Modeling Productivity with the Gradual Learning Algorithm: The Problem of Accidentally Exceptionless Generalizations

Adam Albright
Massachusetts Institute of Technology

Bruce Hayes
University of California, Los Angeles

Prepared for *Gradience in Grammar*, edited by Gisbert Fanselow, Caroline Féry, Matthias Schlesewsky, and Ralf Vogel; Oxford University Press

1. Introduction

Many cases of gradient intuitions reflect conflicting patterns in the data to which the child is exposed during language acquisition. An area in which the learner almost always faces conflicting data is inflectional morphology, where different words in the lexicon often follow different patterns. Thus, for English past tenses, we have *wing* ~ *winged* (the most common pattern in the language) *wring* ~ *wrung* (a widespread [ɪ] ~ [ʌ] pattern), and *sing* ~ *sang* (a less common [ɪ] ~ [æ] pattern). The irreconcilable conflict between these patterns leads English speakers to have ambivalent, gradient intuitions when asked to provide the past tense of a novel verb that fits all of them, such as *spling* (Bybee and Moder 1983; Prasada and Pinker 1993; Albright and Hayes 2003).

In order to get a more precise means of investigating this kind of gradience, we have over the past few years developed an implemented formal model for the acquisition of inflectional paradigms. An earlier version of our model is described in Albright and Hayes (2002), and its application to various empirical problems is laid out in Albright, Andrade, and Hayes (2001), Albright (2002), and Albright and Hayes (2003). Our model abstracts morphological and phonological generalizations from representative learning data and uses them to construct a stochastic grammar that can generate multiple forms for novel stems like *spling*. The model is tested by comparing its "intuitions," which are usually gradient, against human judgments for the same forms.

In modeling gradient productivity of morphological processes, we have focused on the *reliability* of the generalizations: how much of the input data do they cover, and how many exceptions do they involve? In general, greater productivity is found to be correlated with greater reliability, while generalizations covering few forms or entailing many exceptions are relatively unproductive. In this article, we address a puzzling challenge to this way of evaluating

generalizations: the existence of generalizations that are exceptionless and well-instantiated, but are nonetheless either completely invalid, or else not nearly as valid as exceptionlessness would imply (i.e. they give rise to gradient intuitions). We offer a solution for one class of these problems, based on the Gradual Learning Algorithm (Boersma 1997, Boersma and Hayes 2001). We then discuss other cases in which this solution does not work satisfactorily, and we present some of our tentative efforts to find a more general solution.

2. Navajo Sibilant Harmony

The problem of exceptionless but unproductive generalizations arose in our efforts to extend our model to learn non-local rule environments. The first example we discuss comes from Sibilant Harmony in Navajo, a gradient process described in Sapir and Hoijer (1967).

Sibilant harmony can be illustrated by examining the allomorphs of the *s*-perfective prefix. This prefix is realized as shown in (1) (examples from Sapir and Hoijer):

(1)	a.	[šì-]	if the *first segment* of the stem to which it is attached is a [–anterior] sibilant ([č, č', č^{h}, š, ž]), for example in [**šì-čì**d] 'he is stooping over'
	b.	[šì-] or [sì-]	if *somewhere later in the stem* is a [–anterior] sibilant, as in [**šì-téé**ž] ~ [**sì-téé**ž] 'they two are lying' (free variation)
	c.	[sì-]	otherwise, as in [**sì-**fí] 'he is lying'

A fully realistic simulation of the acquisition of Navajo sibilant harmony would require a large corpus of Navajo verb stems along with their *s*-perfectives. Lacking such a corpus, we performed idealized simulations using an artificial language modeled on Navajo: we selected whole Navajo words (rather than stems) at random from the electronic version of Young, Morgan, and Midgette's dictionary (1992), and constructed *s*-perfective forms for them by attaching [šì-] or [sì-] according to the pattern described in (1).

3. How Our Learning Model Works

Our learning system employs some basic assumptions about representations and rule schemata. We assume that words are represented as sequences of phonemes, each consisting of a bundle of distinctive feature specifications, as in Chomsky and Halle (1968). Rules and constraints employ feature matrices that describe natural classes, as well as variables permitting the expression of non-local environments: ([+F]) designates a single skippable segment of the type [+F], while ([+F])* designates any number of skippable [+F] segments. Thus, the environment in (2):

(2) / ___ ([+seg])* [–anterior]

can be read "where a non-anterior segment follows somewhere later in the word" ([+seg] denotes the entire class of segments).

Our learner is given a list of pairs, consisting of bases and inflected forms. For the synthetic version of Navajo we used, such a list would be partially represented by (3):

(3) a. [bàːʔ] [sì-bàːʔ]
b. [č'ìɬ] [šì-č'ìɬ]
c. [čʰòːjìn] [šì-čʰòːjìn]
d. [gàn] [sì-gàn]
e. [k'àz] [sì-k'àz]
f. [kéšgə̀ːˈ] [šì-kéšgə̀ːˈ], [sì-kéšgə̀ːˈ]
g. [síːʔ] [sì-síːʔ]
h. [tə̀š] [šì-tə̀š], [sì-tə̀š]
i. [tə́̃] [sì-tə́̃]
j. [t͡ɬéːž] [šì-t͡ɬéːž], [sì-t͡ɬéːž]

Where free variation occurs, the learner is provided with one copy of each variant; thus, for (3f) both [kéšgə̀ːˈ] ~ [šì-kéšgə̀ːˈ] and [kéšgə̀ːˈ] ~ [sì-kéšgə̀ːˈ] are provided.

The goal of learning is to determine which environments require [sì-], which require [šì-], and which allow both. The learning process involves generalizing bottom-up from the lexicon, using a procedure described below. Generalization creates a large number of candidate environments; an evaluation metric is later employed to select which environments to keep in the final grammar.

Learning begins by parsing the forms into their component morphemes and grouping them by the morphological change they involve. The data in (3) exhibit two changes, as shown in (4); the box surrounds cases of free variation.

(4)

I. Prefix [sì-]		II. Prefix [šì-]	
a. [bàːʔ]	[sì-bàːʔ]	a. [čʰòːjìn]	[šì-čʰòːjìn]
b. [gàn]	[sì-gàn]	b. [č'ìɬ]	[šì-č'ìɬ]
c. [kéšgə̀ːˈ]	[sì-kéšgə̀ːˈ]	c. [kéšgə̀ːˈ]	[šì-kéšgə̀ːˈ]
d. [t͡ɬéːž]	[sì-t͡ɬéːž]	d. [t͡ɬéːž]	[šì-t͡ɬéːž]
e. [tə̀š]	[sì-tə̀š]	e. [tə̀š]	[šì-tə̀š]
f. [k'àz]	[sì-k'àz]		
g. [síːʔ]	[sì-síːʔ]		
h. [tə́̃]	[sì-tə́̃]		

For each change, the system creates hypotheses about which elements in the environment are necessary to condition the change. To do this, it begins by treating each pair as a "word-specific rule," by separating out the changing part from the invariant part. Thus, the first three [ši-] forms in (4) would be construed as in (5):

(5) a. ∅ → ši / [___ čʰòːjìn]
b. ∅ → ši / [___ čʼìɬ]
c. ∅ → ši / [___ kéšgə̀ː]

Next, the system compares pairs of rules that have the same change (e.g., both attach [ši-]), and extracts what their environments have in common to form a generalized rule. Thus, given the two word-specific rules in (6a), the system collapses them together using features, as in (6b).

(6)a. ∅ → ši / [___ tə̀š]
∅ → ši / [___ t͡ɬéːž]

b.

	∅	→	ši	/	[	___	t	ə̀	š	]
+	∅	→	ši	/	[	___	t͡ɬ	éː	ž	]
=	∅	→	ši	/	[	___	[–sonorant, –continuant, –spread gl., +anterior]	[+syllabic, –high, –round]	[–sonorant, +continuant, –anterior, +strident]	]

In this particular case, the two forms being compared are quite similar, so determining which segment should be compared with which is unproblematic. But for forms of different lengths, such as [čʰòːjìn] and [čʼìɬ] above, this is a harder question.[1] We adopt an approach that lines up the segments that are **most similar** to each another. For instance, (7) gives an intuitively good alignment for [čʰòːjìn] and [čʼìɬ]

(7)

	čʰ	òː	j	ì	n
	|			|	|
	čʼ			ì	ɬ

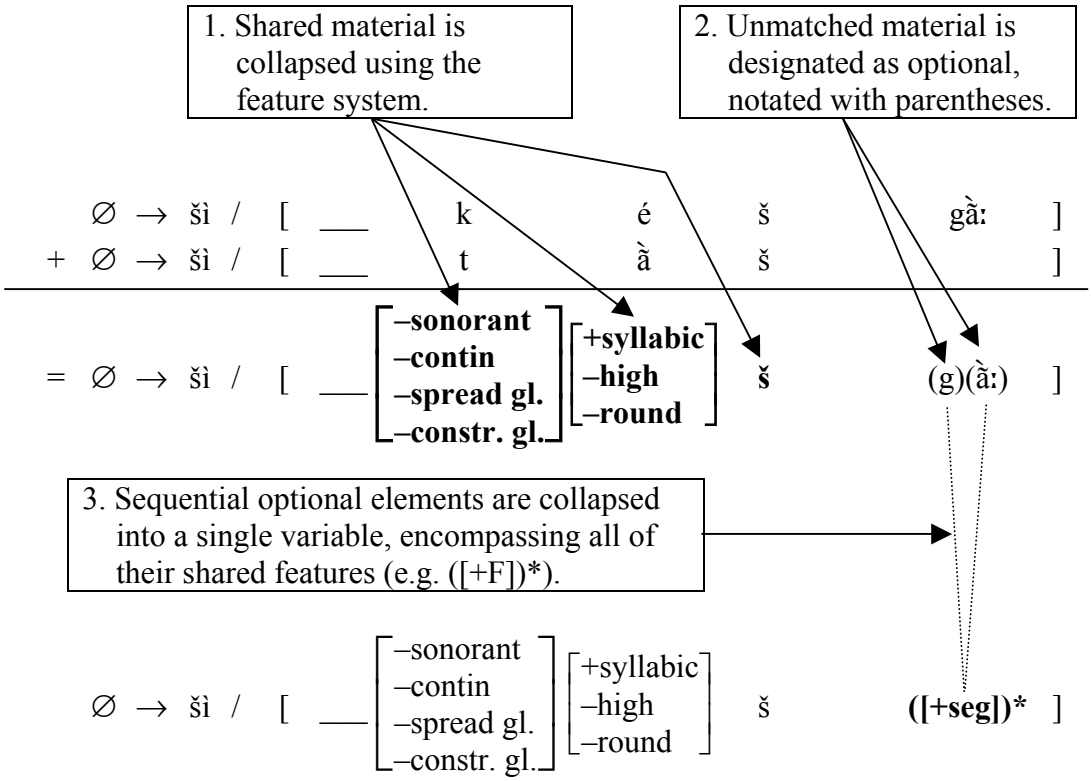
Good alignments have two properties: they match phonetically similar segments like [čʰ] and [čʼ], and they avoid leaving too many segments unmatched. To evaluate the similarity of segments, we employ the similarity metric from Frisch, Pierrehumbert and Broe (2004). To guarantee an optimal pairing, we use a cost-minimizing string alignment algorithm (described in Kruskal 1999) that efficiently searches all possible alignments for best total similarity.

[1] The issue did not arise in an earlier version of our model (Albright and Hayes 2002), which did not aspire to learn non-local environments, and thus could use simple edge-in alignment.

The rationale for using similarity-based alignment is that phonological environments are based on natural classes, and the members of a natural class are phonetically similar.

Seen in detail, the process of collapsing two rules into one is based on three principles, illustrated in (8) with the collapsing of the rules ∅ → ši / [___ kéšgə̃ː] and ∅ → ši / [___ tə̃š].

(8)



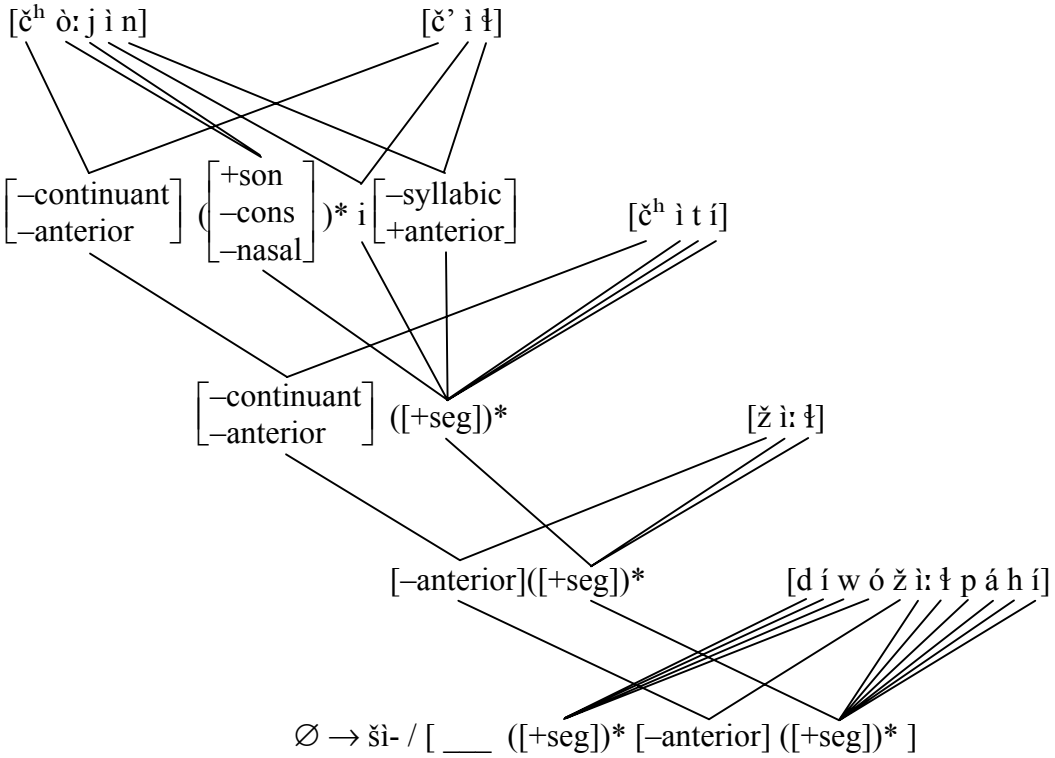
Paired feature matrices are collapsed by constructing a new matrix that contains all of their shared features (see step 1). Next, any material in one rule that is unmatched to the other is designated as optional, represented by parentheses (step 2). Finally, sequences of consecutive optional elements are collapsed together into a single expression of the form ($\mathcal{F}$)*; that is, any number of feature matrices $\mathcal{F}$, where $\mathcal{F}$ is the smallest natural class containing all of the collapsed optional elements (step 3).

The process is iterated, generalizing the new rule with the other words in the learning data; the resulting rules are further generalized, and so on. Due to memory limitations, it is necessary periodically to trim back the hypothesis set, keeping only those rules that perform best.[2] Generalization terminates when no new "keeper" rules are found.

[2] Specifically: (a) for each word in the training set, we keep the most reliable rule (in the sense of Albright and Hayes 2002) that derives it; (b) for each change, we keep the rule that derives more forms than any other.

We find that this procedure, applied to a representative set of words, discovers the environment of non-local sibilant harmony after only a few steps. One path to the correct environment is shown in (9):

(9)



The result can be read: "Prefix [ši-] to any stem that consists of any number of segments followed by a nonanterior segment, followed by any number of segments". (Note that in our feature system, [–anterior] segments in Navajo are necessarily sibilant.) In more standard notation, one could replace ([+seg])* with a free variable X, and follow the standard assumption that non-adjacency to the distal word edge need not be specified; thus the rule would appear as in (10):

(10) ∅ → ši- / ___ X [–anterior]

We emphasize that at this stage, the system is only generating hypotheses; the task of using these hypotheses to construct the final grammar taken up in section 5.

4. Testing the Approach: A Simulation

We will now show that, given representative learning data, the system just described can discover the rule environments needed for Navajo sibilant harmony. As noted above, our learning simulation involved artificial Navajo *s*-perfectives, created by attaching appropriate prefix allomorphs to whole Navajo words (as opposed to stems). Selecting 200 words at

random,[3] we followed Sapir and Hoijer's characterization of the harmony pattern, attaching prefixes to the bases as follows: (a) if the base began with a nonanterior sibilant, we prefixed [šì-] (there were 19 of these in the learning set); (b) if the base contained but did not begin with a nonanterior sibilant, we made two copies, one prefixed with [šì-], the other with [sì-] (37 of each); (c) the remaining 144 bases contained no nonanterior sibilant, and we prefixed [sì-] to each.

Running the algorithm just described, we found that among the 92 environments it learned were three of particular interest: the environment for obligatory local harmony ((11a)); the environment that licenses distal harmony ((11b); note that this includes local harmony as a special case); and the vacuous "environment" specifying the default allomorph [sì-] ((11c)).

(11) a. Obligatory local harmony

$\varnothing \rightarrow$ [šì-] / ___ [–anterior]

b. Optional distal harmony (= (10))

$\varnothing \rightarrow$ [šì-] / ___ X [–anterior]

c. Default [sì-]

$\varnothing \rightarrow$ [sì-] / ___ X

The remaining 89 environments are discussed below.

5. Forming a Grammar

These environments can be incorporated into an effective grammar by treating them not as rules, as just given, but rather as Optimality-theoretic constraints of morphology (Boersma 1998, Russell 1999, Burzio 2002, MacBride 2004). In this approach, rule (11a) is reconstrued as a constraint "Use [šì-] / ___ [–anterior] to form the *s*-perfective". This constraint is violated by forms that begin with a [–anterior] segment, but use something other than [šì-] to form the *s*-perfective. The basic idea is illustrated below:

(12)

Morphological Base	**Candidates that obey USE [šì-] / ___ [–anterior]**	**Candidates that violate USE [šì-] / ___ [–anterior]**
[šáp]	[šì-šáp]	*[sì-šáp], *[mù-šáp], etc.
[táp]	all	none

It is straightforward to rank these constraints in a way that yields the target pattern, as (13) and (14) show:

[3] We repeated our simulation, obtaining similar results, with ten 200-word samples, but report only one of them here.

(13) USE [ši-] / ___ [–ant] >> { USE [sì-] / ___ X , USE [ši-] / ___ X [–ant] } >> all others

ranked in free variation

(14)a.

/ sì-čìd /	USE [ši-] / ___ [–ant]	USE [ši-] / ___ X [–ant]	USE [sì-] / ___ X
☞ ši-čìd			*
* sì-čìd	*!	*	

b.

/ sì-téːž/	USE [ši-] / ___ [–ant]	USE [ši-] / ___ X [–ant]	USE [sì-] / ___ X
☞ ši-téːž			*
☞ sì-téːž		*!	

For (14b), the free ranking of USE [ši-] / ___ X [–ant] and USE [sì-] / ___ X will result in multiple winners generated in free variation (Anttila 1997).

6. Unwanted Constraints

The 89 constraints not discussed so far consist largely of complicated generalizations that happen to hold true of the learning data. One example is the constraint in (15):

(15) USE sì- / ___ ([–round])* $\begin{bmatrix}+\text{anterior}\\+\text{continuant}\end{bmatrix}$ ([–consonantal])*]

As it happens, this constraint works for all 37 forms in the learning data to which it applies.

Such constraints make profoundly incorrect predictions for forms outside the learning data, such as hypothetical /čálá/:

(16) USE sì- / ___ ([–round])* $\begin{bmatrix}+\text{anterior}\\+\text{continuant}\end{bmatrix}$ ([–consonantal])*]

sì- č á l á

If ranked high enough, this constraint will have the detrimental effect of preventing [ši-čálá] from being generated consistently. We will refer to such inappropriate generalizations as "junk" constraints.

One possible response to the problem is to say that the learning method is simply being too liberal, allowing too many generalizations to be projected from the learning data. We acknowledge this as a possibility, and we have experimented with various ways to force the algorithm to stick to more sensible generalizations. Yet we are attracted to the idea that constraint learning could be simplified—and rely on fewer a priori assumptions—by letting constraints be generated rather freely and excluding the bad ones by having an effective

evaluation metric. Below, we lay out such a metric, which makes use of the Gradual Learning Algorithm.

7. The Gradual Learning Algorithm

The Gradual Learning Algorithm (GLA; Boersma 1997, Boersma and Hayes 2001) can rank constraints in a way that derives free variation and matches the frequencies of the learning data; thus it is suited to an attack on the present problem. The GLA assumes a stochastic version of Optimality Theory, whereby each pair of constraints {A, B} is assigned not a strict ranking, but rather a probability: "A dominates B with probability P". Thus, the free ranking given in (13) above would involve assigning to the constraints USE [sì-] / ___ X and USE [šì-] / ___ X [–ant] a 50-50 ranking probability.

Any such theory needs a method to ensure that the pairwise probabilities assigned to the constraints are mutually consistent. In the GLA, this is done by arranging the constraints along a numerical scale, each constraint taking a **ranking value**. On any particular occasion that the grammar is used, a **selection point** is adopted for each constraint, taken from a Gaussian probability distribution with a standard deviation fixed for all constraints. The constraints are sorted by their selection points, and the winning candidate is then determined on the basis of this ranking. In this scheme, pairwise ranking probabilities are determined by the ranking values,[4] and are guaranteed to be mutually consistent.

8. The Need for Generality

Let us now consider the application of the GLA to Navajo. Naively, one might hope that when the constraints are submitted to the GLA for ranking, the junk will settle to the bottom. However, what one actually finds is that the junk constraints get ranked high. Although USE [šì-] / ___ [–ant] does indeed get ranked on top, the crucial constraints USE [šì-] / ___ X [–ant] and USE [sì-] / ___ X end up swamped by higher-ranking junk constraints, and thus rendered largely ineffective. The result is a grammar that performs quite well on the data that trained it (generally producing something close to the right output frequencies for every stem), but fails grossly in generating novel forms. The frequencies generated for novel forms are determined by the number of high ranking junk constraints that happen to fit them, and do not respect the distribution in (11).

The problem at hand is a classic one in inductive learning theory. If a learning algorithm excessively tailors its behavior to the particular forms to which it is exposed, it will learn a patchwork of small generalizations that collectively cover the learning data. This does not suffice to cover new forms, which, after all, is the main purpose of having a grammar in the first place!

Why did the GLA fail here? The reason is that it demotes constraints only when they prefer losing candidates. But within the learning data, our junk constraints generally prefer only

[4] A spreadsheet giving the function that maps ranking value differences to pairwise probabilities is posted at http://www.linguistics.ucla.edu/people/hayes/GLA/.

winners—that is precisely why they emerged from the inductive generalization phase of learning. Accidentally true generalizations thus defeat the GLA as it currently stands. It would seem, then, that what is needed is a way for the GLA to distinguish accidentally true generalizations from linguistically significant generalizations.

9. Initial Rankings Based on Generality

Boersma (1998) suggested that for morphology, initial rankings should be based on generality—the more general the constraint, the higher it is ranked before learning takes place. It turns out that this insight is the key to solving the Navajo problem. What is needed, though, is a way to characterize generality in numerical terms. There are various approaches that could be taken; for example, using the symbol-counting evaluation metric in Chomsky and Halle (1968) (fewer symbols = greater generality). Here, we adopt an *empirical* criterion: a rule or morphological constraint is maximally general if it can be held responsible for all of the forms that exhibit its structural change. We use the fraction in (17):

(17) $$\frac{\text{number of forms that a constraint applies to}}{\text{total number of forms exhibiting the change that the constraint requires}}$$

In the 200-word Navajo simulation discussed above, some representative generality values are given below:

(18)

Constraint	**Relevant forms**	**Forms with this change**	**Generality**
USE [šì-] / ___ [–anterior]	19	56 [šì-] forms	.339
USE [šì-] / ___ X [–anterior]	56		1
USE [sì-] / ___ X	181	181 [sì-] forms	1
Constraint (15) (“junk” constraint)	37		.204

The idea, then, is to assign the constraints initial ranking values that reflect their generality, with more general constraints on top. If the scheme works, we should find that all the data will be explained by the most general applicable constraints, and the others will remain low in the grammar and hence never play a role in deriving output forms.

In order to ensure that differences in initial rankings are large enough to make a difference, the generality values from (17) were rescaled to cover a huge probability range, using the formula in (19):

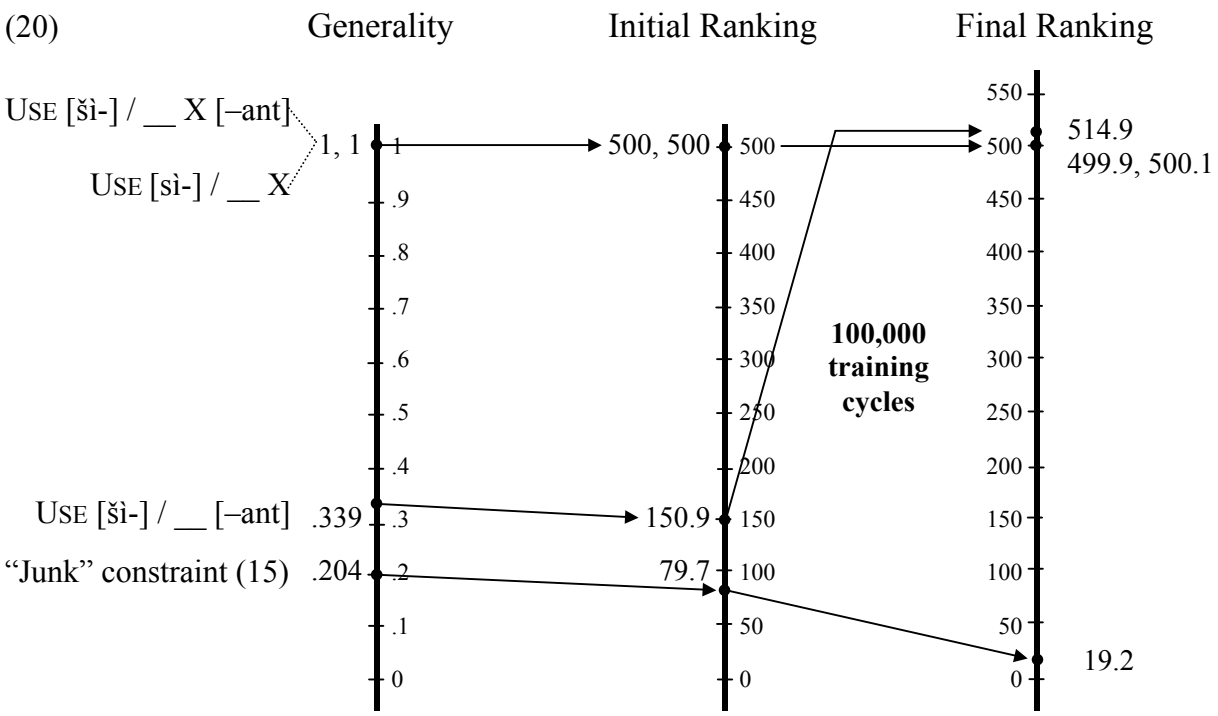
(19) For each constraint c, initial ranking value$_c$ = $500 \times \frac{\text{Generality}_c - \text{Generality}_{min}}{\text{Generality}_{max} - \text{Generality}_{min}}$

where Generality_{min} is the generality of the least general constraint in the system, and Generality_{max} is the generality of the most general constraint.

10. Employing Generality in a Learning Simulation

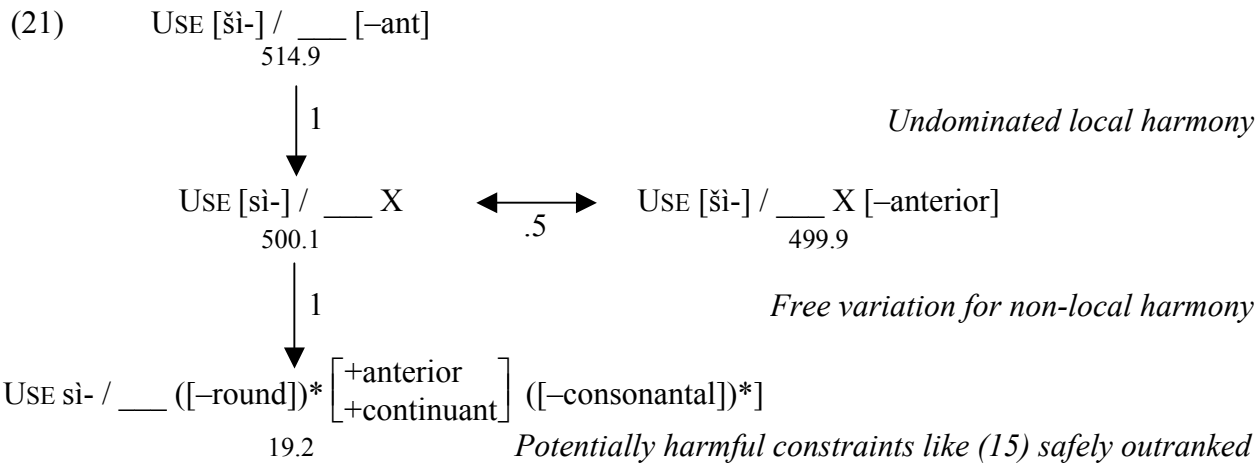
We implemented the scheme just described and ran it (multiple times, to check for consistency) on the Navajo pseudodata described above. For one representative run, it caused the relevant constraints (including here just one representative "junk" constraint (15)), to be ranked as follows:

(20)



The grammar thus learned can be depicted schematically as in (21), where the arrows show the probabilities that one constraint will outrank the other. When the difference in ranking value exceeds about 10, the probability that the ranking will hold is essentially 1 (strict ranking).

(21)



This approach yields the desired grammar: all of the junk constraints (not just (15)) are ranked safely below the top three.

The procedure works because the GLA is error-driven. Thus, junk constraints not only start low, but they stay there, since the general constraint that does the same work has a head start and averts any errors that would promote the junk constraints. Good constraints with specific contexts, on the other hand, like "USE [šì-] / ___ [–ant]", are also nongeneral—but appropriately so. They start low, but they are crucial in averting errors like *[sì-šáp], and thus they are soon promoted by the GLA to the top of the grammar.

We find, then, that a preference for more general statements in grammar induction is not merely an aesthetic bias; it is, in fact, a necessary criterion in distinguishing plausible hypotheses from those which are implausible, but coincidentally hold true in the learning sample.

11. The Realism of the Simulation

In this section we address two possible objections to our model.

11.1 Phonological Rules vs. Allomorph Distribution

Navajo sibilant harmony is typically described as a phonological process, spreading a [–anterior] feature value from right to left within a certain domain. The grammar learned by our model, on the other hand, treats harmony as allomorphy ([sì-] vs. [šì-]), and cannot capture the effect of harmony root-internally. Thus, it may be objected that the model has missed the essential nature of harmony.

In this connection, we note first that harmony processes are often observed primarily through affix allomorphy—either because there is no corresponding root-internal restriction, or because harmony effects are weaker within roots, admitting greater exceptionality. For these cases, an allomorphy analysis may be the only appropriate analysis. For arguments that root-internal and affixal harmony often require separate analyses, see Kiparsky (1968).

More generally, however, the question remains as to how to unify knowledge about allomorphy with knowledge about root-internal phonotactics. Even when affixes and roots show exactly the same harmony patterns, we believe that understanding the distribution of affix allomorphs could constitute an important first step in learning the more general harmony process. Allomorphic alternations provide positive evidence that is frequently lacking for "static" (root-internal) restrictions. What is needed, therefore, is some way of bootstrapping from the constraints on particular morphemes to more general constraints on the distribution of speech sounds. We leave this as a problem for future work.

11.2 Should arbitrary constraints be generated at all?

Another possible objection is that if we had had a more constrained method for hypothesizing constraints, it would never have posited constraints like (15) in the first place. Indeed, if all constraints come from Universal Grammar (that is, are innate), the need to trim back absurd ones would never arise. Against this objection can be cited work from the phonological literature suggesting that environments sometimes really are complex and arbitrary from a synchronic point of view (Bach and Harms 1972; Hale and Reiss 1998; Hayes 1999; Blevins, in press). For instance, in examining patterns of English past tenses, we found that all English verbs ending in voiceless fricatives are regular, and that native speakers are tacitly aware of this generalization (Albright and Hayes 2003). It seems likely that any model powerful enough to handle the full range of attested phonological patterns will need some mechanism to sift through large numbers of possibly irrelevant hypotheses.

12. Analytic Discussion

While the Navajo simulation offers a degree of realism in the complexity of the constraints learned, hand analysis of simpler cases helps in understanding why the simulation came out as it did, and gives greater confidence that the result is a general one.

To this end, we reduce Navajo to three constraints, renamed more generally as follows: (1) USE [sì-], which we will call DEFAULT, (2) the special-context USE [šì-] / ___ X [–ant], which we will call CONTEXTUAL [šì-], and (3) the accidentally-exceptionless (15), which we will call ACCIDENTAL [sì-]. ACCIDENTAL [sì-] is exceptionless because the relevant forms in the training data happen not to contain non-anterior sibilants.

Suppose first that all harmony is optional (50/50 variation) and that the ranking algorithm is the normal GLA. Here, all constraints start out with an equal ranking value, set conventionally at 100. The constraints CONTEXTUAL [šì-] and DEFAULT should be ranked in a tie to match the 50/50 variation. During learning (see Boersma and Hayes 2001, 51-54), these two constraints vacillate slightly as the GLA seeks a frequency match, but end up very close to their original value of 100. ACCIDENTAL [sì-] will remain at exactly 100: this is because the GLA is error driven and none of the three constraints favors an incorrect output for the training data that match ACCIDENTAL [sì-] (DEFAULT and ACCIDENTAL [sì-] both prefer [sì-], which is correct; and CONTEXTUAL [šì-] never matches these forms). Thus, all three constraints end up ranked at or near 100. This grammar is incorrect; when faced with novel forms like (16) that match all three

constraints, CONTEXTUAL [ši-] must compete against two, not one, equally ranked antagonists, deriving [ši-] only a third of the time instead of half.

Initial rankings based on generality (section 9) correct this problem. Given that DEFAULT and CONTEXTUAL [ši-] cover all [si-] and [ši-] forms respectively, they will be assigned initial ranking values of 500. Define the **critical distance** C as the minimum difference in ranking between two constraints that is needed to model strict ranking (informal trials suggest that a value for C of about 10.5, which creates a ranking probability of .9999, is sufficient). It is virtually certain that the initial ranking value for ACCIDENTAL [si-] will be far below 500 – C, because accidentally true constraints cannot have high generality, other than through a freak accident of the learning data. Ranking will proceed as before, with DEFAULT and CONTEXTUAL [ši-] staying close to 500 and ACCIDENTAL [si-] staying where it began. The resulting grammar correctly derives 50/50 variation, because ACCIDENTAL [si-] is too low to be active.

Now consider what happens when the data involve no free variation; i.e. [ši-] is the outcome wherever CONTEXTUAL [ši-] is applicable. When initial rankings are all equal, the [ši-] forms will cause CONTEXTUAL [ši-] to rise and DEFAULT to fall, with their difference ultimately reaching C (CONTEXTUAL [ši-]: 500 + C/2; DEFAULT: 500 – C/2). Just as before, ACCIDENTAL [si-] will remain ranked where it started, at 500. The difference of C/2 between CONTEXTUAL [ši-] and ACCIDENTAL [si-], assuming C = 10.5, will be 5.25, which means that when the grammar is applied to novel forms matching both constraints, [si-] forms will be derived about 3% of the time. This seems unacceptable, given that the target language has no free variation. Again, the incorrect result is avoided under the initial-ranking scheme of section 9, provided that ACCIDENTAL [si-] is initially ranked at or lower than 500 – C/2, which is almost certain to be the case.

In summary, schematized simulations suggest that the patterns seen in our main Navajo simulation are not peculiar to this case. The effect of accidentally true generalizations is seen most strongly when free variation is involved, but they pose a threat even in the absence of optionality. Initial rankings based on generality avoid the problem by keeping such constraints a critical distance lower than the default, so they can never have any effect on the outcome.

13. Small-Scale Exceptionless Generalizations for Irregulars

We conclude by presenting a problem that also involves exceptionless small-scale generalizations, for which we have only a sketchy answer. Since the problem strikes us as a general and important one, we include it here.

The phenomenon is the existence of **small-scale patterns** for irregulars. As Pinker and Prince (1988) point out, when a system includes irregular forms, they characteristically are not arbitrary exceptions, but fall into patterns, e.g. English *cling-clung*, *fling-flung*, *swing-swung*. These patterns have some degree of productivity, as shown by historical change (Pinker 1999) and "wug" (nonce-word) testing (Prasada and Pinker 1993, Albright and Hayes 2003).

The problem at hand is that our algorithm can find environments for these minor changes that are exceptionless. For example, the exceptionless minor change shown in (22) covers the four verbs *dig*, *cling*, *fling*, and *sling*.

(22) ɪ → ʌ / X $\begin{bmatrix} +\text{cor} \\ +\text{ant} \\ +\text{voice} \end{bmatrix}$ —— $\begin{bmatrix} +\text{dorsal} \\ +\text{voice} \end{bmatrix}$]$_{[+\text{past}]}$

The unmodified GLA, when it encounters an exceptionless constraint that conflicts with a more general constraint, inevitably ranks the exceptionless constraint categorically above the general one. For cases like Navajo, where the special constraint was (11a) and the general constraint was (11c), the default constraint for [sì-], this ranking is entirely correct, and corresponds to one of the fundamental purposes of constraint ranking—i.e. to instantiate the concept of default allomorph.

But when exceptionless (22) is ranked categorically above the constraints specifying the regular ending for English (such as USE [-d]), the following prediction is made: novel verbs matching the context of (22) should be exclusively irregular (i.e., *blig* → *blug*, not **bligged*). There is evidence that this prediction is wrong, from wug tests on forms that match (22). For instance, the wug test reported in Albright and Hayes (2003) yielded the following judgments (scale: 1 worst, 7 best):

(23)	**Present stem**	**Choice for Past**	**Rating**
a.	*blig* [blɪg]	*blug* [blʌg]	4.17
		bligged [blɪgd]	5.67
b.	*spling* [splɪŋ]	*splung* [splʌŋ]	5.45
		splinged [splɪŋd]	4.36

The regular forms are almost as good, or better, than the forms derived by the exceptionless rule.

We infer that numbers matter: a poorly attested perfect generalization like (22) is not necessarily taken more seriously than a broadly attested imperfect generalization like USE [-d]. In the Navajo case, strict ranking is appropriate, since the special-environment constraint (11a) that must outrank the default (11c) is robustly attested in the language. In the English case, the special-environment constraint is also exceptionless, but is attested in only four verbs. The GLA —in either version—ranks it on top of the grammar, just as in Navajo, but in this case with incorrect results.

We faced this problem in our earlier work on English past tenses (Albright and Hayes 2003). At the time, we avoided it by simply not using Optimality Theory. Instead, we let each rule have a score corresponding to its overall reliability, and let the predicted well formedness of each candidate output be the score of the best rule that derives it. This worked well for English, permitting both *splung* and *splinged* to be assigned appropriately high scores. However, the cost of abandoning OT was (as we now see) unacceptably high, since it would fail for cases of "special context + default", like Navajo. The Navajo default pattern above has a good overall reliability score (181/237), but that does not means it is appropriate to use it in the special context for [šì-]; that would wrongly derive *[sì-č'ìɬ] and a host of similar forms, as near-perfect options. From the present vantage point, we would judge that right approach should involve

constraint ranking (that is, OT) but have some mechanism to downgrade constraints supported by just a few forms.

The basic principles of the GLA can be supplemented with biases that exert a downward force on morphological constraints that are supported by few data, using statistical smoothing or discounting. As of this writing we do not have a complete solution, but we have experimented with a form of "absolute discounting" (Ney, Essen and Kneser 1994), implemented as follows: for each constraint C, we add to the learning data one artificial datum that violates C and obeys every other constraint with which C is in conflict.

Under this scheme, if C (say, (22) above) is supported by just four forms, then the one artificially-added candidate would have a major effect in downgrading its ranking. But if C is supported by thousands of forms (for example, the constraint for a regular mapping), then the artificially added candidate would be negligible in its effect.

We found that when we implemented this approach, it yielded reasonable results for the English scenario just outlined: in a limited simulated system consisting of the regulars in Albright and Hayes (2003) plus just the four irregulars covered by (22), regular *splinged* was a viable competitor with *splung*, and the relationships among the competing regular allomorphs remained essentially unchanged.

There are of course many ways that small-scale generalizations could be downgraded. We emphasize that the development of a well-motivated algorithm for this problem involves not just issues of computation, but an empirical question about productivity: when real language learners confront the data, what are the relative values that they place on freedom from exceptions vs. size of generalization, and how do they implement these relative values? Both experimental and modeling work will be needed to answer these questions.

14. Conclusion

We have focused on the perils for language learning of the accidentally true, and what the theory of morphological learning might do to cope with these perils.

In the main part of the paper, we argued that accidentally true environments pose a danger to learning, raising the risk of a grammar that forms outputs largely according to which accidentally true environments an input happens to meet. Our remedy for this was to use the Gradual Learning Algorithm to rank the constraints, biasing it with an initial preference for generality.

An unresolved question that we cannot address here is whether a bias for generality can be applied to all types of phonological constraints, or just those that govern allomorph distribution. It is worth noting that for certain other types of constraints, such as faithfulness constraints, it has been argued that specific constraints must have higher initial rankings than more general ones (Smith 2000). At present, we restrict our claim to morphological constraints of the form "USE X".

The second kind of perilous accidentally true generalization is the kind that covers only a few forms, the problem being not to let perfection on such a small scale outweigh the more

significant larger generalizations. Here, our suggestion was that further biases must also be added to constraint ranking that would penalize constraints based on very few forms. The details, and viability, of this proposal remain to be worked out in future research.

References

Albright, Adam (2002) Islands of reliability for regular morphology: Evidence from Italian. *Language* 78, 684-709.

Albright, Adam, Argelia Andrade and Bruce Hayes (2001) Segmental environments of Spanish diphthongization. *UCLA Working Papers in Linguistics* 7, 117-151.

Albright, Adam and Bruce Hayes (2002) Modeling English past tense intuitions with minimal generalization. In Mike Maxwell, ed., *Proceedings of the 2002 Workshop on Morphological Learning, Association of Computational Linguistics*. Philadelphia: Association for Computational Linguistics.

Albright, Adam and Bruce Hayes (2003) Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* 90: 119-161.

Anttila, Arto (1997) Deriving variation from grammar. In Frans Hinskens, Roeland van Hout and Leo Wetzels, eds., *Variation, change and phonological theory*. Amsterdam: John Benjamins, pp. 35-68.

Bach, Emmon and Robert Harms (1972) How do languages get crazy rules? In Stockwell, Robert and Ronald Macauley, eds., *Linguistic Change and Generative Theory*. Bloomington, Indiana: Indiana University Press, pp. 1–21.

Blevins, Juliette (in press) *Evolutionary Phonology*. Cambridge: Cambridge University Press.

Boersma, Paul (1997) How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21: 43–58.

Boersma, Paul (1998) Typology and acquisition in functional and arbitrary phonology. Ms., University of Amsterdam. http://www.fon.hum.uva.nl/paul/papers/typ_acq.pdf.

Boersma, Paul and Bruce Hayes (2001) Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32: 45-86.

Burzio, Luigi (2002) Surface-to-surface morphology: when your representations turn into constraints. In Paul Boucher, ed., *Many Morphologies*. Somerville, MA: Cascadilla Press, pp. 142-177

Bybee, Joan and Carol L. Moder (1983) Morphological classes as natural categories. *Language* 59: 251–270.

Chomsky, Noam and Morris Halle (1968) *The Sound Pattern of English.* New York: Harper and Row.

Frisch, Stefan, Janet Pierrehumbert, Michael Broe (2004) Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22: 179–228.

Hale, Mark and Charles Reiss (1998) Formal and empirical arguments concerning phonological acquisition. *Linguistic Inquiry* 29: 656–683.

Hayes, Bruce (1999) Phonological restructuring in Yidiɲ and its theoretical consequences. In Ben Hermans and Marc van Oostendorp, eds., *The Derivational Residue in Phonological Optimality Theory*. Amsterdam: John Benjamins, pp. 175-205.

Kiparsky, Paul (1968) How abstract is phonology? Bloomington: Indiana University Linguistics Club. Reprinted 1982 in *Explanation in Phonology*. Dordrecht: Foris, pp. 119-163.

Kruskal, Joseph (1999) An overview of sequence comparison. In David Sankoff and Joseph Kruskal, eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, 2nd ed. Reading, MA: Addison-Wesley, pp. 1-44.

MacBride, Alex (2004) *A Constraint-Based Approach to Morphology*. Ph.D. dissertation, UCLA, http://www.linguistics.ucla.edu/faciliti/diss.htm.

Ney, Hermann, Ute Essen, and Reinhard Kneser (1994) On structuring probabilistic dependencies in stochastic language modeling. *Computer Speech and Language* 8: 1–28.

Pinker, Steven (1999) Words *and Rules: The Ingredients of Language*. New York: Basic Books.

Pinker, Steven and Alan Prince (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28: 73–193.

Prasada, Sandeep and Pinker, Steven (1993) Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes* 8: 1–56.

Russell, Kevin (1999) *MOT: Sketch of an Optimality Theoretic Approach to Morphology*. Ms., http://www.umanitoba.ca/linguistics/russell/.

Sapir, Edward, and Harry Hoijer (1967) *The Phonology and Morphology of the Navajo Language*. Berkeley: University of California Press.

Smith, Jennifer (2000) Positional faithfulness and learnability in Optimality Theory. In Rebecca Daly and Anastasia Riehl, eds., *Proceedings of ESCOL 99*. Ithaca: CLC Publications, 203–214.

Young, Robert W., William Morgan Sr., and Sally Midgette (1992) *Analytical Lexicon of Navajo*. Albuquerque: University of New Mexico Press.