ırlier' representa-

ıay well extend to
judgements have
: can imagine that
by speaker-based
ected on the basis
., which may only
ıvestigation.

10

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

ADAM ALBRIGHT AND BRUCE HAYES

10.1 Introduction

Many cases of gradient intuitions reflect conflicting patterns in the data that a child receives during language acquisition.[1] An area in which learners frequently face conflicting data is inflectional morphology, where different words 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). In cases where all of these patterns could apply, such as the novel verb *spling*, the conflict between them leads English speakers to entertain multiple possibilities, with competing outcomes falling along a gradient scale of intermediate well-formedness (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 and implemented a 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 *et al.* (2001), Albright (2002), and Albright and Hayes (2003). Our model abstracts morphological and phonological generalizations from representative learning data and uses

[1] For helpful comments and advice we would like to thank Paul Boersma, Junko Ito, Armin Mester, Jaye Padgett, Hubert Truckenbrodt, the editors, and our two reviewers, absolving them for any shortcomings.

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 judgements for the same forms.

In modelling 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 correlated with greater reliability, while generalizations covering few forms or entailing many exceptions are relatively unproductive. For English past tenses, most generalizations have exceptions, so finding the productive patterns requires finding the generalizations with the fewest exceptions. Intermediate degrees of well-formedness arise when the generalizations covering different patterns suffer from different numbers of exceptions.

The phenomenon of gradient well-formedness shows that speakers do not require rules or constraints to be exceptionless; when the evidence conflicts, they are willing to use less than perfect generalizations. One would expect, however, that when gradience is observed, more reliable generalizations should be favoured over less reliable ones. In this article, we show that, surprisingly, this is not always the case. In particular, we find that there may exist generalizations that are exceptionless and well-instantiated, but are nonetheless either completely invalid, or are valued below other, less reliable generalizations.

The existence of exceptionless, but unproductive patterns is a challenge for current approaches to gradient productivity, which generally attempt to extend patterns in proportion to their strength in the lexicon. We offer a solution for one class of these problems, based on the optimality-theoretic principle of constraint conflict and employing the Gradual Learning Algorithm (Boersma 1997; Boersma and Hayes 2001). In the final section of the paper we return to our earlier work on gradience and discuss the implications of our present findings.

10.2 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 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 (10.1) (examples from Sapir and Hoijer):[2]

[2] We have rendered all transcriptions (including Sapir and Hoijer's) in near-IPA, except that we use [č č^{h} č' š ž] for [tʃ tʃ^{h} tʃ' ʃʒ] in order to depict the class of nonanterior sibilants more saliently.

:e multiple forms for
aring its 'intuitions',
'or the same forms.
l processes, we have
ıch of the input data
e? In general, greater
eneralizations cover-
ly unproductive. For
ions, so finding the
›ns with the fewest
when the generaliza-
nbers of exceptions.
that speakers do not
ˌidence conflicts, they
›uld expect, however,
alizations should be
hat, surprisingly, this
exist generalizations
ıetheless either com-
eralizations.
:rns is a challenge for
ally attempt to extend
/e offer a solution for
heoretic principle of
Algorithm (Boersma
›aper we return to our
'our present findings.

.izations arose in our
ˌironments. The first
n Navajo, a process

ıe allomorphs of the
10.1) (examples from

near-IPA, except that we use
sibilants more saliently.

(10.1) a. [šì-] if the *first segment* of the stem is a [−anterior] sibilant ([č, č', č^h, š, ž]), for example in [šì-č^{h}ìt] 'he is stooping over'

b. [šì-] or [sì-] if *somewhere later in the stem* is a [−anterior] sibilant, as in [šì-t^{h}é:ž] ~ [sì-t^{h}é:ž] 'they two are lying' (free variation)

c. [sì-] otherwise, as in [sì-t^{h}í́] 'he is lying'[3]

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 based on Navajo: we selected whole Navajo words (rather than stems) at random from the electronic version of Young *et al.*'s dictionary (1992), and constructed *s*-perfective forms for them by attaching [šì-] or [sì-] according to the pattern described in (10.1).

10.3 The learning model

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 features, 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 (10.2):

(10.2) /___ ([+seg])* [−anterior]

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

The model is given a list of pairs, consisting of bases and inflected forms. For our synthetic version of Navajo, such a list would be partially represented by (10.3):

(10.3)	a.	[pà:ʔ]	[sì-pà:ʔ]
	b.	[č'ìɬ]	[šì-č'ìɬ]
	c.	[č^{h}ò:jìn]	[šì-č^{h}ò:jìn]

[3] Sapir and Hoijer specifically say (1967: 14–15): 'Assimilation nearly always occurs when the two consonants are close together (e.g. šì-čà:ʔ, from sì-čà:ʔ "a mass lies"; ... but it occurs less often when the two consonants are at a greater distance.'

d.	[kàn]	[sì-kàn]
e.	[k'àz]	[sì-k'àz]
f.	[kʰéškã̀ː]	[šì-kʰéškã̀ː], [sì-kʰéškã̀ː]
g.	[síːɬ]	[sì-síːɬ]
h.	[tʰã̀š]	[šì-tʰã̀š], [sì-tʰã̀š]
i.	[tʰóːʔ]	[sì-tʰóːʔ]
j.	[t͡ɬéːž]	[šì-t͡ɬéːž], [sì-t͡ɬéːž]

Where free variation occurs, the learner is provided with one copy of each variant; thus, for (10.3f) both [kʰéškã̀ː] ~ [šì-kʰéškã̀ː] and [kʰéškã̀ː] ~ [sì-kʰéškã̀ː] are provided.

The goal of learning is to determine which environments require [sì-], which require [šì-], and which allow both. Learning 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 determine how these environments should be employed in the final grammar.

(10.4)

I. PREFIX [sì-]		II. PREFIX [šì-]	
a. [pàːʔ]	[sì-pàːʔ]	a. [čʰòːjìn]	[šì-čʰòːjìn]
b. [kàn]	[sì-kàn]	b. [č'ìɬ]	[šì-č'ìɬ]
c. [kʰéškã̀ː]	[sì-kʰéškã̀ː]	c. [kʰéškã̀ː]	[šì-kʰéškã̀ː]
d. [t͡ɬéːž]	[sì-t͡ɬéːž]	d. [t͡ɬéːž]	[šì-t͡ɬéːž]
e. [tʰã̀š]	[sì-tʰã̀š]	e. [tʰã̀š]	[šì-tʰã̀š]
f. [k'àz]	[sì-k'àz]		
g. [síːɬ]	[sì-síːɬ]		
h. [tʰóːʔ]	[sì-tʰóːʔ]		

Learning begins by parsing forms into their component morphemes and grouping them by the morphological change they involve. The data in (10.3) exhibit two changes, as shown in (10.4); the box surrounds cases of free variation.

For each change, the system creates hypotheses about which elements in the environment crucially condition the change. It begins by treating each pair as a 'word-specific rule', separating out the changing part from the invariant part. Thus, the first three [šì-] forms in (4) would be construed as in (5):

(10.5) a. Ø → šì / [___ č^{h}ò:jìn]

b. Ø → šì / [___ č'ìɬ]

c. Ø → šì / [___ k^{h}éškằ:]

Next, the system compares pairs of rules that have the same change (e.g. both attach [šì-]), and extracts what their environments have in common to form a generalized rule. Thus, given the word-specific rules in (10.6a), the system collapses them together using features, as in (10.6b).

(10.6) a. Ø → šì / [___ t^{h}ằš]

Ø → šì / [___ tɬé:ž]

$$\begin{array}{llllll}
\text{b.} & \varnothing \to \text{šì} \; / \; [\;\_\_\_ & \text{t}^h & \text{ằ} & \text{š} &] \\
+ & \varnothing \to \text{šì} \; / \; [\;\_\_\_ & \widehat{\text{tɬ}} & \text{é:} & \text{ž} &] \\
\hline
= & \varnothing \to \text{šì} \; / \; [\;\_\_\_ & \begin{bmatrix}-\text{sonorant}\\ -\text{continuant}\\ +\text{anterior}\end{bmatrix} & \begin{bmatrix}+\text{syllabic}\\ -\text{high}\\ -\text{round}\end{bmatrix} & \begin{bmatrix}-\text{sonorant}\\ +\text{continuant}\\ -\text{anterior}\\ +\text{strident}\end{bmatrix} &]
\end{array}$$

In this particular case, the 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 [č^{h}ò:jìn] and [č'ìɬ] above, this is a harder question.[4] We adopt an approach that lines up the segments that are **most similar** to each other. For instance, (10.7) gives an intuitively reasonable alignment for [č^{h}ò:jìn] and [č'ìɬ]:

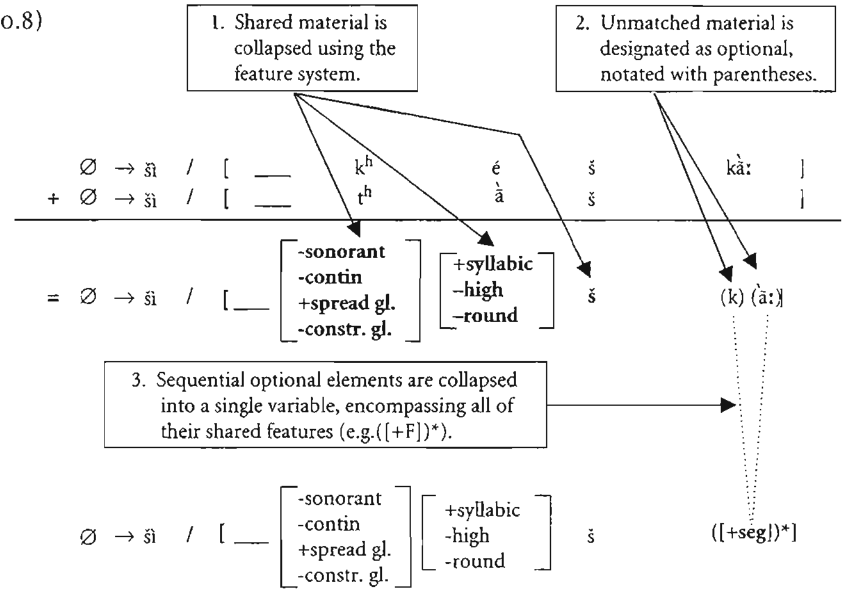
(10.7)
$$\begin{array}{ccccc}
\text{č}^h & \text{ò:} & \text{j} & \text{ì} & \text{n} \\
| & & & | & | \\
\text{č'} & & & \text{ì} & \text{ɬ}
\end{array}$$

[4] 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.

Good alignments have two properties: they match phonetically similar segments such as [$\check{c}^{h}$] and [č’], and they avoid leaving too many segments unmatched. To evaluate the similarity of segments, we employ the similarity metric from Frisch *et al.* (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.

Seen in detail, the process of collapsing rules is based on three principles, illustrated in (10.8) with the collapsing of the rules Ø → šì / [___ k^{h}éškà:] and Ø → šì / [___ t^{h}àš].

(10.8)

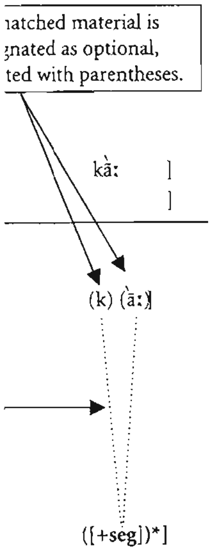

1. Shared material is collapsed using the feature system.
2. Unmatched material is designated as optional, notated with parentheses.
Ø → šì / [___ kʰ é š kà:]
+ Ø → šì / [___ tʰ à š]
= Ø → šì / [___ -sonorant -contin +spread gl. -constr. gl. +syllabic –high –round š (k) (à:)]
3. Sequential optional elements are collapsed into a single variable, encompassing all of their shared features (e.g.([+F])*).
Ø → šì / [___ -sonorant -contin +spread gl. -constr. gl. +syllabic -high -round š ([+seg])*]

Paired feature matrices are collapsed by constructing a new matrix, containing 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 (*F*)*, where *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

onetically similar
o many segments
we employ the
.ntee an optimal
thm (described in
ents for best total

n three principles,
ì / [___ kʰéškà:]

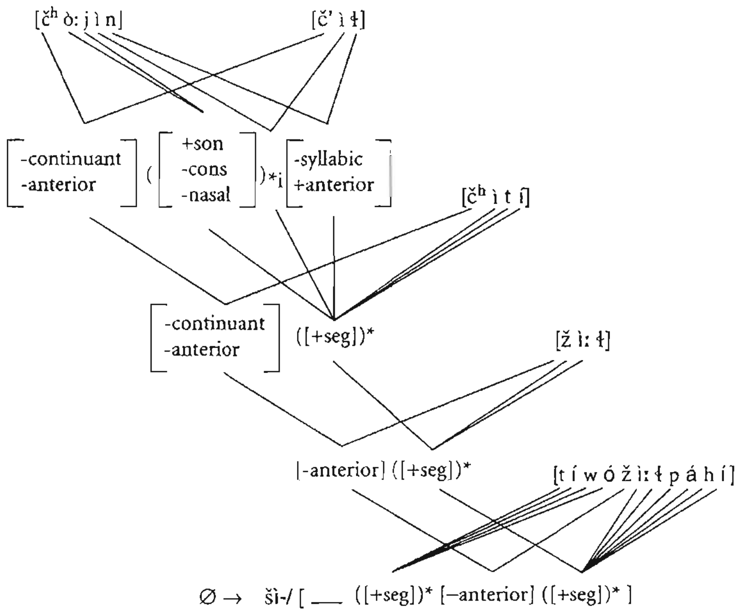


matrix, containing
in one rule that is
ted by parentheses
ents are collapsed
e F is the smallest
ents (step 3).
the other words in
, and so on. Due to
ack the hypothesis

set, keeping only those rules that perform best.[5] Generalization terminates when no new 'keeper' rules are found.

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 (10.9):

(10.9)



The result can be read: 'Prefix [šì-] to a stem consisting of any number of segments followed by a nonanterior segment, followed by any number of segments.' (Note that [−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, as in (10.10):

(10.10) Ø → šì- / ___ 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 is taken up in Section 10.5.

[5] 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.

10.4 Testing the approach: a simulation

We 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,[6] we attached prefixes to the bases as follows, following Sapir and Hoijer's characterization: (a) if the base began with a nonanterior sibilant, we prefixed [šì-] (there were nineteen 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-] (thirty-seven of each); (c) we prefixed [sì-] to the remaining 144 bases.

Running the algorithm just described, we found that among the 92 environments it learned, three were of particular interest: the environment for obligatory local harmony, (10.11a); the environment that licenses distal harmony, ((10.11b); note that this includes local harmony as a special case); and the vacuous 'environment' specifying the default allomorph [sì-], (10.11c).

(10.11) a. Obligatory local harmony

Ø → [šì-] / ___ [−anterior]

b. Optional distal harmony (= (10.10))

Ø → [šì-] / ___ X [−anterior]

c. Default [sì-]

Ø → [sì-] / ___ X

The remaining eighty-nine environments are discussed below.

10.5 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

[6] From the entire database of 3,023 words, we selected 2,000 words at random, dividing this set into ten batches of 200 words each. To confirm the generality of our result, we repeated our simulation on each 200-word training sample. Due to space considerations, we report here the results of only one of the ten trials; the remaining nine were essentially the same in that they all discovered the environments in (10.11). The primary difference between trials was the precise formulation of the other, unwanted constraints (Section 10.6), but in every case, such constraints were correctly ranked below the crucial constraints, as in (10.13).

morphology (Boersma 1998*b*; Russell 1999; Burzio 2002; MacBride 2004). In this approach, rule (10.11a) is reconstrued as a constraint: 'Use [šì-] / ___ [−anterior] to form the *s*-perfective.' This constraint is violated by forms that begin with a nonanterior segment, but use something other than [šì-] to form the *s*-perfective. The basic idea is illustrated below, using hypothetical monosyllabic roots:

(10.12)

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 (10.13) and (10.14) show:

(10.13)

$$\text{Use [šì-]/ \_\_\_ [−ant]} >> \{ \underbrace{\text{Use [sì-] / \_\_\_ X, Use [šì-] / \_\_\_ X [−ant]}}_{\textit{ranked in free variation}} \} >> \text{all others}$$

(10.14) a.

/sì-čìd/	Use [šì-]/___[−ant]	Use [šì-]/___X [−ant]	Use [sì-]/___X
☞ šì-čìd			*
* sì-čìd	*!	*	

b.

/sì-té:ž/	Use [šì-]/___[−ant]	Use [šì-]/___X [−ant]	Use [sì-]/___X
☞ šì-té:ž			*
☞ sì-té:ž		*!	

For (10.14b), the free ranking of Use [šì-] / ___ X [−ant] and Use [sì-] / ___ X produces multiple winners generated in free variation (Anttila 1997).

10.6 Unwanted constraints

The eighty-nine constraints not discussed so far consist largely of complicated generalizations that happen to hold true of the learning data. One example is shown in (10.15):

(10.15) $\text{Use sì- / \_\_\_ }([-\text{round}])^{*} \begin{bmatrix}+\text{anterior}\\ +\text{continuant}\end{bmatrix} ([-\text{consonantal}])^{*}]$

As it happens, this constraint works for all thirty-seven forms that meet its description in the learning data. However, it makes profoundly incorrect predictions for forms outside the learning data, such as the legal but non-existing stem /čálá/ (10.16).

$$
(10.16)\quad
\begin{array}{lccccc}
\text{U}\textsc{se}\ \text{sì-}\ / & \_\_\_ & ([-\text{round}])^{*} & \begin{bmatrix}+\text{anterior}\\ +\text{continuant}\end{bmatrix} & ([-\text{consonantal}])^{*}] \\
 & & \wedge & | & | \\
\text{sì-} & & \text{č}\ \ \text{á} & \text{l} & \text{á}
\end{array}
$$

If ranked high enough, this constraint will have the detrimental effect of preventing [šì-čálá] from being generated consistently. We will call such inappropriate generalizations 'junk' constraints.

One possible response is to say that the learning method is simply 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 restrict the algorithm 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 with an effective evaluation metric. Below, we lay out such a metric, which employs the Gradual Learning Algorithm.[7]

10.7 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. 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 (10.13) above would be captured by assigning 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, assigning each constraint a ranking value. On any particular occasion that the grammar is used, a

[7] A reviewer points out that another approach to weeding out unwanted generalizations is to train the model on various subsets of the data, keeping only those generalizations that are found in all training sets (cross-validation). Although this technique could potentially eliminate unwanted generalizations (since each subset contains a potentially different set of such generalizations), it could not absolutely guarantee that they would not be discovered independently in each subset. Given that such constraints make fatal empirical predictions, we seek a technique that reliably demotes them, should they arise.

s the legal but non-

tal])*]

letrimental effect of
'. We will call such

ethod is simply too
:d from the learning
: experimented with
:neralizations. Yet we
simplified—and rely
be generated rather
:ation metric. Below,
:arning Algorithm.[7]

Boersma and Hayes
tion and matches the
stochastic version of
B} is assigned not a
; with probability P.'
aptured by assigning
ant] a 50–50 ranking

ıairwise probabilities
: GLA, this is done by
ing each constraint a
grammar is used, a

d generalizations is to train
tions that are found in all
eliminate unwanted gener-
neralizations), it could not
ach subset. Given that such
ably demotes them, should

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 determined on the basis of this ranking. Since pairwise ranking probabilities are determined by the ranking values,[8] they are guaranteed to be mutually consistent.

10.8 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, 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 are swamped by higher-ranking junk constraints, and rendered largely ineffective. The result is a grammar that performs quite well on the training data (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 (10.11).

The problem is a classic one in inductive learning theory. If a learning algorithm excessively tailors its behaviour to the training set, it may 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 does the GLA fail? The reason is that it demotes constraints only when they prefer losing candidates. But within the learning data, the junk constraints generally prefer only 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. What is needed is a way for the GLA to distinguish accidentally true generalizations from linguistically significant generalizations.

10.9 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

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

to solving the Navajo problem. What is needed, however, is a way to characterize generality in numerical terms. There are various possible approaches; Chomsky and Halle (1968), for example, propose counting symbols (fewer symbols = greater generality). Here, we adopt an *empirical* criterion: a morphological constraint is maximally general if it encompasses all of the forms that exhibit its structural change. We use the fraction in (10.17):

(10.17)

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

In the 200-word Navajo simulation discussed above, some representative generality values are shown in (10.18).

(10.18)

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 (10.15) ('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, all the data will be explained by the most general applicable constraints, and the others will remain so low that they never play a role in selecting output forms.

In order to ensure that differences in initial rankings are large enough to make a difference, the generality values from (10.17) were rescaled to cover a huge probability range, using the formula in (10.19):

(10.19) For each constraint c, initial ranking value$_c$ =

$$500 \times \frac{\text{Generality}_c - \text{Generality}_{\text{min}}}{\text{Generality}_{\text{max}} - \text{Generality}_{\text{min}}}$$

ever, is a way to
: various possible
propose counting
adopt an *empirical*
.l if it encompasses
ıse the fraction in

s to
:onstraint requires

ome representative

Generality
·339
1
1
.204

ranking values that
ı top. If the scheme
eral applicable con-
never play a role in

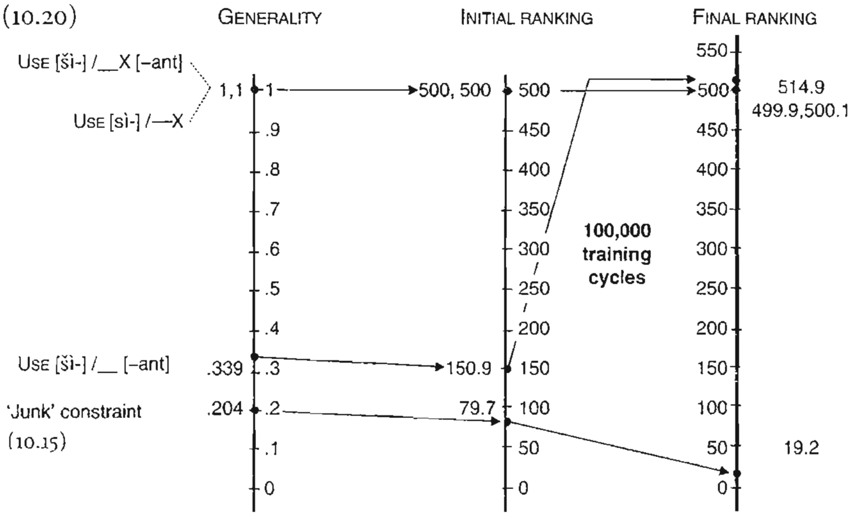
are large enough to
e rescaled to cover a

where Generality$_{min}$ is the generality of the least general constraint, and Generality$_{max}$ is the generality of the most general constraint.

10.10 Employing generality in a learning simulation

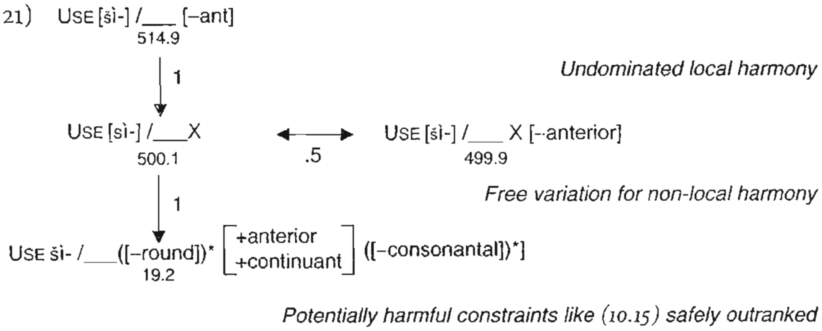
We implemented this scheme and ran it multiple times on the Navajo pseudodata described above. For one representative run, it caused the relevant constraints (including here just one representative 'junk' constraint (10.15)), to be ranked as follows:

(10.20)



The final grammar is depicted schematically in (10.21), 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).

(10.21)


Use [šì-] /___ [–ant]
514.9
1
Undominated local harmony
Use [sì-] /___X
500.1
.5
Use [šì-] /___ X [–anterior]
499.9
1
Free variation for non-local harmony
Use šì- /___([–round])* [+anterior +continuant] ([–consonantal])*]
19.2
Potentially harmful constraints like (10.15) safely outranked

This approach yields the desired grammar: all of the junk constraints (not just (10.15)) are ranked safely below the top three.

The procedure works because the GLA is error-driven. Thus, if junk constraints start low, 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. Even if they start low, 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.

10.11 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 worked, and ensures that the result is a general one.

To this end, we reduce Navajo to three constraints, renamed 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 (10.15), 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). Using the normal GLA, all constraints start out with equal ranking values, 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: 51–4), 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, since the GLA is error driven and none of the three constraints favours 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 are ranked at or near 100. This grammar is incorrect; when faced with novel forms like (10.16) that match all three constraints, Contextual [šì-] competes against two equally ranked antagonists, deriving [šì-] only a third of the time instead of half.

ık constraints (not just

driven. Thus, if junk
al constraint that does
ıat would promote the
xts, on the other hand,
ppropriately so. Even if
[sì-šáp], and thus they
ımar.
tatements in grammar
a necessary criterion in
h are implausible, but

n in the complexity of
helps in understanding
lt is a general one.
renamed as follows: (1)
ontext Use [šì-] /___ X
(3) the accidentally-
sì-]. Accidental [sì-]
ing data happen not to

variation). Using the
ing values, set conven-
ıd Default should be
learning (see Boersma
lightly as the GLA seeks
original value of 100.
ıe GLA is error driven
output for the training
Accidental [sì-] both
] never matches these
ar 100. This grammar is
) that match all three
equally ranked antag-
of half.

Initial rankings based on generality (Section 10.9) correct this problem. Given that Default and Contextual [šì-] cover all [sì-] and [šì-] forms respectively, they will be assigned initial ranking values of 500. Define the **critical distance** C as the minimum difference between two constraints that is needed to model strict ranking. (Informal trials suggest that a value of about 10.5, which creates a ranking probability of .9999, is sufficient.) It is virtually certain that the initial ranking value for Accidental [sì-] will be far below 500−C, because accidentally true constraints cannot have extremely high generality, other than through an unlikely accident of the learning data. Ranking proceeds as before, with Default and Contextual [šì-] staying around 500 and Accidental [sì-] staying where it began. The resulting grammar correctly derives 50/50 variation, because Accidental [sì-] is too low to be active.

Now consider what happens when the data involve no free variation; that is [šì-] is the outcome wherever Contextual [šì-] is applicable. When initial rankings are all equal, [šì-] forms will cause Contextual [šì-] to rise and Default to fall, with their difference ultimately reaching C (Contextual [šì-]: 500+C/2; Default: 500−C/2). Just as before, Accidental [sì-] will remain ranked where it started, at 500. The difference of C/2 between Contextual [šì-] and Accidental [sì-], assuming C = 10.5, will be 5.25, which means that when the grammar is applied to novel forms matching both constraints, [sì-] outputs will be derived about 3 per cent 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 10.9, provided that Accidental [sì-] is initially ranked at or lower than 500−C/2, which is almost certain to be the case.

In summary, schematized analysis suggests that the Navajo result is not peculiar to this case. The effect of accidentally true generalizations is strongest when optionality is involved, but they pose a threat even in its absence. Initial rankings based on generality avoid the problem by keeping such constraints a critical distance lower than the default, so they can never affect the outcome.

10.12 The realism of the simulation

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

10.12.1 *Phonological rules versus allomorph distribution*

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

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

More generally, however, there still remains the question of how to unify knowledge about allomorphy and root-internal phonotactics. Even when affixes and roots show the same harmony patterns, we believe that understanding the distribution of affix allomorphs could constitute an important first step in learning the more general process, provided there is some way of bootstrapping from constraints on particular morphemes to more general constraints on the distribution of speech sounds. We leave this as a problem for future work.

10.12.2 *Should arbitrary constraints be generated at all?*

Another possible objection is that a less powerful generalization procedure would never have posited constraints like (10.15) 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 suggesting that environments sometimes really are complex and synchronically arbitrary (Bach and Harms 1972; Hale and Reiss 1998; Hayes 1999; Blevins 2004). For instance, in examining patterns of English past tenses, we found that all verbs ending in voiceless fricatives are regular, and that speakers are tacitly aware of this generalization (Albright and Hayes 2003). Not only are such patterns arbitrary, but they can also be rather complex (see also Bybee and Moder 1983). Regardless of whether such generalizations are learned or innate, it seems likely that any model powerful enough to handle the full range of attested patterns will need a mechanism to sift through large numbers of possibly irrelevant hypotheses.

10.13 Modelling gradient productivity: the fate of reliability metrics

As noted above, one of our long-term goals is to understand how gradient productivity arises when the learner confronts conflicting data. The results above challenge our earlier views, and in this section we lay out ways in which our previous approach might be revised.

s allomorphy ([sì-]
ffects. Thus, it may
of harmony.
observed primarily
no root-internal
, admitting greater
: only appropriate
mony often require

on of how to unify
tactics. Even when
believe that under-
titute an important
here is some way of
es to more general
/e this as a problem

alization procedure
irst place. Indeed, if
innate), the need to
jection can be cited
: complex and syn-
iss 1998; Hayes 1999;
glish past tenses, we
ar, and that speakers
ayes 2003). Not only
r complex (see also
generalizations are
ul enough to handle
to sift through large

reliability metrics

rstand how gradient
ng data. The results
ay out ways in which

Earlier versions of our model evaluated contexts according to their accuracy, or reliability, defined as the ratio of the number of forms a rule derives correctly, divided by the total number of forms to which the rule is applicable. We have found in many cases that we could model native speaker intuitions of novel forms by using the reliability of the best rule that derives them (adjusted slightly, in a way to be mentioned below). However, the results of our Navajo simulations show that accuracy alone is not an adequate criterion for evaluation, since assiduous rule discovery can sometimes find accidentally-true (and thus perfectly accurate) generalizations which nonetheless lead to disaster if trusted. The Navajo example illustrates why it is not enough to evaluate the accuracy of each generalization independently; we must also consider whether generalizations cover forms that are better handled by a different generalization.[9]

Another possible failing of the reliability approach is that it is ill-suited to capture special case/'elsewhere' relations (Kiparsky 1982). The environment for [šì-] in Navajo is difficult to express by itself, but easy as the complement set of the [sì-] environments. In optimality theory, 'elsewhere' is simply the result of constraint ranking: a context-sensitive constraint outranks the default. Unfortunately for the reliability-based approach, default environments such as (10.11c) often have fairly high reliability (181/237 in this case)—but that does not mean that they should be applied in the special-allomorph context (e.g. of (10.11a)).

In light of this, it is worth considering why we adopted reliability scores in the first place. Ironically, the reason likewise involved accidentally-true generalizations, but of a different kind.

One of the phenomena that compelled us to use reliability scores was the existence of small-scale patterns for irregulars, seen, for example, in English past tenses. 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. *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 (Bybee and Moder 1983; Prasada and Pinker 1993; Albright and Hayes 2003).[10]

[9] A related problem, in which overly broad generalizations appear exaggeratedly accurate because they contain a consistent subset, is discussed in Albright and Hayes (2002).

[10] We restrict our discussion to phonological patterns; for discussion of patterns based possibly on semantic, rather than phonological similarities, see Ramscar (2002). In principle, the approach described here could be easily extended to include constraints that refer to other kinds of information; it is an empirical question what properties allomorphy rules may refer to.

The problem is that our algorithm can often find environments for these minor changes that are exceptionless. For example, the exceptionless minor change in (10.22) covers the four verbs *dig, cling, fling,* and *sling.*[11]

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

The GLA, when comparing an exceptionless constraint against a more general constraint that suffers from exceptions, always ranks the exceptionless constraint categorically above the general one. For cases like Navajo, where the special constraint was (10.11a) and the general constraint was (10.11c), the default constraint for [sì-], this ranking is entirely correct, capturing the special/default relationship. But when exceptionless (10.22) is ranked categorically above the constraints specifying the regular ending for English (such as Use [-d]), the prediction is that novel verbs matching the context of (10.22) 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 (10.22). For instance, the wug test reported in Albright and Hayes (2003) yielded the following judgements (scale: 1 worst, 7 best):

(10.23)	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 such as (10.22) is not necessarily taken more seriously than a broadly attested imperfect generalization such as Use [-d]. For Navajo, strict ranking is appropriate, since the special-environment constraint (10.11a) that must outrank the default (10.11c) is robustly attested in the language. In the English case, the special-environment constraint is also exceptionless, but is attested in only four verbs, yet the GLA—in either version—ranks it on top of the grammar, just as in Navajo.

[11] This is the largest set of ɪ → ʌ verbs that yields an exceptionless generalization. There are other subsets, such as *cling, fling,* and *sling,* that also lead to exceptionless generalizations, and these are also generated by our model. The problem that we discuss below would arise no matter which set is selected, and would not be solved by trying to, for example, exclude *dig* from consideration.

vironments for these
exceptionless minor
nd *sling*.[11]

gainst a more general
he exceptionless con-
ke Navajo, where the
straint was (10.11c),
orrect, capturing the
22) is ranked categor-
g for English (such as
the context of (10.22)
t **bligged*). There is
on forms that match
ht and Hayes (2003)
):

forms derived by the

ct generalization such
n a broadly attested
ajo, strict ranking is
10.11a) that must out-
guage. In the English
ionless, but is attested
anks it on top of the

neralization. There are other
ralizations, and these are also
rise no matter which set is
from consideration.

It can now be seen why in our earlier work we avoided constraint interaction and employed reliability scores instead. With reliability scores, it is simple to impose a penalty on forms derived by rules supported by few data—following Mikheev (1997), we used a statistical lower confidence limit on reliability. Thus, for a wug form like *blig*, two rules of comparable value compete: the regular rule (has exceptions, but vast in scope) versus (10.22) (no exceptions, but tiny in scope). Ambivalence between the two is a natural consequence.

If reliability statistics are not the right answer to this problem, what is? It seems that the basic idea that rules based on fewer forms should be downgraded is sound. But the downgrade need not be carried out based on reliability scores—it might also be made part of the constraint ranking process. In particular, we propose that the basic principles of the GLA 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 *et al.* 1994), implemented as follows: for each constraint C, we add to the learning data an artificial datum that violates C and obeys every other constraint with which C is in conflict. Under this scheme, if C (say, (10.22) above) is supported by just four forms, then an 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 simulation consisting of the regulars in Albright and Hayes (2003) plus just the four irregulars covered by (10.22), regular *splinged* was a viable competitor with *splung*, and the relationships among the competing regular allomorphs remained essentially unchanged.

There are many ways that small-scale generalizations could be downgraded during learning. 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 weights that they place on accuracy versus size of generalization? Both experimental and modelling work will be needed to answer these questions.[12]

[12] 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'.

10.14 Conclusion

The comparison of English and Navajo illustrates an important problem in the study of gradient well-formedness in phonology. On the one hand, there are cases such as English past tenses, in which the learner is confronted with many competing patterns and must trust some generalizations despite some exceptions. In such cases, gradient well-formedness is rampant, and the model must retain generalizations with varying degrees of reliability. On the other hand, there are cases such as Navajo sibilant harmony, in which competition is confined to particular contexts, and the learner has many exceptionless generalizations to choose from. In these cases, the challenge is for the model to choose the 'correct' exceptionless patterns, and refrain from selecting an analysis that predicts greater variation than is found in the target language.

We seek to develop a model that can handle all configurations of gradience and categoricalness, and we believe the key lies in the trade-off between reliability and generality. We have shown here how our previous approach to the problem was insufficient, and proposed a new approach using the GLA, modified to favour more general constraints. The precise details of how generality is calculated, and how severe the bias must be, are left as a matter for future research.

Gradience in Grammar

Generative Perspectives

Edited by
GISBERT FANSELOW, CAROLINE FÉRY,
RALF VOGEL, AND MATTHIAS SCHLESEWSKY

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2006

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd. King's Lynn, Norfolk

ISBN 019–927479–7 978–019–927479–6

1 3 5 7 9 10 8 6 4 2