To appear in TopiCS in Cognitive Science, issue on computational psycholinguistics,
J. Hale & D. Reitter, Eds.

Cognitive biases, linguistic universals, and constraint-based grammar learning[☆]

Jennifer Culbertson[a,b,*], Paul Smolensky[b], Colin Wilson[b]

[a] *Linguistics Program, Department of English, George Mason University, Fairfax, VA 22030, USA*
[b] *Cognitive Science Department, Johns Hopkins University, Baltimore, MD 21218, USA*

Abstract

According to classical arguments, language learning is both facilitated and constrained by cognitive biases. These biases are reflected in linguistic typology—the distribution of linguistic patterns across the world's languages—and can be probed with artificial grammar experiments on child and adult learners. Beginning with a widely successful approach to typology (Optimality Theory), and adapting techniques from computational approaches to statistical learning, we develop a Bayesian model of cognitive biases and show that it accounts for the detailed pattern of results of artificial grammar experiments on noun-phrase word order (Culbertson, Smolensky & Legendre, 2012). Our proposal has several novel properties that distinguish it from prior work in the domains of linguistic theory, computational cognitive science, and machine learning. The paper illustrates how ideas from these domains can be synthesized into a model of language learning in which biases range in strength from hard (absolute) to soft (statistical), and in which language-specific and domain-general biases combine to account for data from the macro-level scale of typological distribution to the micro-level scale of learning by individuals.

Keywords: Bayesian modeling, Optimality Theory, learning biases, artificial language learning, typology, word order

1. Introduction

What are the capabilities and limitations of human language learning? According to classical arguments from linguistics and the theory of learning, answering this question involves discovering the *biases* of human language learners. We propose here a formal model of such biases—set within an existing constraint-based theory of linguistic typology—and apply it to experimental results that connect laboratory language learning with recurring word-order patterns across the world's languages. Our model implements the hypothesis that learners use Bayesian inference to acquire a grammar under the influence of a set of hard (absolute) and soft (statistical) biases; we focus primarily on the soft biases, as their form and implementation are novel. The complete set of biases acts as a *prior probability distribution* for language learning, accounting for statistical structure that has been observed in linguistic typology and in the performance of individual learners. Components of the prior that are language-specific, and others that are plausibly cognition-general, are combined by the simple mathematical operation of multiplication.

1.1. Learning biases and typological asymmetries

A central hypothesis of generative linguistic theory is that human learners bring specialized knowledge to the task of acquiring a language. By restricting the space of hypotheses a learner will entertain, this prior

[☆] The authors would like to thank Géraldine Legendre, Don Mathis, Mark Johnson, John Hale and five anonymous reviewers for helpful comments and discussion.

[*] Corresponding Author: Linguistics Program, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA., Tel: (703) 993-1160, Fax: (703) 993-1161, Email: jculber4@gmu.edu.

knowledge is argued to solve the poverty-of-the-stimulus problem, allowing robust language acquisition from variable and impoverished input (Chomsky, 1965, 1980; although see Pullum & Scholz, 2002).

Striking asymmetries in the frequencies which which different linguistic patterns occur in the world's languages constitute a central body of evidence for the cognitive biases operative in language acquisition. Since language learners have prior assumptions about how linguistic systems are structured, it stands to reason that systems that conform better to those expectations will arise more often, and survive longer, than other systems. Accounting for typological asymmetries has been the explicit goal of generative linguistics work within the classical framework of Principles and Parameters (e.g. Chomsky, 1986; Baker, 2001) and, as discussed further below, within the more recent constraint-based framework of Optimality Theory (Prince & Smolensky, 1993/2004).

An illustrative example, which provides the foundation for the case study developed here, is a generalization about word order in the nominal domain known as Universal 18 (Greenberg, 1963).[1] According to Greenberg's Universal 18, placement of numerals (Num) after the noun (N) asymmetrically implies that adjectives (Adj) must also be post-nominal (N-Num → N-Adj). A language violating this generalization would express 'two trees' as *trees two* but 'green trees' as *green trees*.[2] Greenberg's original sample of 30 languages did not contain even one instance of the prohibited pattern, whence the claim that the generalization is a 'Universal' of language. A modern, much larger sample reveals that some languages do feature the prohibited Adj-N, Num-N order. But this order is nevertheless significantly rarer than the other three possibilities, accounting for only 4% of attested languages (see Table 1, where the prohibited order is designated 'L4'; Dryer, 2008a,b).

Table 1: Sample of languages from WALS illustrating preference for consistent ordering and sub-preference among inconsistent order patterns for N-Adj, N-Num over Adj-N, N-Num (shaded). The labels L1, L2, L3, L4 are used throughout the article for the different language types. English, for example, is a type-1 (or 'L1') language.

	Adj-N	N-Adj
Num-N	**27%** (L1)	**17%** (L3)
N-Num	**4%** (L4)	**52%** (L2)

Typological data of the type illustrated by Table 1 pose a major dilemma, heretofore unresolved, for generative linguistics. It is straightforward to define a set of OT constraints that make L4 an unlearnable pattern, incorrectly predicting that it should not appear in any attested language (see section 2.1). It is also straightforward to define a constraint set that grants L4 the same status as the other three languages, failing to predict that it is much less frequent than the others typologically. The dilemma lies in attempting to account for the finding that L4 is *possible but rare*. This is a statistical finding, a soft universal of language, but generative linguistic theories have previously succeeded in accounting for absolute universals only. Note that the problem is not restricted to constraint-based theories, but applies also to the classical Principles and Parameters approach and its contemporary versions (e.g., Chomsky 1981b; Travis 1989; Kayne 1994; Baker 2001; Cinque 2005; Newmeyer 2010).

Further inspection of the nominal typology, and of similar cross-linguistic tabulations, reveals that the dilemma surrounding 'Universal 18' is far from an isolated problem. Table 1 provides evidence for an-

[1] A great number of universals have been documented by linguists: see for example the on-line Universals Archive (Plank & Filimonova, 2000), which contains 2029 universals at the time of writing. Implicational universals, like Greenberg's Universal 18, of the form "If a language has x, then it will have y" are of particular interest. The majority of these have been shown to be statistical (Bickel, 2007; Evans & Levinson, 2009), and thus are often called generalizations rather than universals. Quite a few, like Universal 18, are stated such that of four logically possible linguistic patterns, three are well-attested and the fourth is rare (e.g. the Final-Over-Final constraint, Biberauer, Holmberg & Roberts, to appear).

[2] Several possible explanations exist as to *why* Universal 18 holds; these are discussed at length in Culbertson et al. (2012).

other well-documented cross-linguistic asymmetry: namely, that languages tend to use consistent ordering of heads relative to their complements (Chomsky, 1981a; Travis, 1984; Baker, 2001). In particular, systems L2 (N-Num, N-Adj) and L1 (Num-N, Adj-N) together account for 79% of attested languages. A constraint or parameter set that allows only those systems will fail to provide learnable grammars for the remaining 21% of languages; a set that allows all four patterns would seem to have nothing to say about the typological asymmetry between consistent and inconsistent systems. In addition, post-nominal placement of both Num and Adj (L2) is slightly more frequent than all other patterns combined (52%) —another statistical generalization that must be lost in a theory with absolute biases only.

To preview our proposed solution, we believe that the problem of possible-but-rare languages disappears under a probabilistic approach to cognitive learning biases, and that previous research in machine learning and computational cognitive science lays much of the groundwork needed to construct a model of soft biases. However, formally stating the probabilistic biases in a simple, effective manner is a non-trivial matter that we take up in the body of the paper. The main idea that we explore with respect to asymmetries in nominal word-order and similar typological patterns is stated in (1).

(1) Con(straint) Bias
   a. The constraint set available to learners is sufficiently rich to generate all attested languages, but
   b. there are soft biases that penalize the grammatical use of particular constraints (e.g., the constraint that is responsible for Num occurring after N in L4).

In our proposal, the language-specific *Con* Bias is combined with a well-established and plausibly domain-general bias discussed below that favors lower-entropy distributions (the Regularization Bias) to form the probabilistic prior for language learning.

1.2. Converging evidence for learning biases

We pause at this point to address a concern that typological asymmetries are a somewhat indirect source of evidence for language learning biases. The indirectness is due to the fact that typological distributions have to a certain extent been influenced by non-cognitive, diachronic factors, such as lineage-specific trends (e.g. Dunn, Greenhill, Levinson & Gray, 2011) and historical or geographic factors that influence how linguistic patterns spread and change (e.g. Dryer, 2012). Precisely which typological asymmetries should be accounted for within cognitive theory proper is therefore not completely clear from cross-linguistic data alone.

It would be desirable to obtain converging evidence for the learning biases from first-language acquisition by children, but the complexity of natural input and absence of experimental control makes this difficult. Fortunately, a number of studies have shown that artificial (or 'miniature') language learning by adults—which can be done under carefully controlled laboratory conditions—reveals biases that align with typological asymmetries (e.g. Pycha, Nowak, Shin & Shosted, 2003; Wilson, 2006; Finley & Badecker, 2008; Moreton, 2008; St. Clair, Monaghan & Ramscar, 2009, and many others) (see Culbertson, 2012; Moreton & Pater, to appear, for reviews of recent literature on this topic).

One such study, previously reported in Culbertson et al. (2012), provides converging evidence for biases on nominal word order. During training learners saw pictures of novel objects, and heard phrases describing them which were uttered by an "informant" and comprised of either an adjective and a noun or a numeral and a noun. The order used by the informant in any particular description depended on the learner's training condition—corresponding to one of the four patterns in Table 1 (the 'dominant' pattern), accompanied by some variation. At test, learners were asked to produce descriptions of pictures using these two-word phrase types.

In this particular study, learners' reaction to the variation in each condition provides the measure of biases by exploiting a general finding in human learning of probabilistic regularities: under certain circumstances, learners will reduce probabilistic variation (e.g. Weir, 1972; Hudson Kam & Newport, 2009;

Reali & Griffiths, 2009). This phenomenon has been called *regularization*.[3] Both this general cognitive bias disfavoring variability *and* the linguistic biases based on the typology in Table 1 were hypothesized to affect learners' behavior. In particular, if the dominant ordering pattern in the input conformed to a learner's linguistic biases, it was expected to be acquired *and regularized*. On the other hand, if the dominant pattern went against a learner's biases it was expected to be less likely to be acquired veridically (much less regularized).

The results from the testing phase, illustrated in Fig. 1, mirror the typology: when the dominant training pattern used consistent head ordering (conditions 1, 2) learners regularized robustly. Significantly less regularization was found in condition 3, where the inconsistent pattern (N-Adj, Num-N) was dominant in the training. Crucially, learners did *not* regularize when the dominant pattern was the typologically rare (Adj-N, N-Num). As is clear from Fig. 1, differences among the conditions were driven mainly by differences in the use of Num majority order, with condition 4 significantly under-producing N-Num order. We will show below that our proposal accounts for the qualitative and quantitative findings of this experiment, in addition to providing the foundation for an explanation of the typological pattern discussed above.

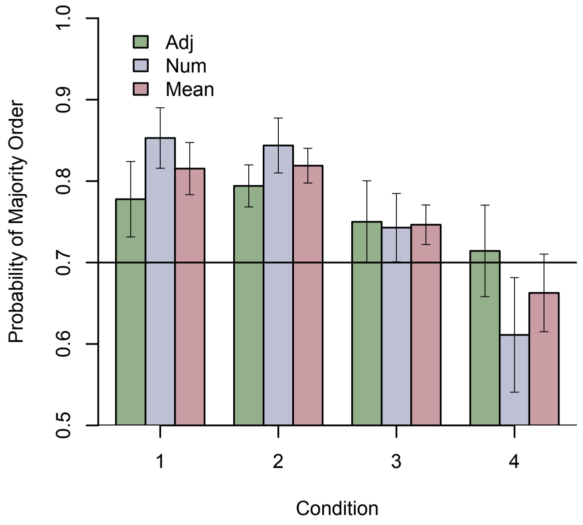


Figure 1: Probability of dominant input order use reported in Culbertson et al. (2012). Bars show, for each condition, probability of dominant order for phrases with adjectives, phrases with numerals, and the mean. The solid line at 0.7 indicates the probability of the dominant order used in the *input*.

1.3. Outline

The rest of the paper is organized as follows. We first introduce non-probabilistic approaches to typology, focusing on Optimality Theory ('OT', section 2). We then show how OT can be adapted to account for language-internal and typological patterns that are stochastic rather than absolute (section 2.2). Language-internal stochasticity results from converting grammatical scores to probabilities in a way that is formally

[3] This notion of regularization should be distinguished from the term 'regularization' as used in machine learning literature. In the latter context regularization refers to the practice of introducing some prior information in order to, for example, prevent overfitting. This practice has been justified as essentially imposing Occam's razor, and in fact in some cases imposes an expectation of *uniformity*, or maximum entropy, whereas our regularization bias favors minimum entropy.

equivalent to the maximum entropy (log-linear) models widely used in computational linguistics. Accounting for gradient typological asymmetries—and artificial-grammar results that mirror them—is the role of the prior distribution, which we express as a product of weighted factors (section 3). We highlight several respects in which our proposal differs from previous work in computational cognitive science and machine learning, and from related work on typological asymmetries within linguistic theory. Holding the *Con* Bias constant, we propose (and evaluate) two possible formulations of the Regularization Bias. Given a space of probabilistic grammars and a prior distribution on that space, language learning is formalized as Bayesian inference (section 4). With a small number of fit prior parameters, our model yields predictions that match the main experimental results and an independent replication (section 5). In the final section, we summarize our proposal, results and conclusions (section 6).

2. Linguistic theories of typology

The ways in which languages vary, and limits on cross-linguistic variation, are central topics in descriptive and theoretical linguistics. Within generative linguistics, discovering limitations on variation has been taken to be central to understanding language acquisition. It is well-known that children generalize beyond their language input, and that all generalizing learning systems must have biases of some form (Gold, 1967; Mitchell, 1980). A simple type of bias in the linguistic domain is an absolute prohibition on certain logically-possible patterns. Therefore, a linguistic theory that prohibits typologically unattested patterns can be viewed as a statement of the (implicit) biases that make language acquisition possible.

To illustrate the typical generative linguistics approach to typology, below we develop a theory that permits all possible orders of Num and Adj with respect to N except for the typologically rare pattern that contravenes Greenberg's Universal 18 (i.e., L4 in Table 1). This theory of nominal word-order will ultimately be relaxed to a soft (statistical) theory, as anticipated in section 1. Formulating the absolute version is worthwhile because it is representative of many theories in the generative literature, it provides grammars for the vast majority (96%) of systems in the survey, and it exposes substantive and formal properties that must be present in any approach to nominal word-order. By beginning with the idealized, absolute theory we can identify exactly which aspects of the data remain to be accounted for by statistical softening.

Recall that the great majority of systems in the typology place Num and Adj on the same side of N (i.e., L1: Num-N, Adj-N and L2: N-Num, N-Adj). This fact taken by itself fits comfortably with a highly influential view of typology in generative linguistics, known as Principles and Parameters (P&P; Chomsky, 1981a, 1986, 1998; Baker, 2001). According to P&P, uniformity across languages results from a set of universal grammar principles, and constrained variation arises from a set of variables or 'parameters' appearing in those principles. Each possible combination of parameter values determines a language type, and these types collectively exhaust the languages that are predicted to be possible. The principles and parameters together serve as a constraining framework for acquisition, with the learner's problem reduced to that of identifying language-specific parameter settings. The locus of language-specificity in this case is the Head Directionality Parameter (Chomsky, 1981a; Travis, 1984), which can be set so that all syntactic heads precede their complements (as in L1) or all syntactic heads follow their complements (as in L2).[4]

The Head Directionality Parameter expresses an important generalization about the typology—nominal heads and complements are consistently ordered in nearly 80% of the languages surveyed—but by itself it provides no grammar for L3 (Num-N, N-Adj). Generative theories have various mechanisms for solving problems of this sort, all of which ensure that a *general* condition (e.g., syntactic heads must precede their complements) can be overridden by a more *specific* one (e.g., the head of an Adj phrase must follow its

[4] Here we call Adj and Num heads (following e.g. Cinque, 2005), however this is essentially a simple expository expedience; no theoretical significance is attached to the term for present purposes as the work here deals in surface word strings only.

complement). For example, consistent ordering may hold at an early stage in the syntactic derivation, with consistency disrupted by subsequent syntactic movements that target specific heads or complements (e.g. as in Travis, 1989; Cinque, 2005). Alternatively, the grammatical structures of a language can be defined by the resolution of conflict among conditions (or constraints) of different strengths, as in Optimality Theory. Under this approach, a weaker general constraint can be violated at all stages of syntactic derivation if there is a stronger specific constraint with which it conflicts.

Constraint-based theories of linguistic typology have proved successful in many empirical domains[5], and have lead to a renewed interest in formal models of language acquisition within generative linguistics. We show below that these hard approaches to typology also provide a rather direct route to a probabilistic model of grammars and grammar learning. For these reasons, in addition to our own areas of expertise, we focus on constraint-based theories in the remainder of the paper, beginning with OT.

2.1. Optimality Theory

OT shares with Principles & Parameters the commitment to a universal set of primitives from which all grammars are constructed. The primitives of OT are members of a universal constraint set (*Con*). A language-specific grammar is a *strict priority ranking* of these universal constraints. The hard typology predicted by a constraint set is determined by all possible priority rankings; if there is no priority ranking of the universal constraints that yields a logically-possible system, that system is predicted to be impossible.

To analyze nominal word-order, we provisionally adopt the constraint set in (2).[6]

(2) *Con* for nominal word-order (provisional)

a. HEAD-L: order all heads to the left of their complements

b. HEAD-R: order all heads to the right of their complements

c. NUM-L: order numeral heads to the left of their complements

The effects of the Head Directionality Parameter discussed earlier can be replicated by ranking a general constraint, either HEAD-L or HEAD-R, above the other constraints. The ranking [HEAD-L $\gg$ HEAD-R, NUM-L] places both Num and Adj before N, while the ranking [HEAD-R $\gg$ HEAD-L, NUM-L] places both Num and Adj after N. The latter case illustrates the central claim of OT that constraint ranking is *strict*. This ranking determines that the order N-Num is grammatical—in spite of the fact that placing Num after N violates both HEAD-L and NUM-L— because the alternative order violates HEAD-R, which takes strict priority over the lower-ranked constraints. A grammaticality calculation of this sort can be conveniently summarized with an OT tableau, as shown in (3).

(3) *Deriving NUM order in L2*

[5] The majority of this work has been in the domain of phonology, however examples in syntax include Legendre, Raymond & Smolensky (1993); Grimshaw (1997); Nagy & Heap (1998); Vogel (2002); Samek-Lodovici (2005); Steddy & Samek-Lodovici (2011); Pater (2011); Philip (to appear). See also the electronic archive `http://roa.rutgers.edu/`.

[6] This constraint set illustrates a number of general guiding meta-principles of OT. One asserts that a family of constraints—pervading grammar at all levels—is the class of 'alignment constraints' (McCarthy & Prince, 1994), which subsumes all the constraints we discuss here. HEAD-L might be formally expressed, e.g., in an X-bar theoretic context, as ALIGN(X^0 head, L, XP phrase, L). Another frequently deployed meta-principle states that constraints come in families targeting elements of different levels of specificity—exactly as we do in (2). In fact, the alignment-style constraints HEAD-L and HEAD-R have independently been proposed to account for other typological patterns in syntax (Grimshaw, 1997; Sells, 2001), and a combination of general and specific head ordering constraints is crucial to the analysis of the typological preference for head-order consistency in Pater (2011). Here, the specific constraints we use target Num, rather than Adj. To derive our simplified analysis of Universal 18, what is crucial is that there be an asymmetry such that NUM-L or ADJ-R, but not NUM-R or ADJ-L be in *Con*, however see note 16. We would expect that other members of this well-motivated class of constraints would be used in a fuller analysis, e.g. to account for ordering differences among adjective classes (Cinque, 1994; Laenzlinger, 2005).

		{Num, N}	Head-R	Head-L	Num-L
a.		Num-N	∗!		
b.	☞	N-Num		∗	∗

In (3), the unordered 'input' expression is '{Num, N}', for which we consider the candidate 'outputs' of the grammar—here, the two possible ordered expressions 'Num-N' and 'N-Num'. The first candidate expression Num-N violates Head-R since its head, Num, is left of its complement, N. That violation is marked by the '∗' in the tableau cell in the candidate's row and the constraint's column. The other candidate, N-Num, satisfies this constraint, so no '∗' appears. The reverse pattern of violation obtains for the second constraint, Head-L. Which candidate is optimal—hence grammatical? *The one preferred by the highest-ranked constraint that has a preference.* Here, that constraint is Head-R, which favors N-Num; the violation of this decisive constraint by Num-N is fatal: that violation is therefore flagged with '!'. The optimal candidate is fingered.

The predictions of the OT analysis differ from those of the Head Directionality Parameter under the ranking [Num-L ≫ Head-R ≫ Head-L]. In this ranking, illustrated in (4), the specific constraint Num-L dominates the general constraint Head-R with which it conflicts. Consequently, while the order of Adj and N is determined by the relative ranking of Head-R and Head-L, the order of Num and N is determined by Num-L. The specific constraint is *active* in this ranking, unlike the two discussed above, and this gives rise to the inconsistent ordering found in languages of type L3. Ranking specific constraints above general constraints is a widely applicable approach within OT of accounting for grammar-internal inconsistencies or 'exceptions' (e.g. Prince & Smolensky, 1993/2004; McCarthy, 2002).

(4) *Deriving L3*

a. *Num position in L3*

		{Num, N}	Num-L	Head-R	Head-L
a.	☞	Num-N		∗	
b.		N-Num	∗!		∗

b. *Adj position in L3*

		{Adj, N}	Num-L	Head-R	Head-L
a.		Adj-N		∗!	
b.	☞	N-Adj			∗

By design, there is no ranking of the constraints in (2) that generates languages of type L4 (N-Num, Adj-N), which are therefore predicted to be typologically impossible. While it appears to be correct to predict that this language type is highly disprefered, the prediction that they are impossible is too strong. Adding another constraint to *Con*, such as Num-R (or Adj-L), would provide rankings that generate L4. However, this would grant the rare language type the same status in the predicted typology as, say, L3. As we discussed in the introduction, this is but one example of a general dilemma for OT and other hard approaches to language typology: tight restrictions on the space of possible grammars leave some attested languages without an analysis; loose restrictions fail to account for quantitative distinctions—as revealed by large typological surveys and converging experimental evidence—among language types.

We think that the solution to this dilemma lies in a Bayesian formulation of language acquisition, according to which both language input and prior biases shape the systems that are learned. A tightly restricted universal set of constraints (or parameters, etc.) is an especially strong type of prior, but the general Bayesian approach allows many softer forms of bias as well. Within OT, it is possible to enrich the constraint set, so as to provide grammars for all attested languages, while simultaneously imposing biases against grammars in which particular constraints are strong. In principle, many degrees of constraint bias could be revealed by sufficiently large typologies and other sources of evidence.

To flesh this proposal out, we first need to extend the preceding analysis to a probabilistic grammar framework that is closely related to Optimality Theory: Probabilistic Harmonic Grammar. This step is desirable on independent grounds, as previous research has established that the linguistic systems of individual speakers are more gradient than can be accounted for with single OT constraint rankings (see Bod, Hay & Jannedy, 2003 and articles therein; Chater & Manning, 2006). This point applies also to speakers' performance in artificial grammar experiments (e.g. Culbertson et al., 2012). In Bayesian terminology, the likelihood of an utterance according to a speaker's grammar is not typically a binary value (grammatical vs. ungrammatical), but rather lies on a continuum. We now show how OT can be extended to encompass gradient grammars of this type.

2.2. *Probabilistic Harmonic Grammar*

We require hypotheses—grammars—that assign probabilities to all possible pieces of data: in our running example, word orders like Num-N and N-Adj. Working with log-probability often proves convenient because while probabilities combine multiplicatively, log-probabilities combine additively. Let us consider the log-probability of an expression x (up to some additive constant z) as a probabilistic measure of the well-formedness of x, higher log-probability interpreted as greater well-formedness. A grammar G, then, assigns a well-formedness value to x, called its *Harmony* $H_G(x)$, which is just the log-probability of x given that it was generated by G: $\log(P(x|G))$.

A numerical counterpart of Optimality Theory's characterization of grammatical computation of well-formedness arises naturally from the assumption that each violation of a constraint C by an expression x contributes a well-formedness penalty $-w_C$ (with $w_C \geq 0$): the magnitude of w_C is the strength of C in the grammar. Since log-probabilities interact additively, we compute the well-formedness of x by adding together the penalties that x incurs from all constraints. In equations (using the natural logarithm, i.e., base e):

$$(5)\quad \log(P(x|G)) + z = H_G(x) = -\textstyle\sum_k w_k C_k(x); \qquad P(x|G) \propto e^{H_G(x)} = e^{-\sum_k w_k C_k(x)}$$

where $C_k(x)$ is the number of times that x violates C_k, and the constant of proportionality is $e^{-z} \equiv 1/Z \equiv 1/\sum_x e^{H_G(x)}$, responsible for ensuring that all probabilities sum to 1.[7] (5) provides the 'likelihood function' of our Bayesian analysis: the probability of data given a hypothesis.

Defining *Probabilistic Harmonic Grammar* (PHG), we propose that the learner's hypothesis space is all probabilistic grammars of the form $P(x|G)$ in Equation (5), where the weights w_k range over the non-negative real numbers, and the constraints in *Con*—and the universe of expressions over which x ranges—are given by a constraint-based grammatical theory such as OT.[8] PHG is a linguistic implementation of a type of model used commonly in machine learning under names including 'log-linear' or 'Maxent' (maximum entropy) models. That OT-style constraint-based grammars could be successfully translated into a probabilistic framework using machine learning techniques for Maximum Entropy models was first shown by Goldwater & Johnson (2003).

How this probabilistic picture changes the evaluation of candidate forms is shown in (6) above the name of each constraint C_k is its weight w_k; the added columns on the right show the Harmony and consequent probability of the alternative expressions for {Num, N}. Observe that $P(x|G)$ may depend on multiple weights for example $P(\text{Num-N}) = 1/(1 + e^{-[w_{\text{HEAD-L}} - w_{\text{NUM-L}}]})$.

(6) *PHG tableau*

a. *Num position in L3* ($Z \equiv e^{-0.85} + e^{-1.7}$)

	{Num, N}	1.7	.85	0	$H_G(x) = -\sum_k w_k C_k(x)$	$P(x\|G) = e^{H_G(x)}/Z$
		NUM-L	HEAD-R	HEAD-L		
a.	☞ Num-N		*		$-[1.7 \cdot 0 + 0.85 \cdot 1 + 0 \cdot 0]$	$e^{-0.85}/Z = 70\%$
b.	N-Num	*!		*	$-[1.7 \cdot 1 + 0.85 \cdot 0 + 0 \cdot 1]$	$e^{-1.7}/Z = 30\%$

[7] Other probabilistic extensions of OT assign probabilities to constraint rankings, each ranking being deployed deterministically as in standard OT (e.g. Anttila, 1997; Boersma, 1998; Jarosz, 2010).

[8] One other such theory is (non-probabilistic) Harmonic Grammar (Legendre, Miyata & Smolensky, 1990; Pater, 2009): a precursor to OT, it uses numerical constraint weights to compute Harmony as in PHG, but differences in Harmony are interpreted as differences in well-formedness, not different probabilities.

The probabilistic model proposed here differs crucially from that developed by Culbertson & Smolensky (2012), which used a formalism—probabilistic context-free grammar—not developed for or used in linguistic typology. The current model can therefore serve as a working example for researchers interested in taking advantage of the progress made in OT and related frameworks for explaining typology.

While the shift from OT to PHG enlarges the space of possible grammars—allowing the probability that a grammar assigns to an expression to range between 0 and 1—it is only the first step in the Bayesian account of the nominal word-order typology. As in OT, the predictions of PHG depend upon the constraint set. If the set of possible constraints remains as in (2), then languages of type L4 are still predicted to be impossible: more precisely, languages in which the probability of N-Num is greater than that of N-Adj are predicted to be unlearnable. Simply allowing grammars to be probabilistic is therefore not sufficient to address the full typology or other data. What we require now is a prior over grammars that both allows all attested systems and properly expresses preference relations among language types.

3. A prior for language acquisition

The particular prior that we propose has two main components, referred to as the *Con*(straint) Bias and the Regularization Bias. We state each component, and discuss its relationship to previous proposals, in the following two subsections.

3.1. Con Bias

Stated at a high level, the *Con* Bias, introduced above and repeated in (7), imposes conditions on the constraint set and the weights of individual constraints.

(7) *Con* Bias

a. The constraint set available to learners is sufficiently rich to generate all attested languages, but

b. there are soft biases that penalize the grammatical use of particular constraints.

For the purpose of analyzing the typology of nominal word-order, we have already seen that the provisional constraint set in (2) is insufficient. In particular, this set cannot generate L4 in either OT or PHG. Therefore, we minimally revise the constraint set as in (8).

(8) *Con* for nominal word-order

a. HEAD-L: order all heads to the left of their complements

b. HEAD-R: order all heads to the right of their complements

c. NUM-L: order numeral heads to the left of their complements

d. NUM-R: order numeral heads to the right of their complements

According to the *Con* Bias, each of these constraints is potentially associated with a penalty. Following prior art on maximum entropy (log-linear) models (Chen & Rosenfeld, 2000; Goldwater & Johnson, 2003; Hayes & Wilson, 2008), we assume that each penalty takes the form of a mean-zero Gaussian distribution on the constraint's weight.[9] As suggested by Goldwater & Johnson (2003), the penalty for a weight that differs from zero can be made stronger for certain constraints by assigning a higher precision (inverse variance) to their Gaussian distributions.

The Gaussian distribution $P(w) \propto e^{-\varphi(w-\mu)^2}$ assigns a lower prior probability—greater penalty—to the weight w, the further it is from the preferred value μ; here $\mu = 0$. The higher the precision parameter φ,

[9] More precisely, we are assuming a *truncated* Gaussian, because it is only defined for non-negative weight values: the left half of the bell-shaped curve, over negative w, is gone. Therefore where we use 'mean' is technically mode.

the more sharply peaked the Gaussian distribution and the more rapidly the probability becomes small—the more rapidly the penalty becomes large—as w moves away from zero. Here we posit a penalty for the use of the specific constraints Num-L and Num-R: these constraints need not be active in the most typologically well-attested languages, those with consistent headedness. The penalty, which we will denote by the precision value κ, is assessed to inconsistent languages, encoding the evident typological dispreference for them. Further, because of the extreme rarity of the inconsistent language L4 which violates Universal 18, we posit an additional penalty—a precision with value λ—to grammars in which the constraint needed to derive L4, Num-R, is active. This gives us 9.

(9) $$P_C(G(\mathbf{w})|\mathbf{b}) \propto [e^{-\kappa w^2_{\textsc{Num-L}}} e^{-\kappa w^2_{\textsc{Num-R}}}] e^{-\lambda w^2_{\textsc{Num-R}}}$$

Here we have introduced $\mathbf{b}$, the vector of parameters determining the prior. The two elements of this vector relevant to P_C are κ and λ.

Apart from the research already cited, there are only limited precedents for our *Con* Bias. Within linguistic theory, it is sometimes claimed that certain parameter settings or other grammatical options are 'marked' (e.g., Chomsky, 1986; Cinque, 2005). But this notion of 'marked' options has not been incorporated into a formal theory of learning, and as such remain inert observations rather than part of the solution to the problem of language acquisition. Hayes, Siptar, Zuraw & Londe (2009) provide evidence that certain grammatical constraints can be 'underlearned', in the sense that their weights are smaller than would be expected given the input to the learner. However, they do not state a formal bias, like ours, that could result in selective underlearning (see Hayes et al. 2009, pp. 853-856 for discussion of various types of bias that might be responsible for their findings). Work on language acquisition within OT and Harmonic Grammar has proposed an initial state of learning in which certain constraints are weaker than others (e.g., all Faithfulness constraints are initially weaker than all Markedness constraints; Smolensky 1996). This idea has often been implemented by assigning smaller initial ranking values to the weaker constraints, and simulation results show that this can be effective in ensuring that certain constraints remain weak at the end point of learning (e.g., Curtin & Zuraw, 2002). Our *Con* Bias is similar, except that the bias applies persistently throughout the learning process rather than only at the initial state; previous work has suggested that persistent biases are more effective than assumptions about the initial state (Prince & Tesar, 2004) but has not formalized it in the way proposed here.[10] Moreover, our notion of bias is not inherently tied to a particular algorithmic approach for learning weights: it states preferences about the outcome of learning, not the starting point of a particular search algorithm.

Finally, we note a connection between the *Con* Bias and the standard generative practice of hard restrictions on the set of primitives from which grammars are constructed: here, the constraint set. If the weight of a constraint were subject to a zero-mean Gaussian prior with *infinite* precision (zero variance), the constraint would necessarily have weight 0 and so would be effectively excluded from Con. For example, the constraint set (8) above becomes equivalent to our smaller, provisional set of (2) in the limit $\lambda \to \infty$. Similarly, in the limit $\kappa \to \infty$ all that remains are the general constraints Head-L and Head-R—a constraint set identical to the classical Head Direction Parameter. It is in this sense that the present proposal is a softening, rather than a wholesale revision, of previous generative approaches to typology. Absolute limits on typological variation correspond to the limiting case of our soft penalties on constraint strength.

3.2. The Regularization Bias: the prior distribution $P_R(G(\boldsymbol{w}))|\boldsymbol{b})$

The Regularization Bias favors grammars with less variation. In the Universal 18 model, this means favoring grammars in which $p_{\text{a}} \equiv P(\text{Adj-N})$ is either close to one (nearly 100% use of pre-nominal Adj-N)

[10] Somewhat closer to our proposal is the idea that constraints have different 'plasticities', which determine how quickly their weights change in response to language input (Jesney & Tessier, 2011).

or close to zero (nearly 100% N-Adj), and similarly for $p_{\mathrm{n}} \equiv P(\text{Num-N})$. We will pursue two approaches to formalizing this bias as a Bayesian prior: one directly targets the probabilities $\{p_{\mathrm{a}}, p_{\mathrm{n}}\}$ themselves, while the other targets the PHG weights which determine those probabilities.

3.2.1. *Prior directly targeting weights: $P_{Rw}(G(\boldsymbol{w}))|\boldsymbol{b})$*

When constraint weights in a PHG are zero, all Harmonies are zero, and, since $p \propto e^{H}$, it follows that all probabilities are equal: both p_{a} and p_{n} are exactly 0.5, i.e. in order varies freely. This suggests the possibility that a Regularization Bias might be definable in terms of constraint weights, punishing those near zero. To achieve this using the formal approach introduced above, the Gaussian prior, we can posit a Gaussian for each weight centered at a *non-zero* value, a Gaussian that assigns low prior probability to weights near zero.[11]

This *weight-based* regularization prior therefore assumes a Gaussian with non-zero mean μ_{H} and precision τ_{H} as the prior probability assigned to the strength w_{H} of the general head-direction constraints, and a non-zero mean μ_{N} and precision τ_{N} for the strength w_{N} of the Num-specific head-direction constraints, as given in (10).

(10) $$P_{\mathrm{Rw}}(G(\mathbf{w})|\mathbf{b}) \propto e^{-\tau_{\mathrm{H}}(w_{\mathrm{H}}-\mu_{\mathrm{H}})^2} e^{-\tau_{\mathrm{N}}(w_{\mathrm{N}}-\mu_{\mathrm{N}})^2}$$

When we present the technicalities of the analysis in section 4.1, we will precisely define what we refer to here by the strengths 'w_{H}' and 'w_{N}'. (As a preview, if $w_{\text{HEAD-L}} > w_{\text{HEAD-R}}$, so that the pressure from HEAD-L for heads to be left dominates the pressure from HEAD-R for heads to be right, then w_{H} will turn out to be $w_{\text{HEAD-L}} - w_{\text{HEAD-R}}$: the net strength of the pressure for heads to be left is the amount by which $w_{\text{HEAD-L}}$ exceeds $w_{\text{HEAD-R}}$.)

The means of the Regularization Bias Gaussians will be determined empirically from the degree of regularization observed in the experimental data: the larger the means, the greater the preferred weight values hence the smaller the preferred degree of variation. In addition, the precision of the Gaussian determines how *strongly* the prior pulls the weight to the mean—how much it penalizes a given degree of deviation from the preferred value.

Combining this Regularization Bias with the *Con* Bias defined above gives the complete weight-based prior P_{Rw}; it is given with its complete parameter vector—the bias vector **b**—in (11).

(11) $$P_{\mathrm{w}}(G(\mathbf{w})|\mathbf{b}) = P_{\mathrm{C}}(G(\mathbf{w})|\mathbf{b}) P_{\mathrm{Rw}}(G(\mathbf{w})|\mathbf{b}); \qquad \mathbf{b} = (\kappa, \lambda; \mu_{\mathrm{H}}, \tau_{\mathrm{H}}; \mu_{\mathrm{N}}, \tau_{\mathrm{N}})$$

However, pulling weights away from zero as this prior does does not suffice to *ensure* regularization. To see this, we return to an observation about tableau (6), where we saw that the probability of the word order N-Num depends on the *difference* between weights $w_{\text{NUM-R}}$ and $w_{\text{HEAD-L}}$ (exerting opposing pressures on the position of Num). A potential problem is that this difference might be small even if the weights themselves are large.

3.2.2. *Prior directly targeting probabilities*

Given that the weight-based prior P_{Rw} may not always succeed in favoring regularization, we also consider a more direct approach which targets the probabilities $p_{\mathrm{a}}, p_{\mathrm{n}}$ themselves.[12]

[11] In Hayes & Wilson (2008), and commonly in the Maxent models of machine learning, the Gaussian priors for weights have mean zero. This enforces the type of preference for uniformity discussed in note 3. The idea of formalizing informative Gaussian priors in a Maxent model by using non-zero means was suggested by Goldwater & Johnson (2003).

[12] The prior developed below over expression-probability space can be transformed to a prior over constraint-weight space but the result is somewhat problematic. The transformation is modeled on (20): $p = 1/[1+e^{-w}]$. The density in probability space of a beta prior, $p^{\alpha-1}(1-p)^{\beta-1}$, with symmetric shape parameters $\alpha = \beta$ that take values between 0 and 1, is maximized near 0 and 1,

Here we follow Reali & Griffiths (2009); Culbertson (2010); Culbertson & Smolensky (2012) in deploying the beta distribution with symmetric shape parameters that assign small prior probability to values of $p_{\mathrm{a}}, p_{\mathrm{n}}$ near zero, with values near 0 and 1 receiving the highest prior probability.

(12) $$P_{\mathrm{Rp}}(G(\mathbf{w})|\mathbf{b}) = [p_{\mathrm{a}}(1-p_{\mathrm{a}})p_{\mathrm{n}}(1-p_{\mathrm{n}})]^{-\gamma}$$

In (12), the beta prior is scaled by $\gamma > 0$,[13] a parameter to be fit against the experimental data (see Smith & Eisner, 2007, for an approach to entropy minimization which also uses such a scaling factor). The complete weight-based prior, and its complete parameter vector—the bias vector **b**—is:

(13) $$P_{\mathrm{p}}(G(\mathbf{w})|\mathbf{b}) = P_{\mathrm{C}}(G(\mathbf{w})|\mathbf{b})P_{\mathrm{Rp}}(G(\mathbf{w})|\mathbf{b}); \qquad \mathbf{b} = (\kappa, \lambda; \gamma)$$

The prior is the main original contribution of the article; it formalizes the central idea of soft biases on linguistic structure combined with a bias to regularize. In section 4, we present the technical details of the model; in section 5, we present the results of fitting the bias parameters to the experimental data.

4. Modeling artificial language learning biases

Because the model proposed here adopts a Bayesian approach to learning, we assume that the learners evaluate grammars based both on experience with linguistic input and prior biases about how that input should be structured. The precise nature of the relationship between these two influences on learning is given by Bayes' Theorem, in (14).

(14) $$P(\text{Grammar}|\text{Data}) \propto P(\text{Data}|\text{Grammar})P(\text{Grammar})$$

In the preceding sections we have formalized the prior—which includes the *Con* Bias and the Reg bias. It is the goal of the model presented here to quantify the strength of these biases based on direct behavioral evidence, namely learning outcomes reported in Culbertson et al. (2012).

4.1. The likelihood that a grammar generates a given corpus of training data: $P(Training_l|G(\boldsymbol{w}))$

The grammar $G(\mathbf{w})$ is defined by a vector **w** consisting of one weight w_k per constraint C_k. Given the input {Adj, N} to express the grammar assigns a probability to each of the two expressions we consider here: Adj-N and N-Adj. According to equation (5) above, these probabilities are exponentially related to the Harmonies of these expressions, which are:

(15) $$H(\text{Adj-N}) = -w_{\text{HEAD-R}} \qquad H(\text{N-Adj}) = -w_{\text{HEAD-L}}$$

because the set of constraints V violated by each of the expressions is: V(Adj-N) = {HEAD-R}; V(N-Adj) = {HEAD-L}. The probability of an expression x is $p(x) = Z^{-1}e^{H(x)}$ where Z is the sum, over all competing expressions y, of $e^{H(y)}$ (ensuring that the sum of all such probabilities is 1). The expression N-Adj is the only competitor to Adj-N for the input {Adj, N} (and N-Num the only competitor to Num-N for {Num, N}). It follows that the probability of the pre-nominal adjective form Adj-N is:

but the corresponding density in weight space is $p^{\alpha}(1-p)^{\beta}$, which peaks at *zero*. The probability mass that is concentrated near 0 and 1 in probability space is spread out to infinity in weight space: while the *density* peaks at weight zero, the bulk of the probability mass is not near zero. Thus maximizing this *density* function in weight space, as in the Maximum A Posteriori procedure we adopt in (24), would have the effect of pushing weights towards zero—exactly the opposite of the desired regularization force.

[13] In the standard parameterization, the beta density is proportional to $p^{\alpha-1}(1-p)^{\beta-1}$. If this is scaled by raising it to a power ψ, and if $\alpha = \beta$, then the form given in (12) results, with $-\gamma \equiv \psi(\alpha - 1)$. The range of interest has $0 < \alpha < 1$, where the favored regions are $p = 0, 1$; then $\gamma > 0$. We assume that the parameters α, ψ have the same value for p_{a} and p_{n}.

$$\begin{aligned}
\text{(16)}\quad p(\text{Adj-N}) &= Z^{-1}e^{H(\text{Adj-N})} = \left[e^{H(\text{Adj-N})}+e^{H(\text{N-Adj})}\right]^{-1}e^{H(\text{Adj-N})}\\
&= \left[e^{-w_{\textsc{Head-R}}}+e^{-w_{\textsc{Head-L}}}\right]^{-1}e^{-w_{\textsc{Head-R}}} = 1/\left[1+e^{w_{\textsc{Head-R}}-w_{\textsc{Head-L}}}\right]\\
&= f(w_{\textsc{Head-L}}-w_{\textsc{Head-R}})\\
&\equiv p_{\text{a}}(\mathbf{w})
\end{aligned}$$

where f is the standard logistic function:

$$\text{(17)}\quad f(u)\equiv 1/(1+e^{-u})$$

The probability of the pre-nominal numeral form Num-N is computed similarly. The constraint violations are now more numerous: V(Num-N) = {HEAD-R, NUM-R}; V(N-Num) = {HEAD-L, NUM-L}. This entails that the Harmonies are

$$\text{(18)}\quad H(\text{Num-N}) = -w_{\textsc{Head-R}}-w_{\textsc{Num-R}} \qquad H(\text{N-Num}) = -w_{\textsc{Head-L}}-w_{\textsc{Num-L}}$$

so that the pre-nominal probability is:

$$\begin{aligned}
\text{(19)}\quad p(\text{Num-N}) &= \left[e^{H(\text{Num-N})}+e^{H(\text{N-Num})}\right]^{-1}e^{H(\text{Num-N})}\\
&= \left[e^{(-w_{\textsc{Head-R}}-w_{\textsc{Num-R}})}+e^{(-w_{\textsc{Head-L}}-w_{\textsc{Num-L}})}\right]^{-1}e^{(-w_{\textsc{Head-R}}-w_{\textsc{Num-R}})}\\
&= 1/\left[1+e^{(-w_{\textsc{Head-L}}-w_{\textsc{Num-L}})-(-w_{\textsc{Head-R}}-w_{\textsc{Num-R}})}\right]\\
&= f([w_{\textsc{Num-L}}-w_{\textsc{Num-R}}]+[w_{\textsc{Head-L}}-w_{\textsc{Head-R}}])\\
&\equiv p_{\text{n}}(\mathbf{w})
\end{aligned}$$

The equations (16) and (19) for p_{a} and p_{n} can be intuitively understood as follows (keeping in mind that the logistic function $f(u)$ rises monotonically from 0 to 1 as u goes from $-\infty$ to ∞, taking the value 0.5 when $u=0$). The pre-nominal adjective probability p_{a}, unsurprisingly, gets larger the stronger HEAD-L is relative to HEAD-R; if HEAD-L is the stronger one ($w_{\textsc{Head-L}} > w_{\textsc{Head-R}}$), the probability that an adjective is positioned left of the noun is greater than 50%. The probability that a numeral is positioned left is governed both by the strength of HEAD-L relative to HEAD-R and of NUM-L relative to NUM-R: $p_{\text{n}} > 0.5$ when the sum $[w_{\textsc{Head-L}}-w_{\textsc{Head-R}}]+[w_{\textsc{Num-L}}-w_{\textsc{Num-R}}] > 0$,

The pre-nominal probabilities $p_{\text{a}}, p_{\text{n}}$ can be written using the abbreviations $(v_H, v_N)\equiv\mathbf{v}$ as follows:

$$\begin{aligned}
\text{(20)}\quad & v_{\text{H}}\equiv w_{\textsc{Head-L}}-w_{\textsc{Head-R}}; \qquad v_{\text{N}}\equiv w_{\textsc{Num-L}}-w_{\textsc{Num-R}}\\
& p(\text{Adj-N}) = f(v_{\text{H}}); \qquad p(\text{Num-N}) = f(v_{\text{H}}+v_{\text{N}})
\end{aligned}$$

Limiting attention to the two-word phrases we have been considering, the probabilities assigned by the grammar $G(\mathbf{w})$ are entirely determined by the two parameters $\mathbf{v}=(v_{\text{H}}, v_{\text{N}})$, and thus for our purposes the hypothesis space of grammars is effectively two-dimensional. All computational results to be described were obtained by computing over the two variables $\mathbf{v}$; for transparency, however, our exposition will continue to use the four weights $\mathbf{w}$. (Note that, while the four elements of $\mathbf{w}$ are constrained to be non-negative real numbers, the two elements of $\mathbf{v}$ can be any real numbers.)

Because the probability distribution defined by $G(\mathbf{w})$ depends only on the difference (v_{H}) between $w_{\textsc{Head-L}}$ and $w_{\textsc{Head-R}}$, and not on the actual values of either, we can restrict these weights by the condition that one or the other be zero. For example, if we consider $w_{\textsc{Head-L}} = 3$, $w_{\textsc{Head-R}} = 5$, then we can shift both down by 3, getting $w_{\textsc{Head-L}} = 0$, $w_{\textsc{Head-R}} = 2$: the shift preserves what matters, the difference $v_{\text{H}} = 3-5 = 0-2 = -2$. In general, shifting both $w_{\textsc{Head-L}}$ and $w_{\textsc{Head-R}}$ down by an amount equal to the smaller of the two preserves their difference while reducing one of them to 0. By the same logic, we can assume the restriction that either $w_{\textsc{Num-L}}$ or $w_{\textsc{Num-R}}$ is zero.

Having determined the probabilities of $G(\mathbf{w})$ generating a single pre-nominal expression, we can now compute the probability of generating a corpus of training utterances. The data produced by a grammar

$G(\mathbf{w})$ is a stream of independently-generated expressions. (Throughout, the probability of the two possible inputs {Adj, N} and {Num, N} are taken to be equal, as they are in the experiment.) The probability that an informant's grammar $G(\mathbf{w})$ would have generated $Training_l$, the training set for condition l—consisting of $t_{l,\mathrm{a}}$ adjective expressions, $c_{l,\mathrm{a}}$ of which are pre-nominal, and $t_{l,\mathrm{n}}$ numeral expressions, $c_{l,\mathrm{n}}$ of which are pre-nominal—is thus given by the product of two binomial distributions:

(21) $P(Training_l|G(\mathbf{w})) = \mathrm{binomial}(c_{l,\mathrm{a}}|p_{\mathrm{a}}(\mathbf{w}), t_{\mathrm{a}})\,\mathrm{binomial}(c_{l,\mathrm{n}}|p_{\mathrm{n}}(\mathbf{w}), t_{\mathrm{n}})$

where $G(\mathbf{w})$'s pre-nominal production probabilities $p_{\mathrm{a}}(\mathbf{w})$, $p_{\mathrm{n}}(\mathbf{w})$ are given above in (16) and (19), and the standard binomial distribution is defined by:

(22) $\mathrm{binomial}(c|p,t) = \binom{t}{c} p^c (p-1)^{t-c}$

The same expressions that relate an informant's grammar to the likelihood of producing a training corpus also relate an experimental participant's grammar to the likelihood of producing a testing corpus, so the preceding equations will also be used in section 4.4 to analyze the learner's own productions.

4.2. The prior $P(G|\boldsymbol{b})$

The weight- and probability-based priors were given in (10) and (12), respectively. All that remains to specify is the weights used in (10); these are:

(23) $w_{\mathrm{H}} = \max(w_{\textsc{Head-L}}, w_{\textsc{Head-R}});\ \ w_{\mathrm{N}} = \max(w_{\textsc{Num-L}}, w_{\textsc{Num-R}})$

That is, w_{H} and w_{N} are the non-zero constraint weights.

4.3. The acquired grammar: the posterior distribution $P(G|Training_l, \boldsymbol{b})$

According to our Bayesian approach, a learner presented with training data $Training_l$ will acquire the grammar with *highest posterior probability*, given these data and the prior.[14] Using Bayes' Theorem, this means the learned grammar $\hat{G}$, specified by the learned constraint weights $\hat{\mathbf{w}}$, is given by equations (24). The likelihood factor $P(Training_l|G)$ is given above in (21) and the prior $P(G|\mathbf{b})$ is determined by equation (11) for the weight-based or equation (13) for the probability-based approach.

(24)
$$\begin{aligned} & P(G|Training_l, \mathbf{b}) \propto P(Training_l|G)\,P(G|\mathbf{b}) \\ & \hat{\mathbf{w}}_l(\mathbf{b}) \equiv \mathrm{argmax}_{\mathbf{w}}\, P(G(\mathbf{w})|Training_l, \mathbf{b}) \\ & \hat{G}_l(\mathbf{b}) \equiv G_l(\hat{\mathbf{w}}_l(\mathbf{b})) \end{aligned}$$

Changing the bias parameters $\mathbf{b}$ changes what grammar gets learned: although the evidence from the data—$P(Training_l|G)$—is unchanged, the conclusion the learner draws from this evidence—which grammar is most likely responsible for generating the evidence—is biased by the prior, which changes with $\mathbf{b}$.

[14] Another Bayesian approach, according to which the learner acquires a grammar that is drawn from the posterior distribution, and predictions thus involve integrating over the entire hypothesis space, proves somewhat problematic because, following the conclusion of note 12, it is expected that the integral will need to weigh all the probability mass spread out to infinite weight values. The present 'MAP' (maximum a posteriori) approach also provides a more transparent analysis.

4.4. *Fitting the prior with the likelihood that a learning bias yields a given corpus of testing data:* $P(Test_l|\boldsymbol{b})$

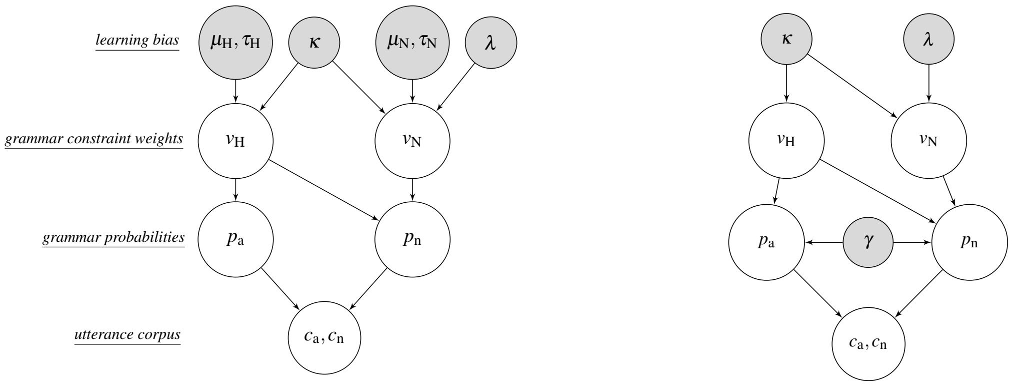
Following equation (24), a given bias $\mathbf{b}$ determines the grammars $\hat{G}_l(\mathbf{b})$ that learners in conditions l will acquire, and these grammars in turn determine what those learners are then likely to produce at test. The corpus of utterances produced during post-training testing by a condition-l learner, $Testing_l$, is characterized by $\bar{c}_{l,\text{a}}$ and $\bar{c}_{l,\text{n}}$, the respective counts of pre-nominal forms Adj-N and Num-N, out of a total of $\bar{t}_{l,\text{a}}$ and $\bar{t}_{l,\text{n}}$ utterances of {Adj, N} and {Num, N}. The equations giving the likelihood of these counts given the weights $\hat{\mathbf{w}}_l(\mathbf{b})$ that learners acquire were already derived in section 4.1, in the context of training data. The probability of the entire set of testing data is just the product of the probabilities of the testing data $Testing_l$ over all the statistically independent conditions l. Thus we have equation (25).

(25) $$P(Test|\mathbf{b}) = \Pi_{l=1}^{4} P(Test_l|\mathbf{b}) = \Pi_{l=1}^{4} \text{binomial}(\bar{c}_{l,\text{a}}|p_\text{a}(\hat{\mathbf{w}}_l(\mathbf{b})), \bar{t}_\text{a})\,\text{binomial}(\bar{c}_{l,\text{n}}|p_\text{n}(\hat{\mathbf{w}}_l(\mathbf{b})), \bar{t}_\text{n})$$

We are now in a position to use the experimentally observed testing data to fit the bias parameters $\mathbf{b}$: we simply find those parameters that make these observations most likely (26). To see the resulting predictions, we then use equations (24) to compute the weights that are learned under that bias for each condition, and see what those weights predict for the pre-nominal probabilities in the testing utterances produced by learners.

(26) $$\hat{\mathbf{b}} = \text{argmax}_{\mathbf{b}}\, P(Test|\mathbf{b}); \qquad \text{Learned weights (condition } l\text{): } \hat{\hat{\mathbf{w}}}_l = \hat{\mathbf{w}}_l(\hat{\mathbf{b}})$$

To summarize concisely the two models whose parameters we will fit, Figure 2 shows graphical representations for both (see Finkel & Manning, 2009, for high-level discussion of a mathematically similar type of model). ($\{v_\text{H}, v_\text{N}\}$ are the weight variables introduced in (20).)



(a) Weight-based Regularization Prior

(b) Probability-based Regularization Prior

Figure 2: Graphical Models. Shaded nodes are prior bias parameters. Text labels for model (a) identify each level of structure. Regularization Bias parameters for Model (a) are $(\mu_\text{H}, \tau_\text{H}), (\mu_\text{N}, \tau_\text{N})$; *Con* Bias parameters are κ, λ. Structure of model (b) differs only in that the Regularization Bias (parameter γ) applies at the level of grammar probabilities.

5. The bias parameters fit to the experimental data

5.1. *Weight-based regularization prior*

The weight-based bias parameters were estimated by fitting the predicted testing data to the average data observed in the Culbertson et al. (2012) experiment.[15] The resulting parameter values and goodness-of-fit

[15] Optimizations were performed using the R function `optim` (R Development Core Team, 2010) with the default method for finding optimal weights, and method L-BFGS-B for optimizing the biases with lower and upper bounds (0, 200). One of the fit parameters reached the upper bound, we found that relaxing the bound led to a fit that was empirically equivalent.

are given in (27). The predictions resulting from that prior are displayed in Fig. 3. We now discuss how the weight-based Regularization Bias, and the *Con* Bias, give rise to these predictions.

(27) *The fit bias parameters, weight-based prior*
$(\mu_{\mathrm{H}}, \tau_{\mathrm{H}}) = (1.17, 200),\ (\mu_{\mathrm{N}}, \tau_{\mathrm{N}}) = (9.49, 0.45),\ \kappa = 1.12,\ \lambda = 2.13$
log-likelihood of the testing data $= -15.01$

Recall the relations between a grammar's constraint weights and the probabilities of pre-nominal adjectives and numerals, given by (16) and (19) in terms of the logistic function f (17). Because the proportion of majority order in all training data is 70%, and $f(0.85) = 0.7$, the general constraint weights required to match the Adj training data have value 0.85 (for HEAD-L in conditions L1 and L4, or for HEAD-R in conditions L2 and L3). These weights maximize the likelihood of the training data, and are the constraint strengths that would be learned in the absence of a bias.

The Gaussian Regularization prior over the general constraint weights has mean $\mu_{\mathrm{H}} = 1.17$; it therefore exerts a force raising the strength of the relevant general constraint from 0.85 towards 1.17. The strength of this bias is high: $\tau_{\mathrm{H}} = 200$. As a consequence, the weights that maximize the posterior probability of the grammar given the training data are close to 1.17. Since p_{a} is determined by v_H alone, this predicts that the probability of the majority Adj order is close to $f(1.17) = 0.76$ for all conditions 1–4, as shown by the light-gray bars in Fig. 3. Since the predicted 76% exceeds the training 70%, *this is regularization*.

Thus the model does not predict any differences across conditions in use of Adj majority order—and indeed these differences are much less pronounced than those among Num order (Culbertson et al., 2012). Importantly, this model *does* capture the rather dramatic quantitative variation over conditions in Num ordering, as shown by the dark-gray bars in Fig. 3.[16] The model also captures the average trend fairly well, as shown by the black bars.

The predictions for Num depend on both the general-constraint weights and the Num-specific constraint weights (19). To match the training data in conditions L1, L2, the specific constraints are not needed and thus have weights of 0. Matching the training data in the inconsistent conditions L3 and L4, on the other hand, require specific-constraint weights that exceed the general-constraint weight of 0.85: the necessary value is in fact 1.7, twice 0.85 (for $w_{\text{NUM-L}}$ in L3, and for $w_{\text{NUM-R}}$ in L4); this is because $0.70 = f(1.7 - 0.85) = f(0.85)$. The Regularization bias for the specific-constraint weights pulls those weights towards their preferred value of $\mu_{\mathrm{N}} = 9.5$—with a relatively weak precision of only 0.45. The overall result is that the posterior-optimizing weights for the specific constraints are, for L1 – L4 respectively: $w_{\text{NUM-L}} = 0.51$, $w_{\text{NUM-R}} = 0.34$, $w_{\text{NUM-L}} = 2.16$, and $w_{\text{NUM-R}} = 1.37$. The consequent predicted values for Num majority order use are shown in Fig. 3. (For example, for L1—in which the general and specific constraints add, both favoring left position—the predicted value is $f(1.15 + 0.51) = 0.84$; the observed value is 0.85.)

These posterior-optimizing weights reflect the influence of not just the Gaussians comprising the Regularization Bias, but also by the Gaussians comprising the *Con* Bias. This latter bias penalizes the use of the specific constraints NUM-L and NUM-R with a mean-zero Gaussian of strength (precision) $\kappa = 1.12$, and inflicts an additional penalty for the use of the dispreferred specific constraint NUM-R through another mean-zero Gaussian with strength $\lambda = 2.13$. The κ factor in the *Con* Bias is responsible for the model predicting a decrement in regularization in the inconsistent conditions L3 and L4 relative to the consistent conditions L1 and L2, and the λ factor leads to the prediction of anti-regularization of N-Num in L4.

To verify that the model is generalizable to a different data set we apply it, with its bias parameters optimized to fit the experiment reported in Culbertson et al. (2012), to a replication experiment reported in

[16] This is essentially the reason that using Num-specific constraints in *Con* leads to a better account of the experimental data than does the corresponding analysis with Adj-specific constraints in *Con* instead.

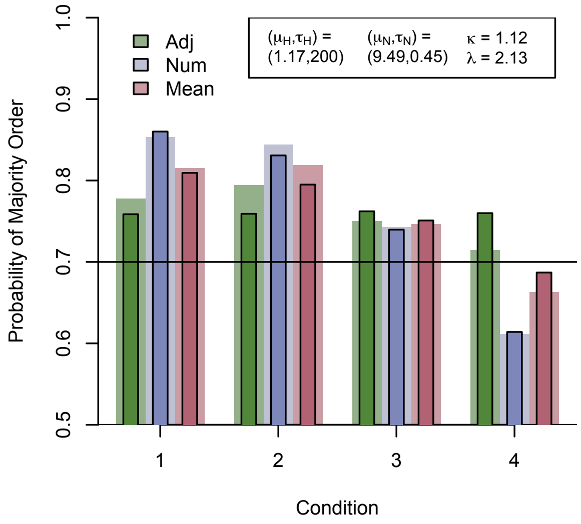


Figure 3: Predictions from the weight-based prior, with fit parameter values. Plotted are the predicted probabilities of the *majority* orders for {Adj, N} (green bars), {Num, N} (blue bars), and their average (red bars), superimposed on the corresponding observed values in the experimental data.

Culbertson (2010). The log-likelihood of the testing data in the replication experiment is -15.85, only a modest decline from -15.01, the result of the optimized fit to the original data.

While research on algorithmic-level issues of Bayesian cognitive models—e.g., appropriate resource-limited approximate optimization principles—is growing, the landscape is currently rather unclear. The results presented above arise from the following assumption: the learner's search for an optimal grammar is restricted: it is a single search starting from the neutral grammar defined by all zero constraint weights—the grammar which assigns equal probability to all expressions. Given this constraint, multiple optimization algorithms prove to find the same results: those reported above. However, this restricted search procedure sometimes finds a local, not a global, optimum: the computationally expensive method of restarting the search from multiple hypothesized initial grammars would be needed to find the global optimum.[17]

5.2. *Probability-based regularization prior*

In section 3.2.2, we presented an alternative formalization of regularization stated directly in terms of the probabilities of expressions, positing a prior that favors probabilities near 0 or 1. The same *Con* Bias is used as in the weight-based model.

The values of the prior parameters found from the optimization (26) are given in (28).[18] The resulting predicted word order probabilities are shown in Figure 4.

(28) *The fit bias parameters, probability-based prior*

[17] No set of bias parameters were found that would provide a satisfactory fit to the experimental data under the assumption that learners manage to compute the globally-optimal grammar; but given the severe instability of the *global* optimum under changes in the bias (note 19) this may simply reflect the extremely erratic surface over which the analyst's search for a satisfactory bias plays out.

[18] All optimizations were performed with the `R` function `optim` using method L-BFGS-B, with lower bounds of 0 and upper bounds of 5 for the constraint weights and (20, 20, 50) for the bias parameters $(\kappa, \lambda, \gamma)$.

$\kappa = 0.41, \lambda = 1.04; \gamma = 7.35$
log-likelihood of the testing data $= -15.52$

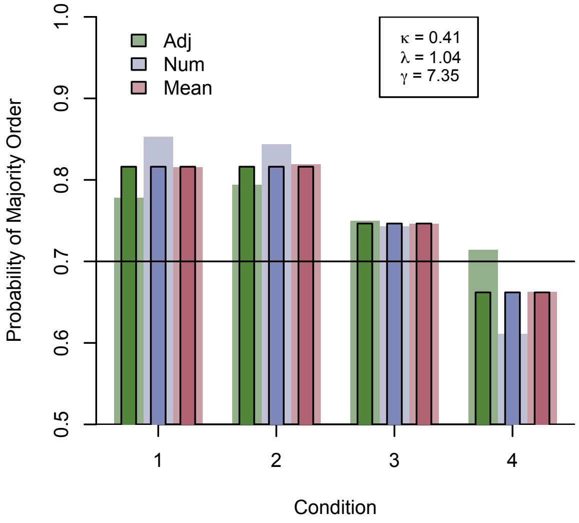


Figure 4: Predictions from the probability-based prior, with fit parameters. Predicted probabilities are shown superimposed on the corresponding observed values.

The model perfectly captures the trend in the averages (black bars), again with the *Con* Prior responsible for lowering the regularization in L3 (through κ) and lowering it still further in L4 (through λ). The model is not however able to capture any of the differences between Adj and Num regularization because the constraint weights are all subject to the same pressures from the prior, as well as from the training data, which is always 30% variation in both Adj and Num. Quantitatively, the fit is only somewhat worse than for the weight-based model, however: the log-likelihoods of the testing data are -15.52 and -15.00. Further, this model is preferable to the weight-based model on complexity grounds since it provides this fit with only 3 bias parameters (compared to 6).

We can again apply this model with its optimized bias parameters to the replication experiment reported in Culbertson (2010). In this case, the log-likelihood of the testing data in the replication experiment is -15.35, actually a slight improvement over -15.52, the result of the optimized fit to the original data. This is due to the fact that differences among conditions are reproduced in the replication experiment but differences among modifier-types less pronounced.

6. Discusssion and conclusions

6.1. *Assessing the models*

The design of the bias presented here was guided by the following desiderata:

(29) *Desiderata for the learning bias*

a. *Formal.* The bias should be formally natural; e.g., the constraint weights (or expression probabilities) are assessed independently—the prior is factorizable into a product of factors, each assessing a single weight (or probability).

b. *Computational.* Search for appropriate weights given training data should not be problematic.

c. *Empirical.* The bias should enable a good fit to the experimental data.

We have encountered several challenges to meeting these desiderata:

(30) *Challenges to desiderata*

a. *Formal: Weight combination.* The probability of an expression is, in general, dependent on multiple weights (those of the constraints it violates). [E.g., (19).]

b. *Computational: Disjunctive search.* Weights are subject to disjunctive criteria, which can make search for appropriate weights difficult.

c. *Empirical.* The experimental data present a challengingly complex pattern.

The *Con* Bias meets the desiderata satisfactorily. A mean-zero Gaussian over prior weights is a natural expression of a bias for small constraint weights, and it has been used successfully in many Maxent models, which have the same formal structure as our Probabilistic Harmonic Grammar. The *Con* Bias factors into a product of priors each targeting a single weight. The precision of the Gaussian assessing the weight of an individual constraint has a clear linguistic interpretation as the strength of the bias against grammars that activate that constraint (by assigning it a relatively high weight). The mean-zero Gaussians present no search problems, and contribute exactly as theoretically expected to capturing the central average tendency in the learning data: decreasing regularization as we go from the languages L1 and L2 with consistent head directionality show the greatest regularization, to the inconsistent L3, and finally to the Universal-18-violating L4.

While modulated by the *Con* Bias, regularization is driven by the Regularization Bias, which poses the greater challenge to formalization. Consider first the weight-based approach to this bias (section 5.1). At the formal level, this prior achieves factorizability: it is the product of two factors, each independently evaluating a single weight, penalizing deviation from a constraint-specific preferred value. However, this prior does not necessarily, even in principle, actually achieve the goal of disfavoring probabilities near 0.5, because probabilities are often determined by *differences* of weights, and a difference may be small even if the individual weights are not. This is the challenge that weight combination presents to an approach to regularization based directly on weights (30a).

In the probability-based approach to the Regularization Bias (section 5.2), the formal status of the prior is largely satisfactory. It is a product of factors each assessing an individual probability, each factor being a scaled beta distribution. The beta distribution is the conjugate prior for our likelihood function, a binomial; it is formally natural in this sense. By construction, this prior favors probabilities close to 0 or 1, providing a regularization force that is not compromised by the problem of weight combination which plagues the weight-based prior.

An important difference between the two forms of the Regularization Bias is the extent to which regularization is formalized as occurring within the grammatical system or outside it, in the following sense. A function of a grammatical analysis is to recast the surface forms of expressions into a more fundamental description; in the PHG case, an expression is analyzed as a set of weighted constraint violations. If this is the correct way to analyze regularization, it should be enlightening to state this bias in terms of constraint weights. This is rejected by the probability-based formalization of the Regularization Bias, which deals with the surface properties of expressions, in particular, their frequencies. To the extent that this formalization of the Regularization Bias is to be preferred, we can conclude that regularization is farther removed from the grammar proper, consistent with the cognition-general status of regularization. However a bias against inconsistency may also be cognition-general, and we have seen that the consistency bias in word order can be naturally expressed within a constraint-based grammar: this may be seen as the reflex, within grammar, of a wider-scope principle of cognition. Whether regularization has the same status is a question addressed by these two formalizations of the prior.

Turning to the computational desideratum, we observed difficulties with the weight-based prior. The learner's search for appropriate weights given their training data faces a challenge arising from the inherently *disjunctive* search space (30b): either $w_{\text{HEAD-L}}$ *or* $w_{\text{HEAD-R}}$ is zero, and similarly for $w_{\text{NUM-L}}, w_{\text{NUM-R}}$, creating a search space with 4 distinct regions. Finding a globally optimal set of weights would essentially require running 4 separate searches, starting in each one of these regions. This is related to the problem of weight combination, since it arises because it is only $[w_{\text{HEAD-L}} - w_{\text{HEAD-R}}]$—and not $w_{\text{HEAD-L}}$ or $w_{\text{HEAD-R}}$ individually—that matters. For the Regularization Bias to disfavor probabilities near 0.5, it must disfavor grammars in which $[w_{\text{HEAD-L}} - w_{\text{HEAD-R}}]$ is near 0, which cannot be done via independent penalties on individual weights—unless we take the step of reducing the weight values so that the smaller is always 0, thereby creating the disjunction problem.[19]

The probability-based Regularization Bias does not appear to suffer from these search problems, and thus is to be favored from the perspective of computational tractability.

With respect to the empirical desideratum, both models present fairly successful fits to the data, capturing the main patterns of interest. The clearest difference along this dimension is in the two models' differentiation of adjectives and numerals. The weight-based prior model predicts (in fact slightly exaggerates) the asymmetry present in the data, while the probability-based prior model fails to predict any such asymmetry (for reasons discussed in section 5.2). However, in sum, the results favor the probability-based prior: while faring somewhat less well on the empirical desideratum, it is less complex and does not suffer the computational problems observed in the weight-based model of the learner's search for an optimal grammar given the training data.

6.2. *Returning to the typology*

We have seen that both models capture the main behavioral asymmetries in the artificial language learning experiment—both models predict greater regularization for consistent ordering patterns (L1, L2) compared to inconsistent patterns (L3, L4), and substantially less use of the majority input order (in fact under-regularization) for the pattern originally ruled out by Greenberg's Universal 18 (L4). Insofar as the behavioral results mirror the typology in Table 1, the models thus also capture, qualitatively, the cross-linguistic distribution of these ordering patterns. The Regularization Bias, exhibited by learners in the experiment, and captured by the modeling results, is also apparent in the typology; in the WALS sample, there are 966 languages determined to have a clearly preferred ordering pattern, but only 117 languages (not shown in Table 1) with no clear preference for either Adj or Num or both. This suggest that linguistic variation is strongly curbed by such a regularizing force.

Nevertheless, the experimental results and the models we have tested here do not capture *every* asymmetry in Table 1. In particular there is no prediction that L2 should differ from L1 despite the clear difference typologically; both are consistent patterns, and no distinction is made here between the constraints HEAD-R and HEAD-L. In this case, the discrepancy between our modeling results and the typology are likely due to the choice of English learners as experimental participants. If English learners have a preference for their native language pattern, L1, then any preference they might have shown for L2 could be masked by this (see Culbertson et al., 2012, for additional discussion).[20]

[19] A further practical difficulty arises because the globally optimal weights are also quite unstable as the bias is varied: the optimal weights instantaneously jump as the bias crosses some critical point at which the global optimum shifts from one disjunctive weight combination to another. This makes quite difficult the analyst's search for a set of prior-parameters which would yield satisfactory results under global weight optimization.

[20] How the bias displayed by artificial-language learners might both reflect native-like priors *and* effects of previously learned languages might be explainable through an assumed hierarchical structure in the Bayesian model—perhaps along the following

6.3. Conclusion

The biases of learners have long been of interest to linguists and psychologists exploring the power and limitations of the human language learning faculty. One source of evidence for such biases comes from linguistic typology—a potential reflection of learners' preferences. We have proposed here a solution to an ongoing challenge in understanding this evidence, namely how to formulate a theory which both accommodates typologically rare languages while at the same time explains their rarity. Within a constraint-based approach, closely related to Optimality Theory—Probabilistic Harmonic Grammar—we explain the relative rarity of a given language type by hypothesizing soft biases which target the constraints needed to derive it. We formalize these soft biases in terms of a Bayesian prior on learning. This solution provides not only a means for qualitatively understanding typology, but also a tool for understanding the behavioral results of language learning experiments. Modeling these results allows us to determine, quantitatively, the prior bias against particular constraints, and how this prior influences what grammars learners acquire.

The major contribution of our solution is a new formalization of the role of *Con*—the set of constraints used to derive language types in our PHG theory. In standard Optimality Theory, the predictive power of the theory of typology comes from the particular set of constraints that the theorist posits are included in *Con*. Language types are either possible or impossible with respect to *Con*. In the theory proposed here, the set of constraints is more inclusive, and the predictive power comes from the cost associated with particular constraints. Some constraints might be so costly that grammars in which they are active are very unlikely to be acquired, and thus the linguistic patterns which would be optimal according to these grammars are effectively impossible. However, languages types which are possible but rare are included in the typology, and their rarity is explained as a result of the cost of activating the constraints needed to derive them.

That the theory we have developed here to analyze learning biases and their relation to typology is formalized within PHG is, we believe, an important asset. First, PHG makes clear contact with a tradition of typological analysis in Optimality Theory—researchers interested in modeling language learning data can make use of the types of constraints already motivated within OT. Second, PHG is a linguistic implementation of Maxent models, and as such, tools developed within the machine learning community can be used to alter or augment what we have proposed here as needed.

The modeling results we have presented illustrate that our PHG theory is able to capture interesting features of the artificial language learning experiment reported in Culbertson et al. (2012). The experiment was designed to test the interaction between (i) learning biases related to the typological asymmetry known as Universal 18, and (ii) a cognitive bias, regularization, which disfavors variation. We showed that, subject to certain starting conditions, a model which encodes the regularization bias in terms of weights of constraints in a grammar (dispreferring values near zero) reveals that learners have a clear bias in favor of consistent-head-ordering languages (derived from a set of general head-alignment constraints) and against inconsistent languages, especially against the particular pattern Adj-N, N-Num. This is as predicted based on the typology. However, encoding regularization in this way comes at a computational cost, making the grammar space quite complex to search and therefore making the inference problem more difficult—clearly at odds with the guiding idea that learning biases serve to aid language acquisition.

Encoding regularization in terms of the probabilities derived from grammars (following Reali & Griffiths, 2009; Culbertson & Smolensky, 2012) solves this computational problem while still providing a good empirical fit to the major trends in the experimental data, revealing a clear quantitative cost of the con-

lines. A learner may acquire multiple languages over a lifetime, possibly including an artificial language learned in a laboratory experiment. Each distinct language has its own learned grammar, and in each case the grammar learning is governed by the type of prior we discuss in this paper. That prior, specified by a set of parameters, is in turn governed by a hyperprior over those parameters. The grammar-learning prior shifts as the languages it accounts for are acquired: learning English will adjust the prior expectation about word orders in languages generally. If the hyperprior is quite strong, however, as new languages are acquired, the prior will only shift slightly away from its initial state.

straints that are hypothesized, based on typological patterns, to be disfavored in our Bayesian prior. In this approach, the force of regularization is in some sense more external to the grammar, as it doesn't directly target constraint weights, but the probabilities that result from them. This is potentially a desirable result since a body of evidence suggest that the regularization bias is cognition-general rather than specific to the language faculty.

References

Anttila, A. (1997). Deriving variation from grammar. In F. Hinskens, R. van Hout, & L. Wetzels (Eds.), *Variation, change and phonological theory* (pp. 35–68). Philadelphia: John Benjamins.

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

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

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, *11*, 239–251.

Bod, R., Hay, J., & Jannedy, S. (Eds.) (2003). *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. Holland Academic Graphics/IFOTT.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*, 335 – 344.

Chen, S., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *Speech and Audio Processing, IEEE Transactions on*, *8*, 37–50.

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

Chomsky, N. (1980). On binding. *Linguistic Inquiry*, *11*, 1–46.

Chomsky, N. (1981a). *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, N. (1981b). Markedness and core grammar. In A. Belletti, L. Brandi, & L. Rizzi (Eds.), *Theory of markedness in core grammar* (pp. 123–146). Pisa: Scuola Normale Superiore di Pisa.

Chomsky, N. (1986). *Barriers*. Cambridge, MA: MIT Press.

Chomsky, N. (1998). *Minimalist inquiries: The framework*. Number 15 in MIT Working Papers in Linguistics. Cambridge, MA: MIT, Dept. of Linguistics.

Cinque, G. (1994). On the evidence for partial n-movement in the Romance DP. In G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi, & R. Zanuttini (Eds.), *Paths towards Universal Grammar: Studies in Honor of Richard S. Kayne* (pp. 85–110). Georgetown: Georgetown University Press.

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

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

Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, *6*, 310–329.

Culbertson, J., & Smolensky, P. (2012). A Bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, (pp. 1–31).

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

Curtin, S., & Zuraw, K. (2002). Explaining constraint demotion in a developing system. In *Proceedings of the Boston University conference on language development* (pp. 118–129). Citeseer volume 26.

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.

Dryer, M. (2012). Are word order correlations lineage-specific? Talk given at the 86th Annual Meeting of the Linguistic Society of America.

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.

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.

Finkel, J. R., & Manning, C. D. (2009). Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* NAACL '09 (pp. 602–610). Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Gold, E. M. (1967). Language identification in the limit. *Information Control*, *10*, 447–474.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the Workshop on Variation within Optimality Theory* (pp. 111–120). Stockholm: Stockholm University.

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.

Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic Inquiry*, *28*, 373–422.

Hayes, B., Siptar, P., Zuraw, K., & Londe, Z. (2009). Natural and unnatural constraints in hungarian vowel harmony. *Language*, *85*.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*, 379–440.

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

Jarosz, G. (2010). Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language*, *37*, 565–606.

Jesney, K., & Tessier, A. (2011). Biases in harmonic grammar: the road to restrictive learning. *Natural Language & Linguistic Theory*, *29*, 251–290.

Kayne, R. (1994). *The Antisymmetry of Syntax*. Cambridge, MA: MIT Press.

Laenzlinger, C. (2005). French adjective ordering: perspectives on dp-internal movement types. *Lingua*, *115*, 645–689.

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic grammar — a formal multi level connectionist theory of linguistic well formedness: Theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 388–395). Cambridge, MA.

Legendre, G., Raymond, W., & Smolensky, P. (1993). An optimality-theoretic typology of case and grammatical voice systems. In *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 464–478). University of California, Berkeley: Berkeley Linguistics Society volume 464.

McCarthy, J. (2002). *A thematic guide to Optimality Theory*. New York: Cambridge University Press.

McCarthy, J., & Prince, A. (1994). The emergence of the unmarked: Optimality in prosodic morphology. In M. Gonzàlez (Ed.), *NELS 24* (pp. 333–379). University of Massachussetts Amherst: GLSA volume 2.

Mitchell, T. (1980). *The need for biases in learning generalizations*. Technical Report CBM-TR-5-110 Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ. New Brunswick, NJ.

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

Moreton, E., & Pater, J. (to appear). Structure and substance in artificial-phonology learning. *Language and Linguistics Compass*, .

Nagy, N., & Heap, D. (1998). Francoprovençal null subjects and constraint interaction. In *CLS 34: The Panels* (pp. 151–166). Chicago: The Chicago Linguistic Society.

Newmeyer, F. J. (2010). Accounting for rare typological features in formal syntax: three strategies and some general remarks. In *Rethinking universals: how rarities affect linguistic theory* (pp. 195–222). New York: Mouton de Gruyter.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, *33*, 999–1035.

Pater, J. (2011). Emergent systemic simplicity (and complexity). *McGill Working Papers in Linguistics*, .

Philip, J. (to appear). (dis)harmony, the head-proximate filter, and linkers. *Journal of Linguistics*, .

Plank, F., & Filimonova, E. (2000). The universals archive. *Sprachtypologie und Universalienforschung*, *53*, 109–123.

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., & Tesar, B. (2004). Learning phonotactic distributions. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in phonological acquisition* (pp. 245–291). Cambridge University Press.

Pullum, G., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50.

Pycha, A., Nowak, P., Shin, E., & Shosted, R. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics* (pp. 101–114). Somerville, MA: Cascadilla Press volume 22.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna.

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

Samek-Lodovici, V. (2005). Prosody-syntax interaction in the expression of focus. *Natural Language and Linguistic Theory*, *23*, 687–755.

Sells, P. (2001). *Structure, Alignment, and Optimality in Swedish*. Stanford, CA: CSLA Publications.

Smith, D., & Eisner, J. (2007). Bootstrapping feature-rich dependency parsers with entropic priors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 667–677).

Smolensky, P. (1996). The initial state and 'richness of the base'in optimality theory. *Rutgers Optimality Archive*, *293*.

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.

Travis, L. (1984). *Parameters and Effects of Word Order Variation*. Ph.D. dissertation MIT.

Travis, L. (1989). Parameters of phrase structure. In M. R. Baltin, & A. S. Kroch (Eds.), *Alternative conceptions of phrase structure* (pp. 263–279). Chicago: University of Chicago Press.

Vogel, R. (2002). Free relative constructions in ot syntax. In G. Fanselow, & C. Féry (Eds.), *Resolving Conflicts in Grammars: Optimality Theory in Syntax, Morphology, and Phonology* (pp. 119–162). Hamburg: Helmut Buske Verlag.

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

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