An Automated Learner for Phonology and Morphology

Adam Albright Bruce P. Hayes

Dept. of Linguistics

UCLA

August 1998, Revised January 1999

1. Purpose

This document is a summary, for ourselves and those who are curious, of the current state of our Phonological Learner. The Learner is the centerpiece of our current research project; it is a computer program whose purposes is to learn morphophonemic systems from input data, and to serve as a tool for modeling phonological and morphological knowledge in humans.

2. Rationale

Linguists of many persuasions take a realist view of linguistic theory as it relates to learning: language learners, in infancy and childhood, are assumed to come equipped with whatever principles of linguistic structure and of language learning are biologically determined in our species. They encounter and process numerous input data, provided essentially at random, from the ambient language. In a few years, this processing results in mental structures—grammar, phonology, lexicon—of extraordinary scope and intricacy. Ideally, linguists seek general principles of language that guide children in their search, and sometimes they attempt to determine the particular algorithms and strategies that children might be using.

But examination of the current literature in theoretical phonology suggests that, with certain significant exceptions,[1] phonological work is not crucially guided by considerations of learnability. The primary goal is usually to characterize phonological patterns in a cogent, theoretically insightful way, by invoking principles that have great generality in the treatment of phonological systems cross-linguistically. On the whole, we think this has been a good strategy, and one which has yielded a great deal of progress. Nevertheless, we question whether such a strategy pursued alone is likely to yield a phonological theory that is fully compatible with considerations of learning. Quite a few phonological analyses we admire incur a tacit intellectual debt, concerning how the rules and representations they posit can be learned.

[1] Among others, Dell 1981, Dresher 1981, McCarthy 1981b, Dresher and Kaye 1990, Tesar and Smolensky 1993, 1998, and Pulleyblank and Turkel 1997.

Our view is that it would be worthwhile for at least a subset of phonological theorists to model speakers instead of languages. In this approach, phonological analyses are not invented by linguists, but must be learned by algorithm, from a representative set of input data. The linguist is responsible for the algorithm, and the algorithm models the language through the grammar it learns. In this way, every analysis produced would address learnability directly, since the algorithm that generates the analysis in effect forms the learnability theory for that analysis.

We anticipate that such a research strategy could benefit the development of phonological theory, in the following ways. First, we have preliminary indications that certain elements of current theory will prove to be essential to a successful learner. If this is so, we would be moving to a novel kind of evidence for theories, a crucial supplement to the traditional research strategy. Second, it is possible that certain *other* proposals in phonological theory would show themselves to be an impediment to learning. These are the proposals that produce a fatal explosion in the size of the hypothesis space, rendering learning impossible. If an otherwise-effective learner is impeded by such proposals, we could get a kind of counterevidence not otherwise obtainable.

In addition, any learner which is successful in modeling the human ability to learn a wide variety of different languages is potentially useful to the field of machine learning, where learners are typically designed to handle a small number of languages at a time. Clearly, for specific tasks the theoretically guided learner of a phonologist could never outperform a learner optimized for a practical application by an engineer. Nevertheless, fields such as artificial intelligence may one day be in a position to use flexible, widely adaptable language learning strategies which are more human-like and less machine-like.

Finally, we think that pursuing learners is worthwhile for its own sake. Our efforts so far lead us to think that producing a good learner is a fascinating challenge, one which requires its own theoretical development, not just appropriate incorporations of existing phonological theory.

3. How Our Current Learner Works: Outline

Our learner was first programmed in early 1997 by Hayes. This Learner relied heavily on strict constraint ranking, and sought to develop deeply-stratified Optimality Theoretic grammars, using the Constraint Demotion Algorithm of Tesar and Smolensky (1998).[2] In Spring 1997 Hayes invited Adam Albright, then a first-year graduate student, to join him as an equal collaborator on the project. Albright ported the program from VisualBasic into Java, and we began testing the Learner on a variety of small pseudo-languages. We found through experience that this approach was too fragile in the face of irregularities in the input data (which, of course, occur frequently in real languages). The early Learner failed to return multiple guesses for forms

[2] This original learner used a greedy top-down algorithm, searching the entire data set for the most general descriptions first and then learning the exceptions, much along the lines of the FOIDL framework (Mooney and Califf 1996, Dzeroski and Erjavec 1997).

where humans return multiple guesses, it failed to capture the Island of Reliability effect, and indeed it often failed to converge to a working grammar at all.[3]

Based on our experiences with this initial learner, in Winter 1998 we began work on a new version of the Learner. This Learner incorporates ideas of others, notably MacWhinney 1978, Pinker and Prince 1988, Ling and Marinov 1993, and Mikheev 1995. More distantly, it owes a debt to the tradition of connectionist models beginning with Rumelhart and McClelland 1986, which gave impetus to the idea of studying linguistic systems by means of learning algorithms.

The leading ideas that guide the Learner so far are:

- Discovery of the truth about a language via **hypothesis growth plus selection**: faced with data, the Learner formulates a quite large number of hypotheses, and then checks how well they do. When asked to perform (for example, to inflect novel stems), the Learner makes use of the best hypotheses it has developed so far.
- Formulation of hypotheses through **minimal generalization**: the Learner generally tries to find the tightest hypothesis that covers the range of relevant data. This helps insure that crucial small-scale hypotheses, that need to be fed to the selection process, are not overlooked. It also means that there is always a working grammar, that covers all the known data, in place at all times.
- The discovery of phonology and morphology are **mutually bootstrapped**: tentative hypotheses for one permit learning in the other, and vice versa.

These ideas will be fleshed out as the exposition proceeds.

To begin: like many of the models just noted, our Learner approaches the task of learning phonology under the idealization of *paradigm mapping*: given one member of an inflectional paradigm (say, an English present tense verb), the Learner must guess the form of another paradigm member (say, the past tense).[4] To do this, the Learner must carry out morphological and phonological analysis.

Following earlier proposals, our Learner begins by construing representative inflectional data pairs as mappings, projecting one from another. In a fictional language (modeled on Finnish), the mappings assumed for the direction "present → past" for four sample verbs might be:

[3] We think this resulted from a kind of anorexia. The earlier Learner constructed grammars in a top-down fashion, so that it digested new data only when it thought that the existing grammar could not handle them. But since noise in the data made it impossible to know at any given point what the "right" constraint ranking was, the early Learner couldn't actually determine with any accuracy whether it really could account for an input datum. When it guessed that it could, but actually couldn't, anorexia result (the Learner refused to eat) and it accordingly died, or at least failed.

[4] Note that other idealizations exist. For example, Touretzky et al. 1991 assume a theory in which the learner is given underlying-surface pairs.

$[\text{istu}]_{\text{present}} \rightarrow [\text{istui}]_{\text{past}}$ $[\text{vipy}]_{\text{present}} \rightarrow [\text{vipyi}]_{\text{past}}$
$[\text{lupu}]_{\text{present}} \rightarrow [\text{lupui}]_{\text{past}}$ $[\text{katap}]_{\text{present}} \rightarrow [\text{katapi}]_{\text{past}}$

These mappings are then factored into **structural change** and **context**. Thus, for example, $[\text{istu}]_{\text{present}} \rightarrow [\text{istui}]_{\text{past}}$ can be represented as ∅ → i / [istu ___]. As such mappings accumulate, the Learner attempts to generalize. When the structural changes of any two mappings are the same, a generalized mapping can be obtained by retaining all shared context material, and replacing all non-shared material with a string variable: (The exact procedure will be covered in more detail in §4.2)

I. ∅ → i / [ist u ___]
+ ∅ → i / [lup u ___]
= ∅ → i / [X u ___] = "Suffix /-i/ to /u/-final stems."

II. ∅ → i / [istu ___]
+ ∅ → i / [katap ___]
= ∅ → i / [X ___] = "Suffix /-i/ to the stem."

The mechanism for detecting context (e.g., the /u/ in I above) is superfluous in the present case, but is crucial for the commonplace phenomenon of arbitrary phonologically-conditioned allomorphy; for example, special allomorphs that attach only to vowel-final stems. Two examples of this are the Korean nominative suffix, and the Yidiɲ ergative suffix:

Korean nominative: ∅ → i / [XC___]
∅ → ka / [XV___]

Yidiɲ ergative: ∅ → du / [XC___]
∅ → ŋgu / [XV___]

In both of these cases, the learner must recognize that one allomorph can attach only to vowel-final stems, while the other can attach only to consonant-final stems.

The choice of what contexts can be searched is of course a central issue of phonological theory, and we are exploring various options concerning this point.

Given the coarse, preliminary morphological analysis that this process yields, further data processing can then lead to phonology. We lay out the method below in stages.

3.1 Finding the Default Mapping

Often there are multiple string mappings that derive the same inflectional type for different stems, e.g. English *creep ~ crept*, *beep ~ beeped*. Research of the past decade by Steven Pinker and his colleagues (e.g. Pinker and Prince 1988) emphasizes the importance of the **default** mapping, which for example is X → Xd for English present → past. The Pinkerian research program has shown that there are important psycholinguistic properties shown by the default. Noting that the default cannot be innate, Marcus, Pinker et al. 1992 suggest a number of ways the default could be learned.

The strategy we adopt is, as noted above, "hypothesis growth plus selection". We first allow a large number of morphological mappings to be generated (by the mechanism given above), then let them compete empirically. The default is what emerges as the winner of this competition. The criterion of victory in the competition is the degree to which the language learner can *trust* the rival mappings in projecting novel inflected forms. In this sense, our system presupposes that people are "rational" learners, who maximize the chance that any novel forms they have to generate will be correct.

We posit that Confidence can be assessed by keeping track of some crucial numbers. The **scope** of a morphological mapping is the number of forms (types, not tokens) in which the structural description of a mapping is met. The **reliability** of a mapping is the fraction of forms where a mapping's structural description is met and it projects the right form. Intuitively, reliability is what makes a mapping trustable. However, reliability based on high scope (say, 990 correct predictions out of 1000) is better than reliability based on low scope (say, 5 out of 5). We follow a formula from Mikheev 1995 (described in section 4.6) that carries out this adjustment by means of lower confidence limit statistics. The result is a single value of **Confidence**, ranging from zero to one. We suggest, tentatively, that the default mapping for any given context is the one with the best Confidence.

Having found the default, our algorithm *retains* the other mappings. We think this is a correct decision. Much experimental work in "Wug" testing (Berko 1958), where speakers are asked to inflect imaginary stems, shows that speakers are willing to provide multiple guesses; thus for English "*spling*" the past tenses [splɪŋd], [splæŋ], or [splʌŋ] are given. In our model, these guesses arise from the conflicting claims of rival mappings, with the better of the smaller-scale mappings competing with the default. Thus, in contrast with the Dual Mechanism Model of Pinker and Prince 1991, our system does the work of the default rule and the hypothesized "associative network" in a single formal component. Our view of the Dual Mechanism Model is as follows: the evidence adduced in favor of it is valid and important, but it is worthwhile to make every effort to explain this evidence with a single mechanism before complicating the system with a second mechanism. Therefore, the research strategy here is to look for ways in which two types of behavior might be derived from a single mechanism. In particular, we conjecture that at some point, perhaps rather late in the game, language learners may cease to memorize inflected forms that can be derived by the default mapping. However, the model would continue to derive defaults through competition within a single mechanism. If this approach is workable (and inspection of the Dual Mechanism Model literature suggests to us that it is), then it would seem desirable to pursue unitary models as far as possible.

3.2 Projecting Phonology

A crucial part of our Learner is that it attempts to “bootstrap” the phonology and the morphology off of one another. Tentative guesses about morphology spawn guesses about phonology. Then, with some phonology in place, the performance of the morphological system is improved. Our specific strategy is to take highly trustable mappings, and try applying them to *all* stems meeting their structural description. For many stems this obtains the right answer, but for others the answer is wrong. The reason, very often, is that the system needs to take phonology into account.

Imagine, for instance, that in the hypothetical language discussed above, the sequence *[ei] is ill-formed, and is adjusted to [i] whenever the morphology would generate it. Our Learner discovers this phonology by applying the default mapping X → Xi to stems ending in /e/, such as /viljele/. By comparing the wrong output, *[viljelei], with correct [viljeli], it is in a position to discover the phonology of /e/ deletion.

A crucial question is: how is the Learner to know that *[ei] is ill-formed? We have tried two approaches. One is to scan the wrong output *[viljelei] for sequences absent in the language in general; the general absence of *[ei], together with the fact that it is “repaired” when it arises from concatenation, constitutes good evidence of its ill-formedness. More recently, we have implemented a separate program to learn the pattern of phonotactic well-formedness independently, prior to morphological analysis.[5]

Armed with the knowledge that *[ei] is illegal, and that the incorrect projected form *[viljelei] contains *[ei], our Learner is then able to discover the phonological mapping e → ∅ / ___ i. Note that although this particular account adopts a parochial phonological rule, our long-range plan is to implement the phonology under the approach of Optimality Theory (Prince and Smolensky 1993). In the present case, this would involve introducing MAX(e) into the system, and ranking it below the ban on *[ei].

Once phonology has been found, our Learner updates the Confidence of the morphological mappings. Typically, the Confidence of the default mapping will rise, since the earlier “counterexamples” are now understood to follow from the application of phonology to the output of morphology.

We believe that learning the morphological mappings and the phonological mappings of a language are both inherently interesting and difficult problems. One weakness of many previous computational models is that they have been interested mainly in one problem or the other. For

[5] While both methods work reasonably well, we currently use the second, under the view that it is more likely to approximate the experience of human children. Our assessment of the acquisition literature (e.g. Werker and Tees 1984, Jusczyk et al. 1993, Jusczyk et al. 1994) is that children probably know a great deal about the system of contrast and phonotactics in a language before they begin to analyze phonological alternations (Hayes, in progress).

example, computational models of morphology typically treat phonology simply as a confounding factor which impedes the morphological parser (for example, see Mikheev 1997). As a result, the mechanisms which are designed to "undo phonology" tend to be either too powerful (i.e., consider many implausible changes) or too weak (i.e., consider only a subset of attested changes). As phonologists, we may be in a position to make a novel contribution to this problem, by suggesting how a learner could search for phonological changes, and what sort of changes are reasonable or not. (See §§4.8-4.9 for more details.)

3.3 Underlying Forms

Our Learner does not reduce stems to a single underlying form; rather, it rationalizes the paradigm by finding how (and to what extent) paradigm members can be projected from one another. Speakers are assumed to memorize not underlying forms, but rather **at least enough surface forms** of each stem so that missing paradigm elements can be projected from what is already known (Burzio 1996; Hayes, in press). Although this is a controversial approach, it is gradually becoming more plausible in light of recent work in theoretical phonology (particularly work on Correspondence constraints within paradigms, by Burzio, Steriade 1996, Kenstowicz 1997, Benua 1997 and others). However, we are not dogmatic on this point, and we are open to the possibility of revising the system to discover underlying forms (or at least, small collections of allomorphs; see section 7.6 below).

3.4 Testing the Learner

An adequate learner captures all generalizations captured by a native speaker exposed to similar data. One way to see what generalizations the computational learner has captured is to manually inspect the rules that it has devised. However, it is impossible to gain such direct access to the rules which the native speaker has devised! Therefore, this approach only lets us compare the model's knowledge against what we, as linguists, think is a reasonable statement of the human speaker's knowledge. We have found that it is extremely useful to be able to do this so we can understand what the learner is doing and whether the results look reasonable, but it is ultimately not the best way to compare the computer's knowledge with the human's knowledge.

A better strategy is to check the learner's behavior against human behavior: we invent forms to elicit that will test whether a learner (either ours, or a human) has internalized a particular generalization, and check if data elicited from the Learner and from humans matches. For example, following a suggestion of Halle 1982, we would regard it as quite fair to ask the Learner, once it had pondered a selection of English plurals, to form the plural of *Bach* [bɑx]. To the extent that English speakers really do consistently come up with [bɑxs], we would require of our Learner that it do the same. We believe that assiduous testing of this type, on both ordinary and exotic word shapes, and with frequent comparison to the judgments of human speakers under careful elicitation conditions, is probably the best way to test the adequacy of a learner.

3.5 Generality

Generality pervades linguistic theorizing. Under the older approach of generative lingusitics, which models the language rather than the speaker, generality actually serves as a means of evaluating the success of an analysis. Other criteria for evaluating analysis include the notion of "insight"; as well as, of course, empirical adequacy.[6]

From the perspective of modeling the speaker, generality takes on a very different role: it is a **means to an end**. If a learning model fails to locate highly general statements of patterns that real speakers apprehend, that learning model will fail. The forms on which it will fail are as follows: they are covered by a generalization that is discovered by people, but it not included in the patchwork of specific hypotheses produced by the failed learner. As a result, native speakers will Wug-test differently on such forms.

This is, in essence, is the heart of the criticism made by Pinker and Prince 1988 against the connectionist past-tense learner of Rumelhart and McClelland (1986) : Pinker and Prince's close inspection showed that the Rumelhart/McClelland learner *failed to generalize* at the level needed to model humans adequately.

3.6 Why Generality is Not Always a Good Thing: Richness in Grammars

Thus, our criterion of generality (i.e., of useful generality) is an operational one. Interestingly, this criterion appears to be a two-edged sword. As we (begin to) document below, linguists have characteristically eschewed formulating rules that are quite specific, particularly when they share the same structural change as a general rule. For example, the economical linguist would not posit an English past tense rule adding [d] specifically to -r final stems, when there is already a general rule adding it to all stems:

specific: $\varnothing \rightarrow$ d / Xr___#
general: $\varnothing \rightarrow$ d / X___#

From a superficial viewpoint of analysis-evaluation, such rules look like warts on the face of the analysis. But in fact, there is reason to think that people learn such rules in abundance. As we try to show below (see also Hayes, forthcoming, and Albright 1998), grammars are rich: learners seize on whatever generalizations will help them to project novel forms acccurately. There is no reason to think that all such generalizations are of very broad scope; the narrow-scope generalizations help, too.

The way our Learner mimics this ability is founded on the particular conception of generality it has. By the process of free combination of existing mappings that share structural changes, our Learner produces essentially all the relevant hypotheses it can, at all levels of generality. Those hypotheses that are maximally effective (have high Confidence values) are

[6] Chomsky and Halle (1968) are very careful to distinguish its notion of generality, which is a technical concept of their system, from the intuitive notion of generality that is used in evaluating scientific hypotheses. However, it is probably fair to say that intuitive generality still plays a role in *SPE*, notably in characterizing the notion of "linguistically significant generalization."

important, no matter what their generality level. Where generality is high, the hypotheses are useful for their ability to project a form even where there is little direct analogy among the existing forms (thus, Pinker and Prince's 1988 *anatomosed* and *ploamphed*). Where generality is low, the mappings enable us to model the "island of reliability effect", laid out in detail in section 6.2 below. Both are important, and if we lacked for either one, our Learner would fail to mimic people.

In summary: we think that grammars should be large, that they should include many mappings of many different levels of generality, and that the criterion for what mappings the grammars most "believes" is the Confidence index mentioned above.

4. How Our Learner Works: Details

4.1 Input Diet

Children hear words more or less at random, one assumes. Plausibly, multiple hearings help to get a clearer idea of what the word means, to be certain as to its phonemic form, and just to lay down enough memory traces.

Normally, we idealize away from this, simply letting the learner process all the data in one sitting, with perfect memory for each form. However, in our simulation of the U-shaped Curve (section 6.3), we do account in a somewhat more realistic way for input order and for frequency effects. Where one is modeling adult competence, it would seem that batch processing is not so far off track: adults plausibly have chewed on the whole dataset as much as is needed, and have extracted essentially all the generalizations that they are ever going to extract.[7]

4.2 Generalization by Collapsing: The Details

As noted above, we discover general morphological mappings by "collapsing" together mappings of less generality. This produces a large "workbench-ful" of mappings, of which the best are then identified by assessing their empirical effectiveness. Here, we describe the collapsing process in greater detail.

4.2.1 Structural Changes

Two morphological mappings can be collapsed into a more general mapping only if they have the same structural change. In certain cases, this is tricky to establish, because a form can be ambiguous with regard to what its structural change is (see section 4.5 below). For now, however, let us assume unambiguous structural changes, and attempt to be fully explicit on how a structural change is located.

[7] Another benefit of this approach is that we may one day be able to model individual differences in how much abstraction is performed over the input data. Although current linguistic theory tends to assume that linguistic abstraction is an automatic and in some sense deterministic behavior, there is some evidence in the psychological literature (Medin, et al 1983) that abstraction is not always automatic, and the amount of abstraction we do may be influenced by task or by temperament.

In most general terms, we have morphological mappings of the form:

$X \rightarrow Y$

where X and Y are strings. Our Learner *parses* X and Y, i.e. breaks them up into relevant chunks. This is done in a way that maximizes a sequence P which is left-justified and identical in X and Y; and a right-justified sequence Q which is right-justified and identical in X and Y. We use “A” to designate what is left over in X, and “B” to designate what is left over in Y; thus:

$X = PAQ$ P, Q maximized
$Y = PBQ$ P, Q maximized
$X \rightarrow Y = PAQ \rightarrow PBQ$

or:

$A \rightarrow B / P \_\_\_ Q$

Note that if P or Q is zero, we have to *say so*: this means that the mapping depends on A being edge-adjacent. For example, if P = null, the notation commonly used is:

$A \rightarrow B / [\_\_\_ Q$

and similarly if Q is null.

Note further that either A or B can be zero. Where A is zero, and edge-adjacent, we are dealing with an affixational mapping. Where B is zero and edge-adjacent, we are dealing with some sort of truncation; e.g. the mapping from English plurals to singulars. Where neither A or B is zero, we are dealing either with two paradigm members that each have their own affix, or cases of ablaut or similar nonconcatenative morphology. Lastly, it is plausible that both A and B are null. Here the variables P and Q collapse to a single string variable X, and we have: $X \rightarrow X$. Such a mapping would cover cases like Latin dative plural → ablative plural; the two can be realized in a variety of ways, but they are always identical (Aronoff 1994).

Structural changes can also involve features, as in umlaut or mutation phenomena. We have not yet implemented this, but logically it seems straightforward. Assuming we have so far arrived at a parse into structural description and change in which A and B are both single segments, we can then “find a bit more context” by looking for *simultaneous* context, using the feature system. Below this is done by parsing the segments A and B into the features for which they share or don’t share values.

PAQ → PBQ

A = [αSharedFeature1, βSharedFeature2, γSharedFeature3 ... δUnsharedFeature1, εUnsharedFeature2]

B = [αSharedFeature1, βSharedFeature2, γSharedFeature3 ... -δUnsharedFeature1, -εUnsharedFeature2]

Thus:

$$\begin{matrix}\delta\text{UnsharedFeature1}\\ \varepsilon\text{UnsharedFeature2}\\ \ldots\end{matrix} \rightarrow \begin{bmatrix}-\delta\text{UnsharedFeature1}\\ -\varepsilon\text{UnsharedFeature2}\\ \ldots\end{bmatrix} / \quad P \underset{\ldots}{\begin{bmatrix}\rule{2.5cm}{0.4pt}\\ \alpha\text{SharedFeature1}\\ \beta\text{SharedFeature2}\\ \gamma\text{SharedFeature3}\end{bmatrix}} \quad Q$$

This maximizes unchanged material, and minimizes the structural change, to the greatest possible degree, under the classical *SPE*-like representations assumed.

4.2.2 Finding Context in Generalized Mappings

Assuming, then, that we've reached the point where two morphological mappings have been located that share a structural change, how do we know how to collapse them? Further, how do we collect the *context* that is crucial to so many morphological mappings (i.e., allomorphy with phonological conditioning)?

What is crucial here is a scheme to factorize mappings into their crucial component parts. Once these parts are located, collapsing is relatively straightforward.

Let us consider first only cases in which contexts are construed as segment sequences, rather than feature bundles; hence PAQ → PBQ. Plausibly, the search for context should proceed *outward* from the locus of the structural change. This makes sense, since morphological and phonological contexts are characteristically local. For more on nonlocal contexts, see section 7.3 below.

Suppose we have parsed two different morphological mappings and found that they share their structural change (A → B):

PAQ → PBQ (or: A → B / P ___ Q)
P′AQ′ → P′BQ′ (or: A → B / P′ ___ Q′)

We can then compare the two mappings further, parsing their context terms (P and Q) as follows:

PAQ	=	Presidue	Pshare	A	Qshare	Qresidue
P′AQ′	=	P′residue	Pshare	A	Qshare	Q′residue

Pshare and P′share are the maximal right-justified substrings of P and P′, respectively, that are identical to each other. Likewise, Qshare and Q′share are maximal left-justified substrings of Q and Q′, respectively, that are identical to each other. Presidue, P′residue, Qresidue, and Q′residue are the material left over from P, P′, Q, and Q′ when the maximal Pshare and Qshare strings have been extracted from them. By maximizing Pshare and Qshare, we seek to formulate the most specific mapping that covers both of the original two mappings.

The remaining task is to specify the actual form of the collapsed constraint. Collapsing requires us to introduce **variables** to cover material that can differ. Specifically, we include a free variable Pvariable in place of where Presidue and P′residue occur; and a free variable Qvariable in place of where Qresidue and Q′residue occur.

Note that the only way that corresponding residues may be the same is if they are both null. This is because if they were non-null but identical, they would have been included in Pshare and Qshare, which are maximized.

The end product of collapsing will fit this schema:

	(Pvariable)	(Pshare)	A	(Qshare)	(Qvariable)
→	(Pvariable)	(Pshare)	B	(Qshare)	(Qvariable)

This can also be stated in a more traditional SPE-style format, using X and Y to respresent variables:

A (B / X[Pshare] [Qshare]Y

It will not be the case that both Pvariable and Qvariable are both null (otherwise, we would have had the very same constraints in the first place; everything else has been maximized under identity). But Pshare and Qshare might both be null; this will be true, for instance, when we've collapsed our way down to straightforward, across-the-board suffixation, which looks like this:

	Pvariable	∅
→	Pvariable	B

4.2.3 Finding Context in Generalized Mappings That Include Features

It is very common for context in morphological mappings to be stated in featural terms; for instance, in section 3.6 we saw the case of the Korean nominative ending, which is "attach the suffix /-i/ to stems that end in a consonant ([-syllabic]), else attach the suffix /-ka/." This pattern also occurs in Japanese,Turkish, Yidiɲ, and others. While in some cases the choice of allomorph might be attributed to general phonological constraints such as *Coda, in other cases there is little choice but to seek the phonological environment directly. Thus, for example, there seems little hope that we can find a principled reason why English ɪ → ʌ in past tenses occurs most productively before /ŋ/ (Bybee and Moder 1983, Prasada and Pinker 1993); instead, we must endow our Learner with the ability to discover this fact from inspection of the data.

What is at issue is, given the extreme number of hypotheses that arise when one looks at features: what is the window within which one looks at the features of the relevant segments?

The potential size of this window (measured in number of hypotheses) is large, as can be seen in the following way. Consider first just one side of the environment bar. Let there be n segments, each with m features. Since each feature can be included or omitted from a generalized structural description, we have roughly 2n*m possible generalized structural descriptions; this number should then be multiplied by whatever obtains for the right-side context, roughly 2*n*m. Taking 3 as a reasonable guess for n and 20 as a reasonable guess for m, this number comes out to be about one trillion, a truly dismaying number of hypotheses. (The exact size of the hypothesis space will be slightly different, owing to the possibility of replacing whole segment strings by variables, but it will be roughly of the same order).

Assuming, then, that utterly unconstrained search of enormous hypothesis spaces is not likely to be feasible, we have arrived at a fundamental question: how large is the phonological hypothesis space that human learners are able to consider, and what is its shape and dimensions?

At present, our Learner incorporates a guess about the answer to this question that is probably not fully correct, but handles a wide array of cases without too much of a hypothesis-space explosion. Since the hypothesis space of structural descriptions is a crucial issue in phonology, we intend to explore further elaborations (see, for instance, section 4.9 below). Our current hypothesis states that constraints may be collapsed together according to the following, seven-position schema. The schema presupposes two mappings PAQ → PBQ' and P'AQ → P'BQ':

Parse of PAQ → PBQ:

PSegment Residue	PShared Features	PShared Segments	A	QShared Segments	QShared Features	QSegment Residue
	PFeature Residue				QFeature Residue	

Parse of P’AQ’ → P’BQ’:

P’Segment Residue	PShared Features	PShared Segments	A	QShared Segments	QShared Features	Q’Segment Residue
	P’Feature Residue				Q’Feature Residue	

Intuitively, this says: “going outward in both directions from the locus of the structural change, find as many shared segments are you can, then in the next segment, find as many shared features as you can.”

An actual case that covers most of the terms in our schema is English present to past [sprɪŋ] → [sprʌŋ], as collapsed with [strɪŋ] → [strʌŋ]:

Parse of PAQ (PBQ for [sprɪŋ] → [sprʌŋ] :

PSegment Residue	PShared Features	PShared Segments	A	QShared Segments	QShared Features	QSegment Residue
	PFeature Residue		(→ B)		QFeature Residue	
s	[-son] [-cont] [-voice] [-del rel] [-nasal]	r	I	N	∅	∅
	[+labial]		(→ ʌ)		∅	

Parse of P′AQ′ → P′BQ′ for [strɪŋ] → [strʌŋ]:

P′Segment Residue	PShared Features	PShared Segments	A	QShared Segments	QShared Features	Q′Segment Residue
	P′Feature Residue		(→ B)		Q′Feature Residue	
s	[-son] [-cont] [-voice] [-del rel] [-nasal]	r	ɪ	ŋ	∅	∅
	[+cor] [+anter]		(→ ʌ)		∅	

Once we have these factorizations, we can do what we did before: produce a generalized mapping, including all shared material, and incorporating variables for all non-shared material.

$$\text{A} \rightarrow \text{B} / [\text{X} \left[\begin{array}{l}\text{SharedFeature1}\\ \text{SharedFeature2}\\ \text{SharedFeature3}\\ \ldots\end{array}\right] \text{PSharedSegments} \_\_\_ \text{QSharedSegments} \left[\begin{array}{l}\text{SharedFeature4}\\ \text{SharedFeature5}\\ \text{SharedFeature6}\\ \ldots\end{array}\right] \text{Y}]$$

Note the presence of variables on either end of the collapsed mapping; a variable here simply indicates that the matrix of features included in the structural description need not be at the edge of the word (indicated here by [ ]).

For our *spring/string* example, the collapsed mapping would look like this:

$$\text{ɪ} \rightarrow \text{ʌ} / [\ \text{X} \left[\begin{array}{l}\text{-sonorant}\\ \text{-continuant}\\ \text{-voice}\\ \text{-delayed release}\\ \text{-nasal}\end{array}\right] \text{r} \_\_\_ \text{ŋ}\ ]$$

This is, in words, "/ɪ/ is replaced by /ʌ/ when preceded by a (not necessarily initial) voiceless stop followed by /r/, and followed by word-final /ŋ/." This is the maximal collapsing of *spring* ~ *sprung* with *string* ~ *strung* that is compatible with our assumptions.

In terms of algorithms, the parsing of mappings into these schemata can be carried out as follows:

- Scan leftward starting at the segment that precede the matched structural changes of the two mappings (A).
- Seek the longest full matches of segments along this scanning direction. This yields PSharedSegments.
- If there remain segments not yet matched, take the next one and collect all features for which the two mappings share a value (the SharedFeature$_{1\text{-}n}$ above).
- All non-matched features are, in effect, replaced by a variable. But because of the ordinary conventions of how feature matrices in general structural descriptions are interpreted, no overt variable is shown in the schema above.
- If there remain any further segments in the scan, use a string variable (X above) to complete the collapsing of the structural descriptions of this side.
- Repeat analogously, in the right-to-left direction, for the strings on the other side of the structural change.

4.3 Is our "Window" Large Enough?

The mappings *spring ~ sprung* and *string ~ strung* are similar in a way that goes beyond what the schema above implies: in both cases, the crucial voiceless stop is preceded by /s/. Thus, ideally, we would have wanted:

$$\text{ɪ} \quad \rightarrow \text{ʌ} \ / \ [\ \text{s} \begin{bmatrix} \text{-sonorant} \\ \text{-continuant} \\ \text{-voice} \\ \text{-delayed release} \\ \text{-nasal} \end{bmatrix} \text{r} \ \_\_\_ \ \text{ŋ}\ ]$$

But capturing this similarity goes beyond the capacity of our classificatory scheme.

This is an alarming point, given that Bybee and Moder have demonstrated some productivity for the *s + stop + liquid* pattern in this form of past tense. A similar phenomenon that has attracted our attention is the projection of "glew" as a past tense of "glow". This is found in quick responses to experimental probes, and indeed appeared as a considered judgment when we once asked it of a six-year-old consultant. The crucial learning data that presumably drive people to this intuition are these forms:

blow - blew
grow - grew
throw - threw[8]

It seems possible that the crucial schema here is: [features for stops/obstruents?][features for liquids] + /ou/. To discover this would require a left-side window consisting of two slots on which featural analysis can be done.

We have not expanded our Learner's structural windows to the point of discovering these generalizations, but anticipate that it will be necessary. What is crucial is to devise a form of expansion that covers that attested cases, but it not so wide as to expand the hypothesis space fatally.

4.4 What Should be the Input To Generalization?

If one is to start with "degenerate mappings" (i.e. rules that handle one stem), and discover usefully general mappings from them by comparison, then one needs to know what gets compared with what.

[8] And perhaps: *draw - drew, fly - flew, slay - slew*; see section 7.8, on product-oriented schemata.

As we write, the Learner is undergoing testing and improvement for phonological features. Features radically change the basis of comparison. First, in case it is interesting, we give a discussion of the old Learner, in which forms are only strings of arbitrary symbols.

4.4.1 Generalization When There are Only Strings

Our view is that in such a Learner, it suffices to compare degenerate mapping with degenerate mapping. It is not necessary to compare degenerate mappings with generalized mappings; nor is it necessary to compare generalized mappings with generalized mappings. To see why, consider the following sequence of degenerate mappings. For clarity, non-shared material, and the variables that replace it, are shown in bold. Variables are in bold, and material that gets collapsed into variables is underlined.

MNPAQR$\underline{\text{U}}$	→	MNPBQR$\underline{\text{U}}$	degenerate mapping
MNPAQR$\underline{\text{S}}$	→	MNPBQR$\underline{\text{S}}$	degenerate mapping
MNPAQR**X**	→	MNPBQR**X**	result of comparison and generalization

PAQ$\underline{\text{T}}$	→	PBQ$\underline{\text{T}}$	another degenerate mapping
$\underline{\text{MN}}$PAQ$\underline{\text{RX}}$	→	$\underline{\text{MN}}$PBQ$\underline{\text{RX}}$	general mapping from line 3 above
YPAQ**X**	→	**Y**PBQ**X**	result of comparison and generalization

Crucially, one could have obtained the same result from comparing two degenerate mappings:

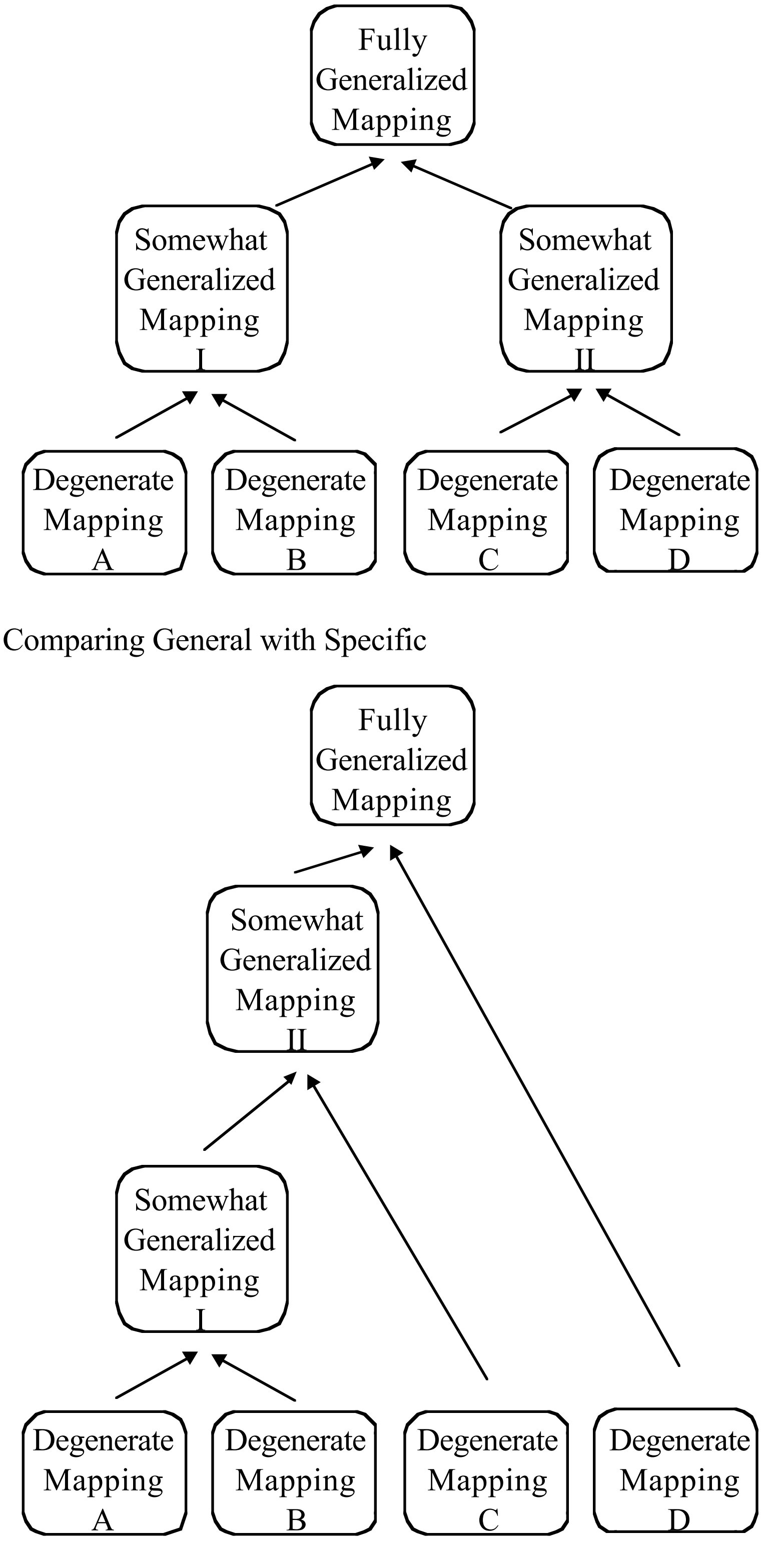
$\underline{\text{MN}}$PAQ$\underline{\text{RU}}$	→	$\underline{\text{MN}}$PBQ$\underline{\text{RU}}$	first degenerate mapping
PAQ$\underline{\text{T}}$	→	PBQ$\underline{\text{T}}$	third degenerate mapping
YPAQ**X**	→	**Y**PBQ**X**	result of comparison and generalization

This is because comparison with a mismatched string yields no different results from comparison with a string variable; in either case it is a string variable that results.

4.4.2 Generalization When Features are Present

To save computational time, we are relying here on a conjecture that seems reasonable, but it not yet proven, namely that it suffices to compare already-existing generalized constraints with the "degenerate" constraints read directly off the input lexical items. The reason to think this is true is as follows. The process of collapsing together constraints to achieve generalized constraints can only delete information from structural descriptions, never add it. Generalized constraints simply embodied the shared information from the structural descriptions that were collectively used in creating them. Thus, every generalized constraint can be formed in an order that merely adds in its individual "degenerate" ancestors one at a time.

The idea is illustrated as follows, with two possible creations of a generalized constraint from four original "degenerate" ancestor constraints.



a. Comparing General with General

b. Comparing General with Specific

Fully Generalized Mapping
Somewhat Generalized Mapping I
Somewhat Generalized Mapping II
Degenerate Mapping A
Degenerate Mapping B
Degenerate Mapping C
Degenerate Mapping D
Fully Generalized Mapping
Somewhat Generalized Mapping II
Somewhat Generalized Mapping I
Degenerate Mapping A
Degenerate Mapping B
Degenerate Mapping C
Degenerate Mapping D

Based on this reasoning, we think it should be possible to find all of the generalized mappings without ever comparing general with general mappings. Since this saves quite a few computational steps, this is what we have currently implemented.

4.4.3 Comparing General and Specific Mappings: The Gruesome Details

The process of comparing a general and a specific mapping is somewhat tedious, because there are so many logically different cases, and often variables must be collapsed out of variegated origins. But since it has to be gotten right, here (for the perusal of interested parties) are the details of what our Learner does.

It should be remembered that the collapsing of environments may, in principle, take place on both sides of the structural change. It is also possible that when a general constraint (i.e., one containing a variable) is collapsed with a specific constraint, it will be "general" only on one side. Therefore, although the following description considers just the left side for expository simplicity, it is important to remember that each side must be considered separately.

The problem with comparing a generalized constraint with a specific one is that they contain different numbers of terms in their factorization; the generalized constraint may contain features and residue, while the specific constraint has only strings:

Generalized constraint:	P residue	P Features	P Segments	A→B	Q Segments	Q Features	Q residue
Specific constraint:			P′ Segments	A→B	Q′ Segments		
Goal:	P residue	PShared Features	PShared Segments	A→B	QShared Segments	QShared Features	Q residue

The scheme, as always, is to start at the change (A→B) and work our way outwards. For the P side, we start by comparing the P Segments of the generalized constraint with the P′ side segments of the specific constraint. Moving outward from the structural change, first find as many shared segments as possible. If you find a mismatch before you run out of segments on either side, then compare the next segment on both sides and find as many shared features as possible.

The plethora of sub-cases comes about when you run out of segments in P Segments or P′ Segments (i.e., if they are different lengths, and one completely subsumes the other). If the P Segments of the generalized constraint is longer than the P′ Segments of the specific constraint, then we have run out of material in the specific constraint; the PShared Segments for the new constraint will be equal to the P′ Segments, the PShared Features will be null, and the P residue will be true (because the generalized constraint still had material left over).

A similar case arises whenthe P Segments of the generalized constraint is the same length as the P′ Segments of the specific constraint, and the two fully match. In this case, PShared features will also be null. If the generalized constraint had active P Features, or if it had P residue, then the P residue of the new constraint will be true; otherwise, it will be false.

The third possibility is that the P′ Segments of the specific constraint is longer than the P Segments of the generalized constraint. In this case, we need to compare the P Features of the generalized constraint with the next segment of the P′ Segments, to build the PShared Features of the new constraint. In this case, the new constraint will have a P residue either if there are still segments left unconsidered in P′ Segments of the specific constraint, or if there was a P residue in the generalized constraint. Otherwise, both constraints have been consumed, and there is no P residue in the new constraint.

The same procedure is also carried out on the right side, to determine the QShared Segments, the QShared Features, and the Q residue of the new constraint.

4.5 Forms Ambiguous for Structural Change

Most form-pairs yield mappings (PAQ → PBQ) that can be factored uniquely into structural description and structural change.[9] But where PAQ or PBQ contain *repeated* sequences that overlap with what is changing, the factorization can be non-unique. This is problematic, since we need to know the structural change to see what the mapping constraint can be collapsed with. In this section, we discuss the problem and possible remedies.

4.5.1 An Example of Ambiguous Structural Change

A simple example arose in German when we mapped infinitives to 1 plural pasts. Where the verb stems ends in an alveolar stop, as in *falten* [faltən] ‘fold-inf.’ → *falteten* [faltətən] ‘we folded’, this is: Xt-ən → Xt-ət-ən.

The three ways of construing this as a structural change are:

P	A	Q	B = tə
fal	∅	tən	

P	A	Q	B = ət
falt	∅	ən	

P	A	Q	B = tə
faltə	∅	n	

[9] At least, most form-pairs that differ in just one location. See §4.5.3 below for other cases.

In the present case, taking all three of these seriously can lead to trouble. In particular, the hypothesis that the mapping involves insertion of /tə/ collides with another misparse, that involving the /-əl/ and /-ər/ stems. Here, the infinitive ending is /n/, not /ən/, as in *klingeln* /klɪŋəl–n/. Given that the 1 plur. suffix is /ən/ after the past ending /t/, we obtain another case that give rise to the appearance of /tə/ insertion:

klɪŋəln → klɪŋəltən ‘ring’

P	A	Q	B = tə
klɪŋəl	∅	n	

This mapping is in and of itself pretty harmless, since it will collapse only with other /əl/ and /ər/ stems, yielding accurate results. But when collapsed with the mapping we obtained with *falten*, we get a problem:

P	A	Q	B = tə
klɪŋəl	∅	n	

P′	A	Q′	B = tə
fal	∅	tən	

yielding:

	(Pvariable)	(Pshare)	A	(Qshare)	(Qvariable)
→	(Pvariable)	(Pshare)	B	(Qshare)	(Qvariable)
	X	l	∅		Y
			tə		

which is a mapping that licenses the insertion of /tə/ after any /l/. Needless to say, this is a very bad mapping, yielding for instance **[foltəgən] from [folgən] ‘to follow’.

There may be ways of surmounting this problem. For example, ∅ → tə / Xl___Y might be spawned but denied admission to the “permanent grammar”, as discussed below. However, ideally we would like to exercise beneficial controls on the process of mapping-generation wherever such controls might help.

4.5.2 Using Typology: Peripheral Changes

Consider English past tenses formed on /Xd/ stems, like *wedded*. The string mapping here is:

wɛd → wɛdəd

which has two possible parses:

P	A	Q	B = əd
wɛd	∅	∅	

P	A	Q	B = də
wɛ	∅	d	

The first is /-əd/ suffixation, the second /-də-/ infixation.

It seems a safe strategy here to assume only the suffixational mapping. Certainly, suffixation is more common than infixation,[10] so the suffixation guess will statistically speaking be right more often than not. Moreover, the strategy of favoring "adfixation" (prefixation or suffixation) is, as far as we can tell, free of the danger of not finding a valid infixational mapping.

Let us consider why, by imagining a true infixation that sometime looks like adfixation. Assuming an infix *-in-*, placed before the first vowel of the stem, we can imagine a stem *nupa*, whose infixational mappings (*nupa* → *n-in-upa*) can be parsed in the following ways:

P	A	Q	B = ni
∅	∅	nupa	

P	A	Q	B = in
n	∅	upa	

Of these, only the first is kept by a stategy that favors adfixation. But if the language in question really has infixation, there will be plenty for forms like *bama* ~ *binama*, *tugu* ~ *tinugu* whose structural changes are completely unambiguous. These will generalize and form the mapping:

P	A	Q	B = in
[-syllabic]	∅	X	

[10] A caution here: our system treats mappings that place one affix between another affix and the stem as "infixation", a case that traditional morphology considers to be simply successive affixation of the inner, then the outer affix. Whether a better approach to morphological parsing will let these cases disappear remains to be seen.

And once this mapping exists, it will correctly derived *ninupa*, even though *ninupa* did not participate in the construction of this mapping.

Let us now dispose of a couple of residual worries. First, suppose there were an infixational mapping that *always* involved an ambiguous structural change; for example, if *-in-* only infixed to bases that started with /n/. But in this case, a grammar that posits only *ni-* prefixation will be **empirically successful,** and the worry reduces to a very philosophical one: the Learner fails to learn one of the two grammars that could in principle yield speaker behavior that is otherwise identical. The grammars are:

I. CX → CinX
II. CX → CinX
  nX → ninX

We are struggling enough with problems that directly impinge on empirical adequacy to worry ourselves about this one.

It might also happen that the consonant of a -VC- infix happens to be the only consonant in the language that has a particular feature value. If, say, the language has only one [-consonantal] consonant, namely /w/, then data like:

pata	piwata
lupu	liwupu
semi	siwemi
wafa	wiwafa
wolu	wiwolu
wume	wiwume

will lead to an oddly bifurcated grammar. Generalization over features finds that /-iw-/ is inserted after the natural class of true consonants (i.e. [+consonantal], which includes all consonants but /w/). However, for /wX/ stems, the preference for adfixation we posit would force the assumption of /wi-/ prefixation before wX stems. This is a bit inelegant, but it too would produce outputs indistinguishable from a general mapping infixing /-iw-/ after the first consonant. Again, we propose to restrict our attention to problems that have some chance of being empirical.

4.5.3 Dealing with Remaining Cases

The previous section noted that we can cut down on ambiguity, with little harm to the final grammars, by assuming a prejudice for adfixational mappings. But even adfixational mappings can have ambiguous structural changes. Consider the following case:

tata → tatata

This mapping is ambiguous between two different adfixational structural changes:[11]

P	A	Q	B = ta	(suffixation of /-ta/)
tata	∅	∅		

P	A	Q	B = ta	(prefixation of /-ta/)
∅	∅	tata		

Therefore, an a priori prejudice for adfixation does not settle the issue of non-unique structural changes. The same point emerges for competing medial changes if one considers a mapping like *patatapa* → *patatatapa.* This mapping is six way ambiguous for its structural change:

p at atatapa
pa ta tatapa
pat at atapa
pata ta tapa
patat at apa
patata ta pa

and all of the changes are medial.

We can see three primary strategies for dealing with such irredeemably ambiguous cases.

- **Many mappings**: insert *all* of the possible mappings into the system (and see how they fare in the competition among mappings).
- **No mappings**: do not posit any new mappings when a form is ambiguous
- **Some mappings**: insert a subset of the possible new mappings, on some principled basis

[11] There are, moreover, three infixational parses, but we have eliminated these already.

All these strategies are risky, in some way. Many mappings risks the possibility that outlandish mappings will spawn constraints that yield “stupid” guesses in Wug testing.[12] No mappings yields the possibility that a mapping will be missed entirely, where it is not recoverable from unambiguous forms. “Some mappings” shares part of the defects of both.

The present version of the Learner, rather arbitrarily, uses a version of the “some mappings” strategy. Specifically, when a new form yields multiple structural changes, and the criterion of favoring changes in peripheral position does not decide, then the Learner posits the single mapping that maximizes the term P in the schema PAQ. Thus for [faltən] → [faltətən], the Learner posits only the structural change ∅ → tə / [faltə___n], and not ∅ → ət / [falt___n] or ∅ → tə / [fal___ tən]. For the mapping [tata] → [tatata], the Learner posits ∅ → ta / [tata___].

We acknowledge the arbitrariness of this solution, whose only merit is that it recognizes the general (if slight) preference for suffixation in languages. We anticipate a better answer will come from improved understanding of how to weed out “stupid” mappings from the system, or perhaps from a better understanding of morphological parsing in general.

The solution does have purely computational advantages: in searching for structural changes, if one has first attempted to maximize P, and finds that Q is null, then the search may terminate, since the result suffixing mapping is the one that will be kept. Also, allowing the system a “bias” towards prefixes or suffixes may be useful if there were a mechanism to identify at a global level whether the language was a primarily prefixing language or a suffixing language, since this would clearly guide the learner to choose one analysis over the other in ambiguous cases.

4.5.4 “Blanket Contexts”

One last pitfall in generalization is discovering contexts which are *too* general. Consider what would happen if we compared the following two forms, one of which adds [ra] in the middle of the word, and one of which adds it at the end of the word:

[12] One wonders if there could ever be a morphological mapping that involve an ambiguous structural change in *all* of its forms. Consider for instance a Bantu-like system where there is always a final vowel suffix, but preceding suffixes are -VC. Imagine that this system involves an affix *-an* that is **potentiated** by another affix *-an*. Potentiation is common enough (cf. *-ity* by *-able*); and our system wants to discover it, to know the areas in which a mapping is productive. If *-an* is potentiated by a homophonous affix, we would have pairs like the following:

muk-an-a	muk-an-an-a
top-an-a	top-an-an-a
sif-an-a	sif-an-an-a

Here, there’s an ambiguity about which *-an* is which. One needed resolve this ambiguity, but one must adopt *at least one analysis*, which wouldn’t happen in the No-Mappings approach.

a.	P	A	Q	B = ra
	ni	∅	ta	

b.	P′	A	Q′	B = ra
	klɪp	∅	∅	

The P sides of these two forms do not share any segments (or, arguably, any features), and therefore the resulting constraint should have just a variable left on the P side. The Q sides also do not share any features, so we should also have a variable on the Q side. This leaves us with the following constraint:

P	A	Q	B = ra
X	∅	Y	

which says: insert [ra] anywhere. How do we apply this constraint to a new form? Do we put the [ra] at the beginning? At the end? Do we generate all possible insertions of [ra]? In order to avoid this problem, we have assumed a convention that mapping constraints must be specified for *some* sort of context; we do not allow both sides to be completely empty.

4.6 The Computation of Confidence

We mentioned in section 3.1 that Confidence cannot be a simple matter of computing what percentage of the time that a mapping succeeds in deriving the right result. The reason is that trust is also based on the amount of **testimony** available that the relevant mapping is trustable. A Reliability value (i.e. the raw fraction of correct predictions made; section 3.1) of 1.000, based on 5 out of 5 cases, is just not as impressive as the lesser Reliability value of .99 when the latter is based on 990/1000 cases. The reason is that five cases are not enough testimony to persuade a rational observer that the fraction really is 1.00 (and we are assuming that the human learning mechanism is indeed quite rational, at a tacit and unconscious level). In other words, what we need here is a way to make the Learner less certain about constraints which look promising, but which are only embodied (so far) by a small number of forms.

Mikheev (1995, 1997) faced exactly this problem when designing a system for the quite practical task of predicting part of speech from orthography. Mikheev's solution was to use the statistic of lower confidence limits, adjusting the empirically observed Reliability to a lower estimate of what the "true" Reliability may be in the real world.[13]

We have made a very rough estimate, based on Wug testing data, that the appropriate confidence limit is about 75%.[14] Here are some representative figures to show how this limit adjusts raw Reliability values to Confidence values, by downgrading Reliability values that are based on fewer data. The chart shows the Confidence that results when the count of Hits is half of the Scope, for a broad variety of Scopes.

Scope	**Hits**	**Raw Ratio**	**Confidence: 75%C.I**	**Confidence: 90%C.I**	**Confidence: 95%C.I**
2	1	0.500	0.146	(not defined)	(not defined)
4	2	0.500	0.310	0.090	(not defined)
6	3	0.500	0.351	0.198	0.089
8	4	0.500	0.374	0.251	0.166
10	5	0.500	0.389	0.282	0.211
20	10	0.500	0.423	0.351	0.307
30	15	0.500	0.438	0.380	0.345
40	20	0.500	0.446	0.397	0.366
60	30	0.500	0.456	0.416	0.392
80	40	0.500	0.462	0.428	0.407
100	50	0.500	0.466	0.435	0.417
200	100	0.500	0.476	0.455	0.442
300	150	0.500	0.481	0.463	0.452

[13] Following Mikheev, we use the following formula to calculate lower confidence limits: first, a particular reliability value ($\hat{p}$) is smoothed to avoid zeros in the numerator or denominator, yielding an adjusted value $\hat{p}^*$:

$$\hat{p}_i^* = \frac{x_i + 0.5}{n_i + 1.0}$$

This adjusted reliability value is then used to estimate the true variance of the sample:

$$\text{estimate of true variance} = \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n}}$$

Finally, this variance is used to calculate the lower confidence limit (π_L), at the confidence level α:

$$\pi_L = \hat{p}^* - z_{(1-\alpha)/2} * \sqrt{\frac{\hat{p}^*(1-\hat{p}^*)}{n}}$$

(The value z for the confidence level α is found by a look-up table.)

[14] This is similar to Mikheev's (1997) finding that the 75% confidence level was also the best criterion for admitting new rules to the grammar for the purpose of tagging unseen words.

400	200	0.500	0.483	0.468	0.459
600	300	0.500	0.486	0.474	0.466
800	400	0.500	0.488	0.477	0.471
1000	500	0.500	0.489	0.480	0.474
2000	1000	0.500	0.493	0.486	0.482
3000	1500	0.500	0.494	0.488	0.485
5000	2500	0.500	0.495	0.491	0.488

As can be seen, the penalty imposed by Mikheev's is negligible for numbers in the thousands, but very substantial for numbers in the single digits.

Plainly, our adoption of the 75% confidence interval is just guesswork; quite sophisticated experimentation will be needed to get any clearer idea whether the 75% confidence limit figures are valid or close to it.

4.7 Types and Tokens

Bybee 1995 suggests that in determining native speaker's intuitions about how novel forms are to be inflected, type frequency is likely to be more rational than token frequency. Our Learner certain works in this way, and to the extent it produces correct predictions about native speaker intuitions, this might be taken as support for Bybee's claim.

We have not tried pursuing a learner based on token frequency, but we are skeptical that a model based on token frequency alone would work well. For example, Marcus et al. 1992 report that roughly 70-75% of the tokens of verbs uttered by parents to children in recorded sessions were irregular, yet the propensity to extend the irregular patterns of English is rather weak. Super-common irregulars seem to have very little influence indeed on their phonetic neighbors; e.g. *play* ~ [plɛd], on the model of *say* ~ *said*, seems very unlikely.

While we hope that our model can help support the claim of priority for type frequency (by making it more explicit what it means to be a similar type), we have nothing to contribute, at present, on the intriguing question of *why* type frequency should be more important than token frequency. Another line of inquiry which might prove fruitful would be to test various ways of taking both type *and* token frequency into account when computing reliability values, to see whether this produces a better fit to the data than simple type or token frequency alone.

4.8 Finding Phonological Structural Descriptions I: Reinvoking the Apparatus

As mentioned earlier, the Learner attempts to bootstrap phonology and morphology off of one another. Improvements in phonology reveal greater generality for morphological mappings, improving their **Confidence**. Conversely, imperfect morphological mappings can used to discover the phonological alternations of the language. Let us return to our earlier Pseudo-

Finnish example, and consider it in more detail. The default mapping X → Xi applied there to *viljele*, yielding **viljelei*, rather than the correct *viljeli*. Now, if we make the hypothesis that **viljelei* is bad (i.e. not the observed output) simply because there is phonology that should apply to it, we can construe

viljelei → viljeli

simply as another kind of mapping. Thus, the search for phonological changes can procede is just like the search for morphological changes — we compare the actual and predicted forms, isolate the change, and construct a mapping. In this case, the comparison would yield:

e → ∅ / [viljel___i]

4.9 Finding Phonological Structural Descriptions II: The Fish

It must be remembered, however, that *not* all morphological errors are appropriately fixed through phonology. Thus, for example, the default mapping for English pasts, X → Xd, may be applied to [sɪŋ] to get *[sɪŋd], but it is by no means true that English has phonology of the type

ɪŋd → ʌŋ / ___ $]_{\text{word}}$

This would derive, for example, **red-wung blackbird* and other monstrosities. Moreover, to the extent that there is some a priori notion of "phonological process", the ɪŋd → ʌŋ mapping presumably falls outside of it.

4.9.1 Phonology Too Must be Assessed for Confidence

A sorry tale from our experience so far emerges from our Learner's encounter with the English present-past pair *make/made*. Here, the application of the default mapping X → Xd produces the predicted form *[meɪkd]. The final cluster, *[kd], is unquestionably ill-formed in English. By doing a comparison of ill-formed with well-formed output, we find a pseudo-remedy:

[m eɪ k d]
[m eɪ ∅ d]

namely, k → ∅ / ___ d. This is fine for *made*, but leads to disaster with any other /Xk/ verb stem (e.g. *take* ~ ***tade*, *slake* ~ ***slade*, etc.).

We anticipate that with further development, we will need to give phonological mappings their own calculated Scope, Hits, Reliability, and Confidence. Such information will make it possible for the Learner to avoid foolish applications of rules that have no real empirical grounding.

4.10 Aids to Faster Computation

We hold little brief that the items that follow match in any way what people do, but they do help make the Learner run at feasible speeds.

- Keep an index, wherein all the mappings of the grammars are coded by their structural change. Since only mappings with the same structural change can be collapsed, this permits the set of relevant comparisons to be located quickly.[15]
- Each mapping constraint stores with it not only the [string1]→[string2] mapping, but also certain shortcut information like the complete factorization, the location of the change, etc., to avoid having to calculate this information over and over again.

5. Tasks Outside the Central Architecture

5.1 The Wug-Tester

The Wug-tester inputs a form and finds all the mappings that apply to it. It is assumed, currently, that the Confidence of a mapping represents a rough approximation of the well-formedness judgement that would be assigned by the Learner to outputs derived by that mapping. Where more than one mapping derives the same output, the mapping that assigns the best rating determines the Learner's "judgment" for that output.

5.2 The Data Presenter

Given a set of forms with frequencies (for both sides of the mapping, not for stems in general), the Data Presenter determines an order in which data pairs are considered by the Learner. It is assumed that there is some learning threshold: the number of times a form must be heard for it to be well installed in memory. For the Learner to make use of an inflectional pair, both of its members must be so installed. The number of times a form must be heard to be properly memorized is a parameter that may be varied.

The Data Presenter is of no value, we think, in studying adult grammars, since adults have digested essentially all of the data for as long as is necessary (and furthermore, they "care" about types, not tokens). However, something like a Data Presenter is needed to study the course of acquisition.

[15] Adam interjects: is this so unreasonable? It seems like a list of all the changes that are used to map from one category to another is a useful kind of summary of the "morphemes" involved. (Or, in the case of ablaut changes, at least a summary of the changes involved, if we don't want to call [ɪ] → [æ] a morpheme, per se.

5.3 The Output Text Generator

A real language learner who produces inflected forms often needs to synthesize them, as Marcus and Pinker (1992) point out: either the correct inflected form has not been encountered yet, or it has been learned imperfectly and cannot be recalled at the moment. The Output Text Generator simulates a long string of produced forms, using the grammars produced by the Learner to estimate the relative frequencies which which correctly remembered, correctly synthesized, and wrongly synthesized forms are produced. It is used in our account of the U-shaped curve in section 6.3 below.

6. Pilot Results

6.1 Simple Phonology

Our present Learner deals readily with the Finnish-like pseudo-language described earlier. We fed it an input file consisting of a few dozen present-past pairs, manifesting the phonological and morphological system. In addition to exemplifying the process of /e/ deletion already described, the synthetic data also included forms in which a stem-final /t/ spirantizes to [s] before /i/ (e.g. *myyt* ~ *myysi*), somewhat as in real Finnish, and also forms where both /e/ deletion and spirantization apply (e.g. *tunte* ~ *tunsi*, with /...tei/ → ti → [si]).

The Learner apprehended these morphological and phonological rules correctly, as we determined by inspecting the grammar it constructed. It also correctly passed Wug tests on appropriate forms, as follows. Queried with present tense [kypy], it replied with the correct past [kypyi], with /-i/ suffixation. Given [pupe], it applied /e/ deletion and responded with correct [pupi]. The input [putut] yielded the correct spirantized reply [putusi]; and [mute] correctly yielded [musi], with both /e/ deletion and spirantization. Thus it appears that the Learner can deal with very simple systems like our "Pseudo-Finnish" without difficulty.

6.2 Modeling Native Speaker Judgments

As noted above, a crucial goal for our Learner is to learn phonology and morphology even in circumstances where there is irregularity. English past tenses are a good place to start. Here, the irregularity has an interesting internal structure (Bybee and Moder 1983, Pinker and Prince 1988), with most irregular forms falling into families like that seen in *fling/flung*, *string/strung* etc. Moreover, a great deal of modeling and psycholinguistic work has already been done, including an extensive "Wug" testing study by Prasada and Pinker 1993. In this study, experimental subjects rated 60 imaginary past tense verbs on a 1-7 scale of well-formedness, both in their regular form (e.g. *spling* ~ *splinged*) and in a suitable irregular form (*spling* ~ *splung*). Previous researchers (MacWhinney and Leinbach 1991, Ling and Marinov 1993, Nakisa, Plunkett and Hahn 1997, Westermann 1997) have used the Prasada/Pinker results to validate their models. Since we plan to apply our model to inflectional systems of considerably greater intricacy than English past tenses, it is clearly important that our pilot Learner show acceptable performance in this task.

The Prasada/Pinker study deployed 12 types of forms, classified as follows. Three groups were devised to resemble existing regulars to a (strong/intermediate/weak) degree. Three other groups were chosen to resemble existing irregulars, also to three degrees. This made six groups, and for each the subjects judged both the regular suffixed form and an alternative irregular form that included a plausible vowel change, making 12 in total.

To teach our Learner about English past tenses, we fed it a corpus of 2181 English present/past pairs, amplified from an file courteously installed by Prof. Brian MacWhinney on his Web site. The Learner developed a grammar for this corpus, and we then Wug-tested it with all 60 imaginary verbs from Prasada and Pinker. For each, we collated the values for the Confidence of the best mapping that generated the regularly-inflected form (*splinged*) and the alternative irregular form (*splung*). We take these numbers to be a reasonable way of stating how good the Learner "judged" these forms to be.

Here are the results. Across the 12 verb categories, the "judgments" of our Learner were highly correlated with the judgments of the Prasada/Pinker subjects: r^2= .889. The following qualitative similarities were apparent.

First, the Learner, like the subjects, judged vowel-change cases like *spling ~ splung* as acceptable only if they **closely resembled** existing irregulars. Thus, across the continuum of (strong/medium/weak) resemblance to existing irregulars, the subjects' ratings were 4.98, 3.85, and 3.41 on the 1-7 scale.[16] The analogous ratings of the Learner, on its own 0-1 scale, were .438, .181, and .028.

The Learner also mimicked a subtle effect in the human judgments that surprised Prasada and Pinker: subjects rated regulars slightly better to the extent that they were **free of resemblance to existing irregulars**. Looking over a continuum of four types that increases in this property (strong-medium-weak resemblance to irregulars, plus strong resemblance to regulars), the values were: subjects 5.46, 5.26, 6.00, 6.22; Learner .861, .878, .923, .933. The reason our Learner judged the forms in this way is that there exist in the vocabulary certain "islands of reliability" for the occurrence of regular verbs. For example, all English verbs ending in /...ɪp/ (our corpus had 16) are regular, so /...ɪp/ is a reliability island. When the mappings for such islands are sufficiently Trustable (in sense of 3.1 above), our Learner used them in producing novel forms for verbs within the island. The Confidence value of the relevant mapping slightly exceeds that of the default. The Learner therefore rated such regular forms slightly better than those not occupying a reliability island.

Although the Learner mimics human behavior qualitatively here, we should note that the size of the effect in the Learner is not as large as in humans. We plan to rerun the simulation as the Learner becomes more sophisticated (notably, in incorporating features), and see if we can achieve a better match.

[16] The lowest rating given by the subjects to any class of forms was 3.3; this may either be the result of a high floor for the judgments, or else of pooling data from subjects with only moderate subject-to-subject agreement.

There was one area in which the Learner's performance patently failed to match that of human subjects. Humans, but not the Learner, seriously downgraded regular forms that didn't resemble existing regulars. Along the continuum of increasing non-resemblance, the values were: humans 6.22, 5.22, 4.99; Learner .933, .924, .907. The reason we are not too concerned about *this* mismatch is that Prasada and Pinker show, with some care, that in humans the effect resulted from the subjects' aversion to phonotactically-bizarre sound sequences in the stems they were given. (Prasada and Pinker had had no choice but to include such stems, in order to construct Wug-verbs that resembled no existing regular.) Thus, the judgment in question was not really a morphological judgment at all, but a phonotactic one. Phonotactic judgments fall outside our Learner's present task.

Summing up, we think the Wug-testing performance of our Learner matches humans fairly well thus far, and we hope to be able to improve it as the Learner is given a more sophisticated internal conception of phonology.

6.3 Modeling the "U-Shaped Curve"

A widely discussed topic in the psycholinguistics of morphology and phonology is the "U-shaped curve," a developmental phenomenon whereby children first implement an inflectional mapping without error (when they inflect at all), then produce overregularization errors like *goed*, then gradually achieve adult performance. Our work here closely follows the account given in Marcus et al. 1992.

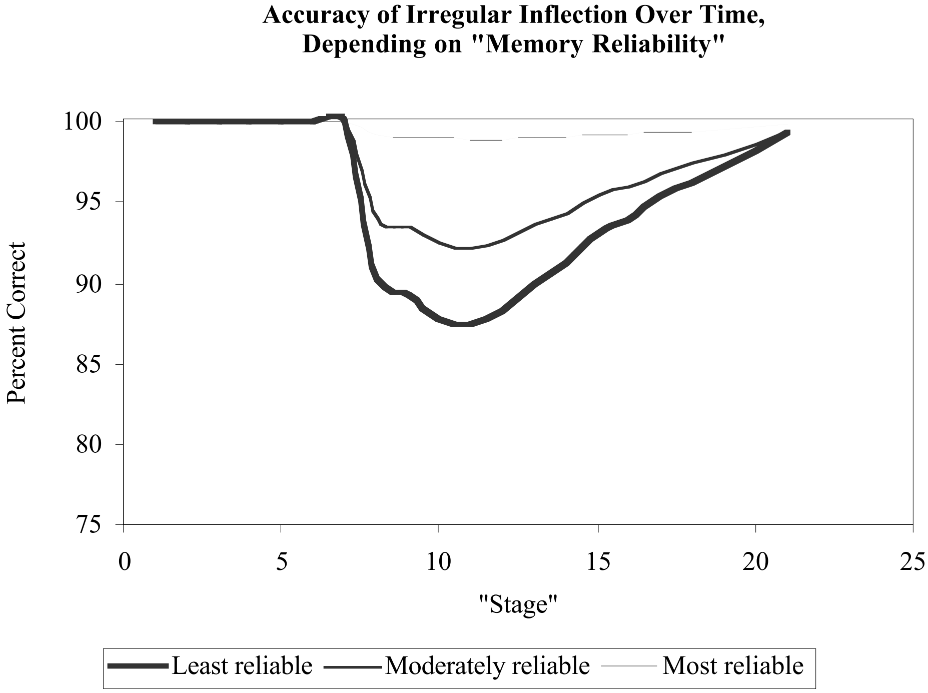
In brief, the early period "no errors, but often not inflected" represents a time when a certain number of inflected forms have been installed in memory, but not enough forms have accumulated to permit learning of the mappings. Overregularization sets in when the default mapping has been learned. Once the default mapping exists, the child can use it to supply a "good-guess" form (as opposed to an uninflected bare stem) when she has not heard the inflected form, or when she has heard it but cannot remember it at the moment. Since cases of "haven't heard it yet" or memory failure are exceptional, it emerges (contrary to what some earlier literature stated) that children overregularize only a small percentage of the time. Finally, after a long terminal phase of learning, going beyond the data period examined in most acquisition studies, the learner has memorized all the irregulars at adult reliability levels, and thus reaches the other side of the "U".

What we wanted to see was whether Marcus et al.'s persuasive qualitative account could be amplified to a quantitative one by making use of our Learner. Therefore, we used the Data Presenter program described in §5.2 to simulate parental input for a child, and we fed this to the Learner program as its input. In particular, we produced 21 "stages" at which the performance of the Learner was to be monitored; each stage represented the point where a certain number of inflected verb pairs have been "learned."[17] We then used the Output Text Generator described

[17] Specifically, we assumed that a past-present pair is learned after both past and present have been heard five times; we plan to examine other possibilities as well.

in §5.3 to simulate the productions of the child in each stage of learning. At each of the 21 stages, this Output Text Generator simulated a child inflecting all of the irregular verbs she knows. Memory failure was modeled as a declining exponential function of the number of times a child has heard a word. Our program also divided up the probabilities for each possible output past tense form of each verb, based on the assumption that a memorized correct form will be used if it is (a) known and (b) remembered; and that otherwise forms generated by the grammar will be used, in proportion to the Confidence values of the mappings that generate them. Finally, the data for all verbs were combined, yielding the probability at each stage that an irregular will be regularized.

What emerges is, indeed, a U-shaped curve, as can be seen in the figure below. The chart shows three different "memory reliability functions" — one representing nearly perfect memory, one representing moderate memory failure, and one representing more substantial memory failure. It is difficult to verify the curves in detail, since as Marcus et al. have shown, we lack enough data to monitor the full shape for any child, especially at later ages. But qualitatively, our curves do seem to match what Marcus et al. were able to find. Varying the "memory reliability function" raises and lowers the curve, much as one would expect.[18]



[18] Our curve is not quite "U-shaped", as it climbs out of the trough very slowly. Discussion of older children, adolescents, and adults in Marcus et al. p. 44-45 indicates that this prediction is probably correct; even adults overregularize on rare occasions.

6.4 Modeling Opaque Phonology

Within Optimality Theory, phonological opacity is a controversial topic for which a wide array of solutions have been proposed. Opacity arises in the attempt to relate members of a paradigm phonologically. A model that apprehends paradigmatic relations directly thus may be an appropriate response. With this in mind, we piloted a small study of Catalan adjectives and nouns, inflected for gender. In these paradigms, where masculines show the bare stem and feminines receive a schwa ending, opaque phonology arises. Here, in outline, is the pattern, analyzed in depth in Mascaró 1975:

Derivation (of masculine)			**Masculine**	**Feminine**	**Gloss**
/...Vnt/	→	[...Vn]	[əˈte**n**]	[əˈte**nt**ə]	‘attentive’
/...Vn/	→	[...V]	[itəlˈja]	[itəlˈja**n**ə]	‘Italian’
/...Vɾt/	→	[...Vɾ]	[uəβeɾ]	[uəβeɾ**t**ə]	‘open’
/...Vɾ/	→	[...V]	[dukəto]	[dukətoɾə]	‘doctor’

Stem-final clusters like /ɾt/ and /nt/ surface as [ɾ] and [n] respectively in masculine forms, where there is no suffix, and illegal codas must be resolved by deletion. Singleton /ɾ/ and /n/ at the end of a bare stem are deleted, even though final /n/ and /ɾ/ are legal on the surface, deriving opaquely from /nt/ and /ɾt/. The schwa ending found in feminines blocks all deletion. Our data at present consist of about 250 stems, all elicited from a native speaker.

Our Learner acquires this opaque mapping without difficulty, giving the right outcomes in a Wug test. However, only the transparent final cluster simplifications are learned as phonology. The opaque patterns of final /n/ and /ɾ/ deletion are treated with feminine-to-masculine morphological mappings: Xnə → X, Xɾə → X. A challenge we assume for future work is to let the Learner generalize this opaque mapping, so that it can apply it to other morphological categories without having to learn them separately (section 7.6 below).

7. Pie in the Sky

The present part of this documented is inherited from its prior status as a grant proposal. This is simply a listing of things we feel it would appropriate to try to do in the future.

It’s worth pointing out a discrepancy between our pilot results and our long-term plans. In our pilot work, we have been interesting in tasks that could be accomplished by a Learner with relatively little sophistication in phonological theory, though perhaps a bit more sophistication in grasping and weighing competing generalizations. These have served primarily as a way of checking the Learner’s basic competence in these areas, and thus are of dubious novelty. For example, the ability of our Learner to model the “island of reliability” effect (section 6.2) merely duplicates an ability earlier demonstrated for a connectionist model by Daugherty and Seidenberg 1994. Similarly, we suspect that other English past tense models might have the ability to model the U-shaped curve if they were embedded in the kind of simulation we deployed.

For future work, our main focus will be questions of phonological theory. We think that as we continue to test the Learner against new datasets[19] we are likely to find that there are quite a few problems that will not be solved until we endow it with crucial ideas from the phonological literature. We anticipate that this research effort will provide support (or on occasion, evidence against) phonological theories. Further, we think that developing a learner can help bring new empirical areas and theoretical proposals into the purview of contemporary phonology.

7.1 What We Will Do

We see a number of diverse research activities emerging from this project.

- Expanding and streamlining the learning algorithm, programming additions and changes, debugging, and testing against data.
- Testing of adult native speaker intuitions. For reasons given below, the main target languages are Korean, Catalan, and Polish. Initially, we will pilot our work with relatively free-form, traditional linguistic elicitation, including Wug-testing. Later, we will gather systematic data following methods such as those of Prasada and Pinker 1993. Software we have already developed will make it possible to obtain judgments in automated fashion, with all words presented to subjects in spoken rather than orthographic form.
- Construction of further learning databases. In addition to the Catalan, Polish, and Korean data, we will develop a Yawelmani database from published sources. The goal is to give the Learner an experience as close as possible to the childhoods of human learners.[20]
- Modeling various acquisition phenomena which have been described in the literature. In §6.3 we presented a brief description of how we have attempted to model "U-shaped curves" for children learning English. The acquisition literature contains many other descriptions of how children handle irregularity, however — for instance, see Orsolini, Fanari and Bowles 1998 for a description of how Italian children use the verb class system. Since the goal of the Learner is not only to simulate adult grammars, but also to explore how children might learn them, it is important to check that our Learner can match human children in the time course of acquisition, nature of errors, etc.
- A related phenomenon which we might try to model is analogy and historical change. This is a matter of interest because historical changes are often cited as another difference between regular and irregular inflections (e.g., Pinker and Prince 1988), although there is no explicit theory of how the synchronic model affects diachrony. By using the Learner to model historical changes, we could tie these informal observations to an explicit theory of language use and acquisition. (It would also provide a novel contribution to the field of historical

[19] Our data sets so far are: 2181 English present/past pairs, about 1600 German present/past pairs, 2000 ten-member Japanese verb paradigms, 3000 Italian 1 sg. pres/infinitive pairs, 250 Catalan adjectives, and the ten-member paradigms of 1700 Latin nouns.

[20] It is important, we think, for researchers in this area to share resources, particularly in order to permit comparative testing of their algorithms. Our completed databases are currently being posted in Excel format on Hayes's Web page (http://www.humnet.ucla.edu/humnet/linguistics/people/hayes/learning/learning.htm), and will be posted in other public locations should these become available.

linguistics, where computational models are rarely provided for theories.) Some historical changes which would be instructive to simulate include: (1) the migration of English verbs from regular → irregular or vice versa, depending on frequency, (2) the simplification of the Latin nominal inflection system (Gaeng 1984), (3) the collapse of opaque vowel harmony in Yawelmani verbs (Hansson 1999).

This is what we see as the best route to improving our Learner: giving it ever greater challenges, and attempting to arrive at an ever-better approximation of human performance.

7.2 Strengthening the Learner

We have also created a version of the Learner with a phonological feature capability in the morphology, enabling it to learn cases where features are crucial in accounting for allomorphy based on natural classes. (We have yet to implement features in the phonological rules, but the infrastructure to do this is now in place.) For maximum flexibility in theorizing, we are letting the choice of feature system occur at the level of individual simulations, rather than "hard-wiring" any particular feature set into the Learner. Our work so far on features has reached the point that (much to our relief) it appears that the addition of featural mappings does not produce an unmanageable explosion of grammar size.

We will also work on strengthening the Learner in other ways, particularly involving the accurate assessment of the Reliability of mappings. For instance, it would be useful to compare various models of how speakers use statistics to evaluate the grammars which our algorithm creates. We have described a model using lower confidence limits, but it would also be possible to use a similarity-based categorization model such as Nosofsky's (1990) Generalized Context Model. Comparing numerous statistical predictions for the best fit with human data is a common practice in the psychological literature, but not very common in linguistics; we see this research as one place where we can incorporate this practice in arriving at the most realistic linguistic analysis.

7.3 Locality

A crucial aspect of phonological theory is the theory of locality: what is the "window" over which structural descriptions are located? The interesting cases are the nonlocal phenomena, where the crucial trigger of a process is separated, perhaps by several segments, from the target. A common case of this sort is vowel-to-vowel processes, which can ignore intervening consonants.

A number of phonological theories posit that a kind of "subrepresentation" consisting solely of the vowels of a string is available to human phonological computations. Such theories include the "vowel projection" of Vergnaud and Halle 1979 and the "minimal scansion" proposed by Archangeli and Pulleyblank 1987. We plan to incorporate such a subrepresentation into our Learner. Operations carried out on the full string will be carried over to the vocalic subrepresentation, and vice versa. Such a subrepresentation has earlier been assumed in Hare's 1990 connectionist learning simulation of Hungarian vowel harmony.

We plan to test the merits of a vocalic subrepresentation by using it in the learning of vowel harmony in Turkish, for which we already have a toy database.[21] The issue of locality also arises with phonologically-conditioned allomorphy, as in, for example, the rounding harmony pattern found in Farsi for (only) the imperative/subjunctive prefix (Thackston 1993). Crucially, for both Turkish and Farsi, we plan to investigate the consequences of *not* having a vocalic subrepresentation. We anticipate that a Learner lacking such a subrepresentation will learn such systems slowly, and will have difficulty in generalizing to the full range of cases (e.g., to cases where the intervening consonants or clusters are rare). Should the difference come out as anticipated, we will have derived a novel argument for a theoretical proposal in phonology, namely that it directly aids learnability.

Much the same can be said for a kind of nonlocal morphology, the CV stem template phenomena found in Semitic and in native languages of Northern California. As a cautious first step in this area, we propose to implement a simple CV-tier model (much as in McCarthy 1981a) of morphological representation, and apply it to the templatic morphology of Yawelmani. Yawelmani is particularly worth the labor of preparing a large database of paradigms, since it also bears on the question of phonological opacity, discussed below. We will test the following hypothesis: that the Learner will be able to grasp the generalizations involved in CV-template morphology only if it possesses appropriate formal representations.

7.4 The Richness of Morphological Mappings

A virtue of an automated learner is that it discovers whatever generalizations are available to it, within the limits of its formal capacities. Checking the results of our Learner, we find it often discovers things that do not appear in standard phonological analyses of a language, but are nevertheless apprehended intuitively by native speakers.

Consider, for instance, the phonological pattern of Catalan noted in section 6.4. Here, a surface masculine form ending in a vowel could in principle have three underlying sources: /...Vn/, /...Vɾ/, and /...V/. Thus, if a speaker is Wug-tested, being given a masculine of the type [...V] and asked for the corresponding feminine, she could in principle provide three alternative answers: [...Vnə], [...Vɾə], or [...Və]. Our pilot work suggests, however, that speakers *don't* freely offer three answers. Rather, they have quite strong preferences, based on the **quality of the final vowel**: /a/ stems characteristically take /n/, /o/ stems take [ɾ], and /u/ stems take a bare suffix. These preferences emerge, in our learning simulations, as the result of existing patterns in the lexicon: each pattern constitutes a "best guess," under the assumption that novel words will inflect like existing ones.

[21] Should it be feasible, we will use instead the extensive, electronically-coded data of the NSF-funded Turkish Electronic Living Lexicon developed by Sharon Inkelas and her colleagues.

Similar patterns occur in Korean, where the way in which neutralized plain stems appear under suffixation is also rule-governed. For example, isolation [t]-final stems strongly prefer presuffixal allomorphs with [s] (Martin 1992, 101-102), even though about a dozen other possibilities would be in principle be possible given the known patterns of stem-final neutralization. The basis of speakers' intuitions seems to be that [t] ~ [s] is by far the most common alternation pattern for coronals.

Strikingly, such intuitions are not just a matter of statistical preferences. Native speakers often judge "Wug" forms that extend existing but rare patterns to be simply *ungrammatical*, even though it would be quite straightforward to generate them under conventional analyses of Korean. For example, on a scale in which 1 is worst and 7 best, we have found that Korean speakers given the Wug-form "[nut]" rate [nut-əl][22] in the 1-2 range, with [nus-əl] at 7.

Our pilot Wug-testing in Catalan and Korean thus suggests to us that these, and probably other phonological systems, are characteristically **underanalyzed**: the native speaker knows more, and has more intuitions, than current phonological analyses provide for. We plan to pin down this conclusion empirically, and also to use our Learner to provide the remedy. The Learner is designed to learn rich grammars, extracting whatever generalizations can be stated within its formal limits. For the case of Catalan, it has in fact already succeeded, we think, in learning the crucial patterns. Our long-term plan is to (a) do more extended and systematic Wug-testing of speakers of Korean and Catalan; (b) develop databases of inflected forms large enough to approximate the childhood experience of real learners of these languages, (c) feed these databases to our Learner; and (d) assess our findings.

7.5 Derived-Environment Phenomena

Derived environment rules (Kiparsky 1973) apply productively when their structural description is set up by morphological concatenation, but not in nonalternating environments. Theorizing about the derived environment phenomenon has produced a number of interesting proposals over the years, though there seems to be no consensus among scholars, particularly among those adopting the non-rule-based approach of Optimality Theory. To our knowledge, no one has worked on the question of how derived environment alternations are learned.

What is needed, we think, is a way of locating the phonological sequences of a language that are illegal in derived forms. Consider a modified version of our Finnish-like language (from 3.2) in which /ei/ is resolved to [i] only where it arises in derived environments. It should be possible to detect a derived-environment pattern in the following way: applying the default mapping, one locates sequences that never arise except when they are "canceled" by appearance in the base form from which the mapping proceeds. Thus in a hypothetical present-to-past mapping *teitu ~ teitui*, the sequence [ei] in the projected past form is also found in the present tense form from which it is projected, and thus is "canceled". In contrast, in *[lupei], projected by the default mapping from [lupe], the sequence *[ei] is *not* canceled in the base. Therefore, derived-

[22] Actually [nud -əl], since plain stops are voiced intervocalically in Korean.

environment phonology must correct it to [lupi]. In this way, it should be possible to locate *[ei] as a constraint that holds only in derived forms—in our terms, it is never violated in uncanceled fashion. We propose to implement this idea explicitly in code and test it against representative examples.

We also anticipate that some of the potential audience for our work (e.g., non-phonologists) might be skeptical that derived-environment effects exist at all. An alternative possibility is that the putative derived environment rule is simply a non-unitary aggregate of individual morphological processes. In fact, we suspected this ourselves, until we found evidence to the contrary. In pilot work, we have discovered that at least some Polish speakers can be Wug-tested on *made-up affixes*. This makes it possible to create entirely novel derived environments, going beyond the existing morphological categories of Polish. For our pilot Polish speakers, novel affixes do in fact trigger derived-environment processes such as palatalization (Rubach 1984), suggesting that the notion of derived-environment rule has linguistic reality. We propose to follow up these initial investigations. If things work out as we anticipate, we will demonstrate the "psychological reality" of derived-environment processes, and also produce a version of our Learner that can learn them.

7.6 Morphemes and Allomorphs

Morphological theory has historically experienced a tension between morpheme-based theories and process-based theories. In a pure morpheme-based theory (for example, McCarthy 1981a), all is concatenation, and special mechanisms are provided to handle the apparent non-concatenative residue. For pure process-based theories, such as Anderson's 1992 "A-Morphous Morphology," concatenation is just one string operation among many.

Our Learner, at present, is a-morphous. However, we suspect it may be better to adopt a mixed strategy: "be morphous and concatenative whenever you can; where this fails, try process morphology." The specific tack we would follow would be to assume, where possible, that an edge-aligned "changing part" in a morphological mapping is in fact an affix, and that the residues are allomorphs of the stem. In cases of alternation, we can then collect stem allomorphs and discover the string mappings and environments that relate them to one another

This strategy could in principle have several payoffs. First, languages show many cases where the string mapping between inflected forms is very complicated, enough to overwhelm our Learner in its present state. An example is German /zɪŋt/ 'sing-3 sg. pres.' ~ /gəzʊŋən/ 'sing-past part.' Here, only a small part of the relevant strings (/...z ...ŋ .../) is shared. The key to this pattern, we think, is to examine the class of past participles as a whole, rather than the individual verb-by-verb mappings. Doing this, one could quickly discover that [gə-] is the normal prefix and [-ən] the normal suffix for this category. Knowing this, the Learner could then treat the rather simple residual pair /zɪŋ/ ~ /zʊŋ/ separately, using a-morphous devices it already possesses. In other words, the strategy "assume concatenativeness, unless forced to do otherwise" could be an effective approach to learning, because concatenation is so simple.

Another possible advantage to morphousness lies in achieving greater generality. Recall that in our Catalan simulation (section 6.4), the treatment of opacity is specific to the feminine-masculine mapping. Ideally, we would like the Learner to notice something better: that in general, stem allomorphs with [nt] before vowel endings appear with [n] in other contexts. If the crucial mappings here are between various allomorphs of the stem (N.B. "stem" is a morphemic concept), rather than between whole inflected forms, then learning the opaque mapping for one paradigm would automatically provide knowledge of phonologically parallel paradigms.

In connection with the second point, we note that it is an empirical question whether the mapping really is general phonology and not a fact about feminine-masculine pairs. We plan also to check out this question by constructing forms of Catalan that require novel applications of phonological deletion to be inflected, and obtain native speaker intuitions on them.

7.7 Opaque Phonology and Derivationalism

Phonology before Optimality Theory commonly treated phonological patterns that are opaque (in the classical sense of Kiparsky 1971) with serial derivation. Opacity is plausibly the most delicate problem faced by classical non-derivational OT. There are indeed current proposals in OT that in various ways bring back derivationalism in a limited way: the Optimality-theoretic version of Lexical Phonology proposed in public lectures by Kiparsky (LSA Institute, 1997), or the Sympathy Theory of McCarthy 1998.

We do not propose to resolve the opacity issue, but we think we can provide some results that might help guide the field to a solution. The question we hope to answer is: can a phonological learner such as ours, which does not make use of derivational or semi-derivational notions, successfully apprehend the data patterns of opaque phonology? The intuition that guides us is that our Learner seems to have rather little difficulty with highly regular data patterns, even if they are rather complex. It is heavily idiosyncratic systems, such as the system of nominal declension in Latin, that seem to pose the toughest learnability issues. We know this already from a Latin simulation we have done on a database of about 1700 stems. Our conjecture, then, is this: any learner with sufficient inductive power to handle Latin nouns would be unfazed by the highly regular and orderly patterns of classic opaque phonology, as in Yawelmani (Archangeli 1997) or Catalan.

To test this, we propose to develop extensive databases for these two languages, the Catalan from native speakers, the Yawelmani from reference sources (primarily Newman 1944). We will feed these databases to our Learner and assess its performance. If the Learner succeeds, we will have learned that the descriptive power of derivationalism is probably not necessary as a way of making phonological systems learnable (but crucially, we recognize that there are other ways by which derivationalism might be justified). If the Learner fails, we will consider how it might be modified, including, possibly, modifications that are derivational in character.

7.8 Product-Oriented Schemata

An observation made by Bybee and Moder 1983 and by Pinker and Prince 1988 suggests modifications to the Learner of a type not often encompassed in generative theories of morphology and phonology: morphological mappings are sometimes coherently defined only in terms of their output targets. For instance, several English verbs form past tenses in [...ɔt], irrespective of their base forms: *buy ~ bought*, *seek ~ sought*, *fight ~ fought*, *bring ~ brought*, *catch ~ caught*. We have seen this in our own data, for example in a 7-year-old learning Catalan for whom feminine forms tend to end in [...anə], irrespective of the vowel of their base. This effect deserves modeling, and we will attempt an account. The basis plausibly would be Optimality Theory, which permits the statement of generalizations about the nature of outputs (e.g., that they end in [...ɔt]). The challenge is to arrange GEN so as to create such outputs, and to keep the grammar in balance so as to avoid extreme overgeneration.

8. Conclusion

Plainly, there is vastly more work to do in this area than the list above includes. To give an obvious example, we have said nothing about prosodic structure (syllables, feet, phonological words, etc.), even though such structure is plainly crucial to understanding phonology in full. We have had to prioritize, and linear phenomena seem to us more tractable in a first approach.

At the most general level, our goals are two. First, we want to show that by studying learning with an explicit model, we can learn things about phonology that we could not otherwise learn. As we have worked on the problem so far, we have constantly been surprised and challenged by how distant our intuitions can be from the actual language data, particularly in the relative frequency of patterns and the generality of processes.

Second, and just as important, we want to show that modeling phonological learning is worthwhile for its own sake. Our own simulations may not always match the data perfectly, but we feel that the only analyses which are demonstrably better are those which have been implemented and can outperform our learner under reasonable learning conditions. We hope that by imposing this standard, we can interest other phonologists in this problem, and ultimately to develop a reasonable-sized research community. If we can do this, we will feel that our project has been more than worthwhile.

References

Albright, Adam (1998) *Phonological sub-regularities in productive and unproductive inflectional classes: Evidence from Italian*, M.A. thesis, University of California, Los Angeles.

Anderson, Stephen R. (1992) *A-Morphous Morphology* , Cambridge University Press.

Archangeli, Diana (1997) "The Yokuts Challenge," in Ignacio Roca, ed., *Derivations and Constraints in Phonology*, Oxford University Press.

Archangeli, Diana and Douglas Pulleyblank (1987) "Maximal and Minimal Rules: Effects of Tier Scansion," in J. McDonough and B. Plunkett (eds.) *Proceedings of the North Eastern Linguistic Society* (*NELS*) 17, Graduate Linguistics Student Association, University of Massachusetts, Amherst, pp. 16-35.

Benua, Laura (1997) *Transderivational Identity: Phonological Relations Between Words*, Ph.D. dissertation, University of Massachusetts, Amherst.

Berko, Jean (1958) "The child's acquisition of English morphology," *Word* 14, 150-177.

Burzio, Luigi (1996) "Multiple Correspondence," paper presented at the BCN Workshop on Conflicting Constraints, Groningen.

Bybee, Joan (1995) "Regular Morphology and the Lexicon," *Language and Cognitive Processes* 10:5, 425-455.

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

Daugherty, Kim G. and Mark Seidenberg (1994) "Beyond rules and exceptions: a connectionist approach to English morphology," in Susan D. Lima, Roberta L. Corrigan, Gregory K. Iverson, eds., *The Reality of Linguistic Rules*, John Benjamins, Amsterdam.

Dell, François (1981) "On the learnability of optional phonological rules," *Linguistic Inquiry* 12, 31-37.

Dresher, B. Elan (1981) "On the learnability of abstract phonology," in C. L. Baker and John McCarthy, eds., *The Logical Problem of Language Acquisition*, MIT Press, Cambridge, MA.

Dresher, B. Elan and Jonathan Kaye (1990) "A computational learning model for metrical phonology," *Cognition* 34, 137-195.

Dzeroski, Saso and Tomaz Erjavec (1997) "Learning Slovene Declensions with FOIDL," http://alibaba.ijs.si/tomaz/Bib/ECML97/

Gaeng, Paul (1984) *Collapse and reorganization of the Latin nominal flection as reflected in epigraphic sources*, Scripta Humanistica, Potomac, MD.

Halle, Morris (1982) "Knowledge unlearned and untaught: what speakers know about the sounds of their language," in Morris Halle, Joan Bresnan, and George A. Miller, eds., *Linguistic Theory and Psychological Reality*, MIT Press, Cambridge, MA.

Hansson, Gunnar O. (1999) "Redefining Phonological Opacity: Yowlumne Vowel Harmony 60 Years Later," Paper presented at the 73rd Annual Meeting of the LSA, Los Angeles, January 11, 1999.

Hare, Mary (1990) "The role similarity in Hungarian vowel harmony: a connectionist account," *Connection Science* 2, 123-150.

Hayes, Bruce (in press) "Phonological Restructuring in Yidiɲ and its Theoretical Consequences," to appear in Ben Hermans and Marc van Oostendorp, eds., *The Derivational Residue*, John Benjamins, Amsterdam.

Hayes, Bruce (in progress) "Should phonological theory recapitulate the course of acquisition?," invited presentation, Third Utrecht Phonology Meeting, June 1988, to appear in the proceedings volume.

Jusczyk, Peter W., Angela D. Friederici, Jeanine M.I. Wessels, Vigdis Y. Svenkerud, and Ann Marie Jusczyk (1993) "Infants sensitivity to the sound patterns of native language words," *Journal of Memory and Language* 32, 402-420.

Jusczyk, Peter W., Paul A. Luce, and Jan Charles-Luce (1994) "Infants' sensitivity to phonotactic patterns in the native language," *Journal of Memory and Language* 33, 630-645.

Kenstowicz, Michael (1997) "Uniform Exponence: exemplification and extension," paper read at the Hopkins Optimality Theory Workshop, Johns Hopkins University, May 1997.

Kiparsky, Paul (1971) "Historical linguistics," in William Orr Dingwall, *A Survey of Linguistic Theory*, University of Maryland, College Park.

Kiparsky, Paul (1973) "Phonological representations," in Osamu Fujimura, ed., *Three Dimensions in Linguistic Theory*, TEC Co., Tokyo.

Ling, Charles X. and Marin Marinov (1993) "Answering the connectionist challenge: a symbolic model of learning the past tense of English verbs," *Cognition* 49, 235-290.

MacWhinney, Brian (1978) *The Acquisition of Morphophonology*, Monographs of the Society for Research in Child Development 174, nos. 1-2.

MacWhinney, Brian and J. Leinbach (1991) "Implementations are not conceptualizations: revising the verb learning model," *Cognition* 49, 235-290.

Marcus, Gary, Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, and Fei Xu (1992) *Overregularization in Language Acquisition*, Monographs of the Society for Research in Child Development 228, University of Chicago Press.

Martin, Samuel (1992) *A Reference Grammar of Modern Korean*, Charles E. Tuttle Company, Rutland, VT.

Mascaró, Joan (1975) *Catalan Phonology and the Phonological Cycle*, Ph.D. dissertation, MIT, Cambridge, MA, distributed by Indiana University Linguistics Club, Bloomington, IN.

McCarthy, John (1981a) "A prosodic theory of nonconcatenative morphology," *Linguistic Inquiry* 12, 373-418.

McCarthy, John (1981b) "The role of the evaluation metric in the acquisition of phonology," in C. L. Baker and John McCarthy, eds., *The Logical Problem of Language Acquisition*, MIT Press, Cambridge, MA.

McCarthy, John (1998) "Sympathy and phonological opacity," Rutgers Optimality Archive 252, http:://ruccs.rutgers.edu/ROA/search.html.

Medin, Douglas L., Gerald I. Dewey, and Timothy D. Murphy (1983) "Relationships between item and category learning: Evidence that abstraction is not automatic," *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9(4), 607-625.

Mikheev, Andrei (1995) "Automatic rule induction for unknown-word guessing," *Computational Linguistics* 23, 405-423.

Mikheev, Andrei (1997) "Unsupervised Learning of Part-of-Speech Guessing Rules," *Natural Language Engineering*.

Mooney, Raymond and Mary Elaine Califf (1996) "Learning the Past Tense of English Verbs Using Inductive Logic Programming," in *Symbolic, Connectionist, and Statistical Approaches to Learning for Natural Language Processing*, Springer Verlag.

Nakisa, Ramin Charles, and Kim Plunkett and Ulrika Hahn (1997) "A Cross-Linguistic Comparison of Single and Dual-Route Models of Inflectional Morphology," in Broeder, P. and J. Murre, eds., *Cognitive Models of Language Acquisition*, MIT Press, Cambridge, MA.

Newman, Stanley (1944) *Yokuts Language of California*, Viking Fund Publications in Anthropology 2, Viking Fund, New York.

Nosofsky, Robert M. (1990) "Relations between Exemplar-Similarity and Likelihood Models of Classification," *Journal of Mathematical Psychology* 34(4), 393-418.

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

Pinker, Steven and Alan S. Prince (1991) "Regular and irregular morphology and the psychological status of rules of grammar," in Susan D. Lima, Roberta L. Corrigan, Gregory K. Iverson, eds., *The Reality of Linguistic Rules*, John Benjamins, Amsterdam.

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

Prince, Alan and Paul Smolensky (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*. Ms.; obtainable from Rutgers University Center for Cognitive Science.

Pulleyblank, Douglas and William Turkel (1997) "Gradient Retreat," in Ignacio Roca, ed., *Derivations and Constraints in Phonology*, Oxford University Press.

Rubach, Jerzy (1984) *Cyclic and Lexical Phonology: The Structure of Polish*, Foris, Dordrecht.

Rumelhart D.E. and J. L. McClelland (1986) “On learning the past tenses of English verbs,” in Rumelhart D.E. and J. L. McClelland , eds., *Parallel Distributed Processing*, Vol. 2, MIT Press, Cambridge, MA.

Steriade, Donca (1996) “Paradigm Uniformity and the phonetics/phonology boundary,” paper given at the Fifth Conference on Laboratory Phonology, Northwestern University, Evanston, IL.

Tesar, Bruce and Paul Smolensky (1993) “The learnability of Optimality Theory: an algorithm and some basic complexity results,” Rutgers Optimality Archive ROA-2, http://ruccs.rutgers.edu/roa.html.

Tesar, Bruce and Paul Smolensky (1998) “Learnability in Optimality Theory,” *Linguistic Inquiry* 29, 229-268.

Thackston, Wheeler M. (1993) *An Introduction to Persian*, Iranbooks, Bethesda, MD.

Touretzky, David S., Gillette Elvgren III, and Deirdre Wheeler (1991) “Phonological rule induction: an architectural solution,” *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 348-355. Lawrence Erlbaum, Hillsdale, NJ.

Vergnaud, Jean-Roger and Morris Halle (1979) “Metrical structures in phonology,” ms., MIT, Cambridge, MA.

Werker, Janet F. and Richard C. Tees (1984) “Cross-language speech perception: evidence for perceptual reorganization during the first year of life,” *Infant Behavior and Development* 7, 49-63.

Westermann, G. (1997) “A Constructivist Neural Network Learns the Past Tense of English Verbs,” *Proceedings of the GALA '97 Conference on Language Acquisition:* Edinburgh, UK: HCRC.

Department of Linguistics
UCLA
Los Angeles, CA 90095-1543

Albright: aalbrigh@ucla.edu
Hayes: bhayes@humnet.ucla.edu