Cognition 122 (2012) 306–329

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Learning biases predict a word order universal

Jennifer Culbertson *, Paul Smolensky, Géraldine Legendre

Cognitive Science Department, Johns Hopkins University, Baltimore, MD 21218, USA

ARTICLE INFO

Article history:
Received 2 November 2010
Revised 15 October 2011
Accepted 20 October 2011
Available online 28 December 2011

Keywords:
Learning biases
Typology
Word order
Artificial Language Learning
Bayesian models
Universals

ABSTRACT

How recurrent typological patterns, or universals, emerge from the extensive diversity found across the world's languages constitutes a central question for linguistics and cognitive science. Recent challenges to a fundamental assumption of generative linguistics—that universal properties of the human language acquisition faculty constrain the types of grammatical systems which can occur—suggest the need for new types of empirical evidence connecting typology to biases of learners. Using an artificial language learning paradigm in which adult subjects are exposed to a mix of grammatical systems (similar to a period of linguistic change), we show that learners' biases mirror a word-order universal, first proposed by Joseph Greenberg, which constrains typological patterns of adjective, numeral, and noun ordering. We briefly summarize the results of a probabilistic model of the hypothesized biases and their effect on learning, and discuss the broader implications of the results for current theories of the origins of cross-linguistic word-order preferences.

1. Introduction

1.1. Language universals and cognitive biases

How do recurrent cross-linguistic patterns—typological universals—emerge from the extensive diversity found across the world's languages? This constitutes a central question for linguistics and cognitive science. One influential view is that these patterns arise largely because of constraints on the grammars that people can (or will) learn. That is, human learning is *biased*: the cognitive processes of learning impose structure on grammars by respecting constraints (hard or soft) that are then reflected in typological regularities.[1] So if a logically-possible grammatical system is not found, or is quite rare cross-linguistically, the explanation offered by such theories is that this system violates a learning bias.

The hypothesis that universal constraints on human language learning strongly shape the space of human grammars has taken many forms, which differ on a number of dimensions including the locus, scope, experience-dependence, and ultimate source of such biases (Christiansen & Devlin, 1997; Chomsky, 1965; Croft, 2001; Hawkins, 2004; Kirby, 1999; Lightfoot, 1991; Lindblom, 1986; Newmeyer, 2005; Newport & Aslin, 2004; Talmy, 2000; Tesar & Smolensky, 1998). However, the general hypothesis that language universals arise from biases in learning stands in contrast to hypotheses that place the source of explanation outside the cognitive system (Bybee, 2009[2]; Dunn, Greenhill, Levinson, & Gray, 2011[3]; Evans & Levinson, 2009[4]).

* Corresponding author. Address: Department of Brain & Cognitive Sciences, University of Rochester, 246 Meliora Hall, Rochester, NY 14627, USA. Tel.: +1 585 275 1844; fax: +1 585 442 9216.

E-mail address: jculbertson@bcs.rochester.edu (J. Culbertson).

[1] 'Learning bias' is used here quite generally to refer to any factor, other than asymmetries in the learning data, that generates asymmetries in learning.

[2] In usage-based theory, general associative cognitive mechanisms are presumed to carry out the grammaticalization processes by which frequent patterns of use enter grammars, but the source of typological asymmetries is asymmetries in use, not asymmetries within cognition.

[3] "Linguistic diversity does not seem to be tightly constrained by universal cognitive factors specialized for language" (p. 82).

[4] "Many generativist approaches … consequently attribute cognitive reality to conditionals of the form 'if structural decision X, then also structural decision Y' … No language typologist would maintain that conditional regularities of this type would be found in speakers' heads." (p. 475).

0010-0277/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2011.10.017

For instance, the non-existence or rarity of certain linguistic patterns could be an accident of history, determined by the "success at survival, propagation, and colonization" (de Lacy, 2006, p. 351) of speakers who happened to utilize these patterns in the past. And commonalities of languages might arise: as a result of descent from a common ancestral language (Dunn et al., 2011; Levinson & Evans, 2010, p. 2743), or spread through language contact (Moravcsik, 1978), under the influence of geographic factors (Atkinson, 2011); or from communicative factors: "language is used in similar ways in different cultures" (Bybee, 2009, p. 18). Alternatively, the probability of errors in transmission between generations may be greater for some patterns than others because of intrinsic properties of the signal: the channel *between* speakers' cognitive systems may explain the rarity of certain grammatical patterns (Ohala, 1993; Steriade, 1997; Stevens, 1972).

Any given typological pattern is likely to be the result of the interaction of many of the factors mentioned above. But for linguists and cognitive scientists, accounts that situate the explanation for typological asymmetries in the mental biases of individual learners are of particular interest, since they potentially shed light on underlying properties of the cognitive system. The question we address in the present work is, do biases that parallel linguistic universals *exist* within the cognitive systems of language learners?

The experimental results we report cannot, we will argue, be explained solely by cognition-external factors. That is not to deny that such factors exist—they surely do; but they are not sufficient to explain the asymmetrical learning patterns that we observe. These asymmetries require a cognition-internal explanation. Although our experiment was designed to test the null hypothesis that cognitive biases paralleling linguistic universals do not exist, our results also speak to some of the issues which distinguish theories affirming their existence. We return to this in closing (Section 5.3).

1.2. Biases in artificial language learning

As the sample of citations above suggests, the extent to which linguistic typology is a reflection of the cognitive system lies at the heart of some of the most prominent debates in the field (for example, a special issue of *Lingua*, Rooryck, Smith, Liptak, & Blakemore, 2010, was recently devoted to continuation of the already extensive debate in *Brain and Behavioral Sciences* of Evans & Levinson, 2009). The longstanding disagreements concerning the explanation(s) for (and even the existence of) robust typological regularities motivate the pursuit of new types of empirical evidence concerning the biases of human language learners and the extent to which these biases parallel typological tendencies.

Recent research has used artificial language learning paradigms with adults to provide direct behavioral evidence for the existence of such biases. This work has focused mainly on laboratory learning of phonological patterns and word segmentation (e.g. Finley & Badecker, 2008; Newport & Aslin, 2004; Saffran, Newport, & Aslin, 1996; Wilson, 2006). Wilson (2006), for example, found that learners were more likely to generalize novel patterns of velar palatalization when the phonological conditions were in line with a purported typological law (Bhat, 1978). It is to explain this type of *asymmetry* in learning outcomes that learning biases are hypothesized.

A few recent studies have targeted typological asymmetries in morphology and syntax (e.g. Christiansen, 2000; Hudson Kam & Newport, 2005; St. Clair, Monaghan, & Ramscar, 2009). Christiansen (2000), for example, shows that artificial languages with inconsistent head order are more difficult to learn than those with consistent head order—in line with a typological preference for the latter.[5] St. Clair et al. (2009) provide evidence of a bias parallel to the cross-linguistic preference for suffixing over prefixing; participants learning an artificial language with suffixes identifying word categories outperformed those learning a language with prefixes.

Although the population of most interest is child learners (that is, learners within the presumed critical period for language acquisition), the results of these studies suggest that at least some relevant biases are present in adult learners (in line with some theories of second-language acquisition—Epstein, Flynn, & Martohardjono, 1996; Smith & Tsimpli, 1995; Vainikka & Young-Scholten, 1996; White, 2003). As we will discuss further in the context of our own results, these findings demonstrate that artificial language learning experiments with adults can provide a valuable source of empirical evidence concerning the existence of learning biases that are congruent with observed typological regularities.

In interpreting results of laboratory learning in adults, we will need to pay due attention to the potentially larger role of problem-solving strategies relative to native-language acquisition, to potential artifacts of the laboratory learning task relative to natural learning environments, and to the role of the adult's first language. But when adult laboratory learning displays asymmetries that cannot be attributed to such extraneous factors, and these asymmetries parallel a typological tendency, there is reason to suspect a common cause: the most parsimonious explanation is that the same bias is at work in both experimental- and native-language learning, and that this bias in native language acquisition—whatever its locus, scope, experience-dependence, and ultimate source—has, over time, driven the languages of the world towards greater conformity with the bias.

The primary objective of this article is therefore to contribute to the (as yet small) body of research examining the relation between asymmetries in artificial language learning and in typological frequencies. Here we investigate learners' biases paralleling typological universals of syntax. We provide experimental evidence for a learning bias in favor of grammars that are in line with a word-order universal, first proposed by Greenberg (1963), that concerns linear ordering of nouns with respect to numerals and adjectives. This particular universal, presented in Section 2, was chosen for several reasons which we elaborate below. First, although not exceptionless, it appears to be typolog-

[5] The consistent language used all head-final phrases, while the inconsistent language had e.g., head-final VP → NP V, but head-initial PP → Prep NP.

ically robust. Second, unlike others that pertain to subtle syntactic distinctions, this universal lends itself straightforwardly to artificial language stimuli. Third, within the universal lies an asymmetry that cannot be attributed to an obvious general learning principle. And fourth, although itself rather narrow in scope, this universal is potentially a sub-case of a general word-order constraint that has been the subject of much recent work in theoretical syntax—the Final-Over-Final Constraint (Biberauer, Holmberg, & Roberts, 2008, in press; Biberauer, Newton, & Sheehan, 2009; Hawkins, 2010; Holmberg, 2000).

The experiment we report uses a methodology developed by Hudson Kam and Newport (2005, 2009) which we call here the *Mixture-Shift Paradigm*. With this paradigm, Newport and colleagues have provided experimental evidence to support the claim that, under certain circumstances, when exposed to input containing unpredictable variability, adults and children tend to acquire more regular rules, increasing the consistency of the system (Meyerhoff, 2000; Sandler, Padden, & Aronoff, 2005; Sankoff & Laberge, 1980; Singleton & Newport, 2004; Slobin, 2004; Smith, Durham, & Fortune, 2007). We use this paradigm to simultaneously investigate the effects of two interacting types of (universal) biases, defined informally in (1).

(1) *Two types of biases*
   a. Regularization bias: acquire grammars that minimize variation [induces reduction of variation that may be present in the input to learning]
   b. Substantive bias: acquire grammars that do not incorporate particular [disfavored] structures

As described in Section 3, learners in our experiment are exposed to a miniature artificial language featuring an inconsistent mixture of word-order patterns. This mix of grammars is an idealization of the input learners are exposed to during a period of linguistic change—thus the experiment makes a clear connection with the idea that learners are agents of change, (implicitly) shifting languages over time to better conform to their biases (e.g. Clark & Roberts, 1993; Croft, 2000; Lightfoot, 1999). In particular, since learners do not always reproduce variation of this sort veridically, as Hudson Kam and Newport (2005), Hudson Kam and Newport (2009) have found, this paradigm makes it possible to observe the conditions under which learners shift the input mixture to produce a less variable language, and to investigate whether these acquired languages conform more closely to purported typological generalizations. This approach is ideally suited to uncovering evidence for the types of learning biases that might underlie typological generalizations since it allows for within-learner comparison of the treatment of typologically attested and unattested (or rare) patterns: in the Mixture-Shift Paradigm, to bring about language change, a learner need only, in effect, choose between patterns present in their input—learners are not required to innovate structures they have never experienced, which is presumably less likely than simply *shifting*, quantitatively, the mixture they were exposed to. Thus this method is potentially more sensitive to learning biases than alternative experimental paradigms in which any effect of a bias depends on learner innovation.

The results of the experiment, we will argue in Section 4, show clear evidence of a learning bias aligning with Greenberg's universal. This bias can therefore explain asymmetries in language change and typological frequencies (Section 5.1). The bias can be quantified in a perspicuous form within a Bayesian model, as summarized in Section 5.2. By formulating our theory of the bias as *probabilistic* we differ from most linguistic theories, which generally treat universals as the result of inviolable constraints specific to the linguistic system.[6] This makes the prediction, which the typological facts suggest is desirable, that even languages that violate the bias are learnable—but less likely to arise (and less likely to be acquired veridically).

In disconfirming the null hypothesis that language learning is unbiased, this paper takes the important first step of providing evidence that a cognitive bias parallel to Greenberg's universal *exists*—contrary to recent claims that biases of individual learners do not shape typology. Having established this, extensive future work is needed to identify the locus, scope, experience-dependence, and ultimate origin of this bias. We close by discussing such larger issues in Section 5.3.

2. Greenberg's Universal 18: a typological generalization in the nominal domain

2.1. The typological data

Perhaps some of the most well-known typological generalizations pertain to word order patterns, and in particular, correlations between the internal orders of different types of sentence constituents (see e.g. Aristar, 1991; Biberauer et al., 2008; Cinque, 2005; Dryer, 1992, 1988; Greenberg, 1963; Hawkins, 1983, 1990; Steddy & Samek-Lodovici, 2011, among many others). In this section we will introduce the typological generalization which will be the focus of the experimental investigation—a word-order universal in the nominal domain, first discussed by Greenberg (1963) under the label 'Universal 18'.

Most languages have ways of quantifying and attributing properties to nouns by combining them with numeral words and attributive adjectives respectively.[7] The four logically possible combinations of {Noun, Adjective} and {Noun, Numeral} ordering are shown in (2).[8] Examples from a language which features each pattern are provided in (3),

[6] Optimality Theory (Prince & Smolensky, 1997) is a general theory of universals, within generative linguistics, in which universal grammatical preferences are formalized as ranked, *violable* well-formedness constraints. But even in Optimality Theory, typological asymmetries of the sort we discuss here are standardly explained by rigid, universal, inviolable requirements on the relative ranking of specified constraints (Prince & Smolensky, 1993/2004, chap. 9).

[7] We will use the term 'modifier' to refer to adjectives and numerals together. This is simply a terminological convenience, not meant to suggest that they be grouped theoretically as elements of the same class (see Section 5.3.3).

[8] The curly-bracket notation refers to the unordered set; a dash indicates a particular order, e.g. Adj-Noun.

(4), (5), (6).[9] (The numbering of the four patterns in (2) will be used consistently throughout the article.)

(2) *Possible patterns of {Noun, Adjective}, {Noun, Numeral} ordering*
1. Adjective-Noun & Numeral-Noun
2. Noun-Adjective & Noun-Numeral
3. Noun-Adjective & Numeral-Noun
4. Adjective-Noun & Noun-Numeral

(3) *Cherokee*, Holmes and Smith (1977)
a. u-wo'-du a-ge-hyu'-tsa
pretty girl
'pretty girl'
b. tso'-i gu:-gu
three bottles
'three bottles'

(4) *Yoruba*, Ward (1956)
a. bata titun
shoes new
'new shoes'
b. awo meje
dishes seven
'seven dishes'

(5) *Basque*, Hualdo and de Urbina (2003)
a. etxe zuri
house white
'white house'
b. bi zuhaitz
two trees
'two trees'

(6) *Sinhala*, Gair (1970)
a. loku pot
big books
'big books'
b. geval tunak
houses three
'three houses'

While all four patterns in (2) are attested, their distribution appears to be constrained cross-linguistically; according to Greenberg (1963) and others (e.g. Hawkins, 1983; Hurford, 2003; Rijkhoff, 1998), only the first three are *well-attested*; the fourth (6) is not. The correlational generalization originally drawn by Greenberg (1963) is the following: Adjective-Noun order implies Numeral-Noun order. The generalization translates to a ban on pattern 4 of (2), that is, the combination of pre-nominal adjectives with post-nominal numerals. Corresponding statistics from an extensive cross-linguistic survey reported in the World Atlas of Language Structures (Dryer, 2008a, 2008b), are shown in Table 1. The distribution of these ordering patterns according to WALS shows that, indeed, very few languages (4%) fall into the gray cell representing the pattern combining the orders Adjective-Noun (henceforth Adj-N) and Noun-Numeral (N-Num).[10]

Based on these data we can confirm that Greenberg's generalization is quite robust; it is a very strong typological tendency. Understanding why such a generalization might exist, even if it is statistical rather than absolute, is therefore desirable, and has the potential to shed light on the language learning faculty.

Greenberg (1963) treated the pattern in the gray cell of Table 1 as containing two independently dispreferred orders, Adj-Noun and Noun-Num, and accordingly, we will call this the *marked* pattern. The pattern containing the two alternative orders, N-Adj and Num-N, we will call *unmarked*. (We do not intend any particular content for the terms "marked" and "unmarked": they are nothing more than names for patterns 4 and 3.)

Table 1
Distribution of languages with each {Noun, Adjective}, {Noun, Numeral} ordering combination in the WALS sample.

	Noun-Adj	Adj-Noun
Num-Noun	149 (17%)	227 (27%)
Noun-Num	443 (52%)	32 (4%)

In addition to the fact that the marked pattern is very rare cross-linguistically, Greenberg noted that patterns which contain orders preserving the position of the noun with respect to both modifiers, i.e. patterns 1 and 2 in (2), are more common cross-linguistically. This is borne out by the data in Table 1 as well. We will follow Greenberg in calling these two patterns 'harmonic'. Based on the typology, we can then rank the ordering patterns in (2) as in (7).

(7) *Ranking of {Adj, Noun}, {Num, Noun} ordering patterns in (2) according to the typology (where 'x ≻ y' means 'x is more common than y')*

1, 2 (harmonic) ≻ 3 (unmarked) ≻ 4 (marked)

Because we are examining the hypothesis that a learning bias exists congruent to Greenberg's Universal 18, we follow Greenberg in not distinguishing between the two harmonic orders 1 and 2. The more recent data from WALS in Table 1 suggests however that among harmonic patterns, 2 is favored over 1. Since our experimental participants speak a pattern 1 language, English, it is problematic to test for a bias of pattern 2 over pattern 1 (although in future work we plan to do so using speakers of other language types; see Sections 5.2 and 5.32).

2.2. Hypothesized learning biases

The question of interest in this article is: do learners show evidence of biases parallel to the constraints respected by the typological pattern outlined here—in particular, the ranking in (7)? If so, then we will have evidence that the ranking is not just representative of cross-linguistic frequency, but also of a cognitive preference scale. The general hypotheses the experiment is designed to test are therefore those listed in (8).

(8) *General hypotheses based on the typology*
HYPOTHESIS 1. Learners have a bias favoring the harmonic patterns (1 and 2) and the unmarked pattern (3).
HYPOTHESIS 2. The bias favoring harmonic patterns [the 'harmonic bias'] is strongest.
HYPOTHESIS 3. Learners have a bias against the marked pattern 4 [the '*L4' bias].

[9] Other examples of the four patterns from WALS (Dryer, 2008a, 2008b) include 1: Modern Greek, Kannada (India), Chamorro (Guam); 2: Thai, Lakhota (Siouan), Arrernte (Australia); 3: Spanish, Modern Hebrew, Welsh; 4: Majang (Ethiopia), Purki (Kashmir), Kewa (Papua New Guinea).

[10] Note that both the statistics and the generalization refer to the typical or most frequently used order in a given language. Some languages have no dominant order for one or both modifier types (WALS lists 110). Further, many languages have exceptions to their general ordering rules; for example, there are some post-nominal adjectives in English, e.g. 'money galore', and some post-nominal uses of numerals, e.g., 'the Brothers Four'.

The cross-linguistic preference for harmonic ordering patterns, which exists beyond the phrase types addressed by Universal 18, is well documented (e.g. Baker, 2001; Dryer, 1992; Greenberg, 1963; Hawkins, 1983), and is of some interest here. However, a preference for consistency of ordering can readily be explained by many types of learning biases—here we will refer to it simply as "the harmonic bias".[11] In contrast, the distinction made by Universal 18 among the two *non-harmonic* patterns cannot be so readily explained. Therefore, if we see a learning asymmetry between the non-harmonic patterns that parallels the asymmetry specified by Universal 18, this will be more informative for the ultimate goal of more precisely characterizing the nature of the bias. For this reason we will be most interested in comparing the learning of patterns 3 and 4, and we will refer to the hypothesized bias against pattern 4 as the substantive **L4* bias.

In an experimental setting, any observed results could be due not to considerations of the type set out in (8), based on typological evidence, but to other cognitive factors. Additional hypotheses based on such other considerations are laid out in (9). First, as we have already hinted at, learners may have a preference for the orders found in their native language. Note that this preference, if sufficiently strong, leads to the predictions that learners' biases will not *exactly* parallel the typology—in particular, as we have said above, they are not likely to prefer pattern 2 over 1. Opposing this preference, however, is the possibility that in production, faster lexical retrieval for nouns could lead to a tendency to utter them first (Oldfield & Wingfield, 1965).

(9) *Additional hypotheses not based on the typology*

HYPOTHESIS 4. Learners are biased to use their native [English] order.

HYPOTHESIS 5. Learners are biased to utter the noun first.

The specific predictions derived from these hypotheses in the context of the experiment will be discussed in Section 3.5, after the experimental design and conditions are described in detail.

3. Experimental design and methodology using the Mixture-Shift Paradigm

3.1. Participants

Participants in this experiment were undergraduate and graduate students at Johns Hopkins University who received either $15 compensation or course credit for completing the experiment. All participants were between the ages of 18 and 30. Data from 65 participants was collected and analyzed (a total of 73 participants were run, but 8 participants failed to pass a basic vocabulary learning criterion—Section 3.3.1). All participants were native speakers of English, however some spoke other languages as well; see Section 4.3 below for discussion of this issue.

[11] In fact, as a reviewer points out, the harmonic bias could be thought of as a type of regularization bias since it favors consistency in ordering.

3.2. The lexicon

The lexicon was comprised of 10 nonce nouns, and 10 nonce modifiers—5 adjectives and 5 numerals. The nouns named novel objects rather than known objects with English names, in order to discourage participants from treating the noun-learning task as one of translation into their native language (some examples are shown in Table 2 below). The 5 adjectives meant "big", "small", "green", "blue", "furry", and the 5 numeral words meant "two", "three", "four", "five", "six".[12] The non-word vocabulary was comprised of di- and trisyllabic nouns all ending in schwa, and monosyllabic C(C)V(C)C modifiers. Non-words were chosen such that onsets (both singletons and clusters) were unique in order to minimize confusability. The neighborhood densities of both adjective and numeral non-words were low (these were calculated by finding the number of English neighbors for a given non-word using Levinshtein (or edit) distance: see Yarkoni, Balota, & Yap (2008)). Mean neighborhood density did not differ significantly across modifier types (mean density for adjectives = 6, mean density for numerals = 5; a *t*-test comparing density for adjective and numeral non-words was not significant, $t(4) = 0.37$, $p = 0.73$). A full list of the nonce words used, along with their neighborhood densities, can be found in Appendix A.

3.3. Experimental procedure

Adult participants were taught a miniature artificial language using a video game interface. Each participant was trained and tested during a single hour-long experimental session. Participants were seated in a private testing room in front of a computer display, and wore headphones through which the experimental audio was played. The experiment began by introducing participants to the learning scenario. They viewed a screen which explained that they would be learning a new language, and that the goal of the task would be to assimilate into a community of aliens living on another planet. Participants were then introduced to a "native speaker" informant whose utterances they learned from. (The artificial language was never presented visually.) The informant's speech was synthetically generated using Apple's text-to-speech software (OS 10.5, speaker "Alex", with pitch altered using Praat). Throughout the entire experiment, the informant uttered phrases in the language stochastically according to the parameters of the grammar specified by experimental condition (explained in detail below). Grammars were probabilistic in the sense that, given a scene (e.g. a picture) and the vocabulary items needed to express it (e.g. *grifta* (a noun) and *fush* ("green")), the particular structural realization was determined only by the relative

[12] The numeral "one" was not used for several reasons, the first pragmatic, the second typological. First, in the experiment each object was initially taught to participants in isolation, so using this same picture later to represent "one" such object might be confusing. Second, the numeral "one" frequently behaves distinctly from other numerals cross-linguistically. For example in Basque, although all other numerals precede the noun, the numeral one follows it. This pattern of distinct behavior is discussed in Hurford (2003).

Table 2
Example picture choices in Training Phase II, the comprehension picture-matching task.

	Correct Adj	Incorrect Adj		Correct Num	Incorrect Num
(a) Adj trial			*(b) Num trial*		
Correct N			Correct N		
Incorrect N			Incorrect N		

probabilities assigned to rule expansions in the grammar (e.g., in conditions 2 and 4, p ($S_{Adj} \rightarrow$ Adj-N) = 70%, p ($S_{Adj} \rightarrow$ N-Adj) = 30% $\Rightarrow$ p(“*fush grifta*”) = 70%, p(“*grifta fush*”) = 30%). Crucially, these probabilities applied to every individual lexical item: order was *not* fixed for any particular noun or modifier.

3.3.1. Phase I

Phase I introduced participants to the set of nonce nouns. During this phase, participants were shown, one at a time, pictures of objects (which were grayscale and roughly the same size), and heard the corresponding object name from the informant. They were told to repeat each object name after they heard it. After 50 exposure trials, participants were tested on what they had learned. In the testing portion of this phase, participants had to provide, within a certain time limit (10 s), the label for the object that appeared on the screen. A correct response within the time window resulted in 5 points. The correct answer was provided by the informant following each trial, regardless of whether the participant's response was correct or incorrect. Participants had to reach a level of 75% correct on 50 test trials, and were given two chances to do so (with no further training). If a participant did not reach the criterion for success on this phase by the second try, they were excluded from all analyses.

3.3.2. Phase II

Phase II introduced participants to the set of modifiers. During this phase participants viewed pictures, and heard the informant describe them using {Noun, Adj} and {Noun, Num} combinations. The objects in this phase were the same set presented in phase I, however here they were presented either singly, in different colors, sizes, or textures (the adjectives), or in groupings of more than one of the same grayscale object (the numerals). In the numeral displays, the multiple objects were arranged in patterns resembling the dots on dice to ease numerosity recognition. Participants were again asked to repeat aloud each phrase they heard.[13] After 80 exposure trials, participants were tested on their comprehension of phrases in the language. Four pictures were presented on the screen, and the informant provided a two-word description of one of the four pictures (using correct lexical items in a word order that was probabilistically selected, independently for each trial). Participants were instructed to choose which of the four pictures matched the informant's description. A correct choice within a specified time window (10 s) earned the participant 10 points. Regardless of the choice, the correct picture was indicated after the response. Participants were not required to achieve a certain score on the 80 test trials in this phase, however their score was visible throughout.

Trials in this task were of two types which were intermixed; one half tested adjective vocabulary, and the other half tested numeral vocabulary. Example pictures for an adjective trial (a) and a numeral trial (b) are shown in Table 2. Note that this task tested participants' ability to comprehend {Noun, Modifier} combinations, but did *not* test their knowledge of the ordering patterns in the language they were learning. This part of phase II will be referred to as the *comprehension picture-matching task*.

3.3.3. Phase III

Phase III assessed participants' knowledge of the word order possibilities in the language they were exposed to. After 20 review trials (of the same form as the basic exposure in phase II), participants were shown pictures and were required to describe them to the informant. The instructions specified that if an answer was accurate and the informant could understand it, they would be awarded 10 points. However they were also told that more than one answer might be possible, and that the informant would provide his own description after the participant's. If the two responses were exactly the same, the participant received 5 bonus points. The informant's responses were generated stochastically according to the parameters of the language condition (ensuring that training and testing probabilities for any given order in a condition were identical) and were therefore independent of the participants'

Table 3
Example trial in production testing phase III.

(1) Picture	(2) Participant responds	(3) Vocab pts.	(4) Informant responds	(5) Order pts.
	“*nerka geej*”	10pts.	“*nerka geej*”	5 pts.

[13] Exposure training stimuli for each participant in this phase were constructed from constrained pseudo-randomized lists, designed so that the stimuli did not include spurious statistics for participants to track, but only the statistics of interest, namely the ordering trend for each modifier type according to the probability-parameters of the language condition. Here and throughout all phases of the experiment, each participant's lists were unique.

Table 4
Experimental conditions; majority orders for conditions 1–4 are in bold.

(a) Condition 3: Unmarked orders	
0.3 S_{Adj} → Adj N	0.7 S_{Adj} → **N Adj**
0.7 S_{Num} → **Num N**	0.3 S_{Num} → N Num
(b) Condition 1: Harmonic (=Eng)	
0.7 S_{Adj} → **Adj N**	0.3 S_{Adj} → N Adj
0.7 S_{Num} → **Num N**	0.3 S_{Num} → N Num
(c) Condition 2: Harmonic (≠Eng)	
0.3 S_{Adj} → Adj N	0.7 S_{Adj} → **N Adj**
0.3 S_{Num} → Num N	0.7 S_{Num} → **N Num**
(d) Condition 4: Marked orders	
0.7 S_{Adj} → **Adj N**	0.3 S_{Adj} → N Adj
0.3 S_{Num} → Num N	0.7 S_{Num} → **N Num**
(e) Condition 0: Equiprobable control	
0.5 S_{Adj} → Adj N	0.5 S_{Adj} → N Adj
0.5 S_{Num} → Num N	0.5 S_{Num} → N Num

responses. An example trial is shown in Table 3. As in the previous phase, although there was no score requirement, the participants were shown their score throughout.

A primary reason for designing the production test in this way was to provide some motivation for participants not to simply adopt English word order. As we will next explain in detail, all orders were *possible* in each input condition, and thus participants could in principle use prenominal order only, ignoring any alternative orders they were exposed to while still producing descriptions that were 'accurate'.[14]

The highest possible score in this task would be achieved by maximizing, for each modifier type, the use of the most frequent order in the input (this would result in the highest frequency of matching between the participant's response order and the informant's response order). Using points and informant utterances diverges from the type of production task typically used in this paradigm (e.g. in Hudson Kam & Newport, 2005, 2009; Wonnacott & Newport, 2005). Several studies of behavior in non-linguistic probability-learning tasks (e.g. Weir, 1972) have found that learners are more likely to match the input probabilities than to maximize the use of the most frequent correct choice in tasks where they are given feedback contingent on a correct response, and less likely to probability match when given input that corrects a wrong answer as well as reinforcing a correct one.[15] Providing participants with informant answers may therefore encourage regularization. Importantly, however, the experiment was repeated without any informant input during Phase III, with results which closely replicate those reported here; see Culbertson (2010, chap. 4), for details. It turns out that, in this task, removing the incentive to match an informant response does not strongly affect the extent to which (adult) participants' productions reflect the frequencies of word orders in their training data; in particular, learners do not simply resort to English order.

[14] Further, a relatively lengthy production testing phase without any informant input might lead to undesirable levels of self-reinforcement; for example, participants could learn from their own incorrect utterances. In the informant's descriptions, the production probabilities of all word orders were the same as during Phase II training, so any additional learning from the informant's utterances that might have occurred during Phase III would merely reinforce the learning in Phase II.

[15] The extent to which feedback during learning is ecologically valid—i.e., occurs during natural language learning—is an issue that has generated much debate (Demetras, Post, & Snow, 1986; Hirsh-Pasek, Treiman, & Schneiderman, 1984; Marcus, 1993; Morgan, Bonamo, & Travis, 1995; Strapp, 1999). For an extensive review, see Saxton (2000).

3.4. Input language conditions

In this experiment, participants were divided equally among five experimental conditions corresponding to different probabilistic input grammars as shown in Table 4. (For each production rule, the conditional probability of the right-hand side given the left-hand side is given; rules generating {Noun, Adj} combinations sum to 1, as do rules generating {Noun, Num} combinations.) For ease of reference, Table 5 shows the majority pattern for conditions 1, 2, 3 and 4 (condition 0 has no majority pattern). (The 2×2 arrangement of Table 5 matches that of displays used below to plot the experimental results. Table 4 is arranged to match as well.) As Table 5 makes clear, the majority word-order patterns for conditions 1–4 correspond to those labeled 1–4 in (2); for example, in condition 1, the majority pattern is the harmonic language '1', with prenominal adjectives and numerals.

The equiprobable condition 0 functions roughly as a control in order to uncover participants' word-order biases in an exposure condition which does not lean in any particular direction. The harmonic condition 1 also functions in part as a control to determine participants' ability to learn and willingness to regularize in a condition that is likely to be relatively easy for them, since it follows English most closely. Conditions 2, 3, and 4 all introduce a probabilistic tendency toward one of the logically possible, but non-English-like, patterns discussed above.

3.5. Behavioral predictions

Recall that we are investigating interacting types of biases—the regularization bias (which favors less variation) and the *L4 bias, disfavoring the non-harmonic pattern Adj-Noun, Noun-Num, in addition to the bias which favors harmonic patterns. The hypotheses (8) and (9) laid out in Section 2.2 yield the following predictions (see (10) below): Regularization in the experiment should be in the direction of a harmonic pattern or the unmarked

Table 5
Summary of majority ordering pattern in experiment conditions 1, 2, 3 and 4 (equiprobable condition 0 has no majority pattern).

	N-Adj	Adj-N
Num-N	3. Unmarked	1. Harmonic
N-Num	2. Harmonic	4. Marked

pattern. Regularization should be strongest in the direction of the two harmonic patterns. No regularization is predicted toward the marked pattern.

To see how these predictions are derived, consider that, all things being equal, the regularization bias will push learners to *boost*, that is, to use the majority pattern more frequently than it is found in the input—in this case, more than 70% of the time.[16] Accordingly, learners in conditions 1, 2 and 3 are predicted to deviate from the input probabilities, boosting in the direction of the current trend of the target grammar, since the hypothesized biases point in mutually compatible directions. Somewhat less boosting is predicted in the non-harmonic condition 3 as compared to the harmonic conditions 1 and 2, since we hypothesize that harmonic languages are most highly favored. Learners in condition 4 are predicted to be at a disadvantage since the regularization bias and substantive bias *L4 are *not compatible*. Participants in this condition are therefore predicted to show less regularization; they should use the majority pattern less frequently than their counterparts in the other conditions.

Given these predictions, the primary dependent measure of interest will be *the proportion of use of the majority pattern* in productions of participants across the different experimental conditions. However, there is no prediction that every participant will behave in precisely the same way; for example individual participants exposed to condition 4 might exhibit probability matching behavior, or they might flatten the input probability distribution (making it more random) or they might shift instead toward a different pattern altogether. For this reason, Section 4.4 will look in detail at individual subject results.

(10) *Experimental predictions (typologically relevant)*

PREDICTION 1. Bias in favor of harmonic and unmarked patterns ⇒ Regularization toward harmonic or unmarked patterns; most regularization in conditions 1, 2, 3.

PREDICTION 2. Stronger bias in favor of harmonic languages ⇒ Most regularization toward harmonic patterns; more regularization in conditions 1, 2.

PREDICTION 3. *L4 bias ⇒ No regularization toward the combination *Adj-Noun* & *Noun-Num*; no regularization in condition 4.

For the equiprobable condition 0, any deviation by learners in the production test from the target probabilities should be in the direction of a harmonic pattern, or the unmarked pattern; it will be particularly important to look for any indications of bias towards the English harmonic pattern in this condition. This relates to additional predictions, laid out in (11), derived from the hypotheses (9) based on non-typological factors that were discussed in Section 2.2. First, learners are predicted to show a general preference for the orders found in their native language. This was part of the motivation for including both the equiprobable condition 0 and condition 1, which is English-like. Second, as mentioned above, the structure of the experiment is such that participants got more exposure to the noun vocabulary than the modifier vocabulary.[17] It is therefore possible that in the production phase, faster lexical retrieval for nouns could lead to a bias to say them first. (This in fact makes the opposite prediction from the hypothesis that learners will favor English order.)

(11) *Additional experimental predictions (not typologically relevant)*

PREDICTION 4. Native-like order preference ⇒ Mod-N order preference

PREDICTION 5. Faster recall of noun vocabulary ⇒ N-Mod order preference

4. Experimental results and analysis

4.1. Vocabulary learning

Recall that, during phase I of the experiment, participants were explicitly tested on their knowledge of the object names in the language and were required to achieve 75% correct (8 of the 73 participants did not pass this criterion). However, it is still possible that vocabulary learning differed across conditions, and this might lead to differential performance on the production test.[18] Fig. 1a shows the average proportion of correct picture choice (accuracy) on the comprehension picture-matching task by condition. Fig. 1b shows the average proportion of correct vocabulary produced by participants in the production test, again by condition. Across conditions, average vocabulary accuracy was generally high: 90–95% on the picture-matching task, and 80–90% on the production test.

To compare vocabulary learning across conditions, accuracy data from the picture-matching task and vocabulary accuracy from the production test were both subjected to logistic regression using mixed effects models.[19] Mixed-effects models allow the use of both fixed effects—covariates

[16] Here we define the regularization bias as necessarily relative to the input—that is, promoting boosting in a direction which is already favored. This differs from the notion of regularization as used in e.g. Hudson Kam and Newport (2005), who treat any kind of consistent or deterministic behavior as regularization.

[17] We designed the experiment in this way to make vocabulary learning more tractable to participants. This need not be considered an artifact relative to first-language acquisition, however, as it is well established that bare nouns are among the earliest elements in a child's lexicon, earlier than adjectives and numerals (Clark, 1995; Gentner, 1982; Thorpe & Fernald, 2006).

[18] For example, if participants in one condition in particular were struggling to learn the vocabulary, they may not have had adequate remaining resources to acquire the ordering pattern in the language, and their use of majority order in that condition might be lower for this reason alone.

[19] We used mixed effects logistic regression rather than ANOVA because the dependent measure, Accuracy, is a proportion, and therefore ANOVA is inappropriate (Agresti, 2002; Jaeger, 2008). A logistic regression for a dependent variable $y = pr(E)$, the probability of some outcome E (ranging between 0 and 1) is simply an ordinary linear regression for the log-odds of E: $log[y/(1-y)]$ (ranging between $-\infty$ and $+\infty$). For recent applications of mixed-effects models to problems in linguistics and cognitive science see Bresnan, Cueni, Nikitina, and Baayen (2007), Jaeger (2008), Baayen (2008), Janda, Nesset, and Baayen (2010). All the regression models reported in this article were analyzed with the lme4 package in R (Bates & Maechler, 2009).

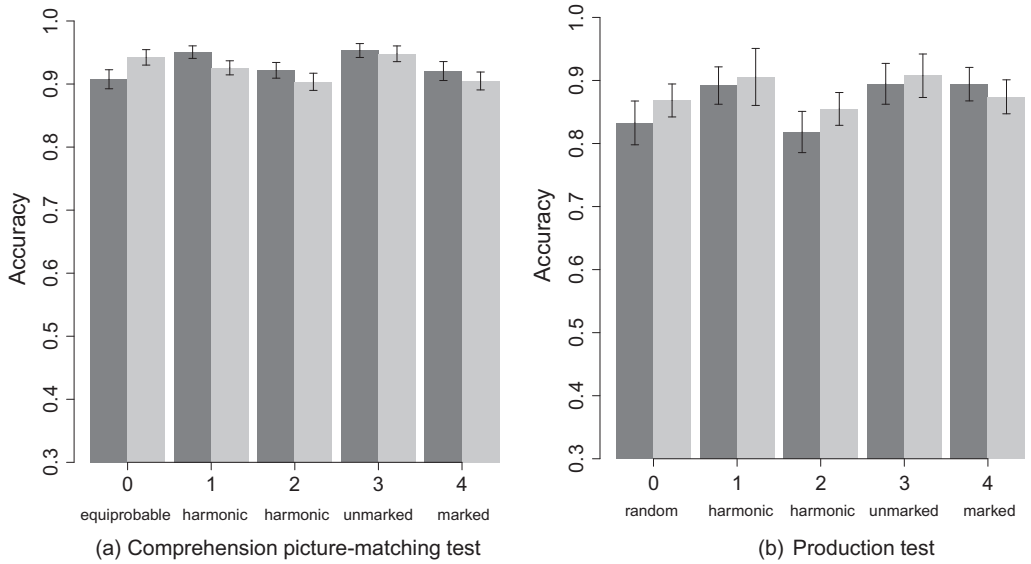


Fig. 1. Vocabulary accuracy by condition; left-hand bars pertain to numerals, right-hand bars to adjectives.

in the model which have fixed and reproducible levels—and random effects—covariates which represent a random sample from the set of all possible levels. For example, we treat individual participants as a randomly selected subset of all possible participants; we would like to generalize the participant effects to the population at large rather than attend to effects associated with particular participants: participants correspond to random effects. We will also treat the items in the artificial lexicon as random effects. In contrast, we do not consider the experiment conditions as a random sample of all possible conditions; we are interested in the particular effects associated with individual conditions: experiment condition is a fixed effect.

To evaluate the effects of a particular covariate (like experiment condition), the fit of two mixed-effects models, one with and one without that covariate, can be compared using the Likelihood Ratio test (Lehmann, 1986), a method for evaluating the fit of nested models which takes into account the added complexity of one model relative to the other. Two Likelihood Ratio tests were conducted to evaluate the effect of condition, first on the production task, and second on the comprehension picture-matching task. For both tasks, we obtained the likelihood values (a measure of the goodness of fit of a model to the data) for each of two mixed-effects models, one which includes condition as a fixed effect and the other which does not (both include participants and items as a random effect). The model which does not include condition as a predictor is called the null model; it is the simpler of the two. The model which *does* include condition as a predictor is called the alternative model.[20] The fit of a more complex model will always be at least as good as the fit of a nested simpler model; the Likelihood Ratio test is used to determine the probability (p-value) of the disparity in likelihood between the two models, given the difference in their degrees of freedom (df1–df2). The test statistic, shown in (12), approximately follows the chi-square distribution (with df1–df2 degrees of freedom; Wilks (1938)), and therefore this distribution is used to calculate a p-value indicating whether the effect of the covariate (here, condition) is significant.

$$(12)\qquad \begin{aligned} LR &= -2\ln\left(\frac{\textit{Likelihood for null model}}{\textit{Likelihood for alternative model}}\right)\\ &= 2([-\ln(\textit{Likelihood for null model})]\\ &\quad -[-\ln(\textit{Likelihood for alternative model})]) \end{aligned}$$

The Likelihood Ratio statistic LR is thus twice the difference between the negative log likelihood values for the null model and the alternative model. For the two models of the production task this gives:

$$(13)\qquad LR = 2(1764.16 - 1757.59) = 13.15$$

The test statistic, calculated in (13), is not significant ($\chi^2(8) = 13.15$, $p = 0.11$), thus adding condition as a predictor in the model does not result in a significantly better fit to the data in this case—in other words, condition is *not* an important factor in explaining vocabulary accuracy in the production task.

The Likelihood Ratio of the two models for the comprehension picture-matching task (using the negative log likelihood values for each model) is:

$$(14)\qquad LR = 2(1288.35 - 1280.73) = 15.23$$

The test statistic, calculated in (14), is marginally significant ($\chi^2(8) = 15.23$, $p = 0.06$). To investigate further the effect of condition on accuracy in the picture-matching task, several post hoc contrast tests were conducted using

[20] Both models include the factor modifier type—either adjective or numeral—as a fixed effect, and the more complex model also includes the interaction between condition and modifier-type; the more complex model has 2 × (number of conditions − 1) additional degrees of freedom.

mixed-effects models. Specifically, vocabulary accuracy for each condition was compared to every other condition. After Bonferroni correction for multiple post hoc comparisons there were no significant differences revealed by any of these contrasts.[21]

The results of these two model comparisons suggest that vocabulary learning did not differ significantly across conditions for either task. Asymmetrical levels of vocabulary learning can therefore be ruled out as a plausible explanation for any differences across conditions in our primary dependent measure of interest, use of the majority word-order pattern.

4.2. Regularization asymmetries: use of majority input order in the production task

For the purpose of analyzing participants' use of the majority order, each utterance was coded as either following the majority pattern in the condition or not. Since there was no majority order for condition 0 (the equiprobable condition), a single pattern—pre-nominal adjectives and numerals (the same as harmonic condition 1)—was arbitrarily chosen to stand in for the majority order. Fig. 2 shows the average proportion of majority order produced by participants in each condition. This figure, and subsequent analyses, include only trials where the correct vocabulary items were produced (12% of trials were excluded). In all graphs, error bars indicate standard error.

Recall that based on the hypotheses derived from the typology, learners in conditions 1, 2 and 3 were predicted to show regularization behavior; they were predicted to boost the use of the majority order over the input level (i.e., 70%, indicated by the dotted line in Fig. 2). By contrast, learners in condition 4 were predicted not to show regularization behavior; they were predicted not to boost, and potentially to use the majority order less than 70% of the time.

A simple way to test these predictions is the one-sample sign test. For each condition, a sign test was computed which compared each subject's average use of the majority order to the input level (70% for conditions 1, 2, 3, 4, and 50% for condition 0). As Fig. 2 suggests, while learners in conditions 1, 2, and 3 showed average levels of majority order production which were significantly greater than 70% ($p < 0.05$ in all cases), learners in condition 4 did *not* differ from 70% ($p = 0.29$). Although there is a trend toward probability-flattening (anti-boosting) in condition 4, a sign test evaluating the hypothesis that average use of the majority order by learners in condition 4 was *below* 70% was not significant ($p = 0.87$). As expected, learners in condition 0 did not differ significantly from the input in either direction ($p = 0.9$).

These results are clearly in line with the predictions; learners in conditions 1, 2, and 3 exhibited boosting behavior, while those in condition 4 did not. Recall that our hypothesis was that learners have a regularization bias which will necessarily apply when not in conflict with another bias; therefore the differences in boosting across conditions support the existence of a learning bias against the majority pattern in condition 4. However, while the analyses presented so far provide evidence for such a bias by revealing a lack of boosting by learners in condition 4, they do not provide evidence of a bias strong enough to produce anti-boosting by those learners.

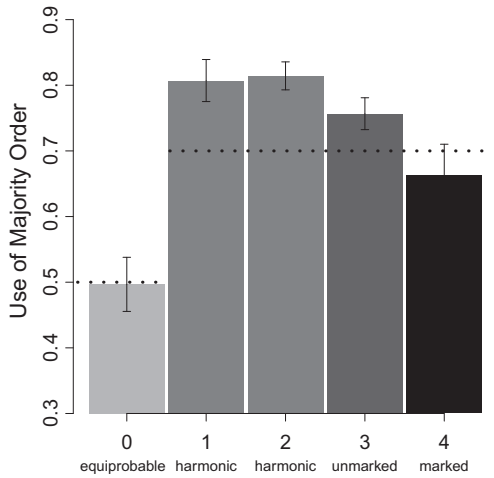


Fig. 2. Average use of the majority order by condition; dotted lines indicated input frequency.

The goal of using logistic regression to analyze these data is to investigate differences between conditions 1–4 in terms of use of the majority orders. However, before reporting individual comparisons of interest, we will first report the results of a Likelihood Ratio test designed to validate the claim that there is in fact an overall effect of condition. Here we compare two mixed-effects models with proportion of use of majority pattern as dependent variable: one which includes condition as a fixed effect (the more complex model) and the other which does not (again both models include modifier type as a fixed effect and participants and items as random effects). A more conservative test would not include condition 0, since it clearly did not induce boosting, thus only conditions 1–4 were included in this model. For these two models, the Likelihood Ratio (using the negative log likelihood values for each model) is the following:

$$(15) \qquad LR = 2(2502.22 - 2470.44) = 63.55$$

The test statistic, calculated in (15), is significant ($\chi^2(8) = 63.55$, $p < 0.0001$), indicating that the condition factor makes a significant contribution in explaining differences among learners' use of the majority order in the production test.

The model described above, which includes both condition and modifier-type as fixed effects, reveals not only a main effect of condition but also a significant interaction (the parameter estimates for this model, model 1, are shown in Table B.8 in Appendix B). In order to test the specific predictions made in (10) and (11), we will investigate these effects further below.

[21] Six comparisons were made (among conditions 1–4), thus the corrected alpha level used was 0.008. The only trend revealed was a general tendency for learners in condition 2 to show lower accuracy than learners in condition 3 ($\beta = -0.43$, $z = -2.31$, $p = 0.02$).

First, in order to confirm that learners in conditions 1, 2, 3, and 4 all showed *some* evidence of learning, we fit another model comparing majority order use in each condition to "majority" use in the equiprobable condition, again using a mixed-effects model (with no factor distinguishing modifier types). Recall the "majority" order for the equiprobable condition was arbitrarily set to pre-nominal adjectives and numerals; this choice has little or no consequence, however: Fig. 2 suggests, as expected, that learners in that condition used both orders almost equally often. This model revealed that all conditions (1, 2, 3, and 4) differed significantly from the equiprobable condition (condition 0 vs. condition 1, $\beta = 1.60$, $z = 6.19$, $p < 0.0001$; condition 0 vs. condition 2, $\beta = 1.59$, $z = 6.15$, $p < 0.0001$; condition 0 vs. condition 3, $\beta = 1.21$, $z = 4.75$, $p < 0.0001$; condition 0 vs. condition 4, $\beta = 0.74$, $z = 2.93$, $p < 0.003$).[22]

4.2.1. Regularization differences across conditions

While learners in conditions 1, 2, 3 and 4 achieved some level of learning—they did not behave like learners in the equiprobable condition—we already know (from the results of the Likelihood Ratio test, and the sign tests reported above) that use of the majority order differed across conditions. We return now to the predictions made in (10), namely that learners in the harmonic conditions should regularize to a greater extent than learners in the non-harmonic conditions, and further that learners in condition 3 should regularize more than those in condition 4. In what follows, we will show that *all these predictions are borne out statistically*. Unless otherwise noted, all models contain one or more levels of condition and modifier-type as fixed effects, participants and items as random effects, and majority order use as the dependent variable.

In order to test these predictions, first the production testing data from conditions 1, 2, 3, and 4 were subjected to a mixed-effects model with Helmert contrast coding—this type of contrast coding allows us to evaluate several key contrasts in a single model by testing each level of the condition factor against the mean of all previous levels. The model also includes interaction terms for modifier-type. The estimated parameters for this model, which we call model 2, are shown in Table 6. Given that we are using Helmert contrasts, the first three coefficients in the model indicate the difference in log odds of producing the majority order (i) between conditions 1 and 2, (ii) between the two harmonic conditions (1, 2) and condition 3, and (iii) between condition 4 and all other conditions, when the level of modifier-type specified by the interaction term is 0, i.e. for adjective trials only. The fourth coefficient indicates the difference in log odds of producing the majority order on numeral trials across all conditions. The last three coefficients are interaction terms, and indicate the additional effect on the difference in use of majority order between conditions for numeral trials.

This fit model provides several important results, spelled out in (16):

Table 6
Parameter estimates for mixed-effects model 2; (∗) indicates marginal significance.

Fixed effects	Estimate	Standard error	*z*-value	*p*-value
(Intercept)	1.286	0.119	10.785	<0.001*
Condition 1 vs. 2	−0.007	0.136	−0.053	0.958
Conditions 1, 2 vs. 3	−0.134	0.077	−1.739	0.082 (∗)
Conditions 1, 2, 3 vs. 4	−0.182	0.054	−3.391	<0.001*
Modifier type = Num	0.023	0.084	0.275	0.783
Condition 1 vs. 2 × Num	−0.034	0.063	−0.533	0.594
Conditions 1, 2 vs. 3 × Num	−0.074	0.033	−2.210	0.027*
Conditions 1, 2.3 vs. 4 × Num	−0.094	0.022	−4.185	<0.001*

(16) *Summary of results of mixed-effect model 2*
- i. Use of majority order (i.e. boosting) does not differ between harmonic conditions 1,2.
- ii. Use of majority order for adjectives differs marginally between the non-harmonic condition 3 and the harmonic conditions.
- iii. Use of majority order for adjectives differs between the non-harmonic condition 4 and all other conditions.
- iv. As indicated by the two significant interaction terms, the difference in use of majority order between the harmonic conditions and condition 3, and between condition 4 and all other conditions, is significantly larger for numeral trials.

4.2.2. Lack of non-typologically-relevant asymmetries

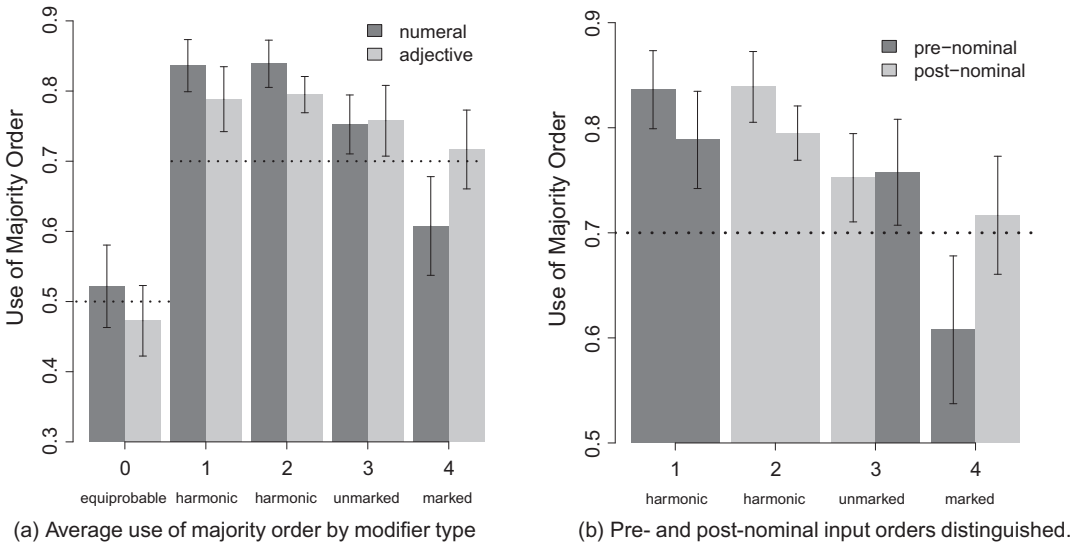
That the harmonic conditions do not differ from one another is important because it suggests, contra prediction 4 of (11), and consistent with the results from the equiprobable condition 0, that there is no general (or net) bias in favor of English orders for either adjectives or numerals (although below we will discuss the effect of native language on individual subject behavior in conditions 3 and 4).

Fig. 3 shows the average use of the majority order broken down by modifier type. It is possible to inspect those bars in Fig. 3 that correspond to English order, and those which correspond to non-English order. Not only were the two harmonic conditions, 1 (English-like) and 2 (anti-English), not significantly different, but across participants in all conditions, English order was used slightly less than half of the time (48%), including in the equiprobable condition—indicating further that there was no significant overall preference for either English or anti-English order. This rules out the possibility that differences between conditions are driven simply by a general preference for English order. Further, we can clearly also rule out the opposite prediction (5 of (11)), that faster lexical retrieval of nouns produces a general preference for noun-first order.

4.2.3. Asymmetry between harmonic and non-harmonic conditions

To further investigate the results in (16) from model 2, several additional models were fit individually comparing

[22] β values are regression coefficients for each contrast in a model. They can be interpreted as the predicted difference in log odds of producing the majority order, in this case between the equiprobable condition and each other condition. Wald Z scores are given by the *z* values.



(a) Average use of majority order by modifier type (b) Pre- and post-nominal input orders distinguished.

Fig. 3. Average use of the majority order by modifier type.

conditions 3 and 4 with the harmonic conditions. These planned comparisons relate, of course, to the prediction that participants would show a bias in favor of harmonic patterns and thus might use the majority order more frequently in conditions 1 and 2 (prediction 3 of (10)). We will first discuss models comparing condition 3 to the harmonic conditions, then we will turn to models comparing condition 4 to the harmonic conditions.

A model comparing conditions 1 and 3 reveals a significant interaction between condition and modifier type ($\beta = -0.12$, $z = -2.12$, $p = 0.034$). Further comparisons by modifier type reveal that the interaction effect is driven by a significant difference between the use of the majority order by learners in conditions 1 and 3 with *numerals* ($\beta = 0.29$, $z = 3.46$, $p < 0.0001$; the difference for adjectives is not significant, $\beta = 0.05$, $z = 0.65$, $p = 0.517$).[23]

Similarly, a model comparing conditions 2 and 3 reveals a marginally significant effect of condition ($\beta = -0.19$, $z = -1.91$, $p = 0.057$), and further comparisons by modifier type reveal a significant difference for numerals ($\beta = 0.28$, $z = 3.28$, $p = 0.001$; the difference for adjectives is not significant, $\beta = 0.09$, $z = 1.17$, $p = 0.240$). These results suggest that learners in the harmonic conditions did in fact use the majority order more than learners in condition 3, however the results are driven by a difference in performance for numerals.

More generally, it is worth noting again here that the effects seem to be due to a preference for harmonic patterns, rather than the effect of native-language; learners in harmonic condition 2 regularized (anti-English) N-Num more than learners in condition 3 regularized (English) Num-N order.

A model was also fit comparing conditions 1 and 4, and another model compared conditions 2 and 4. Both models reveal significant main effects of condition (condition 1 vs. 4, $\beta = 0.45$, $z = 2.77$, $p < 0.0001$; condition 2 vs. 4, $\beta = 0.44$, $z = 3.09$, $p < 0.0001$) as well as significant interactions between condition and modifier type (condition 1 vs. 4, $\beta = -0.25$, $z = -4.29$, $p < 0.0001$; condition 2 vs. 4, $\beta = -0.21$, $z = -3.60$, $p < 0.0004$). To investigate these effects further, additional models were fit which compared conditions by individual modifier type. These analyses reveal significant differences between both conditions 1 and 4 and conditions 2 and 4 for adjectives (condition 1 vs. 4, $\beta = 0.18$, $z = 2.34$, $p < 0.019$; condition 2 vs. 4, $\beta = 0.23$, $z = 2.91$, $p < 0.004$) as well as numerals (condition 1 vs. 4, $\beta = 0.63$, $z = 7.89$, $p < 0.0001$; condition 2 vs. 4, $\beta = 0.62$, $z = 7.49$, $p < 0.0004$). These results suggest that learners in the harmonic conditions used the majority order more than learners in condition 4, and further, in this case the differences exist for both modifier types (although, again the differences appear to be greater for numerals). Here, with adjectives, we see that regularization in non-harmonic condition 4—despite its congruence with English—was less than in harmonic condition 2—with anti-English order N-Adj.

4.2.4. Asymmetry between non-harmonic conditions 3 and 4

We now come to the comparison of greatest interest, between condition 3 and condition 4, where the strong effects of harmonic ordering do not come into play. We have already shown that, unlike learners in all other conditions, those in condition 4 did not boost the majority order, and model 2 already confirms that use of the majority order in condition 4 differs from (the mean of) all other conditions. But a direct comparison revealing a difference between conditions 3 and 4 would provide crucial added support for the hypothesized bias against condition 4 (prediction 2 of (10)) as both 3 and 4 are non-harmonic conditions.

[23] This and subsequent analyses comparing use of majority order across conditions for *single modifier types* use items but not participants as a random effect; because there is only a single trial type (either numeral or adjective), and participant is completely nested under condition, there is effectively only a single number (the proportion of trials which use the majority order) for each subject. This is similar to using *t*-tests for individual contrasts following an ANOVA.

A model comparing the two non-harmonic conditions revealed a marginally significant effect of condition ($\beta = 0.23$, $z = 1.79$, $p = 0.073$), and a significant interaction between condition and modifier type ($\beta = 0.11$, $z = 2.10$, $p = 0.036$). Further comparisons by individual modifier type reveal that, as Fig. 3 suggests, the interaction effect is driven by a significant difference between the use of the majority order with *numerals* by learners in conditions 3 and 4 ($\beta = 0.33$, $z = 4.52$, $p < 0.0001$; the difference for adjectives is only marginally significant, $\beta = 0.13$, $z = 1.17$, $p = 0.08$). Thus the prediction was borne out, but was restricted to numerals; in fact, as suggested by the plots, learners in condition 4 used the majority order significantly less often with numerals than adjectives ($\beta = -0.22$, $z = -3.14$, $p = 0.002$). Note that for numerals, the greater regularization in condition 3 occurs with the anti-English order N-Num, as opposed to the English order Num-N of condition 4. We return to this point in Sections 4.4 and 5.1.

4.2.5. Summary of regression analyses

To summarize the results of our logistic regression analyses: (i) no differences in participants' use of the majority order were found between the two harmonic conditions; (ii) non-harmonic condition 3 differed from the two harmonic conditions: the effect was driven by less frequent use of the majority order for *numerals* by participants in condition 3; (iii) non-harmonic marked condition 4 differed from the two harmonic conditions: participants in condition 4 used the majority order less frequently for each modifier type; and (iv) the two non-harmonic conditions differed from one another: participants in condition 4 were less likely than those in condition 3 to use the majority order for numerals. These results closely follow the predictions made above in (10)—although the contrast in boosting between numerals and adjectives was unanticipated; we will discuss this contrast in Section 5.3.

4.3. A closer look at experience with other languages

Although all participants were native English speakers, some had considerable experience with other languages—operationally defined as a native speaker of, fluent in, or with more than 10 years of study of a language other than English.[24] In order to investigate whether experience with another language impacted boosting levels, this information was entered as a factor in the regression model predicting use of the majority order. Specifically, the model included participants and items as random effects, and condition, modifier type, and "other language experience" as fixed effects. Other language experience had four levels, the first for those participants who did not have considerable non-English experience, and the remaining three corresponding to considerable experience with each of the three non-English patterns represented in the experiment sample. (Not surprisingly, there were no speakers of a pattern-4 language.) No levels of the other-language-experience factor were significant (based on the Wald z-scores, all $p > 0.28$). Further, the other language factor did not contribute significantly to increasing the fit (likelihood) of the model for the testing data ($\chi^2(3) = 2.943$, $p = 0.401$); it was therefore not included in any of the analyses reported above.

Further post hoc analyses were conducted to determine the effect of having taken classes in pattern 3-type languages since, as suggested by an anonymous reviewer, many students in the US study Romance languages which have pattern 3 as their main order (e.g. Spanish, French, Italian). A follow-up survey sent to each participant asked which languages they had studied in school and for how long. Of the 23 responses received from the 60 participants in conditions 1–4, the number of years of study of a pattern-3 language ranged from 0–14. First, a mixed-effects model was fit with participants and items as random effects and with three fixed effects: condition, modifier type, and number of years of classes in a pattern 3 language. This model revealed no significant effect of number of years of study ($\beta = 0.025, z = 0.305, p = 0.760$). A model including the interaction between condition and number of years of classes in a pattern-3 language was also fit, and again resulted in no significant main effect of number of years of classes, nor any significant interactions (based on the Wald z-scores, all $p > 0.12$). For the critical condition 3, with the 9 (of 13) responses received (range: 0–12 years) we computed the correlation between use of majority order and number of years of study. The result was a small, non-significant positive correlation ($r^2 = 0.09$, $p = 0.422$). A test of the correlation between use of majority order and number of years of pattern-3 language study for all participants together also revealed a small, non-significant positive correlation ($r^2 = 0.03$, $p = 0.451$). Because the number of years of foreign-language study was not controlled, and the survey results were incomplete, only limited conclusions can be drawn from these post hoc analyses. The analyses do suggest that future studies should exclude participants with many years of classroom study of languages that could potentially skew the results—the positive correlations in both our analyses were driven by the few participants (5) who had studied a pattern-3 language for more than 8 years.

4.4. Individual learners and within-condition variation

While the mean usage of majority order for each modifier type across conditions provides important evidence confirming the predictions made in Section 3.5, perhaps the most illuminating picture of the results can be seen by plotting estimates of the probabilistic rules each individual learner has acquired for each modifier type. We can achieve this by plotting performance in a two-dimensional space where the conditional probability of producing Adj-Noun appears on the x-axis and the conditional probability of producing Num-Noun appears on the y-axis, as in Fig. 4.[25] In this space, a completely deterministic

[24] Of these participants, for 19/28 the language in question uses the *same order for adjectives and numerals as English* (these included Russian, Korean, Mandarin, Cantonese, Turkish, Hindi and Telugu). Of those remaining, for 8 the language in question used pattern 3 (Spanish, Romanian, French) and for 1 participant pattern 2 (Afaan Oromo).

[25] For suggesting this extremely helpful description, we are most grateful to Don Mathis.

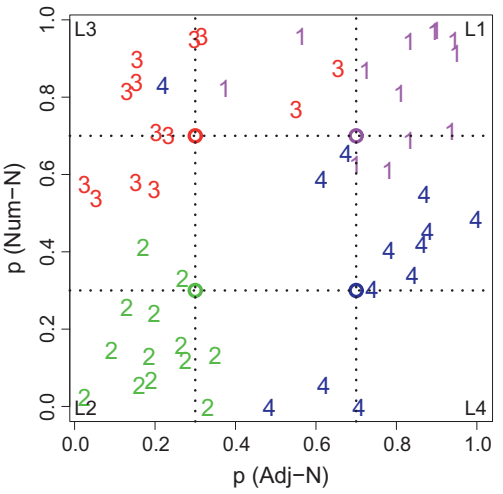


Fig. 4. Plot of p(*N-final*|*Num*) by p(*N-final*|*Adj*) for productions of individual learners in each conditions. Participants in the harmonic conditions 1 and 2 are indicated by purple '1's and green '2's respectively; red '3's indicate non-harmonic condition 3, and blue '4's indicate non-harmonic condition 4. (Training probabilities are open points.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

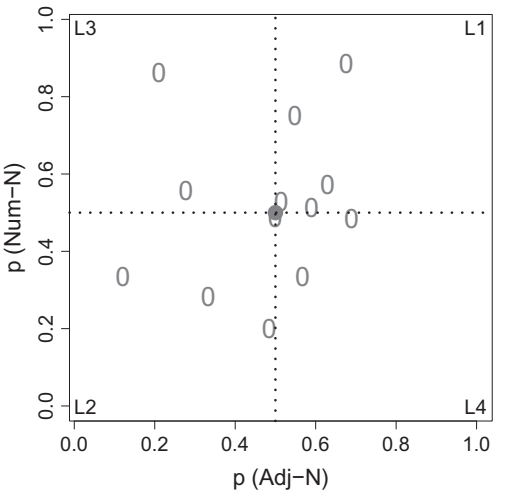


Fig. 5. Plot of p(*N-final*|*Num*) by p(*N-final*|*Adj*) for individual learners in the equiprobable condition 0 (training probabilities are indicated by the gray point).

version of each language type in (2) lies in one of the four corners. The harmonic languages 1 and 2 appear in the upper right and lower left corners, respectively labeled L1 and L2, since the former involves using both Adj-Noun and Num-Noun 100% of the time, while the latter involves using these orders 0% of the time (Noun-Adj and Noun-Num are used categorically). The non-harmonic languages 3 and 4 are then in the upper left and lower right corners, respectively labeled L3 and L4.

Each training condition corresponds to one of the four open circles in Fig. 4 (refer to the descriptions in Tables 4 and 5). For example, training condition 4 corresponds to the lower-right open circle, located at $x = 0.7$ (probability for Adj-N) and $y = 0.3$ (probability for Num-N); this circle is closest to the L4 corner, which constitutes the majority pattern for condition 4. This circle is blue: the plot is colored by condition, and condition 4 corresponds to blue. The probabilities of individual participants' word orders in the production test are indicated by colored digits; the x,y coordinates of each blue '4' gives, for one participant in condition 4, the probabilities of producing Adj-N, Num-N at test. The purple '1's, green '2's, and red '3's give the production probabilities for participants in the other conditions.

This plot makes it quite clear that individual learners in condition 4 behave differently from those in the other conditions: in conditions 1, 2 and 3, learners have typically moved into the corners from the training points for their condition—they have boosted. But *no learner* has moved into the language 4 corner (that is, the area delimited by the dotted lines closest to L4 in Fig. 4). In fact, there is a strikingly empty space surrounding the L4 corner compared to the other corners, and many of the participants in condition 4 appear to have shifted toward one of the harmonic patterns (more of them toward L1, a point we return to below). That is, learners in condition 4 were likely to shift the input toward a pattern which better satisfies the proposed substantive bias, providing further support for the idea that this bias is at play during the language-learning process.[26]

For the equiprobable condition 0, although the input has no general trend, the hypotheses which we set out do in fact lead to a prediction about how learners might shift the input pattern. Specifically, if learners *do* shift the input pattern, they should move towards a harmonic language, or the unmarked language 3, rather than towards language 4. Fig. 5 shows the production output for each individual participant in the equiprobable condition. The dotted lines separate the plot into four quadrants, one for each language, to help visualize how each learner's output is shifted compared to the input. The plot makes it clear that, despite a few who shifted minimally toward the L4 corner, most learners in the equiprobable condition did in fact acquire a language which is shifted toward L1, L2, or L3.

[26] Learners in the other three conditions also tend to cluster more tightly than participants in condition 4, as shown by the increased variance in that condition—roughly twice that of any of the others ($var_{cond4} = 0.064, 0.041$ for adjectives and numerals respectively, compared to $var_{cond1} = 0.018, 0.028$, $var_{cond2} = 0.015, 0.009$, $var_{cond3} = 0.023, 0.033$). For each of conditions 1–4, the mean Euclidean distance of participants' production probabilities from their condition's majority-pattern corner are, respectively, 0.291, 0.279, 0.391, 0.535; since the training data distance is 0.424, this means that in every condition except condition 4, participants on average moved closer to their majority pattern. Most interestingly, the mean distance of condition 4 is significantly greater than that of condition 3 ($p = 0.025$). The same result obtains for a different distance measure, the cross-entropy between the probability distribution of a participant's productions and their corner ($p = 0.043$; both p-values from one-sided permutation test: Good, 2005). Since the pattern violating the universal is a *conjunction of two* orders, neither of which alone constitutes a violation, a two-dimensional metric—distance to the majority corner—is particularly appropriate here.

5. Discussion

5.1. The observed learning bias, language change, and typological universals

The results presented here corroborate previous findings which indicate the existence of a *regularization bias* on the part of the learner (Hudson Kam & Newport, 2005, 2009; Sandler et al., 2005; Sankoff & Laberge, 1980; Singleton & Newport, 2004). Our results also provide evidence for the existence of a harmonic learning bias and—crucially—for **L4*: the *substantive* bias against word-order pattern 4, Adjective-Noun & Noun-Numeral. First, learners' more frequent use of the majority order (boosting) in conditions 1 and 2 provides evidence for a strong bias in favor of harmonic patterns—patterns which preserve the position of the noun with respect to both modifier types. Second, learners' lack of boosting in condition 4, along with the fact that they showed less use of the majority order than learners in *all other conditions*, provides evidence for a bias against the marked pattern specifically. The latter conclusion is illuminated by the picture of individual participant learning provided in Fig. 4: it is not simply that learners in condition 4 do not boost, rather they actually appear to shift the input pattern toward one of the other languages rather than match their training input.

If biases in the cognitive system influence language acquisition, pushing learners in certain directions rather than others, then, all else equal, over generations of learners, languages which (better) satisfy those biases are expected to outnumber those which do not—in this case, resulting in Greenberg's Universal 18. As will be made clear in our formal model below, we believe the biases uncovered in the experiment are best treated as probabilistic. The prediction is then that, generally speaking, although each pattern is possible, pattern 4 is significantly less likely to be acquired veridically, or to be the result of change from another pattern. Computational models of learning biases suggest that even very subtle preferences can be magnified dramatically over generations (Kalish, Griffiths, & Lewandowsky, 2007; Kirby, 1999; Kirby, Smith, & Brighton, 2004; Reali & Griffiths, 2009).

Of course, in the case of specific languages, predictions as to how fast change is expected to occur must take into account numerous other factors which might accelerate or dampen potential changes.[27] What is clear is that if probabilistic learning biases lead learners to change languages in the way we have sketched out, change in particular directions (i.e. from non-harmonic to harmonic, or from pattern 4 to harmonic) should be *more common* than change in other directions, all else equal. In the next section we discuss how these biases can be formally characterized, and briefly summarize the results of a probabilistic model of learning in the experiment. Finally, we return to the issues raised in the introduction about the relevance, for explaining typological universals, of artificial language learning experiments with adults, and the nature of the cognitive bias hypothesized to guide learning towards universally preferred patterns.

5.2. Testing the hypotheses using Bayesian modeling

To strengthen our conclusions about the effect of the hypothesized biases on learner outcomes (8), in work reported elsewhere (Culbertson & Smolensky, submitted for publication; see also Culbertson, 2010, chap. 3), we constructed a formal model of learning in the experiment. Such a model allows us to determine whether we can, in a mathematically precise way, successfully explain the results of the experiment using the types of biases that we have claimed are at work here.

Our model assumes that a learner's initial knowledge about the new language is a probability distribution p over a space of possible grammars: for every grammar G in this space, $p(G)$ is the learner's belief about the probability that G is the grammar of the language to be learned. At the beginning of the experiment, this probability distribution over grammars reflects the learner's biases about how likely each grammar is *a priori* (i.e., prior to seeing any data from the target language).[28] After receiving the training data in a particular experimental condition, the learners' distribution has shifted towards those grammars that best explain that training data (i.e., those grammars that make that training data set most likely). This *a posteriori* distribution results from the combination of the *a priori* distribution and the training data; the formal relation between them is given by Bayes' Theorem,[29] and this framework for analyzing learning constitutes a Bayesian model.

From the perspective of the learners, the prior probabilities of possible grammars constitute their initial knowledge; it is a *given* bias relevant to language learning, assumed to be essentially the same for all experimental participants. From the perspective of the theorist, however, the bias is exactly the *unknown* we are trying to find via the experiment and model. This is done by constructing a parameterized family of prior distributions which we theorize to have a *structure* capable of instantiating biases of the type we hypothesize, and then evaluating which particular parameter values best predict[30] the data that the participants produce in the testing phase of the experiment, assuming that the utterances participants in a particular

[27] It is important to note that although we predict regularization of the type of inconsistent (unconditioned) variation used in our experiment, many languages exhibit variation of another kind—namely lexically conditioned variation. The latter type of variation is in fact often completely deterministic and therefore the pressure to regularize may be quite a bit weaker.

[28] The learner's distribution prior to the experiment is expected to reflect both previous learning, e.g., of the native language, and biases present prior to *any* learning (see Section 5.3).

[29] Generally, $p(A|B) \propto p(A)p(B|A)$. In our case, $p(G|T_k) \propto p(G)p(T_k|G)$, where T_k is the training data for condition k. That is, the posterior probability of a grammar G, given training data T_k, is proportional to the prior probability of G times the probability (or likelihood) that G generated the training data T_k.

[30] Please note that, following the logic of a generative model, we will follow common practice and say that the behavior generated by the model is "predicted" by the model, although it is of course true that since the parameters of the model are fit to the experimental data, these are not *a priori* predictions.

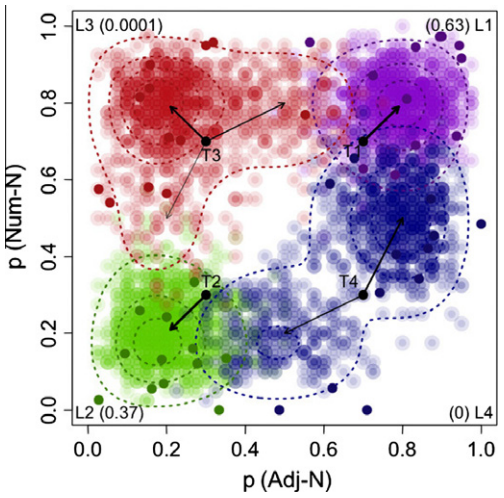


Fig. 6. Plot of *p*(*N-final*|*Num*) by *p*(*N-final*|*Adj*) showing individual learners' data from the experiment as well as predicted data from the model. The arrows represent the learners' biases.

condition produce at test are generated by grammars sampled from the *a posteriori* probabilities they assign based on the training data for their condition. In other words, the structure of the prior represents our theory of learning bias and is built into the model, but its parameter values are fit based on the experimental data. Crucially, the model's structure treats all four language types symmetrically: *no asymmetries are built in*. The model best accounts for the experimental data, however, when the parameter values are strongly asymmetric: that is a result of analyzing the experimental data, not of model stipulation.

For the formal specification of this model and the details of its use to estimate the learners' bias we must refer the reader to Culbertson and Smolensky (submitted for publication).[31] To display the particular language-learning bias—the prior probability distribution over grammars—estimated by the model computations, we directly plot its predicted influence on the learners in Fig. 6. In this plot, *actual* testing data from learners in the experiment is represented by *solid* points (colored by condition), and a sample from the distribution of learning outcomes *predicted* by our model for each condition is shown as superimposed *partially transparent* points. All the learners in condition 2 are trained on the same data mixture T2; but because of the bias, they do not produce the same mixture that they heard. By and large, they move closer to the L2 corner, boosting the proportion of the majority word orders and reducing the proportion of the minority orders relative to their training data T2. The mixture shift predicted to arise from the model's estimated prior bias is shown by the green-shaded area near the lower-left corner: this reflects the *a posteriori knowledge* that condition 2 learners arrive at, according to the model, by combining their condition 2 training data T2 with the prior bias; areas of darker green contain word-order mixtures that these learners are more likely to produce at test. This distribution has moved closer to the corner than the input data T2. This learning-induced shift is encapsulated by the arrow emanating from the T2 point: this points to the most probable word-order mixture according to condition 2 learners' a posterior knowledge of their target language. *These arrows represent the effect of the learning bias.*

The striking property of Fig. 6 is of course the clear differences between the behavior predicted in the harmonic conditions (training data T1 and T2) and each of the other conditions. Unlike the majority of participants in conditions 1 and 2, who are predicted to shift towards the nearest corner, the condition 4 participants are predicted to split, somewhat less than half of them shifting towards the harmonic language L2, somewhat more than half towards the other harmonic language, L1. Such a split in the actual data is shown by the distribution of the solid blue points; this split is captured in the model by the predicted post-training distribution, the blue shading, which is bimodal—reflected in the two arrows emanating from T4, each arrow's tip falling at one of the modes, i.e., a locally-most-probable word-order mixture.

The condition 3 participants display the most complex behavior. In the actual experiment, a number of participants shift in favor of their majority pattern, moving towards the L3 corner. However two other, smaller groups of participants shift towards the harmonic languages L1 and L2. The distribution predicted by the model from its estimated bias is trimodal, shown by the three arrows leaving T3. These three shifts and their relative probabilities are roughly indicated by the dotted curves on the plot. The thickness of an arrow emanating from the training point is a rough indication of the proportion of learners in that condition that the model predicts will shift in that direction.

Evident in Fig. 6 is the repulsion from the L4 corner of the posterior distribution over grammars in all conditions. This arises from an asymmetry in the estimated prior, which is a particular mixture of components each favoring one of the corners of the grammar space. The weight of each component in this estimated prior—computed by optimizing the model's prediction of the observed testing

[31] The learner's hypothesis space of generative models is the set of all probabilistic context-free grammars (PCFGs) of the form shown in Table 4. The theory postulates that the structure of the prior is (i) conjugate to the generative model; (ii) symmetric with respect to regularization of all word orders; and (iii) dependent jointly on the probabilities of orders for Adj and Num phrases. From (i) it follows that the prior is built of beta distributions (conjugate to the binomial distributions specified by PCFGs). From (ii) it follows that the two shape parameters of the beta distributions, which determine the regularization pressure, must be the same for all orders. From (iii) it follows that the prior is built of products of two beta functions, one each for the probabilities of {Adj, N} and {Num, N} orders. The most general conjugate prior of this form is a mixture of four such products, each concentrating probability in a different quadrant of the hypothesis space, i.e., each favoring one of the word-order patterns 1–4. This structure is fully symmetric, treating all word-orders identically. The mixture coefficients (along with beta shape parameters) are computed by maximizing the likelihood of the production-test data from all participants in conditions 1–4; the predicted production probabilities for a given participant are determined, via Bayes' Theorem, by the prior and the training data for that participant; this is a beta-binomial distribution. The model is a batch learner. The data fit excludes outlier participants (1 from each condition) whose use of the majority order for Adj or Num was more than two standard deviations from the condition mean.

data—is shown in Fig. 6, in parentheses, next to each L1–L4 label.[32] The components favoring the harmonic languages have the greatest weights in the prior, 0.63 and 0.37 for L1 and L2 respectively. Next strongest is the L3 component, with weight 0.0001. While this is much smaller than the weights for the harmonic languages, it is nonetheless greater than the weight for the L4 component in the prior, which is the lowest possible: exactly 0. Even a weight of 0.0001 results in predicted behavior for the condition 3 participants that is much different than that for the condition 4 participants.

The results of the Bayesian model strongly support the conclusions drawn from the analyses of the experimental data reported in Section 5.1. We hypothesized a prior structure that treats all word-order patterns symmetrically, and the parameters that best explain the experimental data do in fact exactly encode the asymmetries spelled out in Universal 18. We emphasize that these asymmetries are *not* built into the model; rather, they reflect the empirical facts revealed by the experiment. The learning bias estimated with this model disfavors L4 as much as it possibly can, and forcefully favors the harmonic languages L1, L2 over the unmarked but non-harmonic language L3.

The consequence of this bias is that for the conditions targeting each of the three languages L1–L3 corresponding to typologically-well-attested patterns, for the majority of learners, the proportion of produced utterances displaying the target pattern is predicted to exceed that in their training data. For the language L4 corresponding to a typologically extremely rare pattern, the prediction is entirely different.

A further result of some interest is that the bias component for L1 is the largest, resulting in more shifts from both non-harmonic training conditions towards L1 than towards L2 (particularly from T4). As we noted in Section 2.1, based on the *typological data* we might have predicted a preference for L2 over L1. However, it seems likely that this finding reflects participants' preference for their native language (a preference that is rather subtle, as it is largely not in evidence in the experimental results). The question remains open, then, as to whether a learning bias underlies this particular typological asymmetry—this can be straightforwardly tested in the Mixture-Shift Paradigm by using speakers of a non-harmonic language.

Taken all together, the predictions the model makes are strikingly in line with the observed experimental results; we have successfully provided a formal model of learning in the experiment which relies on the interaction of biases we have hypothesized—the substantive bias parallel to Universal 18, the harmonic bias and the regularization bias.

[32] Because the variance among learners in condition 0 is very high (see Fig. 5), and we are not confident that the experimental results are reliable, the modeling results presented here do not incorporate the condition 0 data. However, when those data are included, the results are quite similar; the weights of the two harmonic patterns are the highest (with L1 higher than L2), the weight for the marked pattern is zero, and the weight for the unmarked pattern falls in between, although it is slightly higher than that shown in Fig. 6.

5.3. Conclusions

5.3.1. Findings

We have shown that adult learners exposed to probabilistic artificial languages characterized by a particular pattern of dominant {Noun, Adjective}, {Noun, Numeral} orders, along with some variation, acquire grammars which do not replicate veridically, nor regularize, the training data. Adopting a Bayesian perspective allowed us to view this learning as an inference process that was systematically affected by learners' (unconscious) prior expectations about which ordering patterns are more likely, and what level of variability should be encoded in probabilistic rules. To precisely characterize these biases and determine how learning was affected by them, we analyzed the experimental data using logistic mixed-effects regression models as well as an explicit Bayesian learning model trained on the same data patterns as the experimental participants.

We found several biases at work during learning in our experiment. The *regularization bias*, favoring grammars with less variation, had the effect of generally pushing the grammars inferred by learners to be more regular, i.e., less variable (learners' output was pushed towards the corners of the mixture space in Fig. 6). Additional biases were also at work: one led learners to prefer *harmonic* ordering patterns (in this case, those which preserve the position of the noun with respect to both modifier types); the other disfavored the *particular non-harmonic pattern* 4 which combines Adj-Noun order with Noun-Num order.

In the best characterization of the bias derived from the Bayesian model of learners in the experiment, the distinction between the two non-harmonic patterns is discrete and decisive: the bias places exactly zero probability on pattern 4—yielding no regularization for that pattern—but a non-zero probability on pattern 3 that is sufficient to generate considerable regularization. This substantive bias mirrors a typological generalization—Greenberg's Universal 18—suggesting that at least some cross-linguistic regularities may have their origin in the biases of learners.

More specifically, the interaction between the regularization bias, the harmonic bias, and the substantive *L4 bias resulted in an asymmetrical pattern of learning outcomes across conditions in our experiment. Despite the fact that learners in each of the conditions were exposed to languages which were, statistically speaking, identical, learners in condition 4—which used predominantly the disfavored pattern 4—behaved strikingly differently from learners in the other conditions: they inferred grammars that were shifted *away* from their predominant pattern 4, towards a harmonic pattern. By contrast, in conditions 1 and 2—which used predominantly harmonic patterns—the output of learners was faithful to the input pattern and was in fact strengthened by the regularization bias. Falling somewhere in between were learners in condition 3, who were exposed to the predominant pattern Noun-Adj, Num-Noun—which lacks the hypothesized advantage of being harmonic, but is not specifically disfavored. Roughly half these learners acquired regularized versions of the input grammar, but the other half shifted toward one of the more preferred harmonic patterns.

The behavior of learners in all conditions revealed the bias favoring harmonicity; the unique behavior of the learners in condition 4 revealed the substantive bias against the word-order pattern violating Universal 18. Learners' behavior in all four experimental conditions is predicted by the Bayesian model's substantive and regularization biases when fit to the experimental data, as shown in Fig. 6.

As we will spell out in the remainder of the article, we believe that the bias against the Universal-18-violating pattern 4—and in particular the contrast between this pattern and the unmarked non-harmonic pattern 3—is a key finding because it cannot be attributed to asymmetries in participants' native language, and because it does not plausibly reflect a domain-general constraint: it therefore constitutes evidence for the existence of cognitive biases specific to language that are relevant to explaining typological asymmetries.

5.3.2. Adult language learning in the laboratory

Any inferences from the observation of asymmetric learning in the experiment to universal biases operating during native-language acquisition require consideration of two types of disparities between learning in the experiment and first-language learning: first, the existence of native language competence in our adult learners, and second, the age of our learners, who are beyond the putative critical period for language learning. These same issues arise, of course, in the study of late second-language acquisition.

The effect of native-language experience. The Bayesian perspective introduced in Section 5.2 suggests the following bigger picture, to a first approximation. A participant's *prior* (i.e., bias) when beginning the experiment is derived from (i) biases applying prior to any linguistic experience—the *universal prior*, combined with (ii) linguistic experience prior to the experiment. As we discuss shortly, this picture of laboratory language learning can be conceived of as parallel to a theory of L2 learning more generally according to which learners retain access to (universal) biases or principles which are at play during native-language learning, but these biases are to some degree impacted (weakened or strengthened) by first-language experience. In our case, pre-experimental experience with English will promote pattern 1, pre-nominal adjectives and numerals, relative to the other patterns.

In fact, the Bayesian experimental prior that best explains our data encodes the preference ordering $1 \succ 2 \succ 3 \succ 4$. If experience with English promotes *only* pattern 1, then we can reasonably conclude that the universal prior must encode $2 \succ 3 \succ 4$. Whether experience with English also promotes other harmonic patterns, i.e. pattern 2, is not clear: pattern 2 is the opposite of English, therefore learning such a pattern requires reversing *two* distinct word-order generalizations (for Adj and Num); English experience plausibly should *demote* pattern 2. In light of previous research revealing harmonic biases (see Section 5.3.3), we are inclined toward the hypothesis that it is not English experience that promotes pattern 2; rather, participants' preference for harmonic patterns derives from the universal prior.

What we would like to argue most strongly, however, is that $3 \succ 4$ *must* be part of the universal prior bias—that is, it is not English experience that promotes pattern 3 relative to pattern 4. For suppose that English experience promotes all patterns with pre-nominal modifiers. Since patterns 3 and 4 each have one pre-nominal modifier, this will not yield any preference for 3 over 4—unless for some reason English experience promotes 3's pre-nominal *numerals* more than it does 4's pre-nominal *adjectives*. Prime facie, this is unexpected given that adjectives are more frequent than numerals in English (by roughly 8:1 in the Brown Corpus).

However, we have noted that learners in condition 4 were particularly unlikely to regularize their predominant N-Num order: perhaps English-speaking learners have a particular problem adapting to the post-nominal numerals in pattern 4, while pattern 3's post-nominal adjectives are not as problematic. An anonymous reviewer suggests the following explanation: "3's N-Adj order is at least *possible* in English, while 4's N-Num order is not". We must ask: (1) Would difficulty of N-Num relative to N-Adj suffice to explain the asymmetries found in the experiment? (2) Could the proposed English source of such a difference in difficulty provide a viable explanation? We do not believe so, for several reasons.

First, learners in condition 2 had no trouble learning N-Num order—in fact, if anything, they were more likely to regularize N-Num than N-Adj. Along similar lines, learners in condition 3 were quite likely to shift toward a harmonic language by *increasing* the use of N-Num. Similarly, several learners in condition 4 moved towards a harmonic pattern by *decreasing* the use of the English order Adj-N. These results would clearly not be predicted if the driving bias were simply to minimize use of orders not well-attested in English.

Second, in English, post-nominal attributive adjectives are restricted to a particular semantic class: stage-level predicates, denoting temporary properties (e.g. "the planets visible (tonight)"). The adjectives used in the experiment belong to a different class: individual-level predicates, denoting permanent properties (cf. "*the planets red"). Further, post-nominal adjectives are in any event extremely rare in English (for example, in the Brown corpus, only 0.05% of adjectives follow the noun they modify).[33]

We believe instead that learners have particular difficulty in condition 4 compared to condition 3 because of the substantive bias against pattern 4. As confirmed by our Bayesian model, learners in condition 4 react to this

[33] Frequency-based explanations are of course highly sensitive to the choice of categories counted. At one extreme, under a deep linguistic analysis according to which the stage-level/individual-level predicate distinction is respected (semantically and/or syntactically: Cinque, 1993), English provides no evidence for the type of N-Adj sequences used in the experiment. At the other extreme, under a superficial linguistic analysis, which fails to make this distinction, English provides evidence that Adj can follow N, but also that Num can follow N. (Not just in the 'Brothers Four' construction, but also 'Bus Four', 'I missed the bus four times', and so forth. In fact, in the Brown corpus, numerals follow nouns with a probability of 5%—100 times that for adjectives.) Absent a plausible level of linguistic analysis that yields the right English counts to clearly favor N-Adj over N-Num, this line of explanation cannot be considered viable.

by shifting toward a harmonic pattern. However, they shift *more* towards the harmonic pattern resembling English than pattern 2—resulting in the particularly low rate of N-Num usage. This explanation makes a clear prediction: if learners were native speakers of a harmonic pattern 2 language (e.g. Basque), in condition 4 they should be more likely to shift toward harmonic pattern 2, and therefore would show lower rates of Adj-N usage.

Critical period. How relevant to the explanation of typological universals are results from our artificial language learning experiment with adults? Such experiments are clearly *directly* relevant to the extent that adult second-language acquisition contributes significantly to typologically-relevant language change. However, even if it should prove true that such change is driven only by native-language acquisition by young children, adult learners would still be relevant to the extent that their biases carry information about the biases of young child learners. It is certainly a logical possibility, since adult learners are outside of the critical period for language acquisition, that their biases are unrelated to those of children, and that they rely on strategies for learning that children do not. But whether learners beyond the critical period in fact learn language in a fundamentally different way is much debated among second-language researchers (Bialystok, 1997; Friederici, Steinhauer, & Pfeifer, 2002; Hakuta, Bialystok, & Wiley, 2003; Johnson & Newport, 1989; Vainikka & Young-Scholten, 1996; White, 2003).

On-going extension to child learners of the experimental work reported here will address this issue most directly. But our key finding already suggests the validity of adult data for uncovering cognitive biases that are relevant to universals and hence relevant to native-language acquisition if that is what drives language change.

In particular, absent a plausible hypothesis predicting that adult problem solving strategies, or particular properties of the experimental learning task, would favor pattern 3 over pattern 4, it would seem that the preference $3 \succ 4$ observed in this experiment should be added to the substantial list of results from experiments showing asymmetries in adult language learning that parallel typological asymmetries and cannot be explained by factors distinguishing the experimental learning task from the native-language acquisition context (Christiansen, 2000; Finley & Badecker, 2008; St. Clair et al., 2009; Wilson, 2006, to name a few). As this list continues to grow still more impressive, it becomes increasingly implausible to suggest that the many and diverse correlations between adult learning and arcane asymmetries in typology are merely coincidental. The case becomes increasingly strong that the most parsimonious explanation is a common cause: commonality between the biases in adult laboratory learning and the biases in the first-language acquisition processes that affect language change and shape the distribution of languages across the world.

5.3.3. Specificity and origin of the biases

The regularization bias. Our experiment uncovered a strong regularization bias on the part of adult learners. A bias in favor of more regular systems of knowledge appears to have very wide scope, operative in learning systems in cognitive domains beyond language.[34] The bias is manifest in children (apparently even more so than adults: Hudson Kam & Newport, 2009), so it is relevant to cognitive development, including language acquisition. The ultimate source of the bias is plausibly functional—the greater ease of use or learning of simpler systems, as well as the increased utility arising from their applicability to a more general class of contexts.

In our work, the regularization bias functions to provide a dependent measure of learning: the degree of regularization, which, we have hypothesized, is modulated by learning biases. Regularization is a means to our primary goal, the study of substantive biases in language learning and the relation of these biases to typological asymmetries.

Word order biases: cognition-internal status and time scales. The results reported here provide evidence for two biases within the cognitive system that are at work during the language-learning process; a bias in favor of harmonic languages, and the *L4 bias parallel to Greenberg (1963)'s typological generalization, Universal 18. Although cross-linguistic typology is certainly the result of many interacting factors, our results suggest that a learning bias should be part of any explanation for this generalization. This conclusion is significant, given that the idea that psychologically real biases in the cognitive system exist, and strongly shape the space of possible human languages, has recently been under assault (see Evans & Levinson, 2009; Tomasello, 2009 among others).

To be more specific, the *existence* of typologically-relevant cognitive biases, and in particular the substantive *L4 bias, is the primary conclusion we draw from the experimental results. Importantly, the finding that such biases exist on the time scale of our experiment—that is, revealed by individual participants in the course of a single experimental session—is not consistent with theories according to which typological asymmetries are the result of factors external to cognition. This includes theories which explain recurrent patterns as resulting from accidental geographic or cultural factors (Bybee, 2009; Dunn et al., 2011; Levinson & Evans, 2010, p. 2743), and those which hypothesize that functional factors induce asymmetries through language change across generations *only* (Bader, 2011, p. 345; Blevins & Garrett, 2004, p. 118; Christiansen & Chater, 2008; Levinson & Evans, 2010, p. 2738). Although a potential explanation of *some* typological asymmetries, these *cannot* account for laboratory language learning asymmetries like those demonstrated by our experiment.

While the experimental results are not consistent with accounts of universals that deny cognition-internal asymmetries, they are consistent with many accounts attributing cognitive reality to the asymmetries manifest in universals. We now elaborate briefly on a number of these possible cognitive accounts (although we will not argue that our experiment favors any one of them) in order to bring out some of the many theoretical dimensions along which viable alternatives must be teased apart in future

[34] See Hudson Kam and Newport (2009) for a proposal concerning the mechanism of regularization which appeals to very general memory limitations.

work. These accounts differ in placing the locus of the universal cognitive biases in the genome, or not.

Harmonic bias. What is the scope of the harmonic bias? A preference for harmonic languages is plausibly the reflex, within the locus of syntax learning, of a preference for simplicity (or generality) with scope spanning all learning in higher cognition. Potential ultimate sources of such a preference include the reduced computational cost of processing simpler hypotheses and the wider utility of more general knowledge (similar to a general regularization bias).

It is worth noting that what counts as "simple" is far from theory-independent, of course, and the case of Universal 18 is not altogether straightforward. A highly superficial linguistic analysis will not do, as encoding the harmony generalization requires at least the syntactic distinction between noun and modifier (or a semantic counterpart). On the other hand, a sophisticated linguistic analysis may not do either, as adjectives and numerals do not form a natural class in many syntactic theories, which assign them different structural roles (Svenonius, 1994; Ticio, 2003; Zamparelli, 2000). Semantic interpretation also requires that they be treated quite differently ('tall girl' creates a single discourse referent with an associated attribute; 'two girls' creates multiple discourse referents which can each be associated with their own attributes). Testing a general simplicity hypothesis will require an adequate general theory of complexity with the right implications for syntax learning; perhaps the Minimum Description Length principle (Grünwald, Myung, & Pitt, 2005) or minimizing Kolmogorov (algorithmic) complexity (Chater & Vitányi, 2003) would suffice.[35]

If the scope of the harmonic bias is taken to be *syntactic* processing, and with respect purely to processing rather than learning efficiency, Hawkins (1990, 1994, 2004) has proposed a symbolic theory of human parsing which is claimed to predict a parallel asymmetry: processing embedded phrases is less costly in grammars with harmonic word orders. (The theory also addresses many other universal asymmetries. See also Frazier, 1985.)

And finally, certain theories in generative linguistics posit a non-functional source, with scope restricted to word order, located within syntactic learning: a bias in the form of an innate principle such as 'heads are ordered [left/right] of their complements'. With a single parameter value (*left* or *right*) for a principle applying to all syntactic heads (e.g., Adj, Num), this generates only harmonic word orders. Departures from harmony then require additional, language-particular rules, so such non-harmonic grammars are 'marked' and thus presumed to be (all else equal) less frequent typologically (cf. Müller, 1999, p. 783).

The experimental results provide strong evidence for the cognitive reality of a bias favoring harmonic word orders: our participants displayed a bias favoring harmonic word-order pattern 2 over both non-harmonic orders 3 and 4, even though pattern 2 had both non-English orders while 3 and 4 each had only one. This result, however, seems equally consistent with all variants of the harmonic bias discussed above.

*The *L4 bias.* What might be the nature and origin of the bias among non-harmonic orders? Why this asymmetry between the preferred ordering of adjectives (post-nominal) and numerals (pre-nominal)? A possible functional source may be located in a semantic asymmetry. Kamp and Partee (1995) propose the (*Semantic-*)*Head Primacy Principle* which states that in N-Adj or Adj-N phrases, the noun (the *semantic* head) is interpreted first, followed by the adjective, regardless of syntactic word order. This is motivated by the existence of adjectives which crucially depend on the noun they modify for interpretation, e.g. gradable adjectives like 'tall'. It follows from the Head Primacy Principle that linear ordering of the noun first allows the following adjective to be interpreted immediately, while having the adjective first involves storing it in a buffer for interpretation after the noun is interpreted—thus making Adj-N less efficient (Hawkins, 2004:51).[36] The crucial asymmetry arises because the Head Primacy Principle presumably does not apply to numerals.[37]

This functional explanation depends on a particular semantic property of adjectives. If the explanatory scope is instead taken to be syntax more generally, then an explanation can be entertained in which adjectives and numerals are both treated simply as syntactic heads; the asymmetry between them arises from a universal according to which adjective phrases are *embedded within* numeral phrases.[38]

Universal 18 can potentially be analyzed as one case of a more general constraint on word order, the Final-Over-Final Constraint, FOFC. (This potential generality is one reason Universal 18 was selected for this study.) FOFC was first formulated by Holmberg (2000) to account for the

[35] In addition to symbolic theories of simplicity, connectionist approaches could be pursued. Christiansen and Devlin (1997) argue that connectionist models of sequential learning (simple recurrent networks: Elman, 1990) display an asymmetry: learning grammars in which embedded phrases order their heads consistently with the matrix phrase yields lower mean-squared error (they are 'more learnable'). The scope claimed for this bias is all sequential learning (although there was no demonstration of a bias favoring some generalization of syntactic head consistency that would apply to sequential learning in all domains).

[36] This reading of the Head Primacy Principle (HPP) assumes the existence of a generalization from the particular class of adjectives whose interpretation depends on the noun, to adjectives in general. Such a generalization would function as a cognitive bias of the type we have argued for. There is, however, an alternative reading of the HPP: only those adjectives whose meaning depends on the noun require postponed interpretation, and therefore such adjectives *alone* incur an online processing cost. Because adjectives in our experiment can all be unambiguously interpreted immediately, this reading of the HPP does not predict any asymmetry between conditions 3 and 4. It is therefore the first reading that we intend.

[37] We thank an anonymous reviewer for suggesting an alternative formulation of the universal: a general preference for harmonic ordering combined with a weaker specific dispreference for pre-nominal adjectives only. This formulation predicts the preference ordering: harmonic N-Adj, N-Num $\succ$ harmonic Adj-N, Num-N $\succ$ non-harmonic N-Adj, Num-N $\succ$ non-harmonic Adj-N, Num-N. This predicted preference $2 \succ 1$ among the harmonic languages accords with the WALS data in Table 1, but cannot describe the experimentally-observed bias, $1 \succ 2$. Furthermore, the differences across conditions are *not* carried by the adjective phrases, as would be predicted if there is a bias against Adj-N but not against N-Num.

[38] For discussion of possible semantic explanations of this asymmetry, and of broader embedding universals of which it is part, see Cinque, 1999, p. 199; Rijkhoff, 2002; Svenonius, 2008; Zamparelli, 2000.

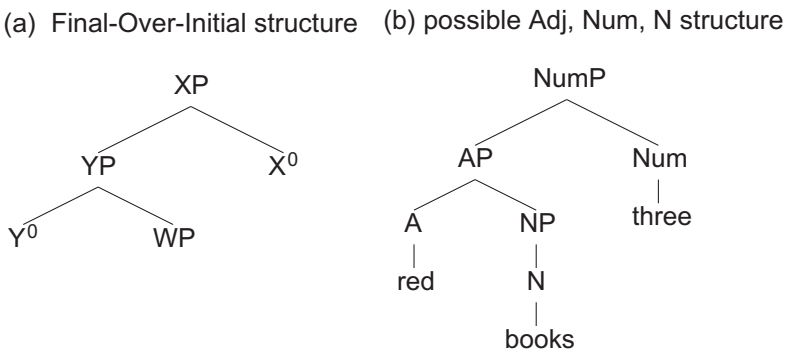


Fig. 7. (a) Final-over-initial structure and (b) possible Adj, Num, N structure.

generalization that while non-harmonic orders (understood now generally as structures in which the position of the head is not consistent across phrases) are overall dispreferred to harmonic systems, certain non-harmonic patterns seem to be particularly avoided cross-linguistically. The FOFC bans structural configurations such as Fig. 7a, in which a phrase (XP) with a final head (X^0) dominates a phrase (YP) with an initial head (Y^0). Universal 18 may in fact be a sub-case of the FOFC: under the analysis shown in Fig. 7b, the pattern-4 grammar generates the FOFC-violating phrase Adj-N-Num.

Biberauer et al. (2008, 2009, in press) propose that the FOFC itself can be derived—from constraints on movement.[39] These constraints dictate that certain heads exert control over other heads within a specified domain (e.g., the nominal or verbal domain): if the controlling head moves (to a phrase-final position), then other (lower) heads in that domain must also move (to a phrase-final position; heads must originate in phrase-initial position). As stated, the FOFC is an inviolable ban on a particular structural configuration, however we have suggested here that Universal 18 is the result of a *probabilistic bias*—if we are correct, and structures parallel to our FOFC-violating pattern (Adj-N, N-Num) are more difficult to learn rather than impossible, the constraints deriving the FOFC would need to be re-stated to accommodate this.

Although the generality of the FOFC puts it beyond the scope of the present experiment, it predicts numerous other asymmetries parallel to that of adjectives and numerals. Many of these are potentially amenable to test using the Mixture-Shift Paradigm.

Final conclusion. For word-order constraints that distinguish among non-harmonic patterns (like our 3 and 4), we know of no proposed explanations that are derived from *general biases* with scope extending beyond language into cognition generally. If Universal 18's substantive bias against a particular type of non-harmonic language is in fact specific to the language system, then the empirical findings reported here constitute clear evidence against recent claims that no such biases exist within cognition (Bybee, 2009; Dunn et al., 2011; Evans & Levinson, 2009; Goldberg, 2006; Levinson & Evans, 2010). Further, we believe these results underscore the utility of artificial language learning in general and the Mixture-Shift Paradigm in particular as a promising method for generating experimental evidence about the role learners play in constraining the typological patterns displayed by the world's languages.

[39] For explanation of the FOFC based on Hawkins (1994)'s processing theory, see Sheehan (2010):151.

Acknowledgments

This work was supported by an NSF graduate research fellowship awarded to the first author, by a Blaise Pascal International Research Chair funded by the Ile-de-France department and the French national government awarded to the second author, and by an NSF IGERT training grant awarded to the Cognitive Science Department at Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or other sponsors. The authors wish to acknowledge the following people for their help with this work: Colin Wilson, Don Mathis, Barbara Landau, Elissa Newport, Simon Fischer-Baum, Kyle Rawlins, Mike McCloskey, Geoff Pullum, Jim Hurford, Simon Kirby, Kenny Smith, Barbara Scholz, Bob Ladd, Antonella Sorace, Thomas Bak, Ian Roberts, Theresa Biberauer, Anders Holmberg, Sharon Peperkamp, Iris Berent, Catherine Sanders, Jacob Sowles, Kessie Alexandre, Andrea Scavelli, the audiences at NELS 40 and LAGB 2010, and the members of the Johns Hopkins Language Acquisition Lab. We would also like to thank Matthew Crocker and three anonymous reviewers for their insightful questions and comments.

Appendix A. Artificial language lexicon

Tables A.7a and b list the nouns and modifiers used in the experiment reported here (including pronunciation and neighborhood density). The orthographic forms are provided simply for illustration; participants were never provided a written representation of the language they were learning.

Appendix B. Parameter estimates for selected mixed-effects models

The parameter estimates for model 1 are shown in Table B.8. This model includes all levels of the condition

Table A.7
Nonce forms, their pronunciation, neighborhood density, and English gloss (for modifiers).

	Pron.	Density		Pron.	Density	Meaning
(a) Nonce nouns			*(b) Nonce modifiers*			
grifta	/'grɪftə/	0	geej	/'giʒ/	1	"blue"
nerka	/'nɛrkə/	0	fush	/'fʌʃ/	14	"green"
blifona	/blɪ'foʊnə/	0	thraf	/'θræf/	1	"big"
slergena	/slɛr'gɛɪnə/	0	jev	/'dʒɛv/	12	"small"
flarma	/'flarmə/	2	cherg	/'tʃɛrg/	2	"furry"
mauga	/'maʊgə/	0	derf	/'dɛrf/	4	"two"
rampeza	/ræm'pɛɪzə/	0	kez	/'kɛz/	7	"three"
wapoga	/wa'poʊgə/	0	glawb	/'glaʊb/	3	"four"
trefunda	/trə'fʌndə/	0	zadge	/'zædʒ/	4	"five"
powarta	/poʊ'warrə/	0	voitch	/'voɪtʃ/	7	"six"

Table B.8
Parameter estimates for mixed-effects model 1.

Fixed effects:	Estimate	Standard error	*z*-value	*p*-value
(Intercept)	−0.135	0.20	−0.673	0.501
Condition 1	1.514	0.279	5.430	<0.001*
Condition 2	1.563	0.280	5.580	<0.001*
Condition 3	1.359	0.276	4.927	<0.001*
Condition 4	1.126	0.276	4.086	<0.001*
Modifier type = Num	0.238	0.164	1.452	0.146
Condition 1 × Num	0.196	0.225	0.872	0.383
Condition 2 × Num	0.070	0.228	0.308	0.758
Condition 3 × Num	−0.308	0.209	−1.476	0.140
Condition 4 × Num	−0.750	0.204	−3.680	<0.001*

factor and modifier-type factors as fixed effects, as well as terms for the interaction between condition and modifier-type. The model used effects coding.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons.

Aristar, A. R. (1991). On diachronic sources and synchronic pattern: An investigation into the origin of linguistic universals. *Language, 67*, 1–33.

Atkinson, Q. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science, 332*, 346.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bader, M. (2011). On being both head-initial and head-final. In H. Yamashita, Y. Hirose, & J. L. Packard (Eds.), *Processing and producing head-final structures* (pp. 325–347). Springer.

Baker, M. (2001). *The atoms of language: The mind's hidden rules of grammar*. New York, NY: Basic Books.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31.

Bhat, D. (1978). A general study of palatalization. In J. Greenberg (Ed.). *Universals of human language* (Vol. 3, pp. 47–92). Stanford, CA: Stanford University Press.

Bialystok, E. (1997). The structure of age: In search of barriers to second language acquisition. *Second Language Research, 13*, 116–137.

Biberauer, T., Holmberg, A., & Roberts, I. (in press). A syntactic universal and its consequences. *Linguistic Inquiry*.

Biberauer, T., Holmberg, A., & Roberts, I. (2008). Structure and linearization in disharmonic word orders. In C. B. Chang & H. J. Haynie (Eds.), *Proceedings of the 26th West Coast Conference on Formal Linguistics* (pp. 96–104). Somerville, MA: Cascadilla Press.

Biberauer, T., Newton, G., & Sheehan, M. (2009). The final-over-final constraint and predictions for diachronic change. *Toronto Working Papers in Linguistics* (Vol. 31). Toronto: University of Toronto Press.

Blevins, J., & Garrett, A. (2004). The evolution of metathesis. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically based phonology* (pp. 117–156).

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In G. Bouma, I. Kraemer, & J. Zwarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bybee, J. (2009). Language universals and usage-based theory. In M. H. Christiansen, C. Collins, & S. Edelman (Eds.), *Language universals* (pp. 17–39). Oxford: Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in Cognitive Science? *Trends in Cognitive Sciences, 7*, 19–22.

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

Christiansen, M. H. (2000). Using artificial language learning to study language evolution: Exploring the emergence of word order universals. In L. Ghadakpour & J. Dessalles (Eds.), *The evolution of language: 3rd international conference* (pp. 45–48). Paris: Ecole Nationale Superieure des Telecommunications.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*, 489–509.

Christiansen, M. H., & Devlin, J. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th annual cognitive science society conference* (pp. 113–118). Mahwah, NJ: Lawrence Erlbaum Associates.

Cinque, G. (1993). On the evidence for partial n-movement in the Romance DP. *University of Venice Working Papers in Linguistics, 3*, 21–40.

Cinque, G. (1999). *Adverbs and functional heads: A cross-linguistic perspective*. USA: Oxford University Press.

Cinque, G. (2005). Deriving Greenberg's Universal 20 and its exceptions. *Linguistic Inquiry, 36*, 315–332.

Clark, E. (1995). *The lexicon in acquisition*. Cambridge: Cambridge University Press.

Clark, A., & Roberts, I. (1993). A computational model of language learnability and language change. *Linguistic Inquiry, 24*, 299–345.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. London: Longman.

Croft, W. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

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

Culbertson, J., & Smolensky, P. (submitted for publication). A Bayesian model of biases in artificial language learning: The case of a word-order universal.

de Lacy, P. (2006). *Markedness: Reduction and preservation in phonology*. Cambridge University Press.

Demetras, M., Post, K., & Snow, C. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language, 13*, 275–292.

Dryer, M. (1988). Object-verb order and adjective-noun order: Dispelling a myth. *Lingua, 74*, 185–217.

Dryer, M. (1992). The Greenbergian word order correlations. *Language, 68*, 81–183.

Dryer, M. (2008a). Order of adjective and noun. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online chapter 87*. Munich: Max Planck Digital Library.

Dryer, M. (2008b). Order of numeral and noun. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The world atlas of language structures online chapter 89*. Munich: Max Planck Digital Library.

Dunn, M., Greenhill, S., Levinson, S., & Gray, R. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature, 473*, 79–82.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Epstein, S. D., Flynn, S., & Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Behavioral and Brain Sciences, 19*, 677–714.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*, 429–448.

Finley, S., & Badecker, W. (2008). Substantive biases for vowel harmony languages. In J. Bishop (Ed.), *Proceedings of West Coast Conference on Formal Linguistics 27* (pp. 168–176).

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189).

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *PNAS, 99*, 529–534.

Gair, J. W. (1970). *Colloquial Sinhalese clause structures. Janua linguarmum*. Mouton: The Hague.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity vs. natural partitioning. In S. Kuczaj (Ed.), *Language development Vol. 2: Language, cognition and culture* (pp. 301–334). Hillsdale, NJ: Erlbaum.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. New York, NY: Springer.

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.

Grünwald, P., Myung, I., & Pitt, M. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. The MIT Press.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical period hypothesis for second language acquisition. *Psychological Science, 14*, 31–38.

Hawkins, J. A. (1983). *Word order universals*. New York: Academic Press.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry, 21*, 223–261.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press.

Hawkins, J. A. (2004). *Complexity and efficiency in grammars*. Oxford: Oxford University Press.

Hawkins, J. A. (2010). FOFC: A typological perspective. Presented at workshop on the final-over-final constraint: Typology, processing and acquisition, University of Cambridge, 29th January 2010.

Hirsh-Pasek, K., Treiman, R., & Schneiderman, M. (1984). Brown and Hanlon revisited: Mothers' sensitivity to ungrammatical forms. *Journal of Child Language, 11*, 81–88.

Holmberg, A. (2000). Scandinavian stylistic fronting: How and category can become an expletive. *Linguistic Inquiry, 31*, 445–483.

Holmes, R. B., & Smith, B. S. (1977). *Beginning Cherokee*. Norman: University of Oklahoma Press.

Hualdo, J. I., & de Urbina, J. O. (2003). *A grammar of Basque*. Berlin: Mouton de Gruyter.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation. *Language Learning and Development, 1*, 151–195.

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

Hurford, J. (2003). The interaction between numerals and nouns. In F. Plank (Ed.), *Noun phrase structure in the languages of Europe* (pp. 561–620). New York: Mouton de Gruyter.

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.

Janda, L. A., Nesset, T., & Baayen, R. H. (2010). Capturing correlational structure in Russian paradigms: A case study in logistic mixed-effects modeling. *Corpus Linguistics and Linguistic Theory, 6*, 29–48.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology, 21*, 60–99.

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

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129–191.

Kirby, S. (1999). *Function, selection, and innateness*. Oxford: Oxford University Press.

Kirby, S., Smith, K., & Brighton, H. (2004). From UG to universals: Linguistic adaptation through iterated learning. *Studies in Language, 28*, 587–607.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York, NY: Wiley.

Levinson, S. C., & Evans, N. (2010). Time for a sea-change in linguistics: Response to comments on 'the myth of language universals'. *Lingua, 120*, 2733–2758.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. The Mit Press.

Lightfoot, D. (1999). *The development of language: Acquisition, change, and evolution*. New York, NY: Blackwell.

Lindblom, B. (1986). Phonetic universals in vowel systems. In J. Jaeger & J. Ohala (Eds.), *Experimental phonology* (pp. 13–44). Orlando, Florida: Academic Press.

Marcus, G. (1993). Negative evidence in language acquisition. *Cognition, 46*, 53–85.

Meyerhoff, M. (2000). The emergence of creole subject–verb agreement and the licensing of null subjects. *Language Variation and Change, 12*, 203–230.

Moravcsik, E. (1978). Language contact. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.). *Universals of human language* (Vol. 1, pp. 93–122). Stanford: Stanford University Press.

Morgan, J. L., Bonamo, K. M., & Travis, L. L. (1995). Negative evidence on negative evidence. *Developmental Psychology, 31*, 180–197.

Müller, G. (1999). Optimality, markedness, and word order in German. *Linguistics, 37*, 777–818.

Newmeyer, F. (2005). *Possible and probable languages*. New York: Oxford University Press.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I: Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Ohala, J. (1993). Sound change as nature's speech perception experiment. *Speech Communication, 13*, 155–161.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *The Quarterly Journal of Experimental Psychology, 17*, 273–281.

Prince, A., & Smolensky, P. (1993/2004). *Optimality theory: Constraint interaction in generative grammar*. New York, NY. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Rutgers Optimality Archive 537, 2002. Revised version published by Blackwell 2004.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*, 1604–1610.

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

Rijkhoff, J. (1998). Word order in the languages of Europe: The simple noun phrase. In A. Siewierska (Ed.), *Constituent order in the languages of Europe* (pp. 321–382). New York: Mouton de Gruyter.

Rijkhoff, J. (2002). *The noun phrase*. USA: Oxford University Press.

Rooryck, J., Smith, N., Liptak, A., & Blakemore, D. (2010). Special issue on Evans & Levinson's "The myth of language universals". *Lingua, 120*, 2651–2758.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606–621.

Sandler, W., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *PNAS, 102*, 2661–2665.

Sankoff, G., & Laberge, S. (1980). The acquisition of native speakers by a language. In G. Sankoff (Ed.), *The social life of languages* (pp. 195–209). Philadelphia, PA: University of Pennsylvania Press.

Saxton, M. (2000). Negative evidence and negative feedback: Immediate effects on the grammaticality of child speech. *First Language, 20*, 221–252.

Sheehan, M. (2010). Formal and functional approaches to disharmonic word orders. *Newcastle Working Papers in Linguistics, 16*, 146–166.

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology, 49*, 370–407.

Slobin, D. (2004). From ontogenesis to phylogenesis: What can child language tell us about language evolution? In J. Langer, S. T. Parker, & C. Milbrath (Eds.), *Biology and knowledge revisited: From neurogenesis to psychogenesis* (pp. 255–286). Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, J., Durham, M., & Fortune, L. (2007). "mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation and Change, 19*, 63–99.

Smith, N. V., & Tsimpli, I.-M. (1995). *The mind of a savant: Language learning and modularity*. Oxford: Blackwell.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*, 1317–1329.

Steddy, S., & Samek-Lodovici, V. (2011). On the ungrammaticality of remnant movement in the derivation of greenberg's universal 20. *Linguistic Inquiry, 42*, 445–469.

Steriade, D. (1997). Phonetics in phonology: The case of laryngeal neutralization. Ms.

Stevens, K. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In P. Denes & E. David, Jr. (Eds.), *Human communication: A unified view* (pp. 51–66). New York: McGraw-Hill.

Strapp, C. (1999). Mothers', fathers', and siblings' responses to children's language errors: Comparing sources of negative evidence. *Journal of Child Language, 26*, 373–391.

Svenonius, P. (1994). The structural location of the attributive adjective. In E. Duncan, D. Farkas, & P. Spaelti (Eds.), *Proceedings of the 12th West Coast Conference on Formal Linguistics* (pp. 439–454). Stanford, CA: CSLI Publications.

Svenonius, P. (2008). The position of adjectives and other phrasal modifiers in the decomposition of DP. In L. McNally & C. Kennedy (Eds.), *Adjectives and adverbs: Syntax, semantics (and discourse)* (pp. 16–42). New York: Oxford University Press.

Talmy, L. (2000). *Toward a cognitive semantics. Concept structuring systems* (Vol. 1). The MIT Press.

Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry, 29*, 229–268.

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds 'listen through' ambiguous adjectives in fluent speech. *Cognition, 100*, 389–433.

Ticio, M. E. (2003). On the structure of DPs. Dissertation, University of Connecticut.

Tomasello, M. (2009). Universal grammar is dead. *Behavioral and Brain Sciences, 32*, 470–471.

Vainikka, A., & Young-Scholten, M. (1996). The initial state in the L2 acquisition of phrase structure. In S. Flynn, G. Martohardjono, & W. A. O'Neil (Eds.), *The generative study of second language acquisition* (pp. 17–34). Mahwah, NJ: Lawrence Erlbaum Associates.

Ward, I. (1956). *An introduction to the Yoruba Language*. Cambridge: W. Heffer and Sons.

Weir, M. W. (1972). Probability performance: Reinforcement procedure and number of alternatives. *The American Journal of Psychology, 85*, 261–270.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

Wilks, S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*, 60–62.

Wilson, C. (2006). An experimental and computational study of velar palatalization. *Cognitive Science, 30*, 945–982.

Wonnacott, E., & Newport, E. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. Clark-Cotton, & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th annual Boston University conference on language development*. Boston, MA: Cascadilla Press.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin and Review, 15*, 971–979.

Zamparelli, R. (2000). *Layers in the noun phrase*. New York, NY: Garland.