Empirical Tests of the Gradual Learning Algorithm

Paul Boersma
University of Amsterdam

Bruce Hayes
UCLA

July 8, 2000

Abstract

The Gradual Learning Algorithm (Boersma 1997) is a constraint ranking algorithm for learning Optimality-theoretic grammars. The purpose of this article is to assess the capabilities of the Gradual Learning Algorithm, particularly in comparison with the Constraint Demotion algorithm of Tesar and Smolensky (1993, 1996, 1998, 2000), which initiated the learnability research program for Optimality Theory. We argue that the Gradual Learning Algorithm has a number of special advantages: it can learn free variation, deal effectively with noisy learning data, and account for gradient well-formedness judgments. The case studies we examine involve Ilokano reduplication and metathesis, Finnish genitive plurals, and the distribution of English light and dark /l/.

Keywords: Learnability, Optimality Theory, variation, Ilokano, Finnish

1 Introduction

Optimality Theory (Prince and Smolensky 1993) has made possible a new and fruitful approach to the problem of phonological learning. If the language learner has access to an appropriate inventory of constraints, then a complete grammar can be derived, provided there is an algorithm available that can rank the constraints on the basis of the input data. This possibility has led to a line of research on ranking algorithms, originating with the work of Tesar and Smolensky (1993, 1996, 1998, 2000; Tesar 1995) who propose an algorithm called Constraint Demotion, reviewed below. Other work on ranking algorithms includes Pulleyblank and Turkel (1995, 1996, 1998, to appear), Broihier (1995), Hayes (1999), and Prince and Tesar (1999).

Our focus here is the Gradual Learning Algorithm, as developed by Boersma (1997, 1998, to appear). This algorithm is in some respects a development of Tesar and Smolensky's proposal: it directly perturbs constraint rankings in response to language data, and, like most previously proposed algorithms, it is error-driven, in that it alters rankings only when the input data conflict with its current ranking hypothesis. What is different about the Gradual Learning Algorithm is the type of Optimality-Theoretic grammar it presupposes: rather than a set of discrete rankings, it assumes a continuous scale of constraint strictness. Also, the grammar is regarded as stochastic: at every evaluation of the candidate set, a small noise component is temporarily

We would like to thank Arto Anttila, Chris Manning, and two anonymous LI reviewers for helpful input in the preparation of this paper. Thanks also to Louis Pols, the University of Utrecht, and the UCLA Academic Senate for material assistance in making our joint work possible. The work of the first author was supported by a grant from the Netherlands Organization for Scientific Research.

added to the ranking value of each constraint, so that the grammar can produce variable outputs if some constraint rankings are close to each other.

The continuous ranking scale implies a different response to input data: rather than a wholesale reranking, the Gradual Learning Algorithm executes only small perturbations to the constraints' locations along the scale. We argue that this more conservative approach yields important advantages in three areas. First, the Gradual Learning Algorithm can fluently handle *optionality*; it readily forms grammars that can generate multiple outputs. Second, the algorithm is *robust*, in the sense that speech errors occurring in the input data do not lead it off course. Third, the algorithm is capable of developing formal analyses of linguistic phenomena in which speakers' judgments involve *intermediate well-formedness*.

A paradoxical aspect of the Gradual Learning Algorithm is that, even though it is statistical and gradient in character, most of the constraint rankings it learns are (for all practical purposes) categorical. These categorical rankings emerge as the limit of gradual learning. Categorical rankings are of course crucial for learning data patterns where there is no optionality.

Learning algorithms can be assessed on both theoretical and empirical grounds. At the purely theoretical level, we want to know if an algorithm can be guaranteed to learn all grammars that possess the formal properties it presupposes. Research results on this question as it concerns the Gradual Learning Algorithm are reported in Boersma (1997, 1998, to appear). On the empirical side, we need to show that natural languages are indeed appropriately analyzed with grammars of the formal type the algorithm can learn.

This paper focuses on the second of these two tasks. We confront the Gradual Learning Algorithm with a variety of representative phonological phenomena, in order to assess its capabilities in various ways. This approach reflects our belief that learning algorithms can be tested just like other proposals in linguistic theory, by checking them out against language data.

A number of our data examples are taken from the work of the second author, who arrived independently at the notion of a continuous ranking scale, and has with colleagues developed a number of hand-crafted grammars that work on this basis (Hayes and MacEachern 1998; Hayes, to appear).

We will begin by reviewing how the Gradual Learning Algorithm works, then present several empirical applications. A study of Ilokano phonology shows how the algorithm can cope with data involving systematic optionality. We also use a restricted subset of the Ilokano data to simulate the response of the algorithm to speech errors. In both cases, we make comparisons with the behavior of the Constraint Demotion Algorithm. We next turn to the study of output frequencies, posed as an additional, stringent empirical test of the Gradual Learning Algorithm. We use the algorithm to replicate the study of Anttila (1997a,b) on Finnish genitive plurals. Lastly we turn to gradient well-formedness, showing that the algorithm can replicate the results on English /l/ derived with a hand-crafted grammar by Hayes (to appear).

2 How the Gradual Learning Algorithm Works

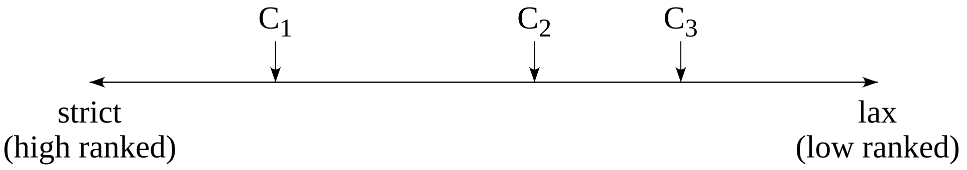
Two concepts crucial to the Gradual Learning Algorithm are the *continuous ranking scale* and *stochastic candidate evaluation*. We cover these first, then turn to the internal workings of the algorithm.

2.1 *The Continuous Ranking Scale*

The algorithm presupposes a linear scale of constraint strictness, in which higher values correspond to higher-ranked constraints. The scale is arranged in arbitrary units, and in principle has no upper or lower bound. Other work that has suggested or adopted a continuous scale includes Liberman (1993:21, cited in Reynolds 1994), Zubritskaya (1997:142–4), Hayes and MacEachern (1998), and Hayes (to appear).

Continuous scales include strict constraint ranking as a special case. For instance, the scale depicted graphically in (1) illustrates the straightforward nonvariable ranking C_1 >> C_2 >> C_3.

(1) *Categorical ranking along a continuous scale*

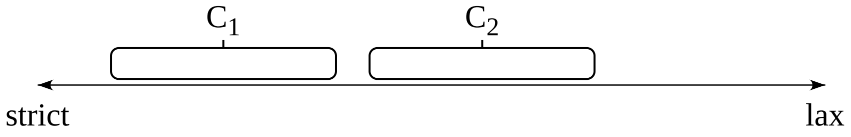


2.2 *How Stochastic Evaluation Generates Variation*

The continuous scale becomes more meaningful when differences in distance have observable consequences, e.g. if the short distance between C_2 and C_3 in (1) tells us that the relative ranking of this constraint pair is less fixed than that of C_1 and C_2. We suggest that in the process of speaking (i.e. at *evaluation time*, when the candidates in a tableau have to be evaluated in order to determine a winner), the position of each constraint is temporarily perturbed by a random positive or negative value. In this way, the constraints act as if they are associated with ranges of values, instead of single points. We will call the value used at evaluation time a *selection point*. The value more permanently associated with the constraint, i.e. the center of the range, will be called the *ranking value*.

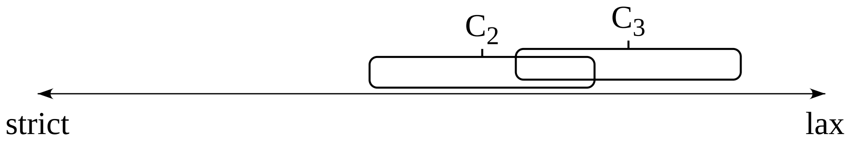
Here there are two main possibilities. If the ranges covered by the selection points do not overlap, the ranking scale again merely recapitulates ordinary categorical ranking:

(2) *Categorical ranking with ranges*



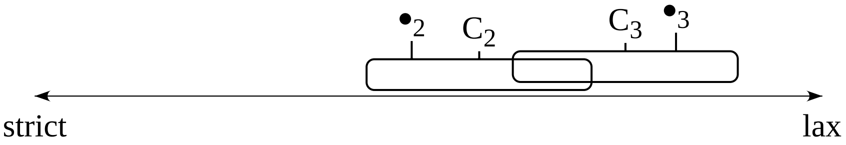
But if the ranges overlap, there will be free (variable) ranking:

(3) *Free ranking*

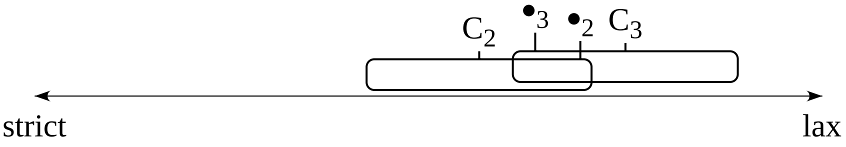


The reason is that, at evaluation time, it is possible to choose the selection points from anywhere within the ranges of the two constraints. In (3), this would most often result in C_2 outranking C_3, but if the selection points are taken from the upper part of C_3's range, and the lower part of C_2's, then C_3 would outrank C_2. The two possibilities are shown below; /$\bullet_2$/ and /$\bullet_3$/ depict the selection points for C_2 and C_3.

(4) a. *Common result*: $C_2 \gg C_3$



b. *Rare result*: $C_3 \gg C_2$



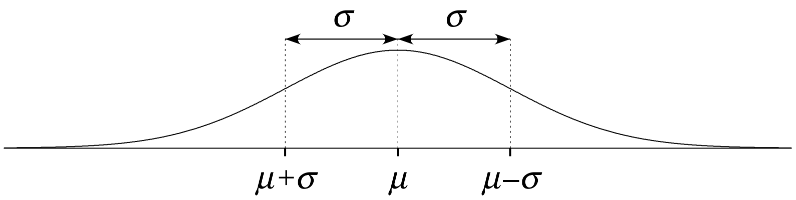
When one sorts all the constraints in the grammar by their selection points, one obtains a total ranking to be employed for a particular evaluation time. With this total ranking, the ordinary competition of candidates (supplied by the GEN function of Optimality Theory) takes place and determines the winning output candidate.[1]

The above description covers how the system in (4) behaves at one single evaluation time. Over a longer sequence of evaluations, the overlapping ranges will often yield an important observable effect: for forms in which $C_2 \gg C_3$ yields a different output than $C_3 \gg C_2$, one will observe *free variation*, i.e. multiple outputs for a single underlying form.

To implement these ideas more precisely, we interpret the constraint ranges as *probability distributions* (Boersma 1997, 1998; Hayes and MacEachern 1998). For each constraint, we assume a function that specifies the probability that the selection point will occur at any given distance above or below the constraint's ranking value at evaluation time. By using probability distributions, one can not only enumerate the set of outputs generated by a grammar, but also make predictions about their relative frequencies, a matter that will turn out to be important below.

Many noisy events in the real world occur with probabilities that are appropriately described with a *normal* (= Gaussian) distribution. A normal distribution has a single peak in the center, which means that values around the center are most probable, and declines gently but swiftly toward zero on each side. Values become less probable the farther they are away from the center, without ever actually becoming zero:

(5) *The normal distribution*



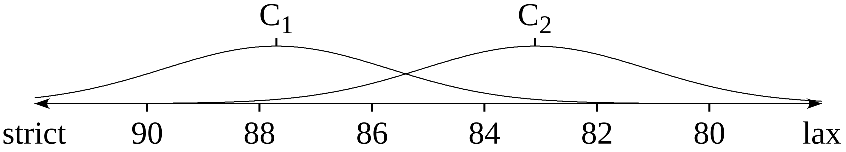
[1] The mechanism for determining the winning output in Optimality Theory, with GEN and a ranked constraint set, will not be reviewed here. For background, see Prince and Smolensky's original work (1993), or textbooks such as Archangeli and Langendoen (1997) and Kager (1999b).

A normal distribution is described by its mean μ, which occurs at its center, and its standard deviation σ, which describes the "breadth" of the curve. Approximately 68 percent of the values drawn from a normal distribution lie within one standard deviation from the mean, i.e. between $\mu-\sigma$ and $\mu+\sigma$. The Gradual Learning Algorithm makes the assumption that selection points for natural language constraints are distributed normally, with the mean of the distribution occurring at the ranking value. The normal distributions are assumed to have the *same* standard deviation for every constraint, for which we typically adopt the arbitrary value of 2.0. [2] In this approach, the behavior of a constraint set depends on its ranking values alone; constraints cannot be individually assigned standard deviations. The process of learning an appropriate constraint ranking therefore consists solely of finding a workable set of ranking values.

When discussing the derivation of forms using a set of constraints, we will use the term *evaluation noise* to designate the standard deviation of the distribution (σ); the term is intended to suggest that this value resides in the evaluation process itself, not in the constraints.

We illustrate these concepts with two hypothetical constraints and their associated normal distributions on an arbitrary scale:

(6) *Overlapping ranking distributions*



In (6), the ranking values for C_1 and C_2 are at the hypothetical values 87.7 and 83.1. Since the evaluation noise is 2.0, the normal distributions assigned to C_1 and C_2 overlap substantially. While the selection points for C_1 and C_2 will most often occur somewhere in the central "hump" of their distributions, they will on occasion be found quite a bit further away. Thus, C_1 will outrank C_2 at evaluation time in most cases, but the opposite ranking will occasionally hold. Simple calculations show that the percentages for these outcomes will tend towards the values 94.8% ($C_1 >> C_2$) and 5.2% ($C_2 >> C_1$).

2.3 *How can there* ***not*** *be variation?*

A worry that may have presented itself to the reader at this point is: how can this scheme depict *obligatory* constraint ranking, if the values of the normal distribution never actually reach zero? The answer is that when two constraints have distributions that are dramatically far apart, the odds of a deviant ranking become vanishingly low. Thus, if two distributions are 5 standard deviations apart, the odds that a "reversed" ranking could emerge are about 1 in 5,000. This frequency would be hard to distinguish empirically, we think, from the background noise of speech errors. If the distributions are 9 standard deviations apart, the chances of a "reversed" ranking are 1 in 10 billion, implying that one would not expect to observe a form derived by this ranking even if one monitored a speaker for an entire lifetime.

[2] Since the units of the ranking scale are themselves arbitrary, it does not matter what standard deviation is used, so long as it is the same for all constraints.

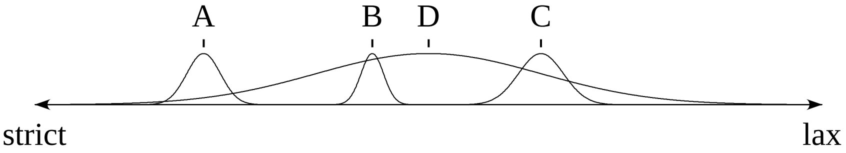
In applying the Gradual Learning Algorithm, we often find that it places constraints at distances of tens or even hundreds of standard deviations apart, giving what is to all intents and purposes nonvariable ranking.

Often, constraints occur ranked in long transitive chains. The ranking scheme depicted here can treat such cases, since the strictness continuum is assumed to have no upper or lower bounds, and the learning algorithm is allowed to take up as much space as it needs to represent all the necessary rankings.

2.4 *Predictions about Ranking*

This concludes our discussion of the model of ranking presupposed by the Gradual Learning Algorithm. Before we move on, it is worth noting that this model is quite restrictive: there are various cases of logically possible free rankings that it excludes. Thus, for example, it would be impossible to have a scheme in which A "strictly" outranks B (i.e., the opposite ranking is vanishingly rare), B "strictly" outranks C, and D is ranked freely with respect to both A and C. This scheme would require a much larger standard deviation for D than for the other constraints. The model does not permit this, since the noise is regarded as a property of evaluation, not of each separate constraint:

(7) *A less restrictive grammar model with different distributions for each constraint*



Thus, while the empirical tests given below are meant primarily to assess the Gradual Learning Algorithm, they also test a general hypothesis about possible free rankings.[3]

2.5 *The Gradual Learning Algorithm*

The Gradual Learning Algorithm tries to locate an empirically appropriate ranking value for every constraint.

The Initial State. The constraints begin with ranking values according to the initial state that is hypothesized by the linguist. In principle, this could give every constraint the same ranking value at the start, or one could incorporate various proposals from the literature for less trivial initial rankings (for example, Faithfulness low: Gnanadesikan 1995, Smolensky 1996, Boersma 1998, Hayes 1999; or Faithfulness high: Hale and Reiss 1998). In the cases considered here, such decisions affect the amount of input data and computation needed, but do not materially affect the final outcome.[4] In our implementations of the algorithm, every constraint starts at the same value, selected arbitrarily to be 100.

[3] Reynolds (1994) and Nagy and Reynolds (1997) adopt a "floating constraint" model, in which a given constraint may be freely ranked against a whole hierarchy of categorically ranked constraints, which is exactly what we just claimed to be impossible. We have undertaken reanalyses of a number of Reynolds's and Nagy's cases, and found that it is possible to account for their data within the model we assume, though never with the same constraint inventory. Some reanalyses are posted at http://www.fon.hum.uva.nl/paul/gla/.

[4] In contrast, for the problem of learning phonotactic distributions from positive evidence only (Smolensky 1996, Hayes 1999, Prince and Tesar 1999), for which no fully adequate algorithm yet exists, the issue of initial

Step 1: A datum. The algorithm is presented with a learning datum, i.e. an adult surface form that the language learner hears in her environment and assumes to be correct. Adopting Tesar and Smolensky's idealization, we assume that the algorithm is also able to access the underlying form for each learning datum.

The idea that the learner obtains access to underlying forms naturally raises questions, since underlying forms are not audible, nor are structures like syllables or feet. We refer the reader to Tesar and Smolensky (1996, 2000) for discussion of how the problem of covert structure might be overcome by embedding the ranking algorithm within a larger learning system.

Step 2: Generation. Since the algorithm is error-driven, the next step is to see what the current grammar generates for the assumed underlying form. Where this yields a mismatch, adjustment of the grammar can then take place.

Generation works as follows. For each constraint, a noise value is taken at random from the normal probability distribution and is added to the constraint's current ranking value to obtain the selection point. Once a selection point has been picked for every constraint, generation proceeds by *sorting* the constraints in descending order of their selection points. This yields a strict constraint ranking, of the traditional kind, which is used only for this particular evaluation. The remainder of the generation process follows the standard mechanisms of Optimality Theory.[5]

Step 3: Comparison. If the form just generated by the grammar is identical to the learning datum, nothing further is done. But if there is a mismatch, the algorithm notices this and takes action. Specifically, it compares the constraint violations of the learning datum with what is currently generated by the grammar. This comparison is illustrated in tableau (8), which depicts a representative grammar with eight schematic constraints.

(8) *A mismatch between the learner's form and the adult form*

/underlying form/	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
√ Candidate 1 (learning datum)	*!	**	*		*			*
☞ Candidate 2 (learner's output)		*	*	*		*		*

As can be seen, Candidate 1, which is the surface form the algorithm has just "heard," failed to emerge as the winner in the overall competition among candidates. That winner happens to be Candidate 2. The algorithm, being error-driven, takes this mismatch as a signal to alter the grammar so that in the future the grammar will be more likely to generate Candidate 1, and not Candidate 2. The alteration will take the form of changes in the ranking values of the schematic constraints C_1-C_8.[6]

rankings may well be crucial. The cases we consider here are more tractable, since the goal is simply to project surface forms from known underlying forms.

[5] It follows that the set of possible outputs that can be generated by a constraint set remain the same under our approach. As noted in §2.4, the theory makes additional predictions about what outputs can occur together in free variation.

[6] Note that in cases of free variation, Candidate 2 might actually be well formed. This case is discussed in §2.6.

The next step is just the same as in the Constraint Demotion Algorithm (Tesar and Smolensky 1998:239): *mark cancellation.* Violations that match in the two rival candidates are ignored, as they make no difference to the outcome.

(9) *Mark cancellation*

/underlying form/	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
√ Candidate 1 (learning datum)	*!	*~~*~~	~~*~~		*			~~*~~
☞ Candidate 2 (learner's output)		~~*~~	~~*~~	*		*		~~*~~

Step 4: Adjustment. In the situation being considered, Candidate 1 (the candidate embodied by the learning datum) should have won, but Candidate 2 was the actual winner. This constitutes evidence for two things. First, it is likely that those constraints for which the learning-datum candidate suffers uncanceled marks are ranked too high. Second, it is likely that those constraints for which the learner's output suffers uncanceled marks are ranked too low. Neither of these conclusions can be taken as a certainty. However, this uncertainty is not crucial, since the ultimate shape of the grammar will be determined by the ranking values that the constraints will take on in the long term, with exposure to a full range of representative forms. The hypothesis behind the Gradual Learning Algorithm is that moderate adjustments of ranking values will ultimately achieve the right grammar. Therefore, the algorithm is set up so as to make a small adjustment to all constraints that involve uncanceled marks.

We define *plasticity* as the numerical quantity by which the algorithm adjusts the constraints' ranking values at any given time. Appropriate values for plasticity are discussed below; for now, the reader should simply assume that the plasticity is reasonably small.

The response of the Gradual Learning Algorithm to data, as processed in the way just given, is as follows:

- For any constraint for which the learning-datum candidate suffers uncanceled marks, *decrease* that constraint's ranking value by an amount equal to the current plasticity.
- For any constraint for which the learner's own output candidate suffers uncanceled marks, *increase* that constraint's ranking value by an amount equal to the current plasticity.

For the schematic case under discussion, these adjustments are shown in tableau (10) as small arrows. Canceled marks are omitted for clarity.

(10) *The learning step: adjusting the ranking values*

/underlying form/	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
√ Candidate 1 (learning datum)	*→	*→			*→			
☞ Candidate 2 (learner's output)				←*		←*		

The adjustments ensure that Candidate 1 becomes *somewhat more likely* to be generated on any future occasion, and Candidate 2 somewhat less likely.

Final state. With further exposure to learning data, the algorithm cycles repeatedly through steps 1 to 4. If for the underlying form under discussion, the adult output is always Candidate 1, then C_4 or C_6 (or both; it depends on the violations of all the other learning pairs) will eventually be ranked at a safe distance above C_1, C_2, and C_5. This distance will be enough that the probability of generating Candidate 2 will become essentially nil, and the resulting grammar will generate Candidate 1 essentially 100% of the time.

2.6 *The Algorithm's Response to Free Variation*

Consider now cases of free variation, where the same underlying form yields more than one possible output. In such cases, it can happen that the current guess of the grammar fails to match the learning datum, yet is well-formed in the target language.

The activity of the Gradual Learning Algorithm in such cases might at first glance seem pointless: as free variants are processed, they induce apparently random small fluctuations in the ranking values of the constraints. Closer inspection, however, shows that the response of the algorithm is in the long run systematic and useful: with sufficient data, the algorithm will produce a grammar that *mimics the relative frequency of free variants* in the learning set. As will be seen in §5 below, the frequency-matching behavior of the Gradual Learning Algorithm has important linguistic consequences.

Intuitively, frequency matching works as follows: a given free variant *F* that is more common than a cooccurring variant *F′* will have more "say" than *F′* in determining the ranking values. However, adjustments induced by *F* will only occur up to the point where the grammar assigns to *F* its fair share of the set of forms derivable from the underlying form in question. Should learning accidentally move the ranking values beyond this point, then tokens of *F′* will get a stronger effect in subsequent learning, resulting in adjustments in the opposite direction. The system eventually stabilizes with ranking values that yield a distribution of generated outputs that mimics the distribution of forms in the learning data. The same mechanism will mimic learning-set frequencies for three, or in principle any number, of free variants.

This concludes the main discussion of the algorithm. In implementing the algorithm, one must select a learning schedule for plasticity and other parameters, ideally in a way that maximizes the speed and accuracy of learning. For now we suppress these details, deferring discussion to Appendix A.

2.7 *Some Alternatives that Don't Work*

We would not want the view attributed to us that use of statistical methods is a panacea in learnability; plainly, it is not. First, the Gradual Learning Algorithm relies on the theory of grammar to which it is coupled (Optimality Theory), along with a specific constraint inventory. If that inventory does not permit the linguistically significant generalizations to be captured, then the grammar learned by the Gradual Learning Algorithm will not capture them.[7] Second, not just any statistically-driven ranking algorithm suffices. We have tried quite a few alternatives and found that they failed on data for which the Gradual Learning Algorithm succeeded. These alternatives include:

[7] We have checked this claim by creating pathological versions of our learning files in which the constraint violations are replaced with random values. We find that even with lengthy exposure, the Gradual Learning Algorithm cannot learn the data pattern when the constraints are rendered into nonsense in this way.

- Decrementing only those constraints that directly cause a wrong guess to win (for example, just C_1 and C_2 in tableau (10) above).
- Decrementing only the highest uncancelled constraint of the learning datum (just C_1 in tableau (10)). This approach is called the "Minimal Gradual Learning Algorithm" in Boersma (1997).
- Decrementing only the highest uncancelled constraint of the learning datum, and promoting only the highest uncancelled mark of the incorrect winner. This is a symmetrized version of the previous algorithm, and was shown to be incorrect in Boersma (1997).

All these learning schemes work correctly for nonvariable data, but crash to various degrees for data involving optionality.

2.8 *Assessing a Learned Grammar*

After the algorithm has learned a grammar, we must assess the validity of that grammar, particularly in cases where the input data exhibit free variation. This can be done straightforwardly simply by repeating the process of stochastic evaluation many times, without further learning. It is quite feasible to run thousands of trials, and thus obtain accurate estimates both of what forms the grammar generates, and of the frequencies with which the forms are generated.

We turn now to empirical applications of the Gradual Learning Algorithm.

3 Free Variation in Ilokano

Hayes and Abad (1989) present and analyze a variety of phonological phenomena of Ilokano, an Austronesian language of the Northern Philippines. The Ilokano data exhibit phonological free variation on a fairly extensive scale. The following will be nothing more than an extract from the language, but we believe it to be representative and faithful enough for the results to be meaningful.

3.1 *The Ilokano Data*

Our interest lies in two areas of free variation: an optional process of metathesis, and variation in the form of reduplicants.

3.1.1 *Metathesis*

Ilokano metathesis permutes precisely one segmental sequence: /ʔw/ optionally becomes [wʔ]. In all cases, the [w] is itself not an underlying segment, but is derived from /o/. Thus, there are forms like those in (11):

(11)	**daʔo**	‘kind of tree’	**/pag-daʔo-an/**	→	**pagdaʔwan**, **pagdawʔan**	‘place where **daʔo**’s are planted’
	baʔo	‘rat’	**/pag-baʔo-an/**	→	**pagbaʔwan**, **pagbawʔan**	‘place where rats live’
	taʔo	‘person’	**/taʔo-en/**	→	**taʔwen**, **tawʔen**	‘to repopulate’
	ʔaggaʔo	‘to dish up rice’	**/pag-gaʔo-an/**	→	**paggaʔwan**, **paggawʔan**	‘place where rice is served’
	ʔagsaʔo	‘to speak’	**/pag-saʔo-en/**	→	**pagsaʔwen**, **pagsawʔen**	‘to cause to speak’

The motivation for metathesis is not hard to find: glottal stop is not generally permitted in Ilokano syllable codas. For example, there are no stems like ***paʔlak**; and special reduplication patterns arise where needed to avoid coda [ʔ] (Hayes and Abad 1989:358). Indeed, the only coda glottal stops of Ilokano are those which arise when glide formation strands a glottal stop, as in the optional variants just given.[8]

The glide formation that is seen in metathetic forms is general in the language, taking place whenever a vowel-initial suffix is added to a stem ending in a nonlow vowel; thus for example **ʔaj*o*** ‘to cheer up’ ~ **ʔaj*w*en** ‘cheer up-GOAL FOCUS’. See Hayes and Abad (1989:337–8) and Hayes (1989:271) for additional examples.

3.1.2 *Reduplication*

Ilokano reduplication comes in two varieties; one in which the reduplicative prefix forms a heavy syllable, and one in which it forms a light syllable. The two reduplication processes are not generally in free variation; rather, each is recruited in the formation of a variety of morphological categories.

The main interest here is a pattern found for heavy reduplication when the stem begins with a consonant + glide cluster. Here, it is an option to form the reduplicated syllable by vocalizing the glide. This vowel is lengthened in order to provide weight.

(12)	**rwa.ŋan**	‘door’	**ruː.rwa.ŋan**	‘doors’
	pja.no	‘piano’	**piː.pja.no**	‘pianos’
	bwa.ja	‘crocodile’	**na.ka.buː.bwa.ja**	‘act like a crocodile’

A second option is to copy the entire C+glide+VC sequence, as in (13):

[8] There are sounds in coda position that Hayes and Abad (1989:340) transcribe as [ʔ]; these are casual-speech lenited forms of /t/. Native speakers tend to hear these as /t/, and sometimes they sound rather intermediate between [t] and [ʔ]. In light of phonetic research on the non-neutralizing character of many phonetic processes (see e.g. Dinnsen 1985), it seems unlikely that these are true glottal stops; rather, they probably contain residual tongue-blade gestures, much as was documented for English by Barry (1985) and Nolan (1992). Thus a constraint banning glottal stops in coda, if restricted to “pure” glottal stops, would not apply to them. Should future phonetic work prove that the glottal stops from /t/ really are straightforward glottal stops (an event we regard as unlikely), then this part of Ilokano phonology must be considered opaque, in the sense of Kiparsky 1973, and the analysis would have to be recast making use of one of the various theoretical devices proposed for treating opacity in Optimality Theory (McCarthy 1996, 1999; Kirchner 1996; Kiparsky 1998).

(13)	**rwaŋ.rwa.ŋan**	‘doors’
	pjan.pja.no	‘pianos’
	na.ka.bwaj.bwa.ja	‘act like a crocodile’

Lastly, there is a possibility of copying the glide as a vowel, as before, but with the heavy syllable created by resyllabifying the first consonant of the stem leftward. The vocalized glide surfaces as short:

(14)	**rur.wa.ŋan**	‘doors’
	pip.ja.no	‘pianos’
	na.ka.bub.wa.ja	‘act like a crocodile’

The evidence that resyllabification has actually taken place is found in stems that begin with /rw/: here, just in case the vowel of the reduplicant is short, the /r/ appears in its voiceless allophone; thus in detailed transcription, **rur.wa.ŋan**.is [ɾuɾ̥.wa.ŋan]. [ɾ̥] is the surface form of /r/ that is generally found in coda position (Hayes and Abad 1989:355).

Variation in Ilokano reduplicants appears to be fairly systematic, though there are individual forms where a particular variant is lexicalized and used more or less obligatorily. We model here the cases of productive formation.

In the following sections, we sketch analyses for Ilokano metathesis and reduplication. For both, we adhere to the general scheme for Correspondence constraints given in McCarthy and Prince (1995), as well as McCarthy and Prince’s account of the mechanisms of reduplication.

3.2 *Analysis of Metathesis*

The basic analysis of metathesis seems fairly straightforward: it reflects a dynamic competition between a constraint that bans glottal stop in coda position (*ʔ]$_\sigma$) with a constraint that requires faithfulness to underlying linear order (LINEARITY). A form like **taw.ʔen** avoids coda [ʔ], whereas a form like **taʔ.wen** preserves the order of /ʔ/ and /o/ (→ [w]), as seen in the underlying form /**taʔo-en**/. Both candidates alter the syllabicity of /o/, thus violating a constraint IDENT-IO(syllabic). The basic idea is summarized in the following tableaux, which derive the two alternative outcomes:

(15) a. *Glide formation*

/**taʔo-en**/	LINEARITY	*ʔ]$_\sigma$	IDENT-IO(syllabic)
☞ **taʔ.wen**		*	*
taw.ʔen	*!		*

b. *Glide formation and metathesis*

/**taʔo-en**/	*ʔ]$_\sigma$	LINEARITY	IDENT-IO(syllabic)
taʔ.wen	*!		*
☞ **taw.ʔen**		*	*

One must also show why underlying **/taʔo-en/** should require its /o/ to appear as [w] in any event. The fully faithful outcome, ***ta.ʔo.en**, is ruled out by Onset (Prince and Smolensky 1993:25), which militates against vowel-initial syllables and is undominated in Ilokano. Vowel deletion, as in ***taʔen** or ***taʔon**, is ruled out by undominated Max-IO(V).

Resolution of hiatus by epenthetic [ʔ] (***taʔoʔen**) is excluded for stems ending in /o/. This fact indicates a fairly high ranking for Dep-IO(ʔ), which forbids glottal epenthesis. However, Dep-IO(ʔ) is not undominated, because glottal epenthesis *is* the normal outcome for /a/ stems, as in **/basa-en/** → **basaʔen** 'read-GOAL FOCUS'. In such stems, glide formation is not a possibility, because of (a) an undominated ban on low glides (*LowGlide, thus ***bas.a̯en**); (b) the undominated faithfulness constraint Ident-IO(low), which requires low segments to remain so (**/basa-en/** → ***baswen**).

Another hypothetical way to avoid [ʔ] in coda position would be to resyllabify it, forming a [ʔw] onset (***ta.ʔwen**). This cannot happen, because Ilokano never permits ʔC onsets in any context. For present purposes we will simply posit a *[$_\sigma$ʔC constraint; noting that other complex onsets (such as stop + liquid or oral consonant + glide) are possible.

The full tableaux in (16) give these additional possibilities, along with the constraints that rule them out.

(16) a. *Glide formation with optional metathesis*

/taʔo-en/	*Low Glide	Ident-IO (low)	*[$_\sigma$ʔC	Max-IO (V)	Onset	Dep-IO (ʔ)	Linea- rity	*ʔ]$_\sigma$	Ident-IO (syllabic)
☞ **taʔ.wen**								*	*
☞ **taw.ʔen**							*		*
ta.ʔo.ʔen						*!			
ta.ʔo.en					*!				
taʔen, **taʔon**				*!					
ta.ʔwen			*!						*

b. *Glottal stop insertion*

/basa-en/	*Low Glide	Ident-IO (low)	*[$_\sigma$ʔC	Max-IO (V)	Onset	Dep-IO (ʔ)	Linea- rity	*ʔ]$_\sigma$	Ident-IO (syllabic)
☞ **ba.sa.ʔen**						*			
ba.sa.en					*!				
ba.sen, **ba.san**				*!					
bas.wen		*!							*
bas.a̯en	*!								*

Note that metathesis is never of any help in the absence of glide formation. A metathesized but unglided [o], as in **ta.o.ʔen**, will incur a violation of Onset in its new location. Since

ta.o.ʔen is additionally burdened with a LINEARITY violation, it can never win out over simpler candidates that avoid metathesis in the first place.

One further constraint is needed to account for the fact that [ʔ] can occur in codas in forms derived by glide formation, but never in underived forms. We find that it is not possible to derive this pattern with a simple ranking of markedness and faithfulness constraints. But the variety of constraints sometimes called “output-to-output correspondence” (Benua 1995, 1997; Burzio 1996; Kager 1999a, 1999b; Kenstowicz 1997; Steriade 2000), which forces uniformity of phonological properties through the paradigm, suffices. The idea is that the ban on glottal stop in coda is outranked by MAX-OO(ʔ), but outranks MAX-IO(ʔ). What this means is that if [ʔ] occurs on the surface of some morphological base form,[9] a corresponding [ʔ] must also occur in forms that are morphologically related to that base. Thus, the [ʔ] of the surface form **taʔo** serves to protect the corresponding [ʔ] of **taʔ.wen.** In contrast, in a hypothetical underived form like **paʔ.lak**, [ʔ] cannot be protected by MAX-OO(ʔ). Therefore, if there were an underlying representation **/paʔlak/** that included a basic coda [ʔ], a [ʔ]-less form created by GEN would win in the competition.[10] The analysis is summarized in the following tableaux:

(17) a. *Paradigm uniformity forces glottal stop in coda*

/taʔo-en/ paradigm includes **taʔo**	MAX-OO(ʔ)	*ʔ]$_\sigma$	MAX-IO(ʔ)
☞ **taʔ.wen**		*	
ta.wen	*!		*

b. *Glottal stop deleted from coda*

/paʔlak/	MAX-OO(ʔ)	*ʔ]$_\sigma$	MAX-IO(ʔ)
paʔ.lak		*!	
☞ **pa.lak**			*

3.3 *Analysis of Variable Reduplication*

We posit that reduplicated forms like **buː.bwa.ja** can win out over **bwaj.bwa.ja** some of the time, because the reduplicant **buː** is a simpler syllable than **bwaj**. More precisely, **buː.bwa.ja** conforms better (one violation instead of two) to the constraint *COMPLEXONSET. Although this constraint does not hold true in general of Ilokano vocabulary, it can be ranked high enough to make **buː.bwa.ja** an option.[11] The possibility of **buː.bwa.ja** is thus an instance of the “Emergence of the Unmarked” effect (McCarthy and Prince 1994).

In return, **bwaj.bwa.ja** can defeat **buː.bwa.ja** some of the time, because it incurs fewer violations of constraints requiring that the reduplicant be a good copy of the base. In particular,

[9] In Ilokano it suffices to assume that the morphological base is the bare stem, which is always a legal isolation form.

[10] In most versions of Optimality Theory, illegal forms are ruled out by the constraint system, rather than by limitations on underlying representations (Prince and Smolensky 1993; Smolensky 1996).

[11] [buː.bwa.ja] also avoids a *CODA violation, but *CODA turns out to reside so low in the constraint hierarchy that it cannot be the responsible factor.

it retains the length of the copied vowel, thus obeying the base-reduplicant identity constraint Ident-BR(long).[12] Moreover, it retains the syllabicity of the glide /w/, respecting Ident-BR(syllabic), and it copies more segments of the base, incurring fewer violations of Max-BR.

The remaining variant **bub.wa.ja** avoids *ComplexOnset violations completely, and can therefore on some occasions beat out its two rivals. It loses, sometimes, because unlike its rivals it fails to display the cross-linguistically favored alignment of syllable and stem boundaries. In terms of the theory of McCarthy and Prince (1993), **bub.wa.ja** violates Align(Stem, L, Syllable, L), whereas **bwaj.bwa.ja** and **buː.bwa.ja** obey this constraint.

Assuming suitable rankings elsewhere, the three-way optionality reduces to variable ranking of just three constraints, as shown in the following tableaux. In the underlying forms, "HRED" stands for the abstract morpheme that is phonologically realized with heavy reduplication.

(18) *Triple variation in heavy reduplication*

a.

/HRED-**bwaja**/	Align	Ident-BR(long)	*ComplexOnset
☞ **bwaj.bwa.ja**			**
buː.bwa.ja		*!	*
bub.wa.ja	*!		

b.

/HRED-**bwaja**/	Align	*ComplexOnset	Ident-BR(long)
☞ **buː.bwa.ja**		*	*
bwaj.bwa.ja		**!	
bub.wa.ja	*!		

c.

/HRED-**bwaja**/	*ComplexOnset	Ident-BR(long)	Align
☞ **bub.wa.ja**			*
buː.bwa.ja	*!	*	
bwaj.bwa.ja	*!*		

Here are a few additional details of the analysis.

- For brevity, we omit the (undominated) constraints that force the reduplicant to be a heavy syllable.

[12] The feature [long] stands here for whatever formal account of length, such as multiple linking, turns out to be appropriate.

- A further candidate ***baː.bwa.ja** manages to avoid a complex onset, just like **buː.bwa.ja**, and moreover avoids miscopying syllabicity (i.e. copying /w/ as [u]). This candidate is completely ill-formed, however; a fact we attribute to its violating CONTIGUITY, the constraint that requires that a contiguous sequence be copied (McCarthy and Prince 1995, 371).
- Both *[$_\sigma$ʔC and *COMPLEXONSET are necessary in Ilokano. [ʔC] onsets are completely impossible, whereas the more general class of complex onsets (including e.g. [bw] or [tr]) are well attested. The role of *COMPLEXONSET lies only in the derivation of correct intervocalic syllabification (e.g. **kwat.ro** ‘four’) and in producing reduplicated forms like **buː.bwa.ja**.

3.4 *Constraint Ranking*

To provide a check on what the Gradual Learning Algorithm does with the Ilokano data, we carried out a hand ranking of the 18 constraints discussed above. The ranking proceeded on the basis of the following representative forms, with the illegal rival candidates shown on the right:

(19) *Legal and illegal forms in Ilokano*

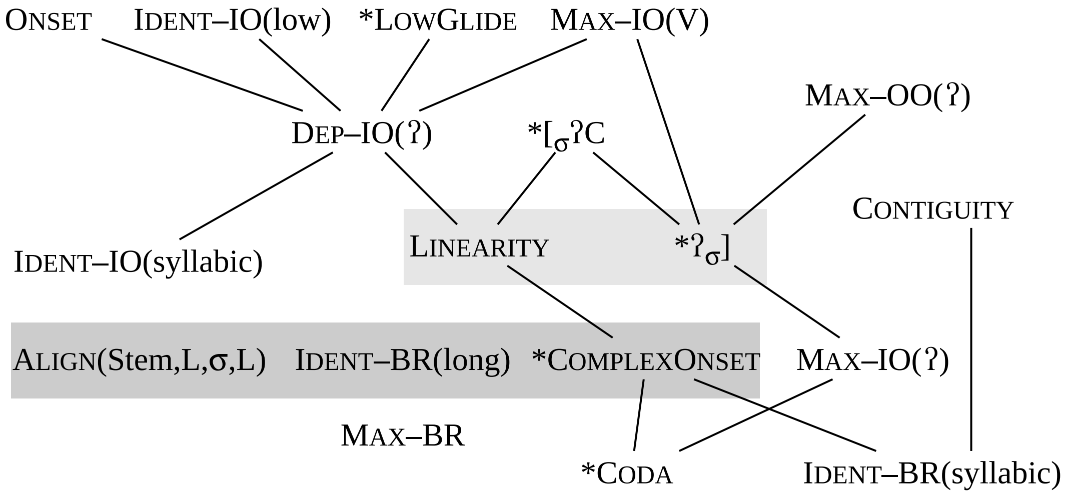
a.	/**taʔo-en**/:	**taw.ʔen**, **taʔ.wen**	***ta.wen**, ***ta.ʔen**, ***ta.ʔo.en**, ***ta.ʔo.ʔen**, ***ta.ʔwen**
b.	/HRED-**bwa.ja**/:	**bwaj.bwa.ja**, **buː.bwa.ja**, **bub.wa.ja**	***bwaː.bwa.ja**, ***baː.bwa.ja**
c.	/**paʔlak**/:	**pa.lak**	***paʔ.lak**, ***pa.ʔlak**
d.	/**labʔaj**/:	**lab.ʔaj**	***la.baj**
e.	/**trabaho**/:	**tra.ba.ho**	***tar.ba.ho**
f.	/**ʔajo-en**/:	**ʔaj.wen**	***ʔa.jen**, ***ʔa.jo.en**, ***ʔa.jo.ʔen**, ***ʔa.jwen**
g.	/**basa-en**/:	**ba.sa.ʔen**	***ba.sen**, ***ba.sa.en**, ***bas.a̯en**, ***bas.wen**

A few forms not yet presented need some annotation here. The fact that **lab.ʔaj** ‘bland’ is well-formed shows that MAX-IO(ʔ) is active in the grammar; else ***la.baj** would be derived.[13] Illegal ***tar.ba.ho** from /**trabaho**/ ‘work’ shows that metathesis is not called for merely to avoid a *COMPLEXONSET violation. /**ʔajo-en**/ and /**basa-en**/, given earlier, illustrate straightforward hiatus resolution where no metathesis configuration arises.

The hand ranking of the constraints was done as follows. We first computed the factorial typology of the constraint set over the candidates given. Then, we gradually added pairwise constraint rankings, recomputing the factorial typology as constrained *a priori* by these rankings, until the output set had shrunk to include only the attested cases of free variation. The ranking that emerged is as follows:

[13] Actually, postconsonantal /ʔ/ *is* optionally deleted in certain forms, but the deletion happens on a sporadic, stem-by-stem basis. We idealize, harmlessly we think, to general non-deletion.

(20) *Hand ranking for Ilokano*



It can be seen that some of the constraints are undominated, and sit at the top of the grammar. Somewhere in the "middle" of the grammar are two sets of constraints that must be freely ranked, in order to derive free variation. These are shown boxed in different shades of gray. Neither of these freely ranked sets is at the bottom of the grammar, as each includes constraints that dominate still others further down. We will show below (§3.7) that the task of learning freely ranked sets in "medial position" is of particular interest in comparing different ranking algorithms.

3.5 *Application of the Gradual Learning Algorithm*

We started with all constraints at a ranking value (selected arbitrarily) of 100. The algorithm was provided with 21,000 underlying/surface pairs. The underlying form for each pair was chosen randomly from the seven forms in (19) with equal probability, so that each underlying form occurred approximately 3,000 times in the data. Where a given underlying form corresponded to more than one legal surface form, we assigned each surface form an equal probability (i.e. 50% each for **taw.?en** and **ta?.wen**, and 33.3% each for **buː.bwa.ja**, **bwaj.bwa.ja**, and **bub.wa.ja**).

It should be noted that each of the forms of (19) is representative of a substantial lexical class, whose members share the same crucial constraint violations. Since it is these violations that drive learning, it suffices to use multiple copies of the forms of (19) to stand in for the full Ilokano lexicon.

The schedules that we employed for plasticity and other crucial values are given in Appendix A.

After 21,000 data, the Gradual Learning Algorithm had assigned the following ranking values to the constraints:

(21) *Machine ranking for Ilokano*

Constraint	Ranking Value
ONSET	164.00
*LOWGLIDE	164.00
IDENT-IO(low)	164.00
MAX-IO(V)	162.00
*$[_\sigma$ʔC	142.00
MAX-OO(ʔ)	138.00
DEP-IO(ʔ)	130.00
*ʔ$]_\sigma$	111.08
LINEARITY	110.92
CONTIGUITY	108.00
ALIGN(Stem, L, Syll, L)	87.96
*COMPLEXONSET	86.64
IDENT-BR(long)	85.64
MAX-IO(ʔ)	80.00
MAX-BR	67.60
IDENT-BR(syllabic)	55.60
*CODA	40.36
IDENT-IO(syllabic)	–56.00

Comparing these with the hand ranking of (20), the reader will see that there is a close resemblance. Where the hand ranking posits strict domination, the machine ranking places the constraints at a considerable distance apart (for example, MAX-IO(ʔ) >> *CODA, needed so that **lab.ʔaj** will defeat ***la.baj**). Where the hand ranking posits free variation, the algorithmic ranking assigns close or near-identical values (for example, free ranking of *ʔ$]_\sigma$ and LINEARITY, needed for optional metathesis). In cases where the ranking makes no difference to the empirical outcome, the machine ranking will harmlessly assign unpredictable values. Thus, IDENT-IO(low) and *$[_\sigma$ʔC are placed 22 units apart, although the hand ranking showed that both are in fact undominated.

The truest test of the algorithm, however, is not the form of the resulting grammar, but what this grammar generates. To determine this, we computed the output probabilities for grammar (21) by running each of the seven underlying forms through the grammar one million times. The results are shown in the final column of table (22).

(22) *Accuracy of predictions made by machine-ranked grammar*

Underlying form	Surface form	Target language	Simulation result
/taʔo-en/	**taw.ʔen**	50%	52.2%
	taʔ.wen	50%	47.9%
	ta.wen	0	0
	ta.ʔen	0	0
	ta.ʔo.en	0	0
	ta.ʔo.ʔen	0	0
	ta.ʔwen	0	0
/HRED-**bwaja/**	**buː.bwa.ja**	33.33%	36.7%
	bwaj.bwa.ja	33.33%	31.2%
	bub.wa.ja	33.33%	32.1%
	bwaː.bwa.ja	0	0
	baː.bwa.ja	0	0
/paʔlak/	**pa.lak**	100%	100%
	paʔ.lak	0	0
	pa.ʔlak	0	0
/labʔaj/	**lab.ʔaj**	100%	100%
	la.baj	0	0
/trabaho/	**tra.ba.ho**	100%	100%
	tar.ba.ho	0	0
/ʔajo-en/	**ʔaj.wen**	100%	100%
	ʔa.jen	0	0
	ʔa.jo.en	0	0
	ʔa.jo.ʔen	0	0
	ʔa.jwen	0	0
/basa-en/	**ba.sa.ʔen**	100%	100%
	ba.sen	0	0
	ba.sa.en	0	0
	bas.a̯en	0	0
	bas.wen	0	0

It can be seen that the grammar generates all and only the correct forms of the language. Moreover, where there is free variation, the grammar does a reasonably good job of matching the frequencies found in the learning data.

Values describing the intermediate stages of learning for this simulation are given in Appendix A.

The grammar described in (21) is the result of just one run of the algorithm. Since the algorithm encounters the data in random order, and itself includes a stochastic component, other runs can produce different results. Therefore, a fully legitimate test must carry out learning many times, checking to see that learning is successful each time. We therefore repeated the entire learning process 1000 times, testing the resulting 1000 grammars with 100,000 trials for each form, and collected statistics on the result.

First, in the entire set of 1000 runs (700 million trials), only 7 illegal forms were generated, all of them ***ta.ʔo.ʔen**. Second, frequency matching was generally good: the mean absolute error in frequency matching had an average value of 0.46% (standard deviation 0.20%). The run in (22) had a fairly typical mean absolute error of 0.39%.

We conclude that the algorithm's attempt to learn the patterns of free variation in Ilokano was successful.

3.6 *The Ubiquity of Free Variation*

Are the Ilokano data, with such abundant free variation, an empirical aberration? We tend to think not. For instance, our experience in working with native speaker consultants is that one virtually always finds more free variation than is given in reference sources. On a less casual basis, the research literature in sociolinguistics (e.g. Labov 1974, 1994) strongly supports the view that free variation is quite normal and characteristic of language. Therefore, in passing a representative test involving free variation, the Gradual Learning Algorithm gives evidence that it possesses a capacity that is crucial to any learning algorithm that seeks to model human abilities.

3.7 *A Comparison with the Constraint Demotion Algorithm*

In this light, we undertake in this section a comparison of the performance of the Gradual Learning Algorithm with that of Tesar and Smolensky's Constraint Demotion Algorithm. The reader should note at the outset that the Constraint Demotion Algorithm was *not designed to treat free variation*; Tesar and Smolensky are quite explicit on this point (Tesar 1995:98–101; Tesar and Smolensky 1996:28–29, 1998:249–51). Taking the Constraint Demotion Algorithm as a starting point, there could in principle be a number of ways that new algorithms inspired by it, or incorporating it, could handle variation. Our goal is to show that the Gradual Learning Algorithm is a workable solution to the research problem that Tesar and Smolensky have posed.

The Constraint Demotion Algorithm comes in a number of versions. For simplicity we focus first on the "Batch" version (Tesar and Smolensky 1993:15); so called because it processes the entire data set repeatedly. We fed to a software implementation of this algorithm the same body of Ilokano underlying forms, candidates, constraints, and violations that we had given to the Gradual Learning Algorithm. Free variation cases were treated by making multiple copies of the relevant underlying forms, assigning each a different output.

The Constraint Demotion Algorithm ranks constraints by forming a hierarchy of strata, such that any constraint in a higher stratum outranks any constraint in a lower stratum. In the batch version of the algorithm, the strata are discovered one by one in decreasing order of strictness. In our simulation, we found that Constraint Demotion began by locating the following three strata:

(23) *Strata found by Constraint Demotion*

Stratum #1: ONSET, *[$_\sigma$ʔC, MAX-IO(V), MAX-OO(ʔ), *LOWGLIDE,
IDENT-IO(low), CONTIGUITY
Stratum #2: DEP-IO(ʔ)
Stratum #3: IDENT-IO(syllabic)

These strata are in agreement with what we found in our hand ranking (20).

The Constraint Demotion algorithm then attempted to select constraints from among the remaining nine for placement in the fourth stratum; it found none. The cause of this was free variation: the multiple outputs derived from the same underlying form led the algorithm to conclude that every one of the as-yet unranked constraints was dominated. Thus, no new stratum could be formed.

The proper interpretation of this situation is partly a matter of choice. One view is that the algorithm simply fails to yield an answer in such a case. Indeed, one good property of the Constraint Demotion algorithm is that it permits a rigorous determination that it has reached this state (Tesar and Smolensky 1993:20).

Another solution that has occurred to us would be to suppose that the residue of unrankable constraints constitutes, en masse, the lowest stratum in the grammar, with stratum-internal free ranking (for definition of this see §4 below). In this approach, the result of Constraint Demotion as applied to the Ilokano case would be a four-stratum grammar, with the unrankable nine constraints construed as occupying a single fourth stratum, placed below the strata of (23). We have calculated the set of outputs that is generated by such a grammar; they are listed in (24).

(24) *Forms with a freely ranked lowest stratum*

Underlying form	Acceptable Outputs	Ill-Formed Outputs
/taʔo-en/	**taw.ʔen**, **taʔ.wen**	(none)
/paʔlak/	**pa.lak**	***paʔ.lak**
/labʔaj/	**lab.ʔaj**	***la.baj**
/trabaho/	**tra.ba.ho**	***tar.ba.ho**
/ʔajo-en/	**ʔaj.wen**	***ʔa.jwen**
/HRED-**bwaja/**	**buː.bwa.ja**, **bwaj.bwa.ja**, **bub.wa.ja**	***bwaː.bwa.ja**
/basa-en/	**ba.sa.ʔen**	(none)

As can be seen, the forms generated include several that are ill-formed in Ilokano. In other words, depending on interpretation, Constraint Demotion either fails to return a grammar for Ilokano, or it returns an incorrect one.

The above discussion holds for the Batch version of Constraint Demotion. We have also submitted the Ilokano data to other versions: On-Line (Tesar and Smolensky 1993:16), and Error Driven (Tesar and Smolensky 1998:247). These versions of Constraint Demotion behave slightly differently than the Batch version when given free variation data: they vacillate eternally in the rankings they develop, with the current state determined by whatever was heard last.

The forms generated under the vacillating sequence of rankings include all the free variants in the learning data. But they also include a substantial fraction of incorrect outputs as well, including all the ill-formed cases in (24). We conjecture that this results from a fundamental property of Constraint Demotion: it responds to input data with a radical change, namely wholesale reranking. In contrast, the Gradual Learning Algorithm responds conservatively to data, especially when plasticity is low. As a result, it avoids trouble that “reckless” Constraint Demotion cannot.

We conclude (as Tesar and Smolensky had anticipated) that none of the various versions of Constraint Demotion is suited to the analysis of free variation.[14]

3.8 *Gradience as a Means to an End*

A point worth making in this connection is that in the Ilokano simulation, the Gradual Learning Algorithm emerges with a grammar that is quite conventional in character: it designates the well-formed outcomes as well-formed and the ill-formed outcomes as ill-formed, insofar as vanishing rarity is considered as essentially equivalent to ill-formedness. Thus, in a certain sense, the end product of the gradient ranking process is not gradient at all. The near-crystalline structure of the finished grammar is created as the limit of a series of ever less-gradient grammars, when the rankings that have to be firm settle into widely separated positions while the crucially free rankings remain free.

It would appear that a statistical learning procedure may indeed be the right approach to learning optionality. As Dell (1981) pointed out, free variation poses very serious learnability problems, because one cannot know in principle whether a particular type of form might not at some point show the free variation seen in other, similar forms. The answer we offer is that a gradual algorithm, given enough time and exposure to the data, has the capacity to distinguish the accidentally missing from the systematically missing.

3.9 *Robustness in the Face of Erroneous Input Data*

It has been argued (Gibson and Wexler 1994:410, Frank and Kapur 1996:625) that learning algorithms should be robust against occasional errors in the input data. Any error that was taken too seriously by a learner might result in permanent “damage” to the grammar, placing it in a lasting state of overgeneration.

We have tested the Gradual Learning Algorithm for this possibility, with a fictional (but plausible) version of Ilokano that abstracts away from free variation, allowing only **taw.ʔen** for **/taʔo-en/** and **bwaj.bwa.ja** for /HRED-**bwa.ja**/. We confronted the Gradual Learning Algorithm with this pseudolanguage, using the same training regimen we had used for real Ilokano. After it had learned a grammar, we gave it a single additional learning token, **taʔ.wen**, which is ill-formed in this hypothetical variety.

This token had an extremely small influence on what the grammar generated, increasing the probability of a ***taʔ.wen** outcome from $1.4 \cdot 10^{-33}$ to $1.6 \cdot 10^{-33}$, and decreasing the probability of a ***tar.ba.ho** outcome from $1.1 \cdot 10^{-17}$ to $1.0 \cdot 10^{-17}$. The ranking values had already been set far apart by a long series of earlier data, so that no single token could change them enough to induce any detectable change in the output pattern.

A stricter test is to include error forms throughout the learning regimen. We tried this by including ***taʔ.wen** at random intervals at 0.1% of the frequency of **taw.ʔen**. The response of the algorithm to these errors was modest at all stages, and culminated in ‘frequency matching’ (§4): the grammar that was eventually learned generated ***taʔ.wen** at a 0.1% rate.

We administered an error test to the other algorithms as well. For the batch version of Constraint Demotion, hearing a single speech error is, of course, instantly fatal: since all data are treated equally, the error causes the algorithm to crash in exactly the way discussed in §3.7.

[14] In examining the ability of On-Line Constraint Demotion to learn freely-ranked strata, we are preceded by Broihier (1995). Broihier’s conclusions are similar to our own.

Error Driven Constraint Demotion is less fragile, but nevertheless responds to errors in rather drastic ways. When this algorithm is given ***taʔ.wen**, it carries out a major constraint demotion in order that ***taʔ.wen** will emerge as more harmonic than **taw.ʔen**. Merely hearing another token of **taw.ʔen** does not suffice to repair this damage to the grammar, because Error Driven Constraint Demotion does not reverse its prior action; instead, it carries out a new constraint demotion that generates ***tar.ba.ho**. ***tar.ba.ho**. is repaired once **tra.ba.ho** is heard, but at the cost of generating ***ʔa.jwen** (which is repaired once **ʔaj.wen** is heard, ending the chain). Moreover, the grammar that arose from hearing ***taʔ.wen** in the first place also generated ***paʔ.lak**, whose repair (by **pa.lak**) sometimes leads to ***la.baj**, depending on the order in which the forms are encountered. The upshot is that a single error can initiate a cascade of damage that is only repaired by reconstructing a large proportion of the original rankings from scratch.

The Gradual Learning Algorithm avoids such cascades by responding modestly to novel forms, merely changing its propensity to generate them, instead of giving them full credence at once. As a result, during recovery time, while the algorithm is readjusting the ranking values back to the optimum, it continues to generate acceptable outputs.

As a final comparison, we fed pseudo-Ilokano forms of the type just described, with randomly selected error forms included at a total rate of 1%, to both the Gradual Learning Algorithm and Error Driven Constraint Demotion. The Gradual Learning Algorithm yielded a stable grammar whose error rate was 0.94% after 21,000 learning data and 1.09% after 100,000 learning data (averaged over 100 replications). In other words, the error rate of the Gradual Learning Algorithm was about equal to the rate of error forms in the learning data. Error Driven Constraint Demotion produced a sequence of rapidly changing grammars, so we tested its output after every 1000 learning data, and averaged the result over a total of 10 million learning data. The average error rate came out to 2.6%, roughly 2.5 times that of the Gradual Learning Algorithm. For 16% of the total duration of the simulation, the grammar learned by Error Driven Constraint Demotion was in a state for which there was at least one correct form that it could not generate.

4 Textual Frequencies: Finnish Genitive Plurals

Anttila (1997a,b) has developed and tested an Optimality-theoretic model intended to predict the relative frequency of forms. His theory and data are of interest here, since as we have already mentioned, it is a property of grammars learned by the Gradual Learning Algorithm to mimic the frequencies of the learning set. In this section, we deploy the Gradual Learning Algorithm against Anttila's data and compare results. We also address the question of which of the two models is more likely to be generalizable across languages.

Anttila's model has a striking simplicity. As a basis for generating free variation, he assumes stratum-internal free ranking. That is, he adopts strata of the type employed by Tesar and Smolensky, but with a different interpretation: a candidate is considered to be a legal output if it emerges as the winner under *any* possible total ranking that respects the domination relations imposed by the strata.[15] To make predictions about frequency, Anttila (following Reynolds

[15] For other interpretations of "tied" constraints in Optimality Theory, see Clements (1997:315), Pesetsky (1998:372), and Tesar and Smolensky (1998:241).

1994) sums the *number of rankings* that can generate each outcome, and posits that this is proportional to the relative frequency with which the outcomes will be observed.

Anttila tests his model against a large data corpus consisting of Finnish genitive plurals. The match between the model’s predictions and the data is remarkably accurate. We attempt to model the same data here.

The Finnish genitive plural can be formed in either of two ways: with a weak ending, typically /**-jen**/, or with a strong ending, typically /**-iden**/. For instance, the stem **naapuri** ‘neighbor’ allows both endings (**náa.pu.ri.en** and **náa.pu.rèi.den**), but many other stems only allow one of the two endings or have a clear preference for one of the two. Since stems ending in a heavy syllable (CVC or CVV) invariably take the strong ending (**puu** ‘tree’ → **púi.den**; **potilas** ‘patient’ → **pó.ti.lài.den**), we will follow Anttila in considering only stems with light final syllables. According to Anttila, the choice between the weak and the strong ending is made on purely phonological grounds.

4.1 *The Constraint Inventory*

We refer the reader to Anttila’s work for full discussion of the constraints assumed. We found that we could derive the corpus frequencies accurately using only a subset of his constraints. Most of the constraints we omitted were constraints that have little support from phonological typology. These include, for example, a requirement that low vowels occur in heavy syllables, or that heavy syllables be stressless. This is not an objection to Anttila’s analysis, but merely reflects a difference of approach: Anttila emphasizes constraint inventories that include all the logical possibilities for the structures under consideration.

The constraints we did include in our replication were as follows. First, there is a correlation between weight and stress, which is given by the constraint below:

(25)

*Weight-to-Stress: “no unstressed heavy syllables” (Prince and Smolensky 1993:59)

Second, Anttila posits that, as in a number of languages, there is a connection between vowel height and stress. Following the method laid out in Prince and Smolensky (1993:67–68), this is implemented by two families of three constraints, each respecting an inherent internal ranking:

(26)

*I′ >> *O′ >> *A′: “no stressed syllables with underlying high (mid, low) vowels”
*Ă >> *Ŏ >> *Ĭ: “no unstressed syllables with underlying low (mid, high) vowels”[16]

In principle, we could have had the algorithm actively maintain these inherent rankings (by demoting a lower-ranked constraint as soon as its sister threatens to overtake it), but in fact they emerged from the data in any event.

Third, Anttila adopts some relatively standard constraints from the analysis of stress (Prince 1983, Selkirk 1984):

[16] These constraints must be interpreted in a particular fashion. In the sense Anttila intends, they refer to vowel height in the *stem*. This height is often altered phonologically in the genitive plural, as in /**kamer**a-**iden**/ → [**kámer**ò**iden**] ‘camera-GEN PL.’. Anttila’s assumption is that in this example *A′, not *O′, is violated.

(27)

*CLASH: “no consecutive stressed syllables”
*LAPSE: “no consecutive unstressed syllables”

Since according to Anttila, *CLASH is undominated, we did not include any candidates that violate it. We likewise followed Anttila in tacitly assuming constraints that ensure the invariant initial main stress of Finnish.

Lastly, Anttila posits constraints that directly regulate the weight sequence, banning consecutive syllables of the same weight:

(28)

*H.H: “no consecutive heavy syllables”
*L.L: “no consecutive light syllables”

4.2 *Anttila’s Account of Variation*

Anttila arranges his constraints (including nine we left out) into five strata. He assumes strict ranking for constraints in separate strata, and free ranking within strata. For each underlying form, there are two candidate outputs, one for each allomorph of the genitive plural. The frequency of the two rival outputs is posited to be proportional to the number of rankings (within the free strata) that generate them.

For instance, the stem **korjaamo** ‘repair shop’ has the candidates **kór.jaa.mo.jen** and **kór.jaa.mòi.den**, which have equal numbers of violations for all constraints in the top three strata. Therefore, the outcome will be determined by the constraints in the fourth stratum. Now, **kór.jaa.mo.jen** has more *LAPSE violations, and **kór.jaa.mòi.den** has more violations of *H.H, *H′, and two other constraints specific to Anttila’s analysis. When *LAPSE is on top, **kór.jaa.mòi.den** emerges; if any of the other four is on top, **kór.jaa.mo.jen** wins. With random ordering within strata, Anttila thus predicts **kór.jaa.mòi.den** in 20 percent of the cases, and **kór.jaa.mo.jen** in 80 percent. These values match well with the attested values in Anttila’s corpus, which are 17.8 and 82.2 percent, respectively. Similar cases in the data work in similar ways.

4.3 *Modeling the Finnish Data with the Gradual Learning Algorithm*

We assembled as many structurally distinct examples as we could find from Anttila’s work. For every underlying form type, we considered two output candidates, one with the weak and one with the strong genitive plural allomorph, and assessed both of these for their violations of the constraints given above. We also arranged to present the algorithm with appropriate tokens of each type, in the relative frequencies with which they occur in Anttila’s corpus. It should be noted that these frequencies often differ greatly; for example, stems of the type $/\sigma\ \sigma\ \text{H H}\ [\text{I}]_\sigma\,/$ occur only twice in Anttila’s data corpus, while stems of the type $/\sigma\ \text{L}\ [\text{A}]_\sigma\,/$ occur 720 times.

We ran the Gradual Learning Algorithm on the data. Given the goal of maximally accurate frequency matching, we felt it appropriate to use a larger number of learning data than for Ilokano. We presented a total of 388,000 data to the algorithm, expecting it to match frequencies in a fairly refined way (see Appendix A for further details of the simulation). In a representative run, the algorithm obtained the following ranking values for the constraints:

(29) *Ranking values from the Finnish simulation*

WEIGHT-TO-STRESS	288.000	*Ŏ	196.754
*I′	207.892	*LAPSE	188.726
*L.L	206.428	*O′	3.246
*Ă	199.864	*A′	0.136
*H.H	199.274	*Ĭ	–7.892

We then tested the algorithm for accuracy in mimicking the input frequencies. As before, we did multiple runs to make sure that individual runs were not yielding idiosyncratic outcomes. The numbers in table (30) reflect an average taken from 100 separate applications of the algorithm, each starting from an initial state with all constraints ranked at 100. After each run, every underlying form was submitted to the resulting grammar 100,000 times to obtain output frequency estimates. Table (30) gives the average predicted frequencies of all the various types, both as they are derived in Anttila's proposal, and as they emerge from the 100 grammars learned by the Gradual Learning Algorithm. Variation across runs of the algorithm is indicated by the standard deviations shown in the final column.

(30) *Results of learning Finnish genitive plurals*

Stem	Example	Candidates	Data	Data	Anttila	GLA	GLA
type				(%)	predicted	mean	s.d.
					(%)	(%)	(%)
XA	**kala**	**ká.lo.jen**	500	100	100	100	0
	‘fish’	**ká.loi.den**	0	0	0	0	0
XI	**lasi**	**lá.si.en**	500	100	100	100	0
	‘glass’	**lá.sei.den**	0	0	0	0	0
XLA	**kamera**	**ká.me.ròi.den**	720	100	100	99.48	0.16
	‘camera’	**ká.me.ro.jen**	0	0	0	0.52	0.16
XLO	**hetero**	**hé.te.ròi.den**	389	99.5	100	99.43	0.19
	‘hetero’	**hé.te.ro.jen**	2	0.5	0	0.57	0.19
XLI	**naapuri**	**náa.pu.ri.en**	368	63.1	67	69.51	1.16
	‘neighbor’	**náa.pu.rèi.den**	215	36.9	33	30.49	1.16
XHA	**maailma**	**máa.il.mo.jen**	45	49.5	50	42.03	2.22
	‘world’	**máa.il.mòi.den**	46	50.5	50	57.97	2.22
XHO	**korjaamo**	**kór.jaa.mo.jen**	350	82.2	80	81.61	0.92
	‘repair shop’	**kór.jaa.mòi.den**	76	17.8	20	18.39	0.92
XHI	**poliisi**	**pó.lii.si.en**	806	98.4	100	100	0
	‘police’	**pó.lii.sèi.den**	13	1.6	0	0	0
XXLA	**taiteilija**	**tái.tei.li.jòi.den**	276	100	100	99.48	0.17
	‘artist’	**tái.tei.li.jo.jen**	0	0	0	0.52	0.17
XXLO	**luettelo**	**lú.et.te.lòi.den**	25	100	100	99.44	0.19
	‘catalogue’	**lú.et.te.lo.jen**	0	0	0	0.56	0.19
XXLI	**ministeri**	**mí.nis.te.ri.en**	234	85.7	67	69.49	1.16
	‘minister’	**mí.nis.te.rèi.den**	39	14.3	33	30.51	1.16
XXHA	**luonnehdinta**	**lúon.neh.dìn.to.jen**	1	100	100	100	0
	‘characterization’	**lúon.neh.dìn.toi.den**	0	0	0	0	0
XXHO	**edustusto**	**é.dus.tùs.to.jen**	84	100	100	100	0
	‘representation’	**é.dus.tùs.toi.den**	0	0	0	0	0
XXHI	**margariini**	**már.ga.ríi.ni.en**	736	100	100	100	0
	‘margarine’	**már.ga.ríi.nei.den**	0	0	0	0	0
XXXLA	**ajattelija**	**á.jat.te.li.jòi.den**	101	100	100	99.48	0.17
	‘thinker’	**á.jat.te.li.jo.jen**	0	0	0	0.52	0.17
XXXLO	**televisio**	**té.le.vi.si.òi.den**	41	100	100	99.43	0.19
	‘television’	**té.le.vi.si.o.jen**	0	0	0	0.57	0.19
XXXLI	**Aleksanteri**	**Á.lek.sàn.te.ri.en**	15	88.2	67	69.51	1.13
	‘Alexander’	**Á.lek.sàn.te.rèi.den**	2	11.8	33	30.49	1.13
XXLHA	**evankelista**	**é.van.ke.lìs.to.jen**	2	100	100	100	0
	‘evangelist’	**é.van.ke.lìs.toi.den**	0	0	0	0	0
XXLHO	**italiaano**	**í.ta.li.àa.no.jen**	1	100	100	100	0
	‘Italian’	**í.ta.li.àa.noi.den**	0	0	0	0	0
XXLHI	**sosialisti**	**só.si.a.lìs.ti.en**	99	100	100	100	0
	‘socialist’	**só.si.a.lìs.tei.den**	0	0	0	0	0
XXHHO	**koordinaatisto**	**kóor.di.nàa.tis.to.jen**	8	80	80	81.61	0.91
	‘coordinate grid’	**kóor.di.nàa.tis.tòi.den**	2	20	20	18.39	0.91
XXHHI	**avantgardisti**	**á.vant.gàr.dis.ti.en**	2	100	100	100	0
	‘avant-gardist’	**á.vant.gàr.dis.tèi.den**	0	0	0	0	0

It can be seen that both models predict the empirical frequencies fairly well. The mean absolute error for the percentage predictions of Anttila’s model is 2.2%, whereas that for the Gradual Learning Algorithm, averaged over 100 runs, is 2.53% (s.d.=0.16%). The models share similar problems, most notably in predicting a zero percentage for **pó.lii.sèi.den**. Anttila has suggested (personal communication) that the constraint system may need amplification to achieve further accuracy. At this stage of research, we think that the performance of our machine-learned grammars may be considered to be roughly at the same level that of Anttila's hand-crafted analysis.

4.4 *Theories of Frequency*

In more general terms, we wish to consider the types of frequency distributions that the two theories (stratal grammars vs. continuous ranking) can treat. We lack firm data to decide this point, but we think we can identify the kind of data that should be considered.

We have in mind cases of free variation in which one free variant is far more common than the other. In our own speech, we have identified possible cases of this sort:

- Dutch words that normally end with final schwa are sometimes pronounced with final [n]; thus *Nijmegen* [ˈnɛimeːɣə, ˈnɛimeːɣən]. In prepausal position, [n]-less forms are far more frequent than forms with [n].
- English words ending in phonemic /...{t,d}ən/ are usually realized with the schwa elided and the /n/ syllabic, thus *Sweden* [ˈswiːdn̩]. Forms in which the schwa surfaces, like [ˈswiːdən], are quite unusual.
- English pronunciations in which /t,d/ are eligible for realization as flaps, but show up unaltered (e.g. *hitting* [ˈhɪtɪŋ]) are possible, but quite infrequent.

Let us assume for purposes of argument that the relative frequencies in these cases are 99 to 1. Now, in the grammatical model assumed by the Gradual Learning Algorithm, it is quite straightforward to model such frequencies. For instance, if each of the two rival outcomes violates just one constraint not violated by the other, a 99 to 1 ratio will be obtained whenever the ranking values of the two constraints differ by 6.58 on the ranking scale (where noise = 2.0). On the other hand, in the model Anttila assumes, such frequencies can be obtained only under very special circumstances. For instance, they would be obtained if in a single stratum 99 constraints favor one outcome and one favors the other, or if within a stratum of five constraints only one of the 120 possible total rankings gives rise to the rare outcome.

We think that for cases of the kind we have mentioned, the analysis is likely to be rather simple, with just a few constraints that do not interact in elaborate ways. Thus in general we anticipate difficulty for Anttila's model in covering cases that involve large disparities among output frequencies. In assessing the empirical adequacy of the two grammar models, careful study of cases like those we have mentioned will be necessary.

5 Intermediate Well-Formedness: Light and Dark /l/

It is a very common experience for linguists gathering intuitions about well-formedness to find that certain forms are felt to be neither impossible nor perfect, but somewhere in between. Dealing with such cases is often a theoretical challenge. When the matter is addressed, it is usually covered in terms specific to the analysis in question. We consider here a much more general explanation.

As we have both noted in earlier work (Boersma 1997; Hayes, to appear), it is likely that many intermediate well-formedness judgments originate in patterns of frequency in the learning data. Here is our reasoning:

- In the limiting case, a language learner encounters certain forms only as *speech errors*. It is clear that such forms ultimately fail to have any real influence on the grammar that emerges, despite the fact that the learner often has no way of identifying them as errors when they are encountered.
- At the opposite extreme, forms that are abundantly attested in the learning data will virtually always lead to a grammar that classifies them as well formed.
- Thus, the interesting cases are the ones that are definitely rare, but not as rare as speech errors. These are likely candidates for emergence in the adult grammar as intermediately well-formed. The language learner lacks the information that would be needed to assign them with confidence to either of the above categories, and thus rationally adopts an intermediate view.

To this basic account of intermediate well-formedness, we wish to add immediately our acknowledgment of a commonplace, namely that speakers frequently produce forms they have never heard before. In the model assumed here, this is because frequencies in the learning data have their influence at the level of grammar construction, rather than in some naive procedure that simply records the frequencies of forms in memory.

Our basic premise, then, is that intermediate well-formedness judgments often result from grammatically encodable patterns in the learning data that are rare, but not vanishingly so, with the degree of ill-formedness related monotonically to the rarity of the pattern. Therefore, with a suitable link-up one can in principle have the Gradual Learning Algorithm learn intermediate well-formedness judgments by having it learn frequencies.

To explore this possibility, we re-examine data concerning the distribution of light and dark /l/ in English from Hayes (to appear). Hayes analyzed these data in the framework of Hayes and MacEachern (1998), which, like the one assumed here, posits a continuous scale of ranking strictness. The Hayes/MacEachern model, however, is considerably less restrictive than the one adopted here: it permits individual constraints to be affiliated with “bands,” each with its own width, that specify the range of selection points. Moreover, this model permits parts of each band to be designated as “fringes,” which lead to intermediate well-formedness if a selection point falls within them. Plainly, if the less powerful theory employed here can account for the same data, this would constitute an advance, particularly since this theory (unlike the Hayes/MacEachern theory) comes with its own learning algorithm.

5.1 *The* /l/ *Data*

The data we model is a set of consultant judgments of light vs. dark /l/ in various English words. We model the means of the judgments of ten consultants, on a scale in which 1 is best and 7 is worst. The forms presented for judgment and the constraints used in modeling the data are presented in detail in Hayes (to appear) and we will only review them briefly here.

In various American English dialects, /l/ is obligatorily light in two positions: initial (*Louanne*, *light*) and pretonic (*allow*, and again *light*). It is obligatorily dark in final and preconsonantal position (*bell*, *help*). In medial, pre-atonic position there is free variation: forms like *Greeley* can have light or dark /l/.

There are also effects of morphology. Where a vowel-initial stressless suffix is added to an /l/-final stem, as in (*touchy*-)*feel-y*, one finds a fairly strong preference for dark [ɫ]. This, we assume, is a gradient effect of output-to-output correspondence: *touchy-feely* strongly prefers the dark [ɫ] inherited from *feel*. Stronger output-to-output effects occur at higher levels of the phonological hierarchy: thus the form *mail it*, with a word break, favors dark [ɫ] even more strongly than *feel-y*. Lastly, there is a paradigmatic effect that goes in the opposite direction: /l/-initial suffixes attached to vowel-final stems rather strongly prefer light [l]. Examples are *grayling, gaily*, and *freely*.

To test these claimed judgments, Hayes asked ten native speaker consultants to rate both the light and dark versions (as Hayes pronounced them) of several representative forms. The ratings that emerged were as follows:

(31) *Acceptability judgments on light and dark* /l/ *in English*

	Mean Rating as Light	Mean Rating as Dark
a. *light*	1.30	6.10
b. *Louanne*	1.10	5.55
c. *gray-ling, gai-ly, free-ly*	1.57	3.34
d. *Mailer, Hayley, Greeley, Daley*	1.90	2.64
e. *mail-er, hail-y, gale-y, feel-y*	3.01	2.01
f. *mail it*	4.40	1.10
g. *bell, help*	6.60	1.12

It can be seen that the general tendencies outlined above do indeed appear in the judgments: initial (31a,b), final (31g), pretonic (31a), and preconsonantal (31g) positions involve near-categorical judgments; medial pre-atonic position in monomorphemes (31d) yields essentially free variation; and gradient preferences, in both directions, are indeed reported where forms are morphologically (31c,e) or syntactically (31f) derived.

5.2 *The* /l/ *Simulation: Constraints*

The constraints adopted by Hayes largely follow current proposals for reference to phonetic principles in Optimality Theory (e.g. Steriade 1997, Boersma 1998, Kirchner 1998).

Dark [ɫ] is assumed to be a lenited form, with a diminished alveolar gesture. It appears as the result of a context-free lenitional constraint, stated informally as /l/ IS DARK. However, dark [ɫ], being identifiable to a large degree from its effect on a preceding vowel, is required to occur postvocalically, by a constraint termed DARK [ɫ] IS POSTVOCALIC.

Two other perceptually-driven constraints govern light [l]. They hold in contexts where the acoustic cues that render [l] identifiable are salient, thus repaying the greater articulatory effort needed. Particular salience is found pretonically, reflected in the constraint PRETONIC /l/ IS LIGHT. Note that in general, it is pretonic position in English that demands fortis allophones, such as aspirated stops, unflapped alveolars, and so on. Failing the optimal pretonic context, the second best context is prevocalic position, which is the most favored licensing position for articulatorily effortful segments crosslinguistically. This principle is reflected in the constraint PREVOCALIC /l/ IS LIGHT. As one would expect, it emerges empirically that PRETONIC /l/ IS LIGHT is ranked higher than PREVOCALIC /l/ IS LIGHT.

The constraints just given dictate the distribution of the allophones of /l/ in monomorphemic forms. PRETONIC /l/ IS LIGHT is undominated, forcing light [l] in *light*. DARK [ɫ] IS POSTVOCALIC is likewise undominated, forcing light [l] in *Louanne*. For intervocalic, pre-atonic /l/, free ranking of /l/ IS DARK and PREVOCALIC /l/ IS LIGHT results in free variation.

To model the effects of morphology, Hayes posits output-to-output correspondence constraints. Paradoxically, these constraints cannot be stated on /l/ per se. This is because in *grayling* and similar forms, the crucial light [l] does not actually occur in the base form *gray*. Hayes's approach is to base the system on the salient vowel allophones that precede a dark [ɫ] (e.g. [e͡ə] in *bail* vs. the normal [e͡ɪ] in *bay*). Since the matchup of vowel allophones to /l/ allophones is quite obligatory (***be͡ɪɫ**, ***be͡ə**), it is possible to impose output-to-output correspondence on vowel quality rather than /l/ darkness. Thus, the diphthong [e͡ɪ] in *gray* **gre͡ɪ** is normally required to appear in derived forms like *grayling* **gre͡ɪlɪŋ**, and due to allophone matching, the following /l/ must therefore be light. Likewise, the [i͡ə] in *feel* **fi͡əɫ** is normally required to appear in derived forms like *feely* **fi͡əɫi**. The output-to-output correspondence constraints that are needed are IDENT-OO(vowel features, morphological), to cover cases like *gray-ling, mail-er,* and *feel-y*; and IDENT-OO(vowel features, phrasal), to cover *mail it*. We will find that, as appears to be the norm cross-linguistically, phrasal output-to-output correspondence is ranked higher.

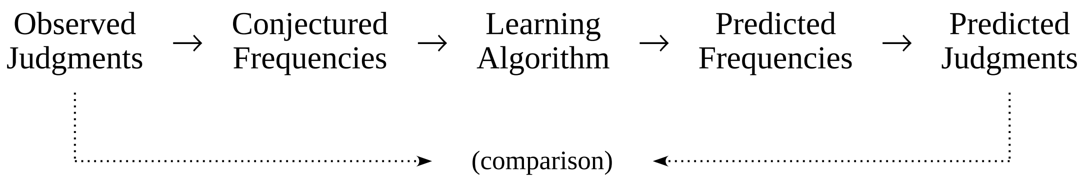
To account for the observed judgments, Hayes (to appear, §3.7) arranges the fringes of the constraints so that dark-[ɫ] ?*gray-ling* and light-[l] ?*mailer* (both taken to be rated as "?") can only be derived by making use of a "?"-class fringe, while light-[l] *mail it* (assumed to be "??") can only be derived by making use of a "??"-class fringe. Intuitively, the various cases result from the following constraint interactions: in monomorphemic forms, competition between an articulatorily-driven lenitional constraint, and various contextual, perceptually-driven constraints derives the basic patterns of obligatory and optional [l] or [ɫ]. In paradigms, the strong (but not invariant) effects of output-to-output correspondence make themselves felt as the result of semi- or near-obligatory constraint rankings, enforced with fringes.

5.3 *Learning the /l/ Pattern with the Gradual Learning Algorithm*

As noted above, our interest is in determining whether the tighter theoretical approach we adopt here, with uniform constraint strictness distributions, can provide an adequate account of the data modeled earlier by Hayes with stipulated strictness bands and fringes. We also hope to make use of the hypothesis stated above: that many intermediate well-formedness judgments are the result of frequency effects in the learning data.

To form a bridge between conjectured frequencies and well-formedness judgments, we adopt a fully explicit hypothesis in the form of equations that relate the two. These equations are presented in Appendix B. We use one equation to convert empirically-gathered gradient well-formedness judgments into conjectured frequencies, for all seven forms. These frequencies are then fed to the Gradual Learning Algorithm, which will produce (if all goes well) a grammar that closely mimics them. Then, by feeding the predicted frequencies into the mathematical inverse of the first equation, we get predicted judgments, which can then be compared with the original data. Summarizing, our simulation takes the following form:

(32)



One further complication must be dealt with before we present how the simulation came out. The frequency percentages of competing forms always sum to 100, but the sum of the light and dark judgments is not constant. This gives rise to difficulties in setting the equations, difficulties which we resolved by modeling the *differences* in judgment for light vs. dark /l/ rather than the raw judgments. This is, of course, a procedure that linguists often follow when presenting delicate cases to real consultants.

We began our simulation by converting the averaged subject data into judgment differences, then converting the differences into conjectured frequencies with equation (40) of Appendix B. The results are shown in table (33).

(33) *Converting well-formedness judgment to conjectured probability of occurrence*

Word type	Judged as light	Judged as dark	Judgment Difference	Conjectured Frequency of Light Variant
a. *light*	1.30	6.10	4.80	99.956%
b. *Louanne*	1.10	5.55	4.45	99.923%
c. *gray-ling, gai-ly, free-ly*	1.57	3.34	1.77	94.53%
d. *Mailer, Hayley, Greeley, Daley*	1.90	2.64	0.74	76.69%
e. *mail-er, hail-y, gale-y, feel-y*	3.01	2.01	–1.00	16.67%
f. *mail it*	4.40	1.10	–3.30	0.49%
g. *bell, help*	6.60	1.12	–5.48	0.0011%

We then submitted seven representative forms from (33), with relative output frequencies as given in the last column, to the Gradual Learning Algorithm. The details of the training schedule are given in Appendix A. The ranking values that emerged are given in (34):

(34) *Ranking values after English simulation*

Constraint	Ranking value
Ident-OO(vowel features, phrasal)	108.146
Dark [ɫ] is Postvocalic	107.760
Pretonic /l/ is Light	103.422
Ident-OO(vowel features, morphological)	103.394
Prevocalic /l/ is Light	100.786
/l/ is Dark	99.084

Running the grammar for 1,000,000 trials, we obtained its predicted frequencies. Lastly, we used equation (41) of Appendix B to convert the predicted frequencies back into predicted judgment differences.[17] Table (35) gives the outcomes; repeated runs of the whole simulation gave very similar results.

(35) *Results of English simulation*

Word type	Observed Judgment Difference	Projected Frequency of Light Variant	Modeled Frequency of Light Variant	Predicted Judgment Difference
a. *light*	**4.80**	99.956%	99.938%	**4.59**
b. *Louanne*	**4.45**	99.923%	99.904%	**4.31**
c. *gray-ling, gai-ly, free-ly*	**1.77**	94.53%	95.76%	**1.94**
d. *Mailer, Hayley, Greeley, Daley*	**0.74**	76.69%	72.62%	**0.61**
e. *mail-er, hail-y, gale-y, feel-y*	**–1.00**	16.67%	16.63%	**–1.00**
f. *mail it*	**–3.30**	0.49%	0.47%	**–3.33**
g. *bell, help*	**–5.48**	0.0011%	0	**–6.00**

From a comparison of the boldfaced columns, it emerges that the algorithm was able to model the well-formedness judgments with considerable accuracy. The values derived for the judgment differences differ from the human originals by an average of only 0.17.

It is also interesting to compare the pattern of ranking values obtained by the algorithm with the pattern of "fringes" posited in Hayes's hand-crafted grammar. In (36) below, the crucial forms are given along with their well-formedness value as assigned by Hayes's grammar. In the same cells is also given the distance in ranking value of the two constraints that must be ranked in "reversed" fashion, as in (4b), in order for the depicted form to be derived.

(36) *Comparison of ranking distances with fringe labels*

	Ident-OO(vowel features, phrasal)	Ident-OO(vowel features, morphological)	Prevocalic /l/ is Light
Prevocalic /l/ is Light	7.360 ??**mai[l] it**	2.608 ?**fee[l]y**	—
/l/ is Dark	—	4.310 ?**gray[ɫ]ing**	1.702 ✓**Gree[ɫ]ey**

For **Greeley**, Hayes allows the strictness bands of Prevocalic /l/ is Light and /l/ is Dark to overlap entirely, predicting two perfect outcomes. The present system essentially repeats this claim, but imposes a rather small difference in their ranking values, namely 1.702. This difference indeed corresponds to a slight preference in the consultants' judgments for light [l] in **Greeley**. The forms ?**fee[l]y** and ?**gray[ɫ]ing** are both derivable in Hayes's system by using a "?" fringe; here, they are derivable from the less-likely ranking of constraint pairs whose ranking values are 2.608 and 4.310 units apart, respectively. Again, this matches to an actual difference,

[17] This raises the question of whether real speakers likewise obtain their judgments by a process of repeated sampling. We are neutral on this point. Our crucial claim is that speakers internalize a grammar that relates well-formedness to frequency, because this is a rational learning strategy. In using their grammar to make judgments, speaker may well use mechanisms other than the Monte Carlo method.

in the predicted direction, in the consultants' judgments: ?**fee[l]y** really was felt to be better (both absolutely and relative to its counterpart) than ?**gray[ɫ]ing**. Lastly, ??**mai[l] it** is derived in Hayes's earlier system by use of a "??" strictness band; here, it is derived using the rather unlikely ranking of two constraints whose ranking values stand 7.360 units apart.

What emerges from these comparisons is that the grammar learned by the Gradual Learning Algorithm is fairly close in form to Hayes's hand-crafted grammar. But it is subtler and captures refined distinctions of judgment that elude the too-coarse categories provided by the fringe system.

The tentative conclusion we draw from this simulation is that, at least for this particular case, the theory of grammar assumed by the Gradual Learning Algorithm slices a Gordian knot. The language-specific arrangement of strictness bands and fringes posited by Hayes in his hand-crafted grammar are unnecessary. Instead, an entirely general system of gradient constraint ranking—all constraints have continuous "fringes," identical for all—suffices to handle the facts.

6 Conclusion

The Gradual Learning Algorithm has here successfully dealt with representative cases chosen to embody important challenges in the theory of learnability: free variation (§3); robustness against speech errors (§3.9); matching corpus frequencies (§4); and gradient well-formedness (§5).

Phonological learning is a difficult area, and many of its problems have yet to be fully solved. Among these are phonotactic learning (fn. 4), discovering hidden structure, relating variation to speaking style (Appendix C), and discovering language-specific constraints, if such exist. An effective constraint-ranking algorithm is likely to be only a part of the theory that ultimately emerges. We think that as it stands, however, the Gradual Learning Algorithm has some potential as a research tool, helping linguists take on new problems, especially cases involving intricate patterns of free variation and intermediate well-formedness.[18]

[18] The Gradual Learning Algorithm is available as part of the Praat speech analysis system, obtainable from http://www.fon.hum.uva.nl/praat/; and also as part of the OTSoft constraint ranking software package available at http://www.humnet.ucla.edu/linguistics/people/hayes/.

Appendix A: Training Schedules

We have little secure knowledge of how schedules for plasticity and other values affect the speed and accuracy of learning. We find that quite a few regimes lead to accurate final results, though they may differ greatly in speed.

Plasticity. A small plasticity value does a better job of matching learning data frequencies in the end, but a large plasticity value nears its goal faster. The virtues of the two approaches can be combined by adopting a learning schedule that decreases the plasticity as learning proceeds. This seems in principle realistic: in humans, grammar apparently stabilizes in adulthood, as nonlexical learning slows or halts.

Evaluation Noise. We also find that letting the evaluation noise (σ in (5) above) diminish during the course of learning improves accuracy, particularly in establishing large differences in ranking values between constraints that ultimately must be ranked categorically. At any given stage of learning, however, the evaluation noise is kept the same for all constraints.

Details of Individual Simulations. All learning schemes involved a sequence of stages, in which the number of forms digested per stage was usually set at 1000 times the number of underlying forms. The scheme for the Ilokano simulation was as follows:

(37) *Training schedule for Ilokano*

Data	Plasticity	Noise
First 7000	2	10
Second 7000	0.2	2
Last 7000	0.02	2

The final noise level of 2 was considered to be the permanent value characteristic of the adult system, and was accordingly used (here as elsewhere) to measure the output distribution of the final grammar.

The training regimes for Finnish genitive plurals and English light and dark /l/ were as in (38):

(38) a. *Training schedule for Finnish*

Data	Plasticity	Noise
First 22,000	2	10
Second 22,000	2	2
Third 22,000	0.2	2
Fourth 22,000	0.02	2
Last 300,000	0.002	2

b. *Training schedule for English*

Data	Plasticity	Noise
First 6,000	0.2	2
Second 300,000	0.02	2
Last 300,000	0.002	2

In these simulations, we used more extended training regimens, since we had a different purpose in mind. For Ilokano, we had been curious to see how few forms it would take for the grammar to achieve a state of high accuracy. For Finnish and English, we sought to model the mature adult state, which occurs after extensive learning has provided sufficient exposure even to very rare forms.

We found that training schedules different from the above produce results that may be less accurate, but only slightly so. For example, when we used the Finnish regimen for Ilokano, the number of illegal forms generated went up from 1 per 100,000,000 (§3.5) to 1 per 2,000,000, though frequency matching improves from 0.46% to 0.10%. When we submitted the English forms to the Finnish regimen, we found that the average error in predicting judgment differences went up from 0.17 to 0.89; the increased error resulted mainly from assigning categorically bad judgments to dark [ɫ] in *light* and *Louanne*, and sometimes to light [l] in *mail it*, i.e., many of the forms that were rated “??” by humans are learned as “*” judgments instead.

Time to Convergence. One can ask whether the amount of data that must be fed to the algorithm to obtain accurate results is excessive in comparison with what real learners are likely to encounter during the acquisition period. We feel that the numbers we used are probably acceptable. It should be recalled that most constraints have considerable generality and are instantiated by large numbers of words. Given the many thousands of words heard by a child in an average day, there is reason to believe that real-life learning data are copious enough to support learning with the Gradual Learning Algorithm. For some estimation of convergence times in general, see Boersma (1998:328).

The Course of Learning. It is also worth considering the route that the Gradual Learning Algorithm takes in arriving at the final set of ranking values. In chart (39), we give the output distributions for the intermediate grammars obtained during the course of the Ilokano simulation. Each output distribution was calculated by running every underlying form through the grammar 1,000,000 times, using the value for noise that was in effect at the relevant stage of learning.

(39) *The stages of learning Ilokano*

Surface form	Target language	Initial state	After 1000 data (noise 10)	After 7000 data (noise 10)	After 7000 data (noise 2)	After 14,000 data (noise 2)	After 21,000 data (noise 2)	After 121,000 data (noise 2)
taw.ʔen	50%	2.7%	38.0%	53.8%	76.0%	50.0%	52.2%	48.9%
taʔ.wen	50%	2.7%	38.2%	42.9%	24.0%	50.0%	47.8%	51.1%
ta.wen	0	3.2%	5.4%	0.5%	0	0	0	0
ta.ʔen	0	29.4%	1.3%	0.0002%	0	0	0	0
ta.ʔo.en	0	29.4%	0.8%	0.0001%	0	0	0	0
ta.ʔo.ʔen	0	29.4%	12.4%	2.7%	0	0	0	0
ta.ʔwen	0	3.2%	3.8%	0.2%	0	0	0	0
buː.bwa.ja	33.3%	6.7%	29.9%	13.6%	0.02%	26.6%	36.7%	32.9%
bwaj.bwa.ja	33.3%	47.1%	28.0%	77.6%	99.98%	44.8%	31.2%	33.7%
bub.wa.ja	33.3%	20.8%	41.3%	8.7%	0	28.6%	32.1%	33.4%
bwaː.bwa.ja	0	16.3%	0.7%	0.004%	0	0	0	0
baː.bwa.ja	0	9.1%	0.1%	0.005%	0	0	0	0
pa.lak	100%	53.3%	79.2%	98.8%	100%	100%	100%	100%
paʔ.lak	0	23.3%	19.2%	1.2%	0	0	0	0
pa.ʔlak	0	23.3%	1.5%	0.0005%	0	0	0	0
lab.ʔaj	100%	50.0%	95.5%	99.8%	100%	100%	100%	100%
la.baj	0	50.0%	4.5%	0.2%	0	0	0	0
tra.ba.ho	100%	66.7%	80.3%	99.2%	100%	100%	100%	100%
tar.ba.ho	0	33.3%	19.7%	0.8%	0	0	0	0
ʔaj.wen	100%	7.5%	95.5%	99.5%	100%	100%	100%	100%
ʔa.jen	0	28.4%	0.001%	0	0	0	0	0
ʔa.jo.en	0	28.4%	0.001%	0	0	0	0	0
ʔa.jo.ʔen	0	28.4%	0.05%	0	0	0	0	0
ʔa.jwen	0	7.5%	4.4%	0.5%	0	0	0	0
ba.sa.ʔen	100%	31.0%	58.0%	96.9%	100%	100%	100%	100%
ba.sen	0	31.0%	11.7%	1.0%	0	0	0	0
ba.sa.en	0	31.0%	8.5%	0.7%	0	0	0	0
bas.a̯en	0	3.5%	6.0%	0.7%	0	0	0	0
bas.wen	0	3.5%	15.9%	0.7%	0	0	0	0
Mean absolute error		36.5%	8.6%	3.9%	6.3%	0.79%	0.39%	0.11%
Outliers		58.8%	16.5%	1.3%	0	0	0	0

The chart illustrates the strategy that was used to obtain quick and accurate learning for Ilokano: the initial stage of high noise (see column 5) created ranking values that, after noise was reduced to 2 (column 6), eliminated the possibility of generating illegal forms. Frequency matching at this stage was poor, but improved after further learning (columns 7 and 8). Learning trials going well beyond what is reported in the main text (column 9) led to further improvement in modeling frequency.

The point of this exercise is to show that, even near the beginning of its efforts, the algorithm generates languages that are already coming to resemble the target language. The data also show that an early "boost period" with high noise can be helpful in excluding outliers; it remains to be seen if evidence will ever be found that humans behave analogously.

Appendix B: Equations Relating Well-Formedness and Frequency

We seek a concrete implementation of the general idea laid out in (32).

First, we need an equation that maps an observed judgment difference to a conjectured fraction of light forms (the conjectured fraction of dark forms is simply one minus this value). Let ΔJ be the observed judgment for dark forms minus the observed judgment for light forms. Average values for ΔJ were shown in column 4 of table (33). To convert this into a conjectured fraction of light forms, we perform a sigmoid transformation:

(40)
$$\mathit{conjectured\ fraction\ of\ light\ forms} = \frac{1}{1 + 0.2^{\Delta J}}$$

Values obtained from this equation appear in the last column of table (33).

Second, we need an equation that maps the frequency F_p of light forms predicted by the Gradual Learning Algorithm (table (35), column 4) to a predicted judgment difference. For this, we carry out the mathematical inverse of the sigmoid transformation (40), which is:

(41)
$$\mathit{predicted\ judgment\ difference} = \frac{\log\left(\dfrac{1}{F_p} - 1\right)}{\log 0.2}$$

The values obtained from this equation appear in the last column of table (35).

Here are examples illustrating what the equations claim. If the judgment of light [l] in some context is a perfect 1, and dark [ɫ] is the opposite extreme of 7, then it is hypothesized that in the learning data that gave rise to these judgments, light [l]'s outnumber dark by a factor of 15,625 to one; a ratio that would likely permit the learner to consider any dark [ɫ] as a speech error. If light and dark /l/ are judged rather closely, say 2 vs. 3, the equations claim that in the learning data they occurred at a conjectured ratio of five to one. Note that this arrangement attributes some reasonable sagacity to language learners: it would be a rash learner that concluded that there is anything seriously wrong with a form that occurs in a sixth of all cases.

Appendix C: Stylistic Variation

The research literature in sociolinguistics clearly shows that variation in language reflects distinctions of casual vs. formal style. A full model of variation would have to reflect this, and in this section we offer a brief speculation on how the model given here could be extended in an appropriate way.

We assume that utterances occur in contexts that can be characterized along a casual/formal continuum. We quantify this continuum with a variable *Style*, such that *Style* equals 0 in maximally casual speech and 1 in maximally formal speech. The selection point for a given constraint C_i is determined by equation (42):

(42)
$$\mathit{selectionPoint}_i = \mathit{rankingValue}_i + \mathit{styleSensitivity}_i \cdot \mathit{Style} + \mathbf{noise}$$

This is the same equation as before, augmented by the term $\mathit{styleSensitivity}_i \cdot \mathit{Style}$, in which $\mathit{styleSensitivity}_i$ is a constraint-specific value. Constraints with positive values for *styleSensitivity* take on higher ranking values in formal speech; constraints with negative values for *styleSensitivity* take on higher ranking values in casual speech, and constraints with zero values for *styleSensitivity* are style-insensitive.

We conjecture that the initial stages of acquisition are insensitive to style. Under this condition, all values of *styleSensitivity$_i$* are zero, and acquisition can proceed with the Gradual Learning Algorithm as described in the main text. Later, as the language learner becomes aware of the stylistic context of utterances, she learns to associate variation in selection points with style. In this view, the appropriate research strategy is to develop a learning algorithm that can learn both the ranking values and the values of *styleSensitivity* for each constraint, given a set of utterances and their affiliated values for *Style*. We defer discussion of such an algorithm to later work.

References

Anttila, Arto. 1997a. Variation in Finnish phonology and morphology. Doctoral dissertation, Stanford University, Stanford, Calif.

Anttila, Arto. 1997b. Deriving variation from grammar: A study of Finnish genitives. In *Variation, change and phonological theory,* ed. Frans Hinskens, Roeland van Hout, and Leo Wetzels. Amsterdam: John Benjamins. Rutgers Optimality Archive ROA-63, http://ruccs.rutgers.edu/roa.html.

Archangeli, Diana, and D. Terence Langendoen. 1997. *Optimality Theory: An overview*. Oxford: Blackwell.

Barry, Martin. 1985. A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics* 4:1–16.

Benua, Laura. 1995. Identity effects in morphological truncation. *Papers in Optimality Theory* [*Occasional Papers* 18], ed. Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, 77–136. Amherst: University of Massachusetts.

Benua, Laura. 1997. *Transderivational identity: Phonological relations between words.* Doctoral dissertation, University of Massachusetts, Amherst. Rutgers Optimality Archive ROA-259, http://ruccs.rutgers.edu/roa.html.

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. *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*. Doctoral dissertation, University of Amsterdam. The Hague: Holland Academic Graphics.

Boersma, Paul. To appear. Learning a grammar in Functional Phonology. In *Optimality Theory: Phonology, syntax, and acquisition,* ed. Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer. Oxford: Oxford University Press.

Broihier, Kevin. 1995. Optimality theoretic rankings with tied constraints: Slavic relatives, resumptive pronouns and learnability. Ms. MIT. Rutgers Optimality Archive ROA-46, http://ruccs.rutgers.edu/roa.html.

Burzio, Luigi. 1996. Multiple correspondence. Ms. Johns Hopkins University, Baltimore, Md. Available from http://www.cog.jhu.edu/~burzio/burzio.html.

Clements, George N. 1997. Berber syllabification: Derivations or constraints? In *Derivations and constraints in phonology,* ed. Iggy Roca, 289–330. Oxford: Clarendon Press.

Dell, François. 1981. On the learnability of optional phonological rules. *Linguistic Inquiry* 12:31–37.

Dinnsen, Daniel A. 1985. A reexamination of phonological neutralization. *Journal of Linguistics* 21:265–279.

Frank, Robert, and Shyam Kapur. 1996. On the use of triggers in parameter setting. *Linguistic Inquiry* 27:623–660.

Gibson, Edward, and Kenneth Wexler. 1994. Triggers. *Linguistic Inquiry* 25:407–454.

Gnanadesikan, Amalia. 1995. Markedness and faithfulness constraints in child phonology. Ms. University of Massachusetts, Amherst. Rutgers Optimality Archive ROA-67, http://ruccs.rutgers.edu/roa.html.

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

Hayes, Bruce. 1989. Compensatory lengthening in moraic phonology. *Linguistic Inquiry* 20:253–306.

Hayes, Bruce. 1999. Phonological acquisition in Optimality Theory: The early stages. Ms. UCLA. Rutgers Optimality Archive ROA-327, http://ruccs.rutgers.edu/roa.html.

Hayes, Bruce. To appear. Gradient well-formedness in Optimality Theory. In *Optimality Theory: Phonology, syntax, and acquisition,* ed. Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer. Oxford: Oxford University Press.

Hayes, Bruce, and May Abad. 1989. Reduplication and syllabification in Ilokano. *Lingua* 77:331–374.

Hayes, Bruce, and Margaret MacEachern. 1998. Quatrain form in English folk verse. *Language* 74:473–507.

Kager, René. 1999a. Surface opacity of metrical structure in Optimality Theory. In *The derivational residue in phonology,* ed. Ben Hermans and Marc van Oostendorp, 207–245. Amsterdam: John Benjamins.

Kager, René. 1999b. *Optimality Theory: A Textbook*. Oxford: Oxford University Press.

Kenstowicz, Michael. 1997. Uniform exponence: Extension and exemplification. In *Selected papers from the Hopkins Optimality Workshop 1997* [*University of Maryland Working Papers in Linguistics* 5], ed. Viola

Miglio and Bruce Morén, 139–154. Revised version to appear in *Bolletino della Societa Linguistica Italiana.* Rutgers Optimality Archive ROA-218, http://ruccs.rutgers.edu/roa.html.

Kiparsky, Paul. 1973. Phonological representations. In *Three dimensions in linguistic theory,* ed. Osamu Fujimura. Tokyo: TEC Co.

Kiparsky, Paul. 1998. Paradigm effects and opacity. Ms. Dept. of Linguistics, Stanford University, Stanford, Calif.

Kirchner, Robert. 1996. Synchronic chain shifts in Optimality Theory. *Linguistic Inquiry* 27:341–350.

Kirchner, Robert. 1998. An effort-based approach to consonant lenition. Doctoral dissertation, UCLA, Los Angeles. Rutgers Optimality Archive ROA-276, http://ruccs.rutgers.edu/roa.html.

Labov, William. 1974. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1994. *Principles of linguistic change*. Vol. 1. *Internal factors*. Oxford: Blackwell.

Liberman, Mark. 1993. Optimality and optionality. Ms. University of Pennsylvania, Philadelphia.

McCarthy, John. 1996. Remarks on phonological opacity in Optimality Theory. In *Studies in Afroasiatic grammar. Proceedings of the Second Colloquium on Afro-Asiatic Linguistics, Sophia Antipolis.* Ed. Jacqueline Lecarme, Jean Lowenstamm, and Ur Shlonsky, 215–243. The Hague: Holland Academic Graphics.

McCarthy, John. 1999. Sympathy and phonological opacity. *Phonology* 16.

McCarthy, John, and Alan Prince. 1993. Generalized alignment. *Yearbook of Morphology 1993*, ed. Geert Booij and Jaap van Marle, 79–153. Dordrecht: Kluwer.

McCarthy, John, and Alan Prince. 1994. The emergence of the unmarked: optimality in prosodic morphology. *Papers of the 24th Annual Meeting of the North Eastern Linguistic Society*, ed. Mercè González, 333–379. Amherst, Mass.: Graduate Linguistic Student Association.

McCarthy, John and Alan Prince. 1995. Faithfulness and reduplicative identity. In *Papers in Optimality Theory* [*Occasional Papers* 18], ed. Jill Beckman, Laura Walsh Dickey, and Suzanne Urbanczyk, 249–384. Amherst: University of Massachusetts.

Nagy, Naomi and Bill Reynolds. 1997. Optimality Theory and word-final deletion in Faetar. *Language Variation and Change* 9:37–55.

Nolan, Francis. 1992. The descriptive role of segments: evidence from assimilation. In *Papers in laboratory phonology II: Gesture, segment, prosody*, ed. Gerard Docherty and D. Robert Ladd, 261–280. Cambridge: Cambridge University Press.

Pesetsky, D. 1998. Some optimality principles of sentence pronunciation. In *Is the best good enough? Optimality and competition in syntax*, ed. Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, 337–383. Cambridge, Mass.: MIT Press.

Prince, Alan. 1983. Relating to the grid. *Linguistic Inquiry* 14:19–100.

Prince, Alan, and Paul Smolensky. 1993. *Optimality Theory: Constraint interaction in generative grammar.* Rutgers University Center for Cognitive Science Technical Report 2.

Prince, Alan, and Bruce Tesar. 1999. Learning phonotactic distributions. Rutgers Center for Cognitive Science, Technical Report TR-54. Rutgers Optimality Archive ROA-353, http://ruccs.rutgers.edu/roa.html.

Pulleyblank, Douglas, and William J. Turkel. 1995. Asymmetries in feature interaction: Learnability and constraint ranking. Ms. University of British Columbia, Vancouver.

Pulleyblank, Douglas, and William J. Turkel. 1996. Optimality Theory and learning algorithms: the representation of recurrent featural asymmetries. In *Current trends in phonology: Models and methods*, ed. Jacques Durand and Bernard Laks, 653–684. University of Salford.

Pulleyblank, Douglas, and William J. Turkel. 1998. The logical problem of language acquisition in Optimality Theory. In *Is the best good enough? Optimality and competition in syntax*, ed. Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky, 399–420. Cambridge, Mass.: MIT Press.

Pulleyblank, Douglas, and William J. Turkel. To appear. Learning phonology: Genetic algorithms and Yoruba tongue root harmony. In *Optimality Theory: Phonology, syntax, and acquisition*, ed. Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer. Oxford: Oxford University Press.

Reynolds, William. 1994. Variation and phonological theory. Doctoral dissertation, University of Pennsylvania, Philadelphia.

Selkirk, Lisa. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, Mass.: MIT Press.

Smolensky, Paul. 1996. The initial state and 'Richness of the Base' in Optimality Theory. Rutgers Optimality Archive ROA-154, http://ruccs.rutgers.edu/roa.html.

Steriade, Donca. 1997. Phonetics in phonology: The case of laryngeal neutralization. Ms. UCLA, Los Angeles.

Steriade, Donca. 2000. Paradigm uniformity and the phonetics-phonology boundary. In *Papers in laboratory phonology V*, ed. Michael B. Broe and Janet B. Pierrehumbert, 313–334. Cambridge: Cambridge University Press.

Tesar, Bruce. 1995. Computational Optimality Theory. Doctoral dissertation, University of Colorado.

Tesar, Bruce, and Paul Smolensky. 1993. The learnability of Optimality Theory: An algorithm and some basic complexity results. Ms. Department of Computer Science and Institute of Cognitive Science, University of Colorado at Boulder. Rutgers Optimality Archive ROA-2, http://ruccs.rutgers.edu/roa.html.

Tesar, Bruce, and Paul Smolensky. 1996. Learnability in Optimality Theory (long version). Technical Report 96-3, Department of Cognitive Science, Johns Hopkins University, Baltimore. Rutgers Optimality Archive ROA-156, http://ruccs.rutgers.edu/roa.html.

Tesar, Bruce, and Paul Smolensky. 1998. Learnability in Optimality Theory. *Linguistic Inquiry* 29:229–268.

Tesar, Bruce, and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.

Zubritskaya, Katya. 1997. Mechanism of sound change in Optimality Theory. *Language Variation and Change* 9:121–148.

Boersma:
Institute of Phonetic Sciences
University of Amsterdam
Herengracht 338
1016CG Amsterdam, The Netherlands
paul.boersma@hum.uva.nl

Hayes:
Department of Linguistics
UCLA
Los Angeles, CA 90095-1543
bhayes@humnet.ucla.edu