, kick, lift, poke, push, throw; see Table 1). Animate nouns served as possible agents or patients, and inanimate nouns served only as possible patients. These studies were not designed to study vocabulary acquisition or the accuracy of lexical transmission. Our goal in the word learning phase was to ensure that all participants knew the lexicon that we had created. We held the number of vocabulary items and the nature of lexical training constant across the experiments so that we could perform direct comparisons where appropriate.

Table 1. Artificial language vocabulary composition across experiments.

Hieroglyph symbols (16)	Verbs (9)	Agent or patient nouns (4)	Patient-only nouns (3)
	chase, drop, eat, hug, kick, lift, poke, push, throw	boy, girl, man, woman	ball, pizza, table

This phase consisted of two parts: (1) exposure and (2) vocabulary test (Fig. 1). During exposure, each hieroglyph and its English translation appeared on screen for three seconds, one at a time in succession. Each word appeared only once, and the order of appearance of the words was randomized across participants. After exposure, participants were tested on their vocabulary retention. For the vocabulary test, participants were presented with the English translations, one at a time, and were asked to select the hieroglyphs that matched. They made their selection for each word by choosing the matching hieroglyph from a grid of all hieroglyphs in the study. All participants were required to repeat the word learning phase (exposure plus vocabulary test) at least once regardless of their performance at first test. Participants were allowed to continue to the next part of the experiment only if they reached 95% accuracy or higher on the second vocabulary test, otherwise they repeated the word learning phase until they reached threshold.

2.2. Syntax acquisition phase

In the syntax acquisition phase, participants were exposed to 28 simple transitive events (videos) and their corresponding descriptions. On each trial, participants would see an event paired with an alien speaker (Fig. 1). The participants would watch the event, and then the alien would produce a description of that event. All descriptions included an animate subject, a verb, and an animate (12 out of 28) or inanimate object (16 out of 28), in one of two word orders: a dominant word order and a non-dominant word order. The dominant word order was the word order that occurred 71% of the time in the input language to the first generation of learners (Gen0) in each experiment. The non-dominant word order was the word order that occurred the remainder of the time in the input language. Across experiments, we manipulated (1) how many sources (alien speakers) the participants learned the language from, (2) what the dominant word order in the input language was, and (3) how much input each source provided (see Table 2). As an engagement check, immediately after each sentence exposure, participants were asked to select the correct description from a set of two sentences for the event they just saw described (Fig. 1). The foil sentence differed from the original either by a single word (e.g., subject replaced with another animate noun, verb replaced with another verb, etc.) or in its word order (which was never one of the two word orders in the Gen0 input). Feedback, with correction if necessary, was provided after each trial. If they answered correctly, participants were told so and allowed to advance directly to the next trial. If they answered incorrectly, participants were shown the correct description (correct hieroglyphs in correct order) before being allowed to move on. The procedure for feedback was the same across all experiments.

Table 2. Summary of experiments.

Experiment	Dominant word order (71%)	Non-dominant word order (29%)	Number of alien speakers	Total input (across all alien speakers)	Tokens per alien speaker (Gen0)
1	SOV	OSV	1	28 sentences	28 sentences (20 SOV, 8 OSV)
2	OSV	SOV	1	28 sentences	28 sentences (20 OSV, 8 SOV)
3	SOV	OSV	4	28 sentences	7 sentences (5 SOV, 2 OSV)
4	OSV	SOV	4	28 sentences	7 sentences (5 OSV, 2 SOV)
5	VSO	SOV	4	28 sentences	7 sentences (5 VSO, 2 SOV)
6	SOV	OSV	4	112 sentences	28 sentences (20 SOV, 8 OSV)

7	VSO	SOV	4	112 sentences	28 sentences (20 VSO, 8 SOV)

2.3. Production phase

Lastly, in the production phrase, participants were asked to produce descriptions of events. Of the 28 events in this phase, 14 were repeated from the syntax acquisition phase (7 animate-animate and 7 animate-inanimate) and 14 were novel events that had never been encountered before (10 animate-animate and 4 animate-inanimate). On each trial, participants saw an event paired with a single alien comprehender (Fig. 1). This alien was a new one that participants had not encountered in the syntax acquisition phase. The participants watched the event and were asked to describe it to the alien. They did so by selecting each word, one at a time, from a grid of all of the hieroglyphs (similar to the word learning phase). Participants were reminded on each trial that they should select only three words per video. This phase also included two practice trials that were excluded from the final analysis. These trials were also repeated from the syntax acquisition phase, and we provided feedback on whether they picked the correct words or not (but not on the word order). Participants received no feedback on the test trials. We were interested in what word orders participants used for the *novel* event trials relative to the input they received; we did not analyze the responses from the *repeated* event trials. This was because participants' productions for repeated event trials could have been influenced by their memory of the particular event-sentence pairings, when instead what we were aiming for was a purer measure of the rules they had learned on the basis of this input.

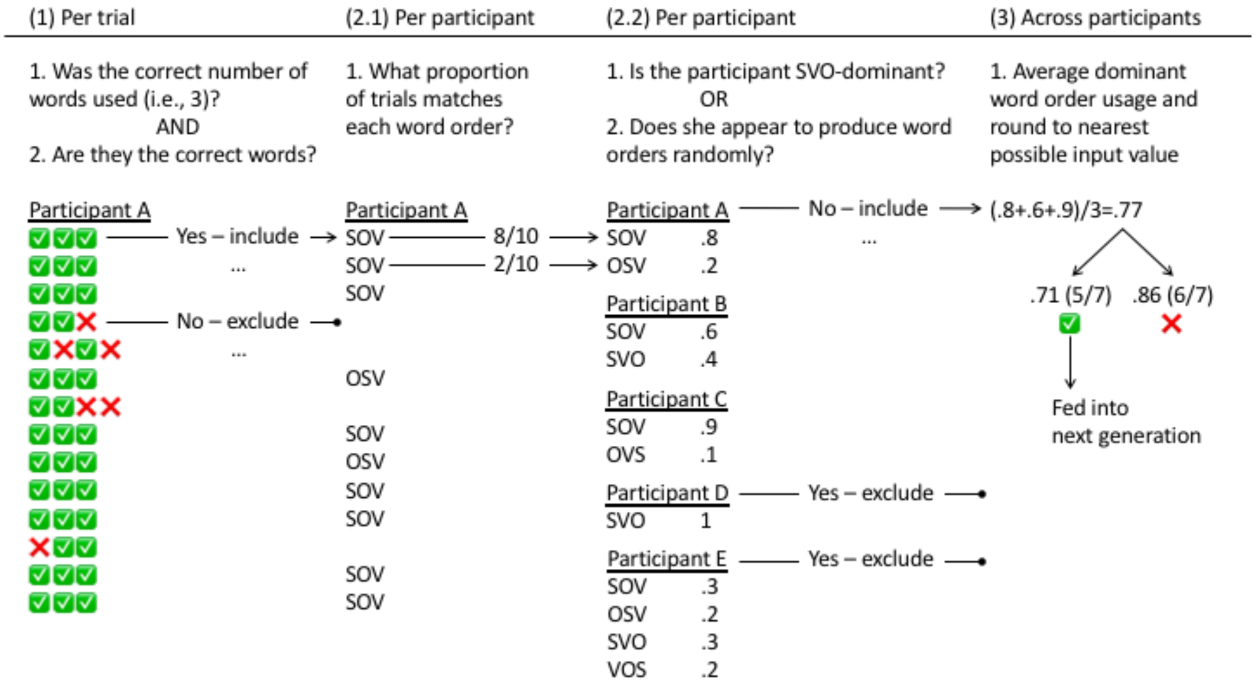
2.4. Input to subsequent generations

To create the input language for the syntax acquisition phase for each generation after the first, we used the output from the previous production phase. To do this, we first filtered out responses in which participants either (1) did not use the correct words or (2) did not provide the correct number of words (coded as NA). From the remaining usable trials, we calculated, separately for each participant, the proportion of their novel event responses that matched the dominant word order (e.g., SOV for SOV-dominant vs. OSV for OSV-dominant). We then averaged these values across the participants in each experiment. This became the proportion of sentences in the dominant word order for the next generation in that experiment, and the reciprocal became the proportion of non-dominant word order sentences. Given the relatively small number of sentences produced by each alien in the syntax acquisition phase (as little as seven), we rounded the resulting proportions to the nearest possible input value (e.g., nearest seventh). For example, if the participants in Gen2 produced an average of 75% dominant word order responses, then the input to Gen3 included 71% (5/7) dominant word order descriptions in the syntax acquisition phase and 29% (2/7) non-dominant word order descriptions. If the resulting proportion fell exactly between two possible input values, we erred on the side of conservatism (closer to the input value that that generation had received).

In creating the input, we excluded responses from participants who met two predetermined exclusion criteria: (1) The participants were SVO-dominant (defined as producing SVO orders 50% of the time or more across all target productions) or (2) they appeared to produce word orders randomly (defined as producing no word order more than 33% of the time across all target productions). However, all participants' responses were included in the statistical analysis of the output at each generation. Fig. 2 shows a schematic of this filtering process.

Note that due to methodological differences the above procedure we used to generate the input for the next generation necessarily differs from previous work comparing learning from one vs. two sources in diffusion chains (Smith et al., 2017). Because our single-speaker chains consisted of multiple learners in each generation (roughly 10 participants who learn from a single speaker) compared to the single-speaker chains in previous work (one participant per generation), the output from all of the participants in each generation was combined to create the new input.

Figure 2. Procedure for creation of input language for subsequent generations of speakers.



3. Experiments 1 and 2

In Experiments 1 and 2, participants are exposed to an artificial language that uses two word orders, SOV and OSV, with the input produced by a single source (one alien speaker). The experiments differ with regard to which of the two orders is dominant in the input: SOV for Exp. 1 and OSV for Exp. 2. We ask whether the participants will frequency match, reproducing the variability in the input, or converge on a single word order.

3.1. Methods

3.1.1. Participants

Sixty-eight native English speakers recruited from Amazon Mechanical Turk participated in Exps. 1 and 2 (29 female, 39 male; mean age=34, SD=7, range=19-56). There were 34 participants in Exp. 1 (Gen1: 12; Gen2: 10; Gen3: 12) and 34 participants in Exp. 2 (Gen1: 10; Gen2: 13; Gen3: 11). An additional 4 participants were excluded for being non-native English speakers (N=1) or for admitting to copying down or taking a picture of (some of) the words during the task (N=3). All participants provided written consent prior to participating and received $3.50 for their participation.

3.1.2. Materials

In Exps. 1 and 2, participants learned from a single alien speaker in the syntax acquisition phase. In Exp. 1, the dominant word order was SOV, and the non-dominant word order was OSV. In Exp. 2, the dominant word order was OSV, and the non-dominant order was SOV. The single alien produced all 28 exposure sentences. Thus, in the syntax acquisition phase for the first generation of participants, she produced 20 sentences of the dominant word order (71%) and 8 sentences of the non-dominant word order (29%).

3.1.3. Coding

Target event productions were scored as SOV, SVO, OSV, OVS, VSO, VOS, and Non-viable. Responses were scored as Non-viable if they contained fewer than or more than three words, or if one or more of the words used was not appropriate for the target video (making it impossible to know what the intended response was). Only responses for novel event videos were included in the analysis. In total, 427 of the 476 novel target productions in Exp. 1 (90%) and 433 of the 476 novel target productions in Exp. 2 (91%) were viable and thus entered into the analysis. For statistical analysis, we coded a response that matched the dominant word order from the initial input (Gen0) as 1, all other viable word orders produced in the production phase were coded as 0, and Non-viable responses were eliminated (i.e., NA). For example, in Exp. 1, we coded SOV orders as 1 and SVO, OSV, OVS, VSO, and VOS orders as 0. In presenting the results (for descriptive purposes), we have aggregated over both participants and items.

3.1.4. Data analysis

In order to determine whether each experiment diverged significantly from the initial input (Gen0), we performed a two-tailed, one-sample Wilcoxon signed-rank test (for non-normal data) on the output of Gen3 against its original starting value (.71). The dependent measure was the proportion of responses, by participant, that used the dominant word order (conditioned on experiment; see above).

In addition, to support more direct comparison with past work, we also report the *entropy* of word order use at each generation (across all utterances and all participants per experiment), as a measure of overall variability of the languages produced. For this, we calculated Shannon entropy (see, e.g., Smith et al., 2017), denoted *H*(word order), which represents variability as the average minimum number of bits required per word order to encode the full set of word orders produced by a population, as a function of their relative frequencies, $p(\text{word order}_i)$:

$$H(\text{word order}) = -\sum_{i=1}^{6} p(\text{word order}_i) \log_2 p(\text{word order}_i)$$

Entropy will be high (≈2.58) when all markers are used with equal frequency (randomly) and 0 when only a single word order is used. Starting entropy (Gen0) was always 0.86.[1]

3.2. Results

Fig. 3 shows the proportions of each word order produced by participants in Gen3 in Exps. 1 and 2. Participants in Exp. 1 produced 75% of the dominant word order (SOV), and participants in Exp. 2 produced 70% of the dominant word order (OSV). Neither of these values was significantly different from the initial input (Gen0): (Exp. 1) V=40, p=.56; (Exp. 2) V=33, p=1. Thus, participants in both experiments frequency matched. The second largest percentage of responses in each case matched the non-dominant order, as expected.

[1] Note that Smith and Wonnacott (2010) and Smith et al. (2017) also calculated *conditional entropy* (of marker use given the noun being marked) to quantify the emergence of conditioned variation. In our experiments, we have a single possible locus of conditioning: animacy of the object noun phrase (N2). It is possible that individual participants conditioned word order on N2 animacy, such that they consistently used one order when the object noun was animate and another when it was inanimate (and these pairings could have been different across participants, washing out at the population level). We explored this possibility and found no evidence for such conditioning (see Appendix A). Across all experiments, participants always saw a new alien in the production phase. As such, we are unable to calculate conditional entropy on speaker identity (for further discussion, see Appendix A).

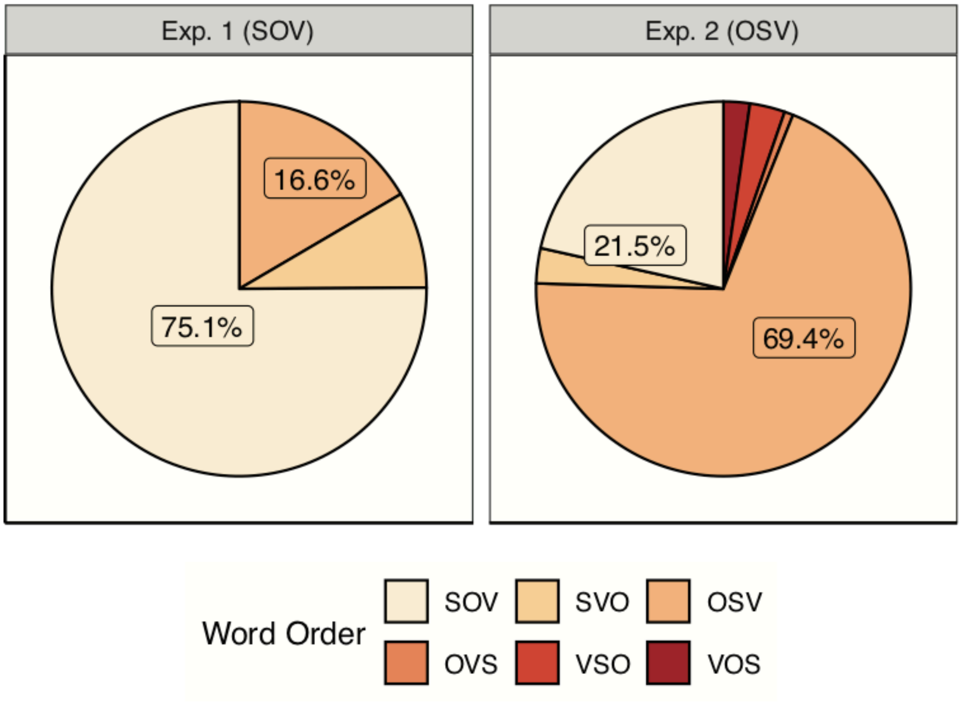


Figure 3. Proportions of word order responses produced in Gen3 in Exps. 1 and 2.

Entropy for all experiments is shown in Fig. 10. There is a slight increase in Exps. 1 and 2 from Gen0, but the two chains remain fairly constant over time and are of equivalently mid-range entropy (1 > *H* > 1.5) by Gen3.

3.3. Discussion

In Exps. 1 and 2, we asked whether participants would converge on the dominant word order or frequency match when provided with input by a single source. In neither case did participants diverge from the initial input by Gen3, as seen in our entropy measures. Thus, participants in these experiments frequency matched across the generations, both when the dominant word order was SOV (Exp. 1) and when it was OSV (Exp. 2). This is consistent with findings of frequency matching in past work when the nature of the variation in the input is relatively simple (Hudson Kam & Newport, 2009). Our study demonstrates that such variability can be maintained over time, across multiple generations of learners. We next ask how frequency matching will be affected over time when we increase the number of sources participants are learning from.

4. Experiments 3-5

In Experiments 3-5, we ask whether participants will frequency match or converge on a single word order when provided with input from four different sources (four alien speakers) instead of one. Tracking input from four sources is more complex than tracking input from a single source. We know that added complexity leads to more regularization (e.g., Hudson-Kam & Newport, 2009). Thus, increasing the size of the language community--and therefore the complexity of the input--might accelerate regularization. Exps. 3 and 4 have the same two dominant word orders as Exps. 1 and 2 (SOV and OSV), respectively. Comparing the four experiments (1-4) will allow us to address the question of how the complexity of the language community (one vs. many input sources) influences learning.

In Experiment 5, we use a different dominant word order--VSO (with SOV non-dominant)--and four alien speakers. By comparing Exps. 3-5 we can ask whether and how convergence patterns differ across the word orders. Will participants always converge on the dominant order (SOV in Exp. 3, OSV in Exp. 4, VSO in Exp. 5), or will they instead converge on the more typologically common word order (SOV)? Since speakers tend to regularize towards typologically preferred patterns (e.g., Culbertson & Newport, 2015; Culbertson et al., 2012; Fedzechkina et al., 2012), it is possible that our participants will be more likely to converge on the more typologically common word order (SOV, Baker, 2001; Dryer, 1992, 2013a; 2013b; Greenberg, 1963) even when it is not the dominant one (Exps. 4 and 5).

4.1. Methods

4.1.1. Participants

Ninety-eight native English speakers recruited from Amazon Mechanical Turk participated in Exps. 3-5 (54 female, 43 male, 1 trans; mean age=35, SD=10, range=19-70, 1 age unknown). There were 33 participants in Exp. 3 (Gen1: 11; Gen2: 12; Gen3: 10), 32 participants in Exp. 4 (Gen1: 9; Gen2: 11; Gen3: 12), and 33 participants in Exp. 5 (Gen1: 12; Gen2: 11; Gen3: 10). An additional 7 participants were excluded for being non-native English speakers (N=6) or for admitting to copying down or taking a picture of (some of) the words during the task (N=1). All participants provided written consent prior to participating and received $3.50 for their participation.

4.1.2. Materials

In Exps. 3-5, participants learned from four different alien speakers in the syntax acquisition phase rather than one. In Exp. 3, the dominant word order was SOV, and the non-dominant word order was OSV. In Exp. 4, the dominant word order was OSV, and the non-dominant order was SOV. In Exp. 5, the dominant word order was VSO, and the non-dominant order was SOV. Each of the four alien speakers produced 7 of the 28 exposure sentences, and the proportion of dominant and non-dominant word orders was always the same across speakers. Thus, in the syntax acquisition phase for the first generation of participants, each alien produced

5 sentences of the dominant word order (71%) and 2 sentences of the non-dominant word order (29%).

4.1.3. Coding

Target event productions were coded as in the previous experiments. Only responses for novel event videos were included in the analysis. In total, 430 of the 462 novel target productions in Exp. 3 (93%), 396 of the 448 novel target productions in Exp. 4 (88%), and 406 of the 462 novel target productions in Exp. 5 (88%) were viable and thus entered into the analysis.

4.1.4. Data analysis

Once again, we used Wilcoxon signed-rank tests to determine whether each experiment diverged significantly from the initial input (Gen0) by Gen3. As we will see, in all three cases, Gen3 diverged from Gen0. In addition, to investigate whether and how convergence varies with the number of input sources, we combined the data from Exps. 1-4 into a logistic mixed-effects model (Baayen, Davidson, & Bates, 2008; Jaeger, 2008) in the lme4 package in R (Bates, 2010), with Word Order (SOV-dominant, OSV-dominant), Number of Speakers (1-alien, 4-alien), Generation (Gen1, Gen2, Gen3), and their interaction as fixed effects. All fixed effects were effect coded.[2] The dependent measure was the proportion of responses that used the dominant word order (conditioned on experiment; see above). We performed backward model comparisons using likelihood-ratio tests (`anova` function in R) to determine the significance of our fixed effects. We first tested for significance of the three-way interaction by comparing the full model with the three-way interaction term (Word Order × Number of Speakers × Generation) against an

[2] Specifically, here and throughout, effects with two levels contained a single term coded as [1, -1], while effects with three levels contained two terms coded as [1, 0, -1] and [0, 1, -1], respectively.

interaction model without the three-way interaction (Word Order × Number of Speakers + Word Order × Generation + Number of Speakers × Generation). We then tested for significance of the three two-way interactions by comparing this same interaction model against models without one of the three two-way interactions (Word Order × Number of Speakers, Word Order × Generation, Number of Speakers × Generation), respectively. We finally tested for significance of the three main effects by comparing a main effects model (Word Order + Number of Speakers + Generation) against models without Word Order, Number of Speakers, or Generation, respectively. The random effects structure for these models included intercepts for participant and item (target event) and slopes for both Word Order and Number of Speakers within items.[3]

Finally, to better understand how convergence varies with word order, we combined the data from Exps. 3-5 into a logistic mixed-effects model in the lme4 package in R, with Word Order (SOV-dominant, OSV-dominant, VSO-dominant), Generation (Gen1, Gen2, Gen3), and their interaction as fixed effects. Both fixed effects were effect coded. The dependent measure was the proportion of responses that used the dominant word order (conditioned on experiment; see above). We performed backward model comparisons using likelihood-ratio tests (`anova` function in R) to determine the significance of our fixed effects. Specifically, we first tested for significance of the interaction by comparing the full model with the interaction term (Word Order × Generation) against a main effects model without the interaction (Word Order + Generation). We then tested for significance of the two main effects by comparing this same main effects model against models without Word Order or Generation, respectively. The random

[3] For the model comparisons to be interpretable, we held the random effects structure constant across the set of models being compared (see, e.g., Barr, Levy, Scheepers, & Tily, 2013). We started by using the maximal random effects structure appropriate for the experimental design (Barr et al., 2013), including, in this case, random intercepts for participant and item and random slopes for Word Order, Number of Speakers, Generation, and their interaction within items. However, these models failed to converge. We therefore pruned each model, one random effect term at a time, until it reached convergence. The reported models include only those random effects that appeared in *all* of the final models. Nevertheless, the pattern of results is similar in the maximal model that converged.

effects structure for these models included intercepts for participant and item (target event) and a slope for Word Order within items.[4]

4.2. Results

Fig. 4 shows the proportions of each word order produced by participants in Gen3 in Exps. 3-5. Participants in Exp. 3 produced 97% of the dominant word order (SOV), participants in Exp. 4 produced 23% of the dominant word order (OSV), and participants in Exp. 5 produced 29% of the dominant word order (VSO). Each of these values was significantly different from the initial input (Gen0): (Exp. 3) V=55, p=.005; (Exp. 4) V=1, p=.003; (Exp. 5) V=3, p=.01. Thus, participants in Exp. 3 converged on the dominant word order, while participants in Exps. 4 and 5 did not. If anything, participants in Exps. 4 and 5 produced more subject-initial responses (SOV and SVO) than appeared in the initial input (Gen0): 72% and 54%, respectively (cf. 29%). Post hoc Wilcoxon signed-rank tests confirmed this impression: (Exp. 4) V=77, p=.003; (Exp. 5) V=47, p=.049. Critically, this tendency was not driven entirely by an increase in SVO responses consistent with their native language of English: Among responses with subject-initial orders produced in Exps. 4 and 5, the proportion of SOV responses was greater than that of SVO responses (62% vs. 38% and 63% vs. 37%, respectively). Interestingly, participants in Exp. 5 also produced a fair number of OSV responses despite this order being absent in the input: 12%.

[4] See Fn. 2.

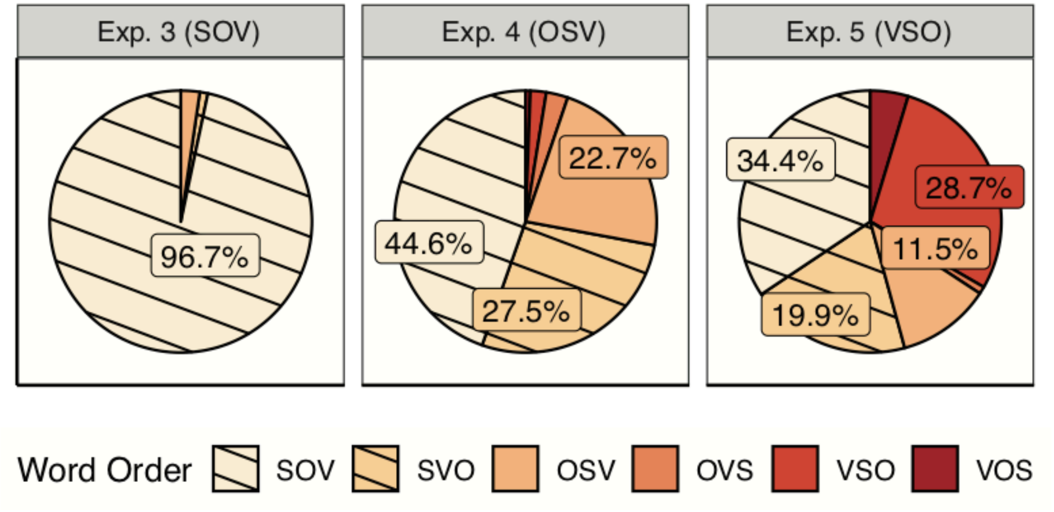


Figure 4. Proportions of word order responses produced in Gen3 in Exps. 3-5. Subject-initial responses are textured with diagonal lines.

4.2.1. Model comparisons

Fig. 5 shows the trajectories of dominant word order responses across the three generations in Exps. 3-5. The model comparisons revealed a significant main effect of Word Order, $\chi^2(2)=35.86$, $p<.001$. Follow-up pairwise models (full model minus the relevant level of Word Order) revealed that the number of dominant word order responses produced by participants in the SOV-dominant chain (Exp. 3) was significantly greater than those both in the OSV-dominant chain (Exp. 4; 87% vs. 38%), $\beta=2.28$ (SE=.41), $z=5.55$, $p<.001$, and in the VSO-dominant chain (Exp. 5; 87% vs. 36%), $\beta=2.43$ (SE=.41), $z=5.91$, $p<.001$, but there was no difference in dominant word order responses between the OSV-dominant chain (Exp. 4) and the VSO-dominant chain (Exp. 5; 38% vs. 36%), $\beta=.18$ (SE=.37), $z=.50$, $p=.62$.

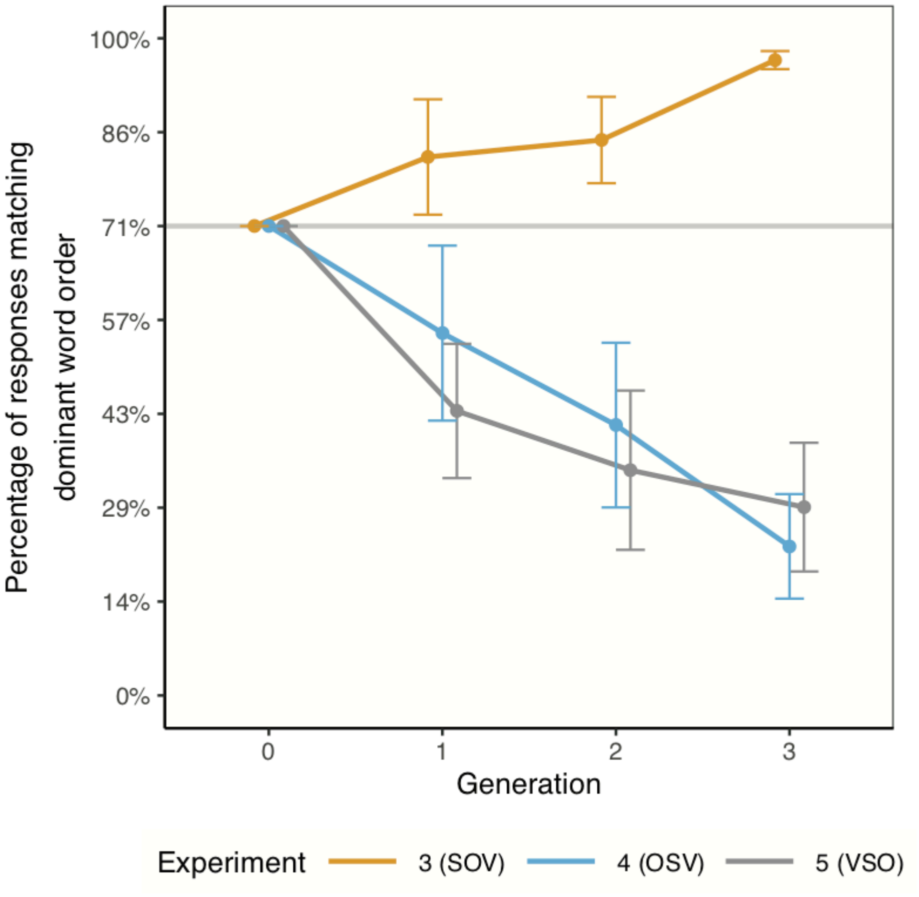


Figure 5. Trajectories of dominant word order responses in Exps. 3-5. Error bars reflect by-subject standard errors.

Fig. 6 shows the trajectories of dominant word order responses across the three generations in Exps. 1-4. The model comparisons revealed a significant main effect of Word Order, such that the number of dominant word order responses produced by participants in the SOV-dominant chains (Exps. 1 and 3) was significantly greater than that in the OSV-dominant chains (Exps. 2 and 4; 79% vs. 53%), $\chi^2(1)=19.08$, $p<.001$. However, this was in the context of a significant Word Order by Number of Speakers interaction, $\chi^2(1)=18.73$, $p<.001$. Follow-up pairwise models within each level of Word Order (with full random effects structure, including random intercepts for participant and item and random slopes for Number of Speakers, Generation, and their interaction within items) revealed that the number of dominant word order responses produced by participants in the SOV-dominant chains (Exps. 1 and 3) was greater in the 4-alien case (Exp. 3) than in the 1-alien case (Exp. 1; 87% vs. 71%), β=-1.13 (SE=.38), z=-2.96, p=.003, while the number of dominant word order responses produced by participants in the OSV-dominant chains (Exps. 2 and 4) was *less* in the 4-alien case (Exp. 4) than in the 1-alien case (Exp. 2; 38% vs. 67%), β=1.06 (SE=.33), z=3.20, p=.001. There were no other main effects or interactions in either model.

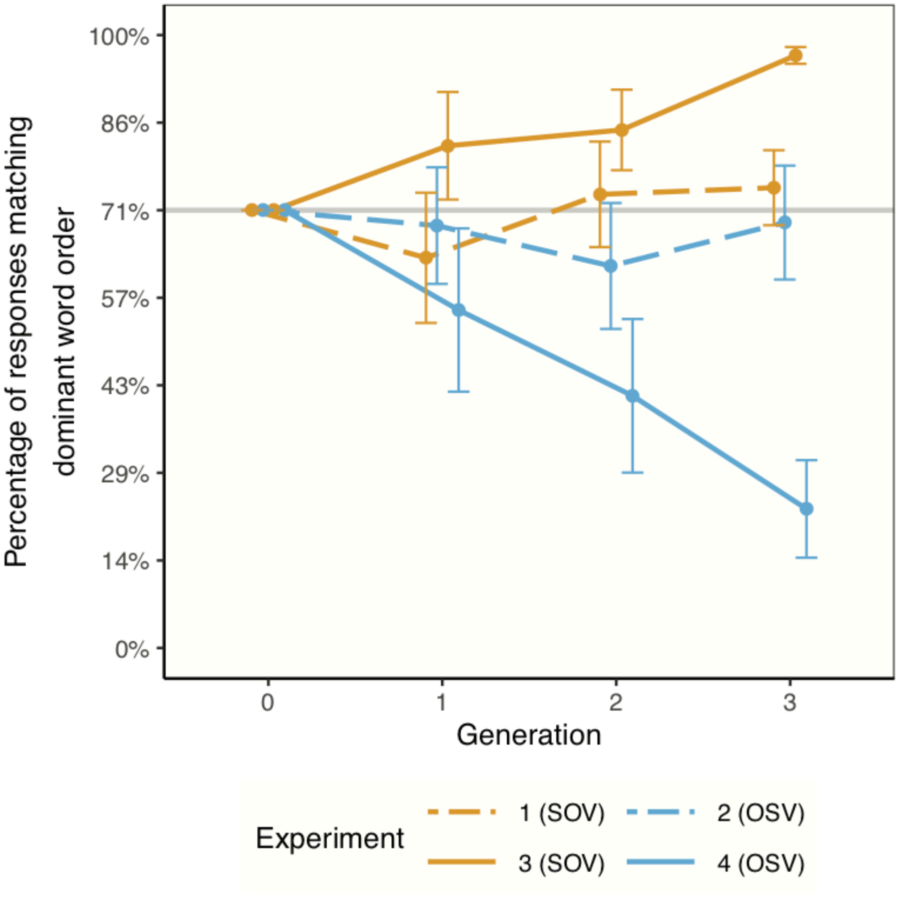


Figure 6. Trajectories of dominant word order responses in Exps. 1-4. Error bars reflect by-subject standard errors.

These results are corroborated by the entropy measures (see Fig. 10). Entropy goes down in Exp. 3 by Gen3, where we see convergence on SOV, to below 0.25. In contrast, entropy starts high and continues to rise in Exps. 4 and 5, where we see much greater response variability and no consistent pattern, and by Gen3 is relatively high ($1.75 > H > 2.5$).

4.3. Discussion

In Exps. 3-5, we asked whether participants would converge on the dominant word order or frequency match when provided input by four different sources instead of one. We found no evidence of frequency matching, in contrast to Exps. 1 and 2. Instead, participants in all three experiments moved away from their input. However, they did so in different ways, as observed in the entropy calculations. Participants in Exp. 3 converged on the dominant word order by Gen3, while participants in Exps. 4 and 5 moved *away from* the dominant word order (Fig. 6). Chains that received OSV- or VSO-dominant input produced a wider range of word order responses, with a general tendency toward typologically more common subject-initial responses (SOV and SVO). This tendency was not due solely to participants resorting to SVO order, consistent with their native language of English, as the majority of these subject-initial responses were SOV. We will discuss the implications of this pattern in the general discussion.

Overall, our results suggest that it is difficult to track variation in the input when learning from multiple speakers relative to learning from a single speaker: When provided with input by one speaker, participants frequency matched (Exps. 1 and 2), but not when provided with input by four speakers (Exps. 3-5). However, the relevant contrasts (Exps. 1 vs. 3 and 2 vs. 4) also differed in another way. Not only were there fewer speakers in Exps. 1 and 2, but there were also more tokens per speaker. Specifically, the single alien speaker in Exps. 1 and 2 provided four times as much input than each of the four alien speakers in Exps. 3 and 4. In Exp. 6, we equate for the amount of input provided by each speaker across the two community sizes (one vs. four aliens) to assess how these two factors trade off in catalyzing language change.

5. Experiments 6 and 7

In Experiments 6 and 7, we again ask whether participants will frequency match or converge on a single word order when provided input by four different sources (four alien speakers, as in Exps. 3-5) but at four times as much input per source (28 sentences each, as in Exps. 1 and 2). Perhaps by giving participants sufficient input per source will we reduce the complexity of the tracking problem, allowing them to frequency match even when learning from multiple sources. Exp. 6 has the same dominant word order as Exps. 1 and 3 (SOV), and Exp. 7 has the same dominant word order as Exp. 5 (VSO). Comparing these experiments will allow us to address the question of how both the number of speakers and the amount of input trade off to influence convergence.

5.1. Methods

5.1.1. Participants

Fifty-nine native English speakers recruited from Amazon Mechanical Turk participated in Exps. 6 and 7 (34 female, 25 male; mean age=38, SD=10, range=23-67). There were 30 participants in Exp. 6 (Gen1: 11; Gen2: 9; Gen3: 10) and 29 participants in Exp. 7 (Gen1: 9; Gen2: 11; Gen3: 9). An additional 4 participants were excluded for being non-native English speakers (N=1) or for admitting to copying down or taking a picture of (some of) the words during the task (N=3). All participants provided written consent prior to participating and received $5.00 for their participation.

5.1.2. Materials

In Exps. 6 and 7, participants again learned from four different alien speakers in the syntax acquisition phase, as in Exps. 3-5. However, the amount of input provided by each alien was quadrupled so that each alien produced the same number of sentences as the single alien in Exps. 1 and 2. Specifically, rather than producing 7 sentences, each alien produced the same 28 exposure sentences (20 dominant, 8 non-dominant), for a total of 112 sentences overall. Exp. 6 was designed to be compared to Exps. 1 and 3, and thus used the same dominant and non-dominant word orders (SOV and OSV, respectively); Exp. 7 was designed to be compared to Exp. 5, and thus used the same dominant and non-dominant word orders (VSO and SOV, respectively; see Table 2).

5.1.3. Coding

Target event productions were coded as in the previous experiments. Only responses for novel event videos were included in the analysis. In total, 368 of the 420 novel target productions (88%) in Exp. 6 and 372 of the 406 novel target productions in Exp. 7 (92%) were viable and thus entered into the analysis.

5.1.4. Data analysis

Again, we used Wilcoxon signed-rank tests to determine whether Exps. 6 and 7 diverged significantly from the initial input (Gen0) by Gen3. In addition, in order to investigate whether and how convergence varied by both the number of input sources and amount of input per source, we built three separate logistic mixed-effects models in the lme4 package in R. The first two models investigated how the amount of input per source affected convergence by holding the number of input sources constant. For this, we separately compared Exp. 3 with Exp. 6 and Exp. 5 with Exp. 7, using Input Amount (7-sentence, 28-sentence), Generation (Gen1, Gen2, Gen3), and their interaction as fixed effects. Both fixed effects were effect coded.[5] The dependent measure was the proportion of responses that used the dominant word order (conditioned on experiment; see above). We performed backward model comparisons using likelihood-ratio tests (`anova` function in R) to determine the significance of our fixed effects. We first tested for significance of the interaction by comparing the full model with the interaction term (Input Amount × Generation) against a main effects model without the interaction (Input Amount + Generation). We then tested for significance of the two main effects by comparing this same main effects model against models without Input Amount or Generation, respectively. The random effects structure for the models comparing Exp. 3 and Exp. 6 included intercepts for participant and item (target event) and slopes for both Input Amount and Generation within items.[6] The random effects structure for the models comparing Exp. 5 and Exp. 7 included intercepts for participant and item (target event) and slopes for Input Amount, Generation, and their interaction within items.

The third model investigated how the number of input sources affected convergence when the amount of input per source was held constant. For this model, we compared Exps. 1 and 6, with Number of Speakers (1-alien, 4-alien), Generation (Gen1, Gen2, Gen3), and their

[5] Exp. 7 was collected at a later time than Exps. 1-6. Thus, comparisons between it and the other experiments should be interpreted with caution.

[6] See Fn. 2.

interaction as effect-coded fixed effects. We followed the same backward model comparison procedure as above. The random effects structure for these models included intercepts for participant and item (target event) and a slope for Number of Speakers within items.[7]

5.2. Results

Fig. 7 shows the proportions of each word order produced by participants in Gen3 in Exps. 6 and 7. Participants in Exp. 6 produced 65% of the dominant word order (SOV), and participants in Exp. 7 produced 63% of the dominant word order (VSO). Neither of these values was significantly different from the initial input (Gen0): (Exp. 6) V=17, p=.31; (Exp. 7) V=12, p=.25. Thus, participants in both experiments frequency matched. In Exp. 6, the second largest percentage of responses matched the non-dominant order (OSV). In Exp. 7, the second largest percentage of responses had VOS order, followed by the non-dominant order (SOV) in third.

[7] See Fn. 2. In this case, the models converged with random intercepts for participant and item and a random slope for either Number of Speakers within items or Generation within items, but not both. However, the two modeling strategies yielded similar results. We therefore report only the output of the former.

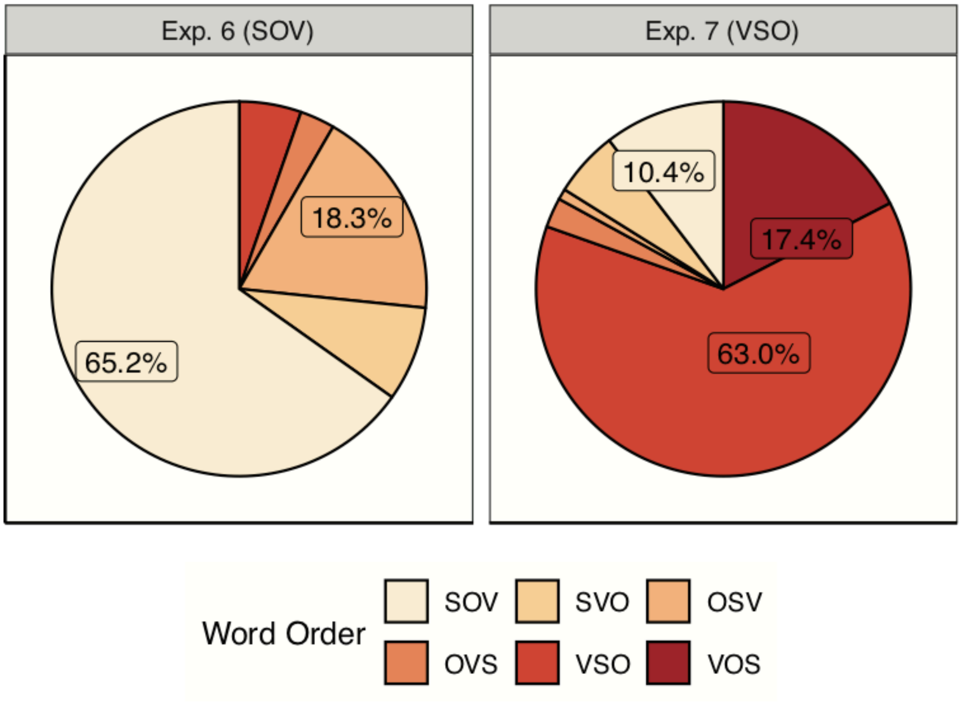


Figure 7. Proportions of word order responses produced in Gen3 in Exps. 6 and 7.

5.2.1. Model comparisons

Fig. 8 shows the trajectories of dominant word order responses across the three generations in Exps. 1, 3, and 6. The model comparisons revealed a significant main effect of Input Amount, such that the number of dominant word order responses produced by participants in the 4-alien chain with 7 sentences per alien (Exp. 3) was significantly greater than that in the 4-alien chain with 28 sentences per alien (Exp. 6; 87% vs. 68%), $\chi^2(1)=15.01$, $p<.001$.

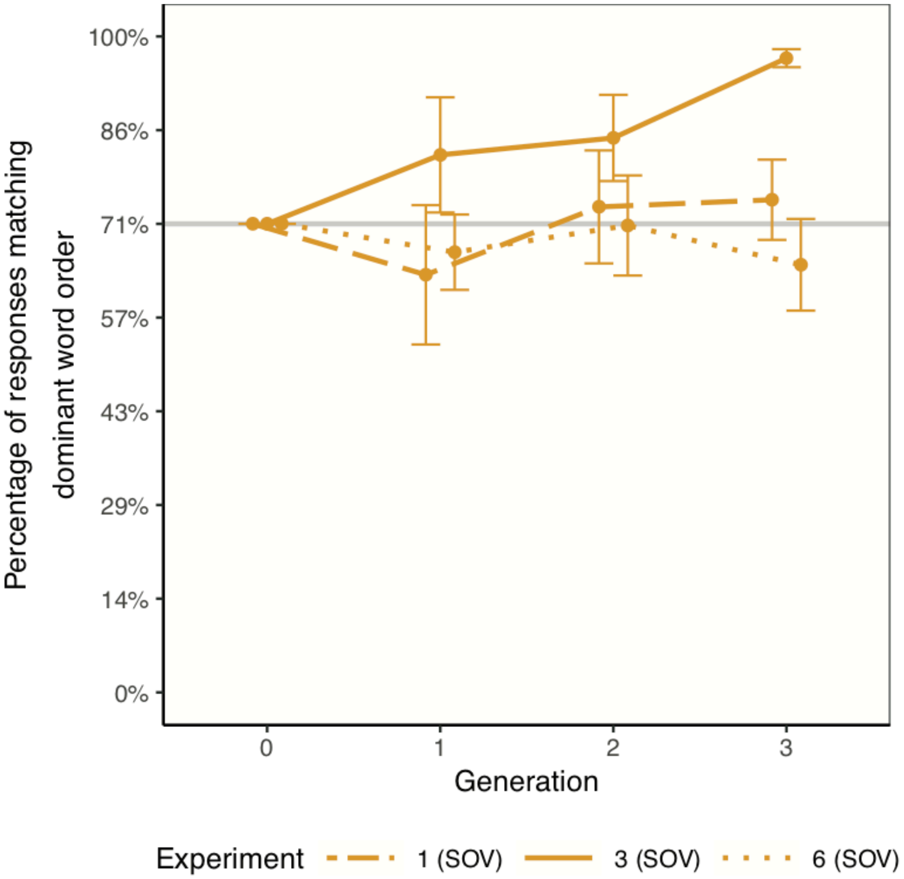


Figure 8. Trajectories of dominant word order responses in Exps. 1, 3, and 6. Error bars reflect by-subject standard errors.

Fig. 9 shows the trajectories of dominant word order responses across the three generations in Exps. 5 and 7. The model comparisons revealed a significant main effect of Input Amount, such that the number of dominant word order responses produced by participants in the 4-alien chain with 7 sentences per alien (Exp. 5) was significantly less than that in the 4-alien chain with 28 sentences per alien (Exp. 7; 36% vs. 63%), $\chi^2(1)=11.08$, $p<.001$. There were no other main effects or interactions in the models.

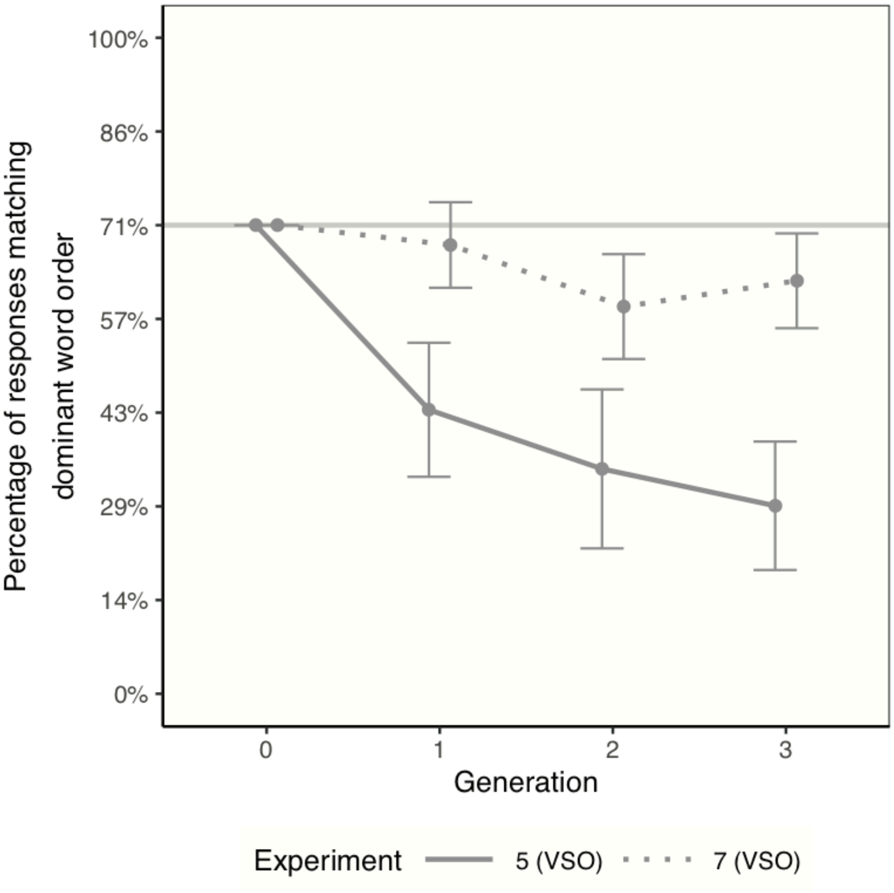


Figure 9. Trajectories of dominant word order responses in Exps. 5 and 7. Error bars reflect by-subject standard errors.

As before, entropy remains fairly constant over time, within mid-range levels (1 > *H* > 1.5), in cases of frequency matching (Exps. 6 and 7; see Fig. 10).

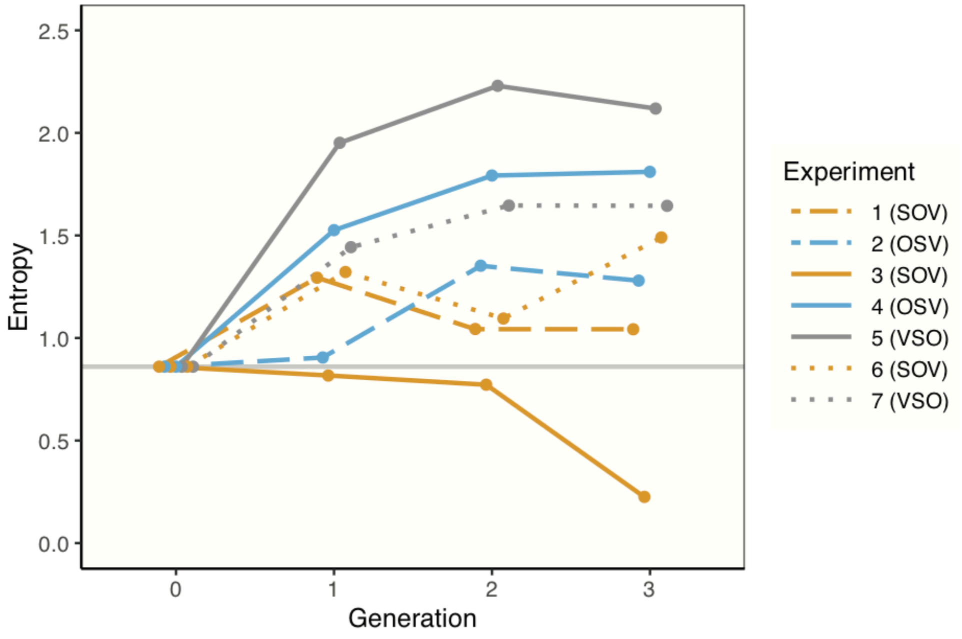


Figure 10. Entropy of word order by generation in Exps. 1-7.

5.3. Discussion

In Exps. 6 and 7, we asked whether participants would converge on the dominant SOV word order when provided input by four different sources (as in Exps. 3-5) but when the amount of input per source was quadrupled (as in Exps. 1 and 2). We find that under these conditions participants frequency matched instead, as seen in the entropy measures, even when the dominant word order was one that with four aliens and sparser input participants could not previously converge on (cf. Exp. 5). Thus, participants *are* able to reproduce variation in the input even when there are multiple speakers providing that input but only when they are given enough evidence about the usage pattern of each speaker (Exps. 1, 2, 6, and 7). This suggests that the occurrence of linguistic regularization depends not only on the number of input sources but crucially also on the amount of input that each source provides: When there is not enough evidence to appropriately track the variability within an individual, speakers regularize. We discuss this idea more below.

6. General Discussion

We used artificial grammar learning in a paradigm with diffusion chains to investigate the transmission of word order variation across multiple generations of learners. Across seven experiments, we manipulated (1) whether the input was provided by one speaker or four speakers, (2) what the dominant word order was in each chain, and (3) how much input each speaker provided (see Table 2). Three clear findings emerged.

First, the size of the language community affected regularization. When there was just one speaker, participants frequency matched, both when the dominant order in the input was SOV (Exp. 1) and when it was OSV (Exp. 2). When there were multiple speakers providing the same amount of input, participants no longer reproduced the variability that was present in the input.

Second, when participants did not frequency match, their behavior depended on the dominant word order in the input. When the dominant order was SOV (Exp. 3), participants quickly converged on this order, effectively eliminating the subordinate order by the third generation. When the dominant order was OSV (Exp. 4) or VSO (Exp. 5), the data pattern was more complex. By the third generation, there was a sharp decline in the use of the dominant order, a moderate increase in the frequency of the subordinate order (SOV), and the appearance of orders not provided in the input. This pattern is not due to an inability to learn or produce a non-subject-initial word order: Participants preserved the use of OSV word order for three generations when the input came from a single speaker (Exp. 2).

Finally, we found that community size interacted with the amount of input provided. When there were multiple speakers *but* more data was provided per speaker, participants again frequency matched (Exps. 6 and 7).

We discuss each of these three findings in more detail below.

6.1. Frequency matching breaks down in the face of complexity

When the input came from a single source, learners frequency matched, reproducing the variability from the input in their own output (Exps. 1 and 2). Critically, while SOV was dominant in the input for Exp. 1, OSV was dominant in Exp. 2. OSV is extremely rare in spoken languages: Just 4 out of 1377 languages in the World Atlas of Language Structures (WALS) have a dominant OSV order. In comparison, SOV is the dominant order in 565 languages (Dryer, 2013). These experiments demonstrate that learners *can* learn and produce different orders, faithfully reproducing the relative frequency of a form over multiple generations, even if it is typologically marked.

When the same amount of input is distributed across multiple speakers (Exps. 3-5), learners fail to frequency match, perhaps because they are unable to track the relative frequency of the two structures. Curiously, their behavior differs depending on the word orders provided in the input. When the dominant order is SOV, learners rapidly converge on that order (Exp. 3). When the dominant order is OSV (Exp. 4) or VSO (Exp. 5), participants produce more variable word orders. There are two features of this pattern that necessitate further discussion. The first is why convergence occurs at all, given the capacity of adult learners to frequency match. The second is why it occurs only when SOV order is dominant.

We interpret the convergence we saw in light of the prior literature on artificial language learning and variability. Taken as a whole, this literature suggests that when complexity of the input is high, relative to the processing capacity of the learner, learners simplify the language by eliminating exceptions or creating new regularities rather than faithfully reproducing the variation that they encounter (Culbertson et al., 2012; Culbertson & Newport, 2015; Fedzechkina et al., 2012; Hudson Kam & Newport, 2005; 2009). This generalization explains why frequency matching in adults breaks down when the input contains a very large number of different forms, and why young children, who have less processing capacity, are more likely to regularize than adults (Hudson Kam & Newport, 2005; 2009). What our results demonstrate is that there is another form of complexity – the complexity of the community – that can interfere with frequency matching. At first blush, this finding is surprising: Participants in our study could easily have ignored the four different aliens who provided the input. There was no difference in how each alien used the two word orders and no systematic difference in the content of each alien's messages. Why should the mere presence of different speakers be so disruptive?

We see one plausible explanation for this disruption. To frequency match, learners must: 1) define the set of things they are tracking (e.g., SOV vs. OSV word order), 2) define the context over which they will generalize (e.g., the language vs. the dialect), and then 3) track those items within that context. Typically, learning problems are ambiguous and thus learners may have different hypotheses about the first two steps. The learner may select the wrong structures (e.g., searching for generalizations like four-legged creatures appear first in sentences) or they may select the wrong context (e.g., focusing on sentences spoken inside vs. outside). We suspect that by introducing four aliens, we disrupted this second step of the process, leading participants to define each alien as a separate context. In essence, they were trying to infer the structure of the input for each speaker from a sample of 7 sentences. Participants in Exps. 3-5 were unable to do this, and what they did instead depended on the dominant word order in the input. We turn to this idea next.

6.2. The effect of complexity depends on the dominant word order

In the multiple-speaker, less data experiments, performance depended largely on the dominant word order. When participants encountered many instances of SOV word order, they latched onto it and eliminated the non-dominant order, much as children do when confronted with unpredictable variation. This pattern suggests that, even with limited input, learners were able to correctly define one of the forms to be tracked (SOV) and observe that it was commonly used in the input. However, when the dominant order was OSV or VSO, participants failed to converge or frequency match. Under these conditions, it looks like they had difficulty constructing a stable hypothesis about word order and instead produced orders they did not receive in the input and gradually used fewer instances of the dominant order. There at least two ways to interpret this performance. First, it could reflect a difficulty in determining the description under which to track the utterances. Second, it could be the result of integrating a small amount of data with a strong prior belief about the nature of languages. If a learner believes that SOV (or SVO) word order is far more likely than OSV or VSO, then seven sentences may be insufficient to override this hypothesis. Both explanations raise the question of why OSV and VSO orders would be different from SOV. There are at least two possible reasons. The first is rooted in the native language of our participants, the second in typological patterns.

We suspect that our findings of difficulties with OSV and VSO compared to SOV do *not* solely reflect native language biases. Our participants were English speakers, who have spent their lives using an SVO language. While the SOV order is like English in one respect (the subject is at the beginning of the sentence), it is dissimilar from SVO in other ways. Neither of the two local orderings in SVO, an SV unit and a VO unit (Carroll, 1978; MacWhinney, Bates, & Kliegl, 1984), is present in the SOV order. There is evidence that English speakers are sensitive to these local groupings and rely on them when confronted with non-SVO input. Specifically, MacWhinney and colleagues (1984) presented English speakers with sentences containing an agent, a patient, and an action, following one of the six possible orders (e.g., "The eraser the pig chases," "Licks the cow the goat," etc.), and asked them to report which noun was the agent, or the one doing the action. English adults showed a strong bias to interpret NNV sentences as OSV (not SOV) and VNN sentences as VOS (MacWhinney, Bates, & Kliegl, 1984).[8] Thus, if participants were treating the new language in our experiments as a variant on English, we might expect them to succeed in converging on the OSV language, which contains an SV unit (Exp. 4), but not the SOV language (Exp. 3).

The second class of explanations is rooted in typological patterns across languages. The SOV word order is the most common order in the languages of the world (48% in WALS), closely followed by the SVO order (41%). In contrast, OSV and VSO orders are uncommon (7% and <1%, respectively). The origins of this typological pattern are unclear. It could reflect biases in the way in which linguistic structure is built, interpreted, or learned. Or it could reflect properties of our conceptual structures that have existed prior to language (in development and in human history). For example, in both SOV and SVO orders, the subject occurs at the beginning of the sentence. This could reflect the fact that causal agents generate the action and thus exist prior to the event, or it could reflect the salience of animate actors in human thought (Bever, 1970; Dowty, 1991; Halliday, 1967; MacWhinney, 1977; Osgood, 1980; Prince, 1981). Our findings closely mirror those of single-generation studies which find that adults will reorganize variable input and make it more consistent with typological generalizations, even when this means abandoning the form that is most frequent in the input (Culbertson et al., 2012;

[8] This bias for OSV observed in MacWhinney et al. (1984) might partially explain the increase in OSV responses in Exp. 5 despite its being absent in the input.

Fedzechkina et al., 2012). Thus, cognitive biases seem to manifest themselves robustly in intergenerational transmission, particularly in learning situations where the language community is more complex and there is greater variability in the input. Future work with users of a non-S-initial language would provide a stronger test of the interaction between learning biases and transmission.

Of course, it is possible that we might have observed different patterns had we not averaged the responses of the participants to generate the new input (see also section 6.3 below). However, we suspect that this did not change our findings. Pilot studies in our lab ran a modified version of our paradigm, where we maintained inter-speaker variability by turning each participant's output into one alien speaker's input language. We observed the same broad patterns (see Ziegler, Kocab, & Snedeker, 2017; data linked in Appendix B). Specifically, we find frequency matching in a low-complexity environment (when the input comes from one speaker), a failure to frequency match in a high-complexity environment (when the input comes from multiple speakers), and rapid convergence over three generations to the SOV order. Future studies could explore this variant of the paradigm in more detail.

6.3. Limitations of the present study

One limitation of our experiments concerns the filtering procedure for how the input was generated for each generation. In all experiments, we presented participants with two word orders, one dominant and one non-dominant. In our statistical analyses, we coded the output in terms of the order that the participant actually used. However, in creating the input for the subsequent generation, we collapsed into one category all responses that did not follow the dominant order, but met the inclusion criteria, and presented those as instances of the non-dominant word order. This re-coding was done for two reasons: (1) to be able to tightly control the proportions of the different word orders in each generation's input and (2) to minimize the likelihood of our English monolingual participants shifting toward the SVO word order across generations.

As a result, our findings cannot address the degree to which *new* word orders may arise over multiple generations of language users. In addition, our artificial language did not include any morphological markers (languages with case marking systems tend to have freer word orders), further constraining any generalizations to word order patterns in the world's languages. The breakdown of responses for Gen3 in Exps. 1 and 2 (Fig. 3) shows that these alternate word orders (not including SOV or OSV) were quite rare (~8%). They were also infrequent in Gen1 and Gen2. However, it is possible that these other orders would have increased over time, if included in the input. In fact, given the ability of the adults to probabilistically produce a large number of alternate forms in single-generation artificial grammar studies (Hudson Kam & Newport, 2009), there is no reason to think that these new forms in the input would be suppressed. This limitation in our filtering procedure, however, cannot explain the primary findings of this paper. The same procedure was used for both Exp. 1 and Exp. 2, and yet the final output in these experiments was different, reflecting the relative probability of each form in the original input to the chain. Thus, we know that participants were differentially faithful to the input, frequency matching in some cases and not in others. Similarly, the filtering procedure cannot account for the differences between the single- and multiple-speaker conditions, since the filtering procedure was identical in all experiments.

6.3.1. Previous findings

In a recent paper, Smith et al. (2017) extended the paradigm used in Smith and Wonnacott (2010) looking at plural marking of nouns to test the effect of adding multiple individuals to each link in the diffusion chain. Adult participants were tested in one of three conditions: (1) *one-person*, (2) *two-person identity-known*, and (3) *two-person identity-unknown*. In the one-person condition, each participant served as a single link in the diffusion chain. The first participant in each chain was trained on the initial input language and their output served as the input for the next participant in the chain. In both of the two-person conditions, each generation in the chain consisted of two participants who learned the same input language. Their individual output was then combined to generate the new input language for the next generation. In the two-person identity-known condition, participants were provided with a picture of one of two aliens alongside a speech bubble during the training. Half of the examples came from one alien and the other half from the other alien. During the testing phase, one of the participants was asked to provide the label produced by alien 1 and the other participant was asked to provide the label produced by alien 2. In the two-person identity-unknown condition, descriptions were presented in a dislocated speech bubble with no information about the speaker; thus, participants were unaware that there were multiple speakers.

Smith et al. (2017) observe that a system of conditioned variance emerges more quickly in the one-person condition compared to the two-person conditions. Conditioned variance is a form of regularization that occurs when a marker is used in a consistent way contingent on some other aspect of the stimulus. In this study, use of the plural marker was contingent on the noun it appeared with. At first blush, these results may appear to be at odds with the results of our experiments: (1) Smith and colleagues observe a form of faster convergence in their single-speaker condition whereas we observe frequency matching (preserved variation) in our single-speaker experiments; (2) Smith and colleagues observe slower convergence in their multiple-speaker conditions and we observe faster convergence in our multiple-speaker experiments. However, there are several important differences between Smith et al. (2017) and our study, which we detail below.

First, there may be a temptation to draw parallels between the single-speaker chains in Smith et al. (2017) and our single-alien case, and between their two-speaker chains and our multiple-speaker chains, which might lead to the conclusion that the two papers present diverging results. However, in our single-alien case, the provided input is the averaged output of many learners. Thus, our single-alien case is more similar to the Smith et al. (2017) two-person identity-unknown condition, where variation is preserved for longer compared to the single-speaker chains. Note also that the effect of a single participant's output is less strong in our single-speaker chains compared to previous experiments (because each individual's output does not make up the entirety of the new input, and is instead combined with the output of other participants in that generation).

The two studies also vary in how regularization manifests. Smith et al. (2017) find that variation gradually comes to be conditioned upon the nouns that are being marked. In contrast, regularization in our study consists of the gradual disappearance of alternate word orders over time. This reflects differences between the two learning problems. While systems of conditioned variance are frequently found in morphology (and thus Smith et al. are well-positioned to observe lexically-conditioned variance), this is not the case for word order. Mixed word order systems are relatively rare in the world's languages (Dryer, 2013) and when they do occur, the basis for variation is typically syntactic rather than lexical (e.g., whether the verb is in a main or embedded clause or whether there is an auxiliary verb, as in seen in German and Dutch). The studies in this paper were not designed to test these forms of conditioned variance.

Our study does not have a clear parallel to the Smith et al. (2017) multiple-speaker identity-known condition (we did not assign to any of our participants a single alien to track). We note, however, that the Smith et al. (2017) experimental study treats the input provided by the multiple speakers as separate streams. In other words, one participant is assigned to track one speaker, and the other participant the second speaker. One way of conceptualizing this condition is one where the two learners are in separate communities, where in each one a single speaker provides the input. This design is very different from our multiple-speaker chains. These methodological differences, as well as the difference in domain (as discussed above) may drive the difference in speed of convergence we observe in the two studies (faster in ours with word order, slower in Smith et al. with plural marking).

In pilot work (Ziegler, Kocab, & Snedeker, 2017), we ran a version of our paradigm that is more similar to the Smith et al. (2017) multiple-speaker identity-known condition (data linked in Appendix B). In these experiments, participants learn from six alien speakers who provide the input. The output of each participant becomes the input of one alien speaker. Thus, there is a one-to-one mapping between each participant speaker and each alien speaker. We observe the same broad pattern of results, where variability in the input is not preserved when it comes from multiple speakers.

Our results have clear implications for artificial grammar learning studies. The number of speakers who produce the input affects what is learned. When confronted with a more complex language community, participants may find it more challenging to track the relative frequency of each form. As a result, they may eliminate variation or fail to reproduce typologically marked forms.

6.4. The effect of the amount of data per speaker

The presence of multiple speakers prevented participants from being able to frequency match (Exps. 3-5). But when the amount of data per speaker was increased (Exps. 6-7), participants were able to frequency match, as in Exps. 1 and 2, correctly reproducing the proportions of the two orders in the input: SOV and OSV (Exp. 6) and VSO and SOV (Exp. 7).

In section 6.1, we provided one possible explanation for why learners failed to frequency match when there were four speakers: The presence of multiple speakers may have led our participants to treat each alien’s utterances as a separate context. Given only a small set of sentences produced by each speaker, participants may not have been able to track the input sufficiently well to frequency match. If this hypothesis is correct, then we might expect that when participants had as much data on each speaker as they had in Exps. 1 and 2, they would again be able to frequency match. This is exactly what we found. When we increased the tokens per speaker (from 7 to 28 sentences), our participants faithfully reproduced the variability in the input. In sum, the presence of complexity leads frequency matching to break down, with more marked word orders dropping out of the language. As our findings from Exps. 6-7 demonstrate, we can rescue frequency matching in adults by providing more data, even when the dominant order is more marked (Exp. 7).

An open question is how these forces -- complexity and the amount of data -- play out for actual learners in different language communities acquiring different kinds of regularities. Answering this question in each case would require knowing how much data the learner receives relative to the complexity of the input. This is further complicated by the varying forms of complexity: the number of speakers (and their uniformity or variability), the range of grammatical contexts which must be tracked (e.g., declaratives vs. questions, or transitives vs. intransitives vs. datives), the lexicalized exceptions to many rules, and the varying capacities of the learners (adults vs. children).

There is some evidence suggesting that both child and adult learners can identify speakers. A recent study taught adults and children a semi-artificial language with known nouns, and novel verbs and two particles. The variants were conditioned on speaker identity, where one speaker produced particle 1 and the other speaker produced particle 2, either deterministically or probabilistically. Both children and adults were able to track speaker identity cues. However, when regularization was observed, it was not conditioned on speaker identity. Children frequently increased their use of one particle over the other but did so equally for the two speakers. Adults tended to condition particle use on lexical items rather than the speaker (Samara et al., 2017). While this finding suggests that learners are sensitive to speaker identity, due to the small number of speakers in this paradigm (and overlap in lexical items), learners may not be receiving sufficient evidence to learn that the variation in the two speakers' production should be treated as two distinct dialects or languages. Nonetheless, the results, as well as the ones presented in this paper, suggest that speakers (their number, their identity and variability) present a source of complexity that participants track and which shapes learning.

How might these results on bear on language acquisition? If children rely on a small number of sources for their input (e.g., their parents), these findings might predict that children will probability match to their input. Indeed, some have argued that children do this kind of probability matching: When adults produce variable structures, children tend to match this variability (e.g., Ambridge, Pine, Rowland, Freudenthal, & Chang, 2014; Ambridge et al., 2013; Ambridge et al., 2012).

However, many children, particularly those in non-WEIRD (Western, educated, industrialized, rich, and democratic) contexts, receive input from multiple speakers. Children often encounter many speakers in their socially complex environments. Presumably, in these circumstances, children may also receive sparser data about the distribution of variable forms in each individual. As such, another prediction one could make is that, given this social complexity, children should regularize. This prediction is in line with findings from creolization and other natural experiments, such as the case of Simon and Nicaraguan Sign Language, where child learners tend to regularize variable input (e.g., Bickerton, 1981; 1984; Gleitman & Newport, 1995; Hudson Kam & Newport, 2009; Newport, 1990; Pinker, 1994; Romaine, 1988; Senghas & Coppola, 2001; Senghas, Kita, & Özyürek, 2004; Singleton & Newport, 2004).

We suspect that natural environments provide cases of both types: dense input from a single speaker and sparser input from multiple speakers. Linking the actual input children receive and their output to experimental manipulations like ours would require decades of more work that systematically investigates complexity (along with a definition of complexity that would allow us to equate across different complexity types), token-type frequency, speaker variability, learner characteristics, and how they interact.

7. Conclusion

In this series of experiments, we implemented an iterated learning paradigm to explore the preservation of word order variation across multiple generations of learners. We went beyond prior studies of multi-generational diffusion chains by looking at a new phenomenon, variability in word order, and adding a new manipulation, the number of speakers who provide the input. Our findings are consistent with the prior artificial grammar single-generation studies that find that frequency matching is affected by complexity. When complexity is low relative to the amount of data in the input, adult learners frequency match. When complexity is high relative to the amount of data in the input, adult learners regularize. We observe strong effects of cognitive bias for word order, such that convergence occurred when the dominant order in the input was SOV, which is the most typologically prevalent word order observed in the world's languages.

While our experiments looked at just a single domain, the syntactic ordering of arguments, our findings regarding the effect of complexity on regularization have implications for natural cases of language learning. Real language learners receive far more input than the learners in these studies. However, real languages contain multiple levels of representation, and they are used in very complex social environments and are acquired by children who have more limited processing capacities. Studies of artificial language learning can help us understand how these pressures conspire to produce the peculiar mix of regularity and variation that characterizes natural languages.

Acknowledgements

This work was supported by a Pershing Square Fund grant from Harvard’s Foundations of Human Behavior initiative (to JZ). Special thanks to Melissa Kline for sharing the codebase, to Hannah Lam for stim creation, and to SnedLab/LDS and audiences at BUCLD 2017 for helpful discussion of this work.

References

Ambridge, B., Pine, J. M., Rowland, C. F., & Chang, F. (2012). The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language*, *88*, 45-81.

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & Bidgood, A. (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*, 47-62.

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., & Chang, F. (2014). Avoiding dative overgeneralisation errors: semantics, statistics or both? *Language, Cognition and Neuroscience*, *29*, 218-243.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Baker, M. (2001). The atoms of language: The mind's hidden rules of grammar.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley. Bickerton, D. (1981). *Roots of language*. Ann Arbor: Karoma Press.

Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, *112*, 2097-2102.

Carroll, J.M. (1978). Sentence perception units and levels of syntactic structure. *Perception and Psychophysics*, *23*, 506-514.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Comrie, B. (1989). *Language universals and linguistic typology,* 2nd edition. Chicago: University of Chicago Press.

Croft, W. (2003). *Typology and universals*, 2nd edition. Cambridge: Cambridge University Press.

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

Culbertson. J., & Newport, E. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71-82.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1-12. doi:10.3758/s13428-014-0458-y

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*,*67*, 547–619.

Dryer, M. S. (2005). The order of subject, object and verb. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures* (pp. 330–333). Oxford, UK: Oxford University Press.

Dryer, M. S. (2013). Order of Subject, Object and Verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online*. Leipzig: Max Planck Institute for evolutionary anthropology. Available at http://wals.info/chapter/81. Accessed February 8, 2017.

Erguvanli, E. E., & Taylan, E. E. (1984). *The function of word order in Turkish grammar* (Vol. 106). University of California Press.

Fedzechkina, M., Jaeger, T.F., & Newport, E.L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*, 17897-17902.

Flaherty, M., & Senghas, A. (2011). Numerosity and number signs in deaf Nicaraguan adults. *Cognition*, *121*427-436.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word order variation. *Psychological Science*, *24*, 1079–1088.

Givón, T. (1985). Iconicity, isomorphism and non-arbitrary coding in syntax. *Iconicity in syntax*, 187-219.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3-55.

Gleitman, L. R., & Newport, E. L. (1995). The invention of language by children: Environmental and biological influences on the acquisition of language. In L.R. Gleitman & M. Liberman (Eds.), *An invitation to cognitive science, Vol 1: Language* (pp. 1-25). Cambridge, MA: MIT Press.

Göksun, T., Hirsh-Pasek, K., & Michnick Golinkoff, R. (2010). Trading spaces: Carving up events for learning language. *Perspectives on Psychological Science*, *5*, 33-42.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*, 341-363.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.

Griffiths, T. L., & Kalish, M.L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441-480.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Science*, *105*, 9163–9168.

Grimshaw, J. (1981). Form, function and the language acquisition device. In C.L. Baker, J. McCarthy (Eds.), *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods, 48*(3), 829-842. doi:10.3758/s13428-015-0642-8

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition*, *129*, 1–17.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, *3*, 199– 244.

Hartshorne, J. K., O'Donnell, T. J., Sudo, Y., Uruwashi, M., Lee, M., & Snedeker, J. (2016). Psych verbs, the linking problem, and the acquisition of language. *Cognition*, *157*, 268-288.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

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*, 10681-10686.

Kocab, A., Lam, H., & Snedeker, J. (2017). When cars hit trucks and boys hug girls: The effect of animacy on word order in gestural language creation. *Cognitive Science*, *42*, 918-938.

Kocab, A., Senghas, A., & Snedeker, J. (2016). The emergence of temporal language in Nicaraguan Sign Language. *Cognition*, *156*, 147-163.

Lakusta, L., Wagner, L., O'Hearn, K., & Landau, B. (2007). Conceptual foundations of spatial language: Evidence for a goal bias in infants. *Language Learning and Development*, *3*, 179-197.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, Vol. 71. Cambridge, UK: Cambridge University Press.

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology*, *60*, 291–318.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS One*, *5*(1), e8559.

MacWhinney, B. (1977). Starting points. *Language*, *53*, 152–168.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Behavior*, *3*, 127-150.

MacWhorter, J. (2002). What happened to English? *Diachronica*, *19*, 217–272.

Maslova, E. (2003). A case for implicational universals. *Linguistic Typology*, *7*, 101- 108.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Phil. Trans. R. Soc. B*, *367*, 1829-1836.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11-28.

Osgood, C. E. (1980). *Lectures on language performance*. New York: Springer-Verlag.

Pawley, A. (2006). On the size of the lexicon in preliterate language communities: Comparing dictionaries of Australian, Austronesian and Papuan languages. In *Favete linguis: Studies in honour of Viktor Krupa*. Institute of Oriental Studies.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*, 138-155.

Pinker, S. (1984). *Language learnability and language learning*. Cambridge, MA: Harvard University Press.

Pinker, S. (1994). *The language instinct.* New York, NY: Harper Perennial Modern Classics.

Pyers, J. E., Shusterman, A., Senghas, A., Spelke, E. S., & Emmorey, K. (2010). Evidence from an emerging sign language reveals that language supports spatial cognition. *Proceedings of the National Academy of Sciences*, *107*, 12116-12120.

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

Richie, R., Yang, C., & Coppola, M. (2014). Modeling the emergence of lexicons in homesign systems. *Topics in Cognitive Science, 6*, 183-195.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85-114.

Senghas, A. (1995). *Children's contribution to the birth of Nicaraguan Sign Language*. Doctoral dissertation, Massachusetts Institute of Technology.

Senghas, A. (2003). Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. *Cognitive Development, 18*, 511-531.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological Science*, *12*, 323-328.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*, 1779-1782.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Phil. Trans. R. Soc. B*, *372*: 20160051. http://dx.doi.org/10.1098/rstb.2016.0051

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

Strickland, B. (2017). Language reflects “core” cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive Science*, *41*, 70-101.

Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, MA: MIT Press.

Tomasello, M. (1999). *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press.

Tomasello, T., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*, 495-552.

Trudgill, P. (2011). Sociolinguistic typology. *Social determinants of linguistic complexity*.

West, R., & Stanovich, K. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition*, *31*, 243–251.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*, 1-14.

Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial language. *Journal of Memory and Language*, *95*, 36-48.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*, 543-578.

Ziegler, J., Kocab, A., & Snedeker, J. (2017). The effect of population size on intergenerational language convergence: An artificial language learning paradigm. Poster presented at the 42nd Boston University Conference on Language Development, Boston, MA.

Appendix A. Conditional entropy.

In addition to Shannon entropy, we also calculated the *conditional entropy* of word order use given the animacy of the object noun phrase (N2; see Smith & Wonnacott, 2010; Smith et al., 2017). This was calculated as follows:

$$H(\text{word order}|\text{N2 animacy}) =$$

$$-\frac{1}{2}\sum_{j=1}^{2}\sum_{i=1}^{6} p(\text{word order}_i|\text{N2 animacy}_j)\log_2 p((\text{word order}_i|\text{N2 animacy}_j))$$

P(word order$_i$|N2 animacy$_j$) is the frequency with which word order i (for each of the six orders) is used when the object noun has animacy j (is animate or animate). N2 conditional entropy will be high when the language exhibits variability and that variability is unconditioned on the animacy of N2. Low N2 conditional entropy indicates either the absence of variability in word order or the conditioning of variation, such that sentences with animate objects are usually encoded using one word order and sentences with inanimate objects are usually encoded using another. For reference, N2 conditional entropy in the input (Gen0) was always ≈0.85.

We first calculated N2 conditional entropy collapsed across participants, depicted in Fig. S1. This pattern is nearly identical to that depicted in Fig. 10, suggesting that there was no conditioned variation by N2 animacy *at the population level*.

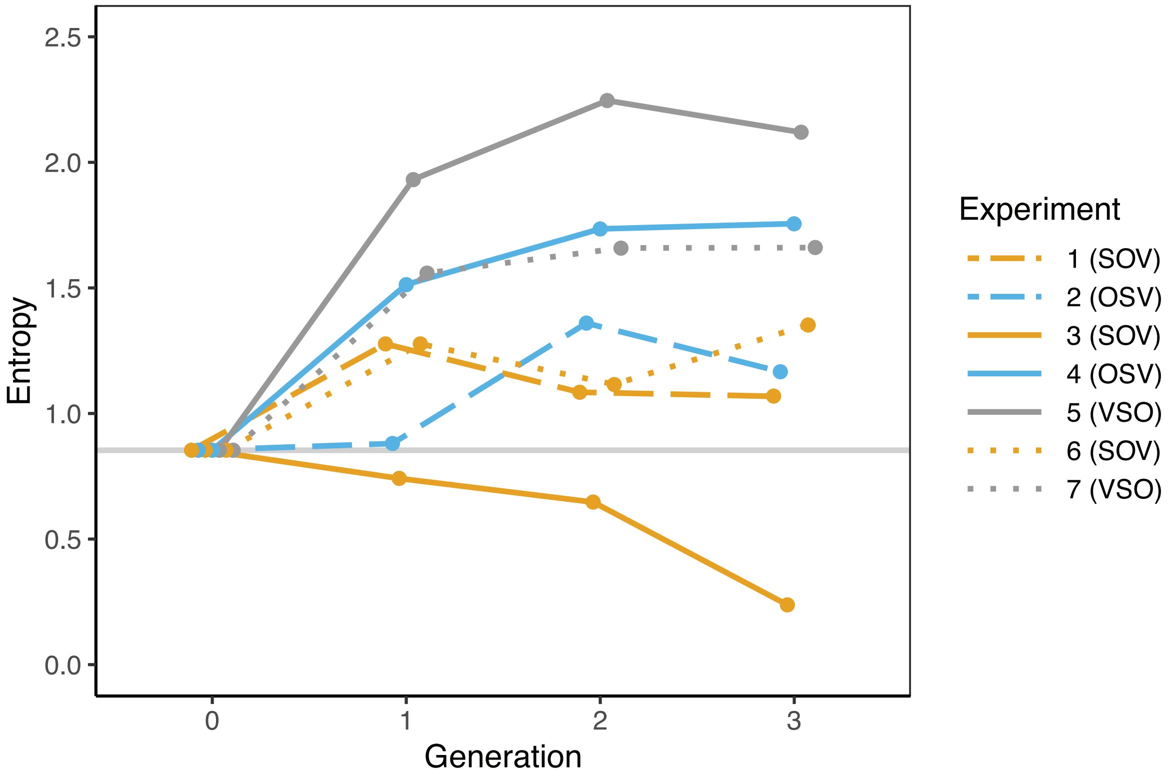


Figure S1. Conditional entropy of word order use given animacy of object noun phrase (N2) across participants by generation in Exps. 1-7.

Nevertheless it is possible that *individual* participants may have landed on different pairings of word orders with N2 animacy. To explore this possibility, we also calculated N2 conditional entropy for each participant separately, depicted in Fig. S2.

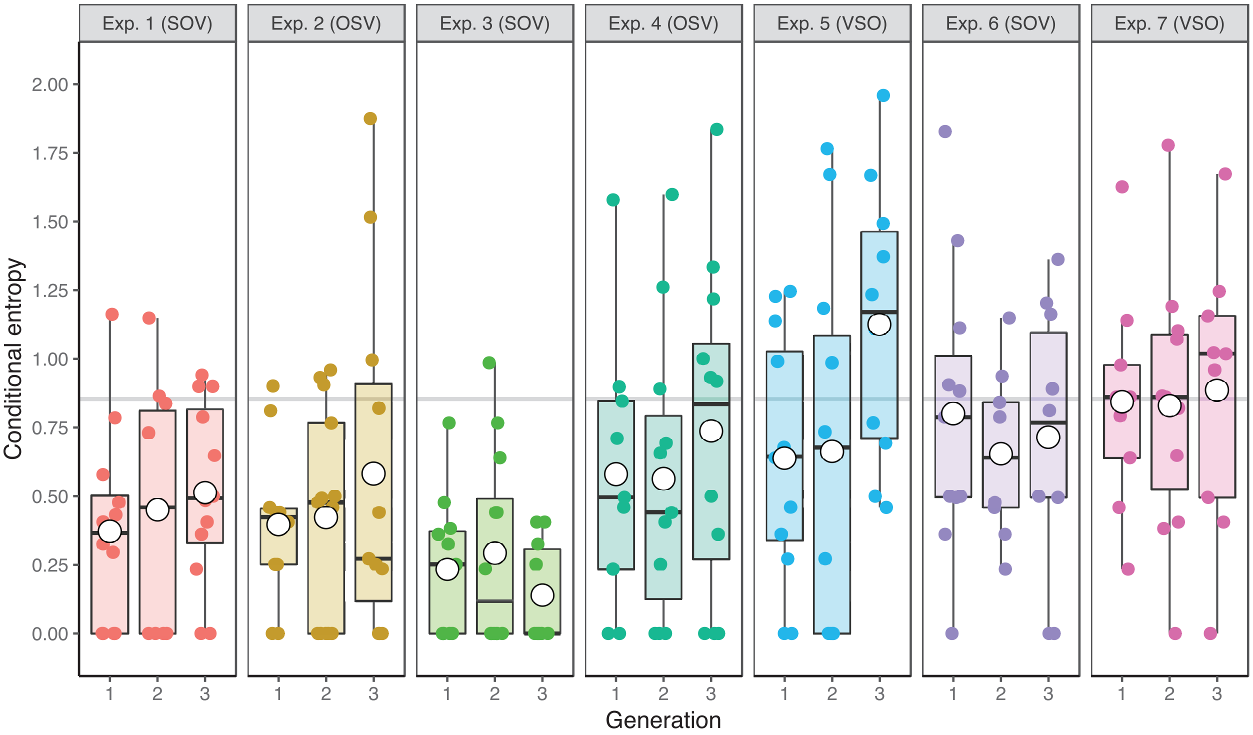


Figure S2. Conditional entropy of word order use given animacy of object noun phrase (N2) by participant by generation in Exps. 1-7 (full data set; N=225). Colored circles represent individual participants; white circles represent means across participants.

If the language was showing decreased entropy over time at the level of the individual, then we would expect the mean values to get lower from generation 1 to generation 3. Figure S2 shows no evidence of such a trend. However, there are a number of data points at 0 even in generation 1. This *could* reflect the strong conditioning of variation in the first generation, but it could also signal the absence of variability altogether. Indeed, of these 56 individuals, 55 produced only a single word order and thus did not display any variability (much less conditioned variation). When we remove these participants from the plot, we observe the pattern in Fig. S3. Here again there is no systematic trend toward decreasing conditional entropy.

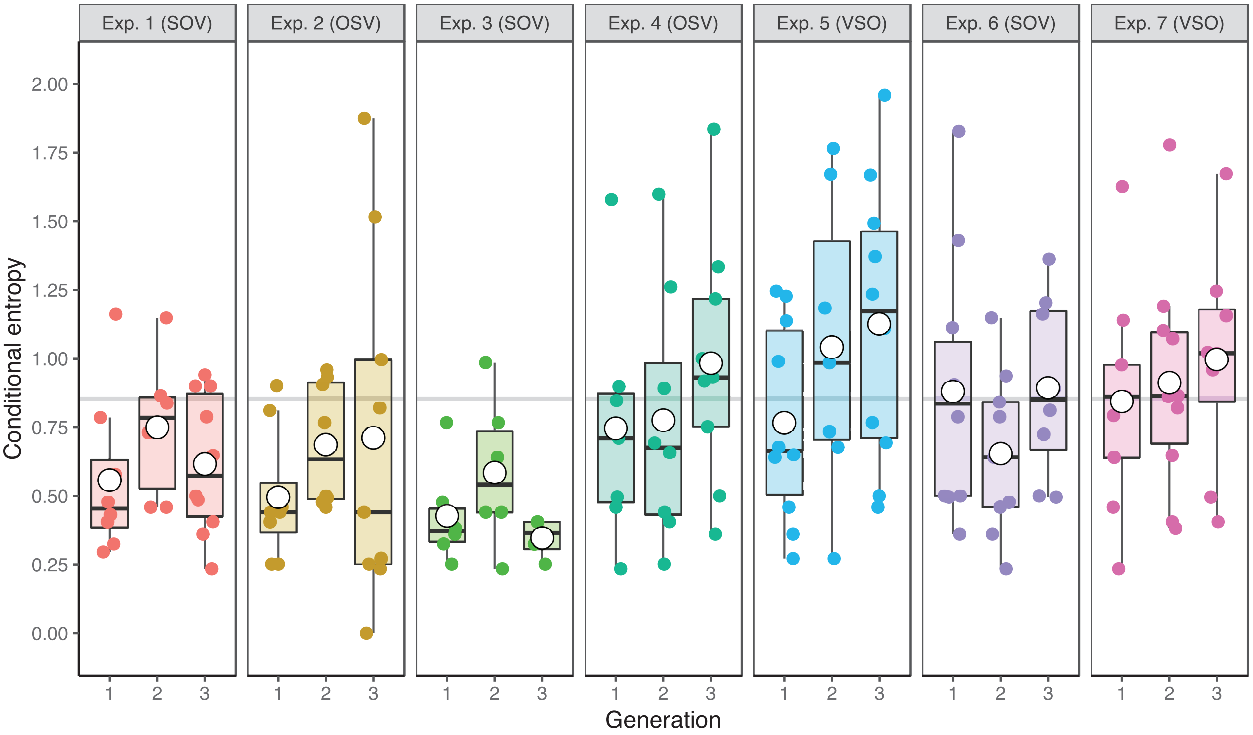


Figure S3. Conditional entropy of word order use given animacy of object noun phrase (N2) by participant by generation in Exps. 1-7 (minus participants without variability; N=170). Colored circles represent individual participants; white circles represent means across participants.

In short, we find no strong evidence for conditioning on N2 animacy *at the participant level* either. When and under what circumstances participants condition their learning on aspects of the input is an interesting question for follow-up work, but ultimately one that our experiments were not designed to test.

Note that Smith et al. (2017) also looked at conditioning by *speaker identity*. Our experimental design does not allow for this type of conditioning. Although our participants were exposed to four different aliens at learning, they produced output for only a *single* alien at test; and critically, this alien was a completely new one that they had never before seen. Pilot results from a modified version of our paradigm, reported in Appendix B, suggest that our participants were aware of the different speakers. Future work using this paradigm should explore whether learners may condition word order on speaker identity.

Appendix B. Supplementary material.

The data associated with this article, including pilot results, can be found at https://doi.org/10.17605/OSF.IO/5QMR2.