LEARNING BIASES, REGULARIZATION, AND THE EMERGENCE OF TYPOLOGICAL UNIVERSALS IN SYNTAX

Submitted for the Robert J. Glushko Dissertation Prize *January, 2012*

by Jennifer Culbertson

Overview of the dissertation and its significance for the cognitive science of language

Understanding language from the perspective of cognitive science involves studying two things: the *representations* that underlie grammatical knowledge, and the nature and content of any *constraints* on these representations that impact how they are acquired and processed. In this dissertation, I tackle key questions that bear on these issues by investigating *recurrent typological patterns*—linguistic patterns which appear to be systematically favored (or disfavored) across languages—and the role that learning biases play in constraining these patterns, which are often called typological 'universals'.

The importance of typological universals lies in their potential to reveal deep properties of the human language faculty—however this depends on the claim that such patterns are in fact the result of *constraints on the cognitive system*. A link between typology and biases in the learner is thus both integral to generative linguistics and of critical interest to the cognitive science of language—however evidence for this connection remains minimal.

Combining new evidence generated with experimental, theoretical, and mathematical tools, I provide support for two types of biases, in the domain of syntax. The first type are *substantive biases*, which (dis)favor particular structures. The second type is a *formal regularization bias* favoring the minimization of variation. A central source of new experimental evidence I provide in the dissertation comes from the Mixture-Shift Paradigm for artificial language learning, based on Hudson Kam & Newport (2005), which I develop to answer the types of questions of interest in the dissertation.

The paradigm is motivated by the idea that typological patterns evolve from gradual changes effected by generations of learners. The method introduces learners to variable input in order to observe the extent to which they *alter the language to bring it in line with hypothesized biases*. If learners in these experiments show evidence of sensitivity to a hypothesized bias—a bias which can also explain the cross-linguistic distribution of a given pattern—then we will have convincing evidence that a typological universal is indeed the result of an underlying constraint on the cognitive system.

The goal of the dissertation is to complement more traditional sources of evidence from theoretical linguistics with new experimental and mathematical tools to answer the following:

(1) Do constraints on learning shape how languages change and therefore what typological universals emerge?
(2) What is the form and substance of those constraints?
(3) Can we see how ongoing change in a language is driven by learners and constrained by their biases?

Answering these questions *requires an interdisciplinary approach*, and accordingly, the dissertation research builds on previous work in linguistics (including language acquisition, creolization, theoretical syntax and morphology, the syntax-prosody interface, grammaticalization theory, typology), psycholinguistics, experimental psychology, and statistical modeling, and uses a number of methodologies familiar from these fields (e.g. corpus analysis, grammaticality judgments, prosodic analysis, artificial language learning, Bayesian modeling).

The dissertation is organized as follows. Chapter 2 reports the first use of the Mixture-Shift Paradigm, showing that adult learners exposed to variable noun, adjective, and numeral ordering patterns exhibit *regularization constrained by a substantive bias* in line with cross-linguistic typology. Chapter 3 develops a Bayesian model of the learning in the experiment. In chapter 4, I investigate the effect of feedback on learners,

and discuss broader issues concerning the nature of learning biases. Chapters 5 and 6 investigate a recurrent pathway of change involving grammaticalization of pronouns into agreement affixes. I provide evidence of this process in French, arguing that the newly-evolved agreement system follows proposed typological constraints. I then report an experiment using the Mixture-Shift Paradigm which reveals learners' sensitivity to these constraints—variable agreement is regularized only when the system conforms to attested patterns. This experiment offers further evidence that (formal and substantive) learning biases constrain linguistic change, resulting in recurrent typological patterns that provide insight into the language acquisition mechanism.

The existence of typological universals as an interdisciplinary question

The debate

How typological universals emerge from the extensive diversity found across the world's languages constitutes a central question for linguistics. One mainstream view is that these patterns arise largely because of (hard or soft) constraints on the grammars people can learn. These constraints may in principle be innate, learned, or emergent from processing; they may be domain-general constraints applied to language, or have no strict counterpart outside the domain of language.

The assumption that universal constraints on language learning strongly shape the space of human grammars has been a fundamental principle of generative linguistics; it has also not gone unchallenged. Proposed alternative explanations for typological universals include shaping by cognition-external functional pressures, genetic relationships between languages, and accidental geographic or cultural factors. Some of the most prominent debates in the field have centered around whether constraints on the cognitive system (either specific to language or domain general) are a major explanatory force (Chomsky 1965; Saffran et al 1996; Kirby 1999; Baker 2001; Newmeyer 2005; Bybee 2008; Evans & Levinson 2009; many others). Although a given typological pattern may result from the interaction of several factors, explanations based on learning biases are of particular interest to linguists and cognitive scientists, since they potentially shed light on underlying properties of the cognitive system (e.g. which structures or feature combinations are more costly, and why).

New sources of evidence

Traditionally, linguists have used cross-linguistic data along with theoretical analysis to argue for particular constraints on linguistic representations. However, longstanding disagreements concerning the existence of and explanation for statistically robust typological regularities suggest the need for new types of empirical evidence on the biases of learners and the extent to which these biases parallel typological tendencies. Recent research has used artificial language learning paradigms with adults to provide *direct behavioral evidence* for the existence of such biases. This work has focused mainly on laboratory learning of phonology and word segmentation (e.g. Saffran et al 1996; Newport & Aslin 2004; Wilson 2006; Finley & Badecker 2008), although a few studies have targeted typological patterns in morphology and syntax (Christiansen 2000; Hudson Kam & Newport 2005; St. Clair et al 2009).

A primary objective of this dissertation is therefore to contribute to this (as yet small) body of research; that is, to provide experimental evidence that learning biases parallel typological patterns. In focusing on the role of the learner, this work does *not* rely on the typical view that evidence to learners is impoverished, or that the learning mechanism is not powerful enough to glean necessary rules or constraints from the input. Rather I try to show that learning biases can help explain *why the input looks the way it does*—why certain patterns are common, while others are not found. The two universals I target concern (i) *word order in the nominal domain*, and (ii) *patterns of subject-verb agreement*. Both have been claimed to hold on the basis of cross-linguistic findings, but as mentioned above, additional evidence is needed to show that these recurrent typological patterns have as their underlying cause some psychologically real constraint on learning. That such evidence is needed is highlighted in the case of the word order universal, known as Greenberg's Universal 18.

Greenberg's Universal 18

First proposed by Greenberg (1963) in his seminal work on typology, Universal 18 concerns linear ordering of nouns with respect to numerals and adjectives; of the four possible ordering combinations in (1), Universal 18 bans pattern 4, which combines pre-nominal adjectives with post-nominal numerals. In Greenberg's sample of

30 languages, none were found to use this pattern, however according to a larger sample of languages, pattern 4 is in fact attested (Table 1).

(1) *Possible patterns of {Noun, Adjective}, {Noun, Numeral} ordering*:
1. Adj-N, Num-N 'harmonic'
2. N-Adj, N-Num 'harmonic'
3. N-Adj, Num-N 'unmarked'
4. Adj-N, N-Num 'marked'

	Noun-Adjective	Adjective-Noun
Noun-Numeral	443 (52%)	32 (4%)
Numeral-Noun	149 (17%)	227 (27%)

Table 1. Distribution of {Noun, Adjective}, {Noun, Numeral} orders from WALS (Dryer 2008a, 2008b).

What Table 1 shows is that the 'marked' pattern 4, is extremely rare compared to the other three patterns; only 4% of languages use it. In addition, it shows that most languages use the '*harmonic*' patterns 1 and 2. These are patterns which preserve the position of the noun as either always preceding, or always following. The cross-linguistic data therefore suggest the possibility of a universal which is quite complex ((2) summarizes), and as such is of particular interest. First, the universal is not absolute, but instead appears to be a strong tendency. This is precisely the type of universal which has been criticized by various researchers as not providing convincing evidence for cognitive constraints on the language system (e.g. Evans & Levinson 2009). Second, it combines a general preference for harmonic (over non-harmonic) patterns with a dispreference for a *particular* non-harmonic pattern. This makes it a possible sub-case of a more general word-order constraint which has been the subject of much recent work in theoretical syntax—namely the Final-Over-Final Constraint (Holmberg 2000; Biberauer et al 2008; Philip 2010).

(2) *Greenberg's Universal 18 reformulated*: Ranking of {Noun, Adjective}, {Noun, Numeral} ordering patterns in (1) according to the typology (where `x ≻ y' means `x is preferred to y'):

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

Providing evidence that learners' biases are parallel to this typological universal thus impacts the *content* of the constraints linguists posit as part of speakers' grammatical knowledge, and the *form* they take—whether they can be treated as absolute (hard) constraints or should be treated as (soft) biases which are real but nevertheless can be overcome. The latter is critical since many mainstream syntactic theories cannot straightforwardly instantiate soft constraints. In addition, although it has been claimed that a preference for harmonic patterns can be treated as a bias in the language *processing* system (e.g. Hawkins 1994), the preference for one *non-harmonic* pattern over another one is more difficult to explain in these terms and may therefore provide evidence for a grammar-internal constraint. In the next section, I will summarize the experimental method used in this dissertation, and the results obtained when applied to Universal 18.

The Mixture-Shift Paradigm, and a summary of results on word order

The artificial language learning experiments reported in this dissertation use a methodology first developed by Hudson Kam & Newport (2005) to provide evidence for the claim that when exposed to input containing unpredictable variability, learners tend to acquire more *regular* rules, increasing the consistency of the system (Sankoff & Laberge 1980; Singleton & Newport 2004; Sandler et al. 2005). *What is novel in this dissertation* is the use of this paradigm to simultaneously investigate the effects of *two interacting types of biases*, defined informally in (3).

(3) *Two types of biases:*

a. Regularization bias: acquire a grammar which minimizes variation present in the input

b. Substantive bias: acquire grammars which do not incorporate disfavored structures

These two types of biases are hypothesized to interact such that learners will only regularize variable patterns which satisfy substantive biases. In chapter 2, I use this paradigm to investigate whether the likelihood of regularization by learners parallels (2) above. Adult learners are exposed to a miniature artificial language with a variable pattern of {Noun, Adjective} and {Noun, Numeral} ordering. Each of four language conditions *tends toward* one of the patterns described above—i.e. uses that pattern the majority of the time. During testing, learners produce utterances, and the dependent measure of interest is whether they regularize the majority pattern in their training condition. Figure 1 illustrates the results.

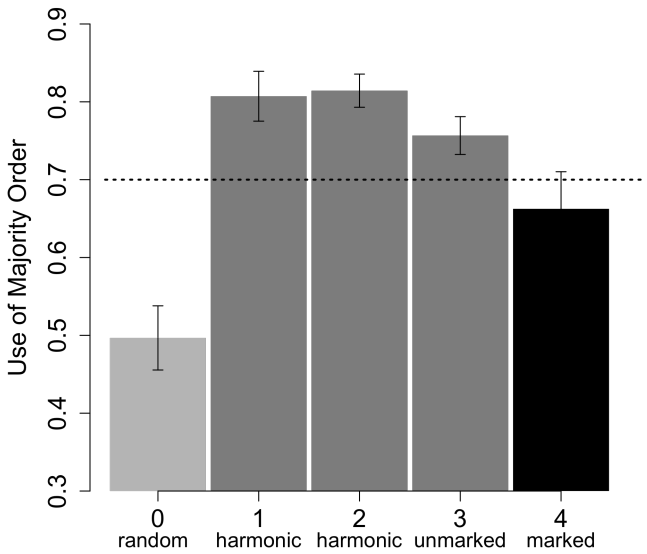

Figure 1. Average use of the majority order by learners in Experiment 1; dotted line shows proportion of time majority order was used in the input for each (non-random) condition.

Learners' behavior in this experiment precisely bears out the predictions made—learners regularize variable patterns that are favored by the proposed substantive bias (patterns 1, 2), but don't regularize the disfavored pattern 4. Regularization of pattern 3—not favored or disfavored according to (2)—falls in the middle. Learners' behavior in this task *and* the typological data in Table 1 are therefore explained if we posit a constraint on the cognitive system, active during learning, which encodes the preferences in Universal 18.

The role of feedback, and processing explanations of word order constraints

In chapter 4, I discuss a replication of Experiment 1 which manipulates the amount of feedback learners receive and shows that the critical results still hold (§4.3). Understanding how feedback affects the learning process is an issue of ongoing debate in language acquisition (e.g. Chomsky 1965; Marcus 1993; Hirsch-Pasek et al 1984; Strapp 1999) and probability learning research (e.g. the famous studies conducted by Estes 1976; Weir 1972). It is also of methodological interest since this paradigm has not been widely used and therefore the conditions under which certain results (e.g. regularization) obtain are not understood.

Chapter 4 (§4.5) also evaluates two opposing analyses of word order constraints—the first claims that such constraints result from biases in the processing system (Hawkins 1994, subsequent work), and the second claims that these constraints operate on linguistic representations in the grammar (Biberauer et al 2008; Holmberg 2000). I argue that although both approaches may account for the typological data, the experimental data is more amenable to analysis as a constraint on linguistic representations. Briefly, this is because proposed processing-related constraints necessarily operate on a structure larger than what learners in the experiment are exposed to; namely phrases including both an adjective *and* a numeral phrase (e.g. Numeral-Noun-Adjective). Learners in the experiment only heard two-word phrases containing *either* an adjective *or* a numeral. This suggests that the relevant constraints operate on higher-level representations in the grammar learners inferred rather than the phrases they were required to actively process.

Bayesian modeling of learning biases

The goal of using the Mixture-Shift Paradigm is to explore how biases affect learning. Data from this paradigm is thus particularly well suited to Bayesian probabilistic modeling since the framework assumes that learners combine *experience* and *prior biases*—probabilistic constraints on the hypothesis space—in order to make inferences, e.g. about what grammars are most likely to have generated the input. Such models provide a formal specification of hypothesized prior biases, and validate claims made about their influence on learning. In chapter 3, I propose a Bayesian model of learning in the experiment described above. The model explains the results through the interaction of two prior biases: one preferring more regular grammars, and the other favoring harmonic patterns, and disfavoring the non-harmonic pattern Adj-Noun, Noun-Num. This model not only successfully captures the experimental data, it reveals an intriguing pattern of individual learner behavior in each condition. Figure 2 illustrates the distribution of learning outcomes predicted by the model (after parameter fitting)—that is, how the grammars learners infer are predicted to shift compared to training *as a result of the interaction between the regularization and substantive biases.*

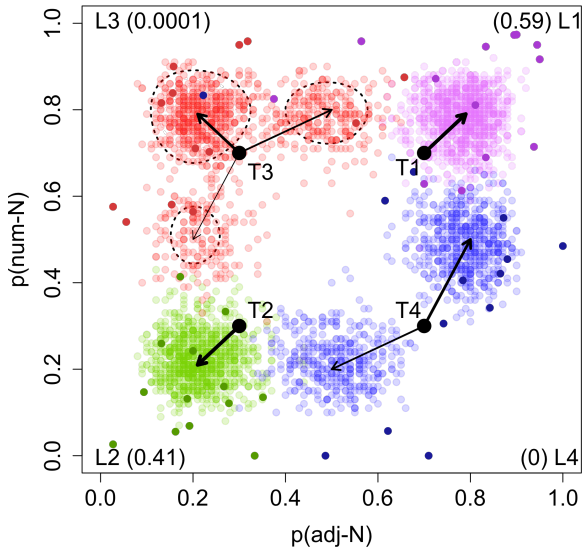

Figure 2. 2D plot of *grammar-space* (x-axis is probability of producing Adj-N, y-axis is probability of producing Num-N). Corners labeled L1, L2, L3, L4 correspond to deterministic versions of each pattern in (1). Black points labeled T1, T2, T3, T4 correspond to experiment training conditions. Opaque colored points are *actual* testing probabilities for learners in each condition. Transparent colored points are *predicted* probabilities according to the model. Arrows indicate *likely shifts made by learners*. This plot is an enhanced version of Figure 22 of the dissertation.

Language change and grammaticalization as a window in learners' biases

Language change is often considered a source of evidence for imperfect transmission of grammar from one generation to the next (Lightfoot 1999; Yang 2000; Niyogi 2006). An extension of this idea is that common types of changes reveal biases of learners—if a particular structure is commonly altered, this might indicate that it is difficult to learn. Thus not only typological distributions across *synchronic* grammars, but pathways of *diachronic* grammar change serve as a window into learners' biases. The study of *grammaticalization*—the process by which lexical items become grammatical items—documents and attempts to explain these common pathways of change (Hopper & Traugott 1993; Bybee 2006).

One such pathway, called a grammaticalization cline, illustrates how new agreement affixes evolves from pronominals. The cline in (4) has been documented in many languages, and thus is a potential source of evidence for how acquisition preferences continuously shape language.

(4) *Pronoun → Agreement cline:*

Independent pronoun → weak (clitic) pronoun → agreement affix

The endpoints on this cline are relatively clear, but intermediate points, like the clitic stage, are

considerably more turbid. The dissertation work targets Modern French subject clitics, whose status has attracted substantial debate since the 1970's (e.g. Kayne 1975; Auger 1994; de Cat 2007); on the cline in (4), these elements seem to lie somewhere between weak pronouns and agreement affixes, and thus one view is that it is in principle impossible to categorize them. I argue that this is not the case, using *converging evidence* from morphophonology, quantitative corpus analysis, grammaticality judgments, prosodic analysis, and language acquisition to produce a clear answer.

Bringing together sources of evidence from diverse subfields in linguistics is crucial in characterizing these elements, and revealing sources of information present in the input to learners. A full picture of this input makes it possible to form testable hypotheses about which features might impact how learners analyze these elements, and how these features might interact with learning biases.

The Mixture-Shift Paradigm and the evolution of new agreement systems

One distinctive property of Colloquial French subject clitics, which suggests that they are in fact agreement markers, is their participation in so-called 'clitic doubling' constructions. These constructions involve a lexical subject co-occuring with a subject clitic (e.g. *le garçon il parle*, 'The boy is talking'; where *il* is the subject clitic). On the surface, clitic doubling resembles subject-verb agreement, and in fact I show that it follows typological constraints on agreement—in particular, the implicational hierarchy in (5) which has been proposed to govern agreement systems (Croft 2000; Siewierska 2004). The hierarchy states that if agreement is triggered by a subject type on the right, it must also be triggered by all types to the left (e.g. if a language has agreement triggered by definite subjects, it must also have agreement triggered by pronoun subjects). Clitic doubling in Colloquial French obeys this hierarchy—agreement is triggered by definite noun and pronoun subjects, but no other subject types.

(5) *Implicational hierarchy of definiteness governing agreement systems:*

pronoun → definite noun → indefinite noun → *wh*-phrase (e.g. *who*, *what*)

This hierarchy also governs how agreement systems change, therefore it *makes a clear prediction* about how the new system of agreement in Colloquial French might continue to evolve; following along the hierarchy, agreement should be extended next to indefinites. If, as I have argued, language change is constrained by biases internal to learners, then we expect acquisition to be sensitive to the implicational hierarchy in (5). The hypothesis, as with Universal 18, is that this sensitivity interacts with the hypothesized regularization bias. Learners of Colloquial French are in fact exposed to a variable system of agreement; not only does agreement only occur with certain subject types, agreement with definites does not always surface—it is *not obligatory*. Thus, we predict learners of such a language will acquire a system of agreement that is more regular (less variable) than the input.

In chapter 6, I show that in the case of French, *corpus evidence* suggests that new generations of speakers are driving the change from pronoun to agreement affix. However, direct experimental evidence would significantly strengthen this claim. I therefore investigate whether the predictions made above are borne out *by using the Mixture-Shift Paradigm* to expose adult learners to systems of variable agreement modeled after what children acquiring Colloquial French might hear. Learners are trained on a language in which agreement does not occur with all subject types, and in contexts where it *can* occur, does not do so obligatorily. This manipulation makes it possible to see whether learners tend to regularize the type of variable agreement which characterizes the evolving systems documented in the grammaticalization literature and in the case study of French reported in chapter 5.

In order to test learners' sensitivity to the proposed substantive bias—the implicational hierarchy in (5)—the experiment has two language conditions, shown in Table 2. The first condition most closely resembles Colloquial French, and is allowed by the hierarchy, while the second is precisely the type of language *banned* by the hierarchy.

	Agreement (optional)	*No agreement*
NATURAL CONDITION	definite subjects	indefinite subjects
UNNATURAL CONDITION	indefinite subjects	definite subjects

Table 2. Conditions in Experiment 3.

The results of this experiment are shown in Figure 3. As in the experiments on Universal 18, learners show evidence of a *regularization bias which interacts with a hypothesized substantive bias*. Learners in both conditions extend agreement to new subject types. Critically however, learners in the natural condition tend to regularize agreement with definites, but learners in the unnatural condition *do not* regularize agreement with indefinites. The results suggest that learners' regularization bias can drive systems of agreement to progress in a way which is constrained by the implicational hierarchy.

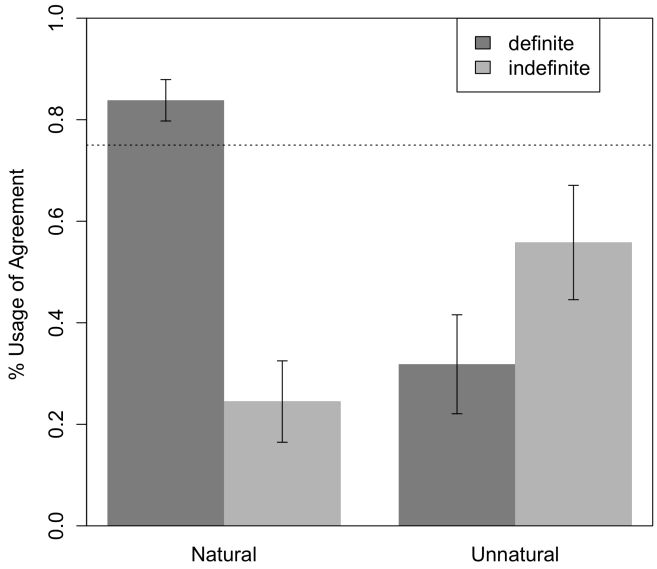

Figure 3. Average use of agreement with definite and indefinite subjects in Experiment 3; dotted line shows proportion agreement used in the input with the subject type that triggered it for each condition (definites for natural condition, indefinites for unnatural condition).

Conclusion

This dissertation brings together evidence from linguistic theory, laboratory learning of language, and mathematical modeling to support a central role for the learner in constraining language change and typology. The idea that typological regularities stem from learning biases is a foundational claim of generative linguistics which has recently drawn skepticism from the cognitive science community. Showing that such biases—that is, *biases in the cognitive systems of individual learners*—in fact exist and can explain well-documented typological regularities is the main contribution of the dissertation. The two universals targeted are Greenberg's Universal 18—a constraint on word order in the nominal domain—and an implicational hierarchy of definiteness governing agreement. These serve as test cases for the Mixture-Shift Paradigm, an artificial language learning paradigm I develop, showing that it can be productively applied to a range of typological patterns, providing a new source of evidence for underlying biases.

In chapters 2–4, I show that two interacting learning biases—a regularization bias favoring minimization of variation, and a substantive bias parallel to Universal 18—offer an explanation for how learners alter languages in the laboratory *and* how such shifts might have resulted in the typological asymmetries documented by Greenberg. The conclusions I argue for are based on results using the Mixture-Shift Paradigm, strengthened by state-of-the-art inferential statistics, including Bayesian modeling (ch. 3), and informed by current linguistic theory.

Chapter 5 introduces, using an in-depth case study of ongoing language change in French, the second typological pattern targeted in the dissertation. I provide converging evidence from, quantitative corpus analysis, grammaticality judgments, prosody, and typology to argue that a new series of agreement markers has developed in the Colloquial register, in part driven by new generations of learners. In chapter 6, I test specific hypotheses about how learning biases both drive and constrain the evolution of new agreement systems. Again using the Mixture-Shift Paradigm, I show that learners tend to regularize patterns of variable agreement

(modeled after what French-learning children hear), but only when such patterns are predicted possible by the implicational hierarchy proposed to govern agreement systems.

Although I have provided some preliminary evidence along the way, the nature and origin of substantive constraints on learning remain to be fully understood, and present an important area for future research. However, I hope to have shown how innovative experimental methods, and multiple sources of evidence, can help to advance important debates in cognitive science—in this case whether typological patterns are the result of constraints on the cognitive system.