Modeling English Past Tense Intuitions with Minimal Generalization

Adam Albright
aalbrigh@ucla.edu

Bruce Hayes
bhayes@ucla.edu

Department of Linguistics
University of California, Los Angeles
Los Angeles, CA 90025-1543

Abstract

We describe here a supervised learning model that, given paradigms of related words, learns the morphological and phonological rules needed to derive the paradigm. The model can use its rules to make guesses about how novel forms would be inflected, and has been tested experimentally against the intuitions of human speakers.

1 Introduction

In recent years, linguists have explored theoretical models of how speakers discover the rules of their language. Automated learning systems can be of great use in developing such models. The predictions of a theoretical model can be fully tested only when it is formalized explicitly enough to be implemented.

In our research, we have developed and implemented a model for discovering rules of morphology and phonology. The model is trained on pairs of morphologically related words, and learns the rules by which one form of a pair can be derived from the other. We have tested our model by comparing its predictions against intuitions gathered experimentally from human speakers.

2 Criteria for Evaluating Models

A number of properties are desirable in a learning model whose goal is to mimic human intuition. We have been motivated to develop our own model in part because these criteria have rarely been met by previous models. Such models include, for example, connectionist models (Rumelhart and McClelland 1986, Daugherty and Seidenberg 1994, MacWhinney and Leinbach 1991), neighborhood similarity models (Nakisa, Plunkett and Hahn 2001), decision tree/ILP models (Ling and Marinov 1993, Mooney and Califf 1996, Dzeroski and Erjavec 1997), and other rule-based models (Neuvel, to appear).[1]

Our first criterion is that a model should be able to generate complete output forms, rather than just grouping the outputs into (possibly arbitrary) categories such as "regular," "irregular," "vowel change," etc. The reason is that people likewise generate fully specified forms, and a model's predictions can be fully tested only at this level of detail.

Second, a model should be able to make multiple guesses for each word and assign numerical well-formedness scores to each guess. People, too, often favor multiple outcomes, and they also have gradient preferences among the various possibilities (Prasada and Pinker 1993).[2]

Third, a model should be able to locate detailed generalizations. Here is an example: English past tenses are often formed by changing [ɪ] to [ʌ] when the final consonant of the word is [ŋ] (*fling-flung*, *cling-clung*, *sting-stung*). As experiments show, such generalizations are learned by speakers of English (that is, speakers do more than just memo-

[1] The Analogical Model of Language (Skousen 1989, Eddington 2002) satisfies all of our criteria. However, in our use of this model so far, we have been unable to find any setting of its parameters that can achieve good correlations to our experimental data, reported below in section 4.

[2] On a practical level, an ability to consider multiple outputs would also improve the performance of a recognition system. For example, a system not told that *spelt* is a dialectal past tense for *spell* should be able to interpret it as such, even if *spelled* were its first choice.

rize each irregular verb). For example, experimental participants often volunteer *splung* as the past tense of *spling*, extending the generalization to a novel verb.

The importance of detailed generalizations is not limited to irregular forms. We have found that speakers are often sensitive to detailed generalizations even among regulars. For example, verbs in English ending in voiceless fricatives ([f, θ, s, ʃ]) are always regular. Our experiments indicate that English speakers are tacitly aware of this pattern. Thus, an accurate model of their linguistic intuitions must be able to detect and learn the pattern in the training data.

Although detailed generalizations are important, it is also crucial for a learning model to be able to form very broad generalizations. The reason is that general morphological patterns cannot be learned simply as the aggregation of detailed patterns. Speakers can generate novel inflected forms even for words that don't fit any of the detailed patterns (Pinker and Prince 1988, Prasada and Pinker 1993). Thus, a general rule is needed to derive an output where no close analogues occur in the training set. A special case of this sort is where the base form ends in a segment that is not phonologically legal in the language (Halle 1978). Thus, the German name *Bach* can be pronounced by some English speakers with a final voiceless velar fricative [x]. Speakers who can pronounce this sound agree firmly that the past tense of *to out-Bach* must be [aʊtbaxt] (Pinker 1999), following a generalization which is apparently learned on the basis of ordinary English words.

In summary, we believe it is important that a learning model for morphology and phonology should produce complete output forms, generate multiple outputs, assign each output a well-formedness score, and discover both specific and broad generalizations.

3 Description of the Model

3.1 Rule induction by minimal generalization

Our model employs a bottom-up approach to learning, iteratively comparing pairs of surface forms to yield ever more general rules. It takes as its input ordered pairs of forms which stand in a particular morphological relation – e.g., (present, past) – and compares the members of each pair to construct rules that derive one from the other. As an example, consider the pairs of forms in (1).

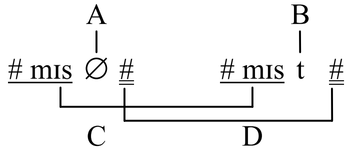
(1)	([mɪs]$_{\text{pres.}}$, [mɪst]$_{\text{past}}$)	'*miss*(*ed*)'
	([prɛs]$_{\text{pres.}}$, [prɛst]$_{\text{past}}$)	'*press*(*ed*)'
	([læf]$_{\text{pres.}}$, [læft]$_{\text{past}}$)	'*laugh*(*ed*)'
	([hʌg]$_{\text{pres.}}$, [hʌgd]$_{\text{past}}$)	'*hug*(*ged*)'
	([rʌb]$_{\text{pres.}}$, [rʌbd]$_{\text{past}}$)	'*rub*(*bed*)'
	([nid]$_{\text{pres.}}$, [nidəd]$_{\text{past}}$)	'*need*(*ed*)'
	([dʒʌmp]$_{\text{pres.}}$, [dʒʌmpt]$_{\text{past}}$)	'*jump*(*ed*)'
	([plæn]$_{\text{pres.}}$, [plænd]$_{\text{past}}$)	'*plan*(*ned*)'

When we compare the present and past forms of each word, we see that the relation between them can be expressed as a *structural change* (in this case, adding [-t], [-d], or [-əd]) in a particular *context* (after [mɪs], after [hʌg], etc.). Formally, the structural change can be represented in the format A → B, and the context in the format / C__D, to yield word-specific rules like those in (2). (The symbol '#' stands for a word boundary.)

(2) ∅ → t / # mɪs __ #
∅ → t / # prɛs __ #
∅ → t / # læf __ #
∅ → d / # hʌg __ #
∅ → d / # rʌb __ #
∅ → əd / # nid __ #
∅ → t / # dʒʌmp __ #
∅ → d / # plæn __ #

The exact procedure for finding a word-specific rule is as follows: given an input pair (X, Y), the model first finds the maximal left-side substring shared by the two forms (e.g., #mɪs), to create the C term (left side context). The model then examines the remaining material and finds the maximal substring shared on the right side, to create the D term (right side context). The remaining material is the change; the non-shared string from the first form is the A term, and from the second form is the B term.

(3)



Note that either A or B can be zero. When A is zero and edge-adjacent, we are dealing with an affixational mapping. When B is zero and edge-adjacent, we are dealing with some sort of truncation; e.g. the mapping from English plurals to singulars. When neither A nor 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.

As such word-specific rules accumulate, the model attempts to generalize. As soon as two rules with the same structural change have been discovered, their contexts are compared to yield a more general rule, retaining all shared context material, and replacing all non-shared material with a variable. Here is the generalization process as applied to *miss* and *press*:

(4)
$$\begin{array}{llllllll}
 & \varnothing \rightarrow \mathrm{t} & / & \mathrm{m} & \text{ɪ} & \mathrm{s} & \_\_ & \# \\
+ & \varnothing \rightarrow \mathrm{t} & / & \mathrm{pr} & \text{ɛ} & \mathrm{s} & \_\_ & \# \\
\hline
= & \varnothing \rightarrow \mathrm{t} & / & \mathrm{X} & \begin{bmatrix} \text{+syllabic} \\ \text{–low} \\ \text{–back} \\ \text{–tense} \\ \text{–round} \end{bmatrix} & \mathrm{s} & \_\_ & \#
\end{array}$$

The procedure for comparing contexts of two rules is much like the procedure for creating a word-specific rule. The general scheme is as shown in (5):

(5)
$$\begin{tikzcd}
\mathrm{A}\rightarrow \mathrm{B}\ / & & C_1 & & \_\_ & & D_1 & \\
+\ \mathrm{A}\rightarrow \mathrm{B}\ / & & C_2 \arrow[dl] \arrow[d] \arrow[dr] & & \_\_ & & D_2 \arrow[dl] \arrow[d] \arrow[dr] & \\
=\ \mathrm{A}\rightarrow \mathrm{B}\ / & \mathrm{X} & C'_{\mathrm{feat}} & C' & \_\_ & D' & D'_{\mathrm{feat}} & \mathrm{Y}
\end{tikzcd}$$

Given two rules that share the same structural change (Rule 1: $A \rightarrow B$ / C_1 __ D_1, Rule 2: $A \rightarrow B$ / C_2 __ D_2), the model compares C_1 with C_2, and D_1 with D_2. Working outwards from the structural change, it first locates the maximal right-side substring shared by C_1 and C_2; this shared substring forms part of the context for the new rule (C') — in this case, [s]. If C_1 and C_2 both contain additional unmatched material, then the segments immediately to the left of C' (here, [ɪ] and [ɛ]) are compared to see what features they have in common. If they share any feature specifications, these are retained as a left-side featural term (C'_{feat}), in this case, [+syllabic, –low, –back, –tense, –round]. Finally, if either C_1 or C_2 contains any additional material that has not been included in C' or C'_{feat}, this is converted into a free variable (X). The same procedure is carried out in mirror image on the right, yielding shared D' and D'_{feat} terms, and a right-side variable Y. Any of these terms may be null.

This generalization procedure retains as much shared material as possible, yielding the most specific rule that will cover both input forms. For this reason, we call it *minimal generalization*.

Minimal generalization is iterated over the data set. Iteration consists of comparing word-specific rules against other word-specific rules, and also against generalized rules.[3] The procedure for comparing a word-specific rule with a generalized rule is much the same as in (5), but with the complication that it is often necessary to compare a segment in the word-specific rule with a featural term (C'_{feat}, D'_{feat}) in the generalized rule.

The result of this procedure is a large list of rules, describing all of the phonological contexts in which each change applies. The fact that the model retains rules for each change means that it has the potential to generate multiple outputs for a novel input, satisfying one of the criteria we proposed in section 2.

In some learning models, the goal of rule induction is to find the most general possible rule for each change. However, as noted above, we also require our model to assign gradient well-formedness scores to each output. To do this, we evaluate the reliability of rules, then evaluate outputs on the basis of the rules that derive them.

3.2 Calculating reliability and confidence

The reliability of rules is calculated as follows. First, we determine the number of forms in the training data that meet the structural description of the rule (for $A \rightarrow B$ / C__D, these are the forms that contain CAD). This number is the *scope* of the rule. The *hits* of the rule is the number of forms that it actually derives correctly. The reliability of a rule is simply the ratio of its hits to its scope.

Intuitively, reliability is what makes a rule trustable. However, reliability based on high scope (for example, 990 correct predictions out of 1000) is better than reliability based on low scope (for example, 5 out of 5). Following Mikheev (1997), we therefore adjust reliability using lower confi-

[3] We believe, but have not proven, that no additional rules are discovered by comparing generalized rules against generalized rules.

dence limit statistics.[4] The amount of the adjustment is a parameter (α), which ranges from $.5 < \alpha < 1$; the higher the value of α, the more drastic the adjustment. The result of this adjustment value, which ranges from 0 to 1, we call *confidence*. Confidence values are calculated for each generalized rule, as soon as it is discovered. As each new input pair is processed, it is compared against previously discovered generalized rules to see whether it adds to their hits or scope. If so, their confidence values are updated.

The list of rules, annotated for confidence, can be used to derive outputs for novel (unknown) inputs. In some systems, rules are applied in order of decreasing specificity; the particular rule that is used to derive an output is the most specific one available. In our system, rules are applied in order of decreasing confidence. The novel form is compared against each known change $A_i \rightarrow B_i$ to see if it contains the input to the change (A_i). If so, the rules for that change are examined, in order of decreasing confidence, checking each rule to see if it is applicable. Once an applicable rule has been found, it is applied to create a novel output, and the next change ($A_{i+1} \rightarrow B_{i+1}$) is considered. Each output is assigned a well-formedness score, which is the confidence value of the rule that derives it; that is, the confidence value of the best available rule. These well-formedness scores allow the model to satisfy the second criterion laid out in section 2.

Minimal generalization and confidence values provide an effective method of discovering the phonological context in which a particular morphological change applies. Rules that describe productive processes in the correct context will have a very high confidence, whereas rules that describe exceptional processes or the wrong contexts will have lower confidence.

Moreover, when a change applies with especially high reliability in some particular context, the rule that the model discovers for this context will have especially high confidence. Thus, for example, the rule that suffixes [-t] in the context of final voiceless fricatives (§2), which is exceptionless and abundantly attested, is assigned an extremely high confidence value by our model.

3.3 Improving confidence with phonology

In many cases, it is possible to improve the confidence of morphological rules, and even expand their context, by discovering phonological rules. To continue with the example from (1) above, consider the rule that the model will generalize from the items [hʌg] and [rʌb]. In the feature system we use, the minimal natural class that covers both [g] and [b] is the set of voiced stops [b,d,g], so the model constructs a generalized rule that attaches [-d] after any member of this class.

Suppose that the model is presented next with the input pair ([nid], [nidəd]). It first attempts to update the confidence of the previously discovered generalized rules, including the rule adding [d] after voiced stops. Specifically, it tries to apply each rule to [nid], checking to see if the rule can derive the correct output [nidəd]. When it does this, it discovers that the [-d] affixation rule fails, producing instead the incorrect output *[nidd]. What we want the model to do in this situation is to recognize that [nidəd] is in fact an instance of [-d] affixation, but that there is an additional phonological process of [ə] insertion that obscures this generalization.

We allow the model to recognize this in the following way: first, we provide it ahead of time with a list of sequences that are illegal in English: *dd#, *td#, *fd#, *pd#, *bt#, and so on. (We believe that it is not unrealistic to do this, because experimental work (Jusczyk et al., 1993; Friederici and Wessels, 1993) suggests that children have a good notion of what sound sequences are legal in their language well before they begin to learn alternations.) When the learning model assesses the reliability of a rule and finds that it yields an incorrect output, it compares the incorrect output against the actual form, and hypothesizes a phonological rule of the form $A \rightarrow B / C \_\_ D$ that would

[4] 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:

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

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

change the incorrect form into the correct one. In this case, applying the [-d] suffixation rule to [nid] yields incorrect *[nidd], which is compared against correct [nidəd], and the phonological rule that is hypothesized is ∅ → ə / d__d.[5] Finally, the model examines the target of the phonological rule (CAD, in this case [dd]) to see if it contains a member of the list of known illegal sequences. If so, then the model has discovered a phonological rule that can help the morphological rule to produce the correct output, by fixing a phonologically illegal sequence. In the present case, the phonological rule allows [nid] to be counted as a hit for the morphological rule of [-d] suffixation, thus increasing the latter rule's reliability.

3.4 Overcoming complementary distribution

Unfortunately, not all phonological rules can be discovered by waiting for morphological rules to produce incorrect outputs. Consider how our model would analyze the pair ([mɪs], [mɪst]) 'miss(ed)'. Using the mechanisms described above, this would initially be treated as a case of [-t] suffixation. However, a more general analysis can be found if we realize that [-t] can be the result of /-d/ suffixation, with a phonological rule of devoicing that converts /-d/ to [-t] after a voiceless consonant. This could be achieved by having the model try attaching [-d] to [mɪs], yielding incorrect *[mɪsd], from which the devoicing rule could be discovered using the procedure described in the previous section. However, under the assumption of strictly minimal generalization, the opportunity to try [-d] after [mɪs] would never arise. The reason is that [-d] suffixation was learned solely on the basis of voiced stems, so it would never apply to a voiceless stem like [mɪs]. More generally, the [-d] and [-t] allomorphs of the English past tense suffix occur in complementary distribution, so a system that uses minimal generalization would never construct rules that attempt to use one allomorph in the environment of the other.

Our solution to this problem involves a slight relaxation of minimal generalization. The intuition is that when a new change is discovered (A → B, in this case ∅ → d), we should check to see if there are any potentially related changes that have already been discovered (A → B′, here ∅ → t) that take the same input (A), but yield a different output. The idea is that B and B′ might be the result of the same morphological rule, obscured by a phonological change.

To do this, we take every context that appears in a rule with change A → B and pair it with the change A → B′, creating a new set of rules, which we will call *cross-context rules*. For example, when the model encounters the first pair employing the ∅ → d change, it takes all of the existing ∅ → t rules and creates cross-context ∅ → d variants of them. The result is, among other things, a rule affixing [-d] after voiceless fricatives, mirroring the previously generalized rule affixing [-t] in the same environment.

The model then assesses the reliability of this cross-context rule, applying it to (among others) [mɪs] and deriving incorrect *[mɪsd]. By comparing this with the actual output [mɪst], the model posits a phonological rule for devoicing, in the same manner as described in the previous section. It then checks to see if the proposed phonological rule will enable the cross-context rule to produce the same output as the rule from which it was cloned in all cases. If so, the cross-context rule is kept, and can serve as the input for further generalization. Thus, the phonological rule is able to extend the set of contexts in which [-d] affixation successfully applies.

With these procedures in place, our model is able to discover a single rule that covers all English regular past tenses, namely ∅ → d / ___ #. The various regular past tense allomorphs are derived from /-d/ by phonological rules of voicing assimilation (deriving [-t]) and [ə] insertion (deriving [-əd]). We would guess that these are the rules that are assumed by most linguists; see Pinker and Prince (1988) for a detailed presentation. However, we discuss evidence below suggesting that simple [-d] affixation is not the only rule that derives regulars.

3.5 The grammar so far

We summarize here the grammar that is learned by our model (as described up until this point) when exposed to a representative corpus of English present-past pairs. The most general rule of the grammar is the noncontextual suffixation rule ∅ → d /

[5] The set of possible phonological rules is restricted to inserting a segment, deleting a segment, altering a segment, converting one segment into two (diphthongization), converting two segments into one (simple coalescence), or converting two segments into two others (length-preserving coalescence (/XZ/ → [YY]) and metathesis).

___ #; with the help of phonology this rule can derive all regulars. In addition, the model also discovers a large number of rules with lower generality. Many of these rules describe subgeneralizations about the regular process, for example, the highly reliable rule suffixing [-t] (or its underlying counterpart /-d/) after voiceless fricatives. Other rules describe exceptional processes, such as ɪ → ʌ before [ŋ] (*fling-flung*, *wring-wrung*, etc.), i → ɛ between a liquid and [d] (*bleed-bled*, *read-read*, etc.), and no change after [t] (*hit-hit*, *cut-cut*, etc.). In general, such exceptional processes will have much lower confidence than the regular rules, partly because they are based on fewer forms, and partly because there are regular forms that fail to obey them (*need-needed*, not **ned*).

Lastly, the model learns a large number of rules that could fairly be described as detritus, because they are never used in deriving any form (other, more reliable rules take precedence over them). In principle, we could prune these rules from the finished grammar, though we have not taken this step in our current implementation.

3.6 The distributional encroachment problem

Exceptional forms are easy to identify as such when they involve a change that occurs in only a few words, such as ɪ → ʌ. Not all exceptions have this property, however; sometimes exceptions are disguised by the fact that they involve a change that is regular, but in a different environment.

An example of this type of exception is seen in the past tense forms in (6), which occur in some dialects of English:

(6)	([bɚn]$_{\text{pres.}}$, [bɚnt]$_{\text{past}}$)	*'burn(t)'*
	([lɚn]$_{\text{pres.}}$, [lɚnt]$_{\text{past}}$)	*'learn(t)'*
	([dwɛl]$_{\text{pres.}}$, [dwɛlt]$_{\text{past}}$)	*'dwell(t)'*
	([spɛl]$_{\text{pres.}}$, [spɛlt]$_{\text{past}}$)	*'spell(t)'*
	([smɛl]$_{\text{pres.}}$, [smɛlt]$_{\text{past}}$)	*'smell(t)'*

These words form their past tense using one of the regular changes (∅ → t), but in the wrong environment (after sonorant consonants, rather than after voiceless ones). We call this type of exception *distributional encroachment*, because one morphological change is encroaching on the phonological context in which another change regularly occurs.

Distributional encroachment appears to be a major problem for all morphological learning systems that attempt to find large-scale generalizations. In what follows, we will explain why the example in (6) is problematic, then propose a method for coping with distributional encroachment in general.

Assume that prior to hearing any of the forms in (6), the model has already processed a fair number of regular stems ending in voiceless obstruents.[6] Comparing forms like [mɪs]-[mɪst] 'miss(ed)', [læf]-[læft] 'laugh(ed)', and [dʒʌmp]-[dʒʌmpt] 'jumped', the model would learn a number of rules of [t]-suffixation. Since [t] suffixation after voiceless obstruents is the regular outcome in English, these rules will achieve quite high confidence scores. Moreover, if we are willing to have a phonological rule that voices /-t/ to [-d] after a voiced obstruent, the context of /-t/ suffixation could be expanded to all obstruents. Under this analysis, past tense forms like *hugged* can now be derived as /hʌg/ → hʌgt → [hʌgd], so the confidence for this generalized rule would be even higher.

The distributional encroachment problem is encountered when the model, having reached this state, is confronted with one of the exceptional forms in (6). The result will be a serious overgeneralization. Suppose that the first such form encountered is [bɚn]-[bɚnt] 'burn(t)'. The model would first posit a single-form rule adding [-t] after the stem [bɚn]. Then, the generalization procedure would compare it with the other known [-t] affixation rules, all of which apply after obstruents. This comparison would lead to a generalized rule adding [-t] after any consonant at all: ∅ → t / [–syllabic]__#.

Let us now estimate the reliability of this generalized rule. Corpus counts show that the final segments of verb stems occur in roughly the following proportion in English:

(7)	Obstruents	60%
	Sonorant consonants	25%
	Vowels	15%

Suppose that prior to learning the form *burnt*, the model has learned 600 input pairs, of which 500 are regular and 100 are irregular exceptions, none

[6] The voiceless obstruents of English are [p, t, tʃ, k, f, θ, s, ʃ, h], and the voiced obstruents are [b, d, dʒ, g, v, ð, z, ʒ].

of them of the *burnt* type. Assume for simplicity that the distribution of final segments in both regulars and irregulars follows the proportions of (7). Thus, there will be 300 regular obstruent-final stems, and 60 obstruent-final exceptions, giving the rule attaching [-t] after obstruents a reliability of 300/360 = .83. Since 500 of the verbs are regular and 100 are irregular, the confidence of the rule attaching [-d] after any segment will be 500/600, which is also .83.

When the model encounters the pair ([bɚn], [bɚnt]), this adds a sonorant-final stem employing the ∅ → t change. The first step the model takes is to update reliability scores. Rules attaching [-t] after obstruents will be unaffected, since [n] is not an obstruent. The reliability of the rule attaching [-d] everywhere drops a minuscule amount, from 500/600 to 500/601. The second step is the fatal one: generalization with [bɚnt] gives rise to the new rule ∅ → t / [–syllabic]__#. This rule works correctly for 301 verbs (the 300 regular obstruent-final stems plus *burnt*), and fails for 210 verbs (the 60 obstruent-final exceptions, plus 150 verbs other than *burnt* that end in a sonorant consonant). Thus, its reliability would be 301/511, or .59. The prediction therefore is that for novel verbs that end in sonorant consonants, such as *pran* [præn], pasts with [-t] (*prant* [prænt]) should be at least moderately acceptable as a second choice, after the regular *pranned* [prænd]. We believe that this prediction is wrong; *prant* strikes us as absurd.

3.7 Impugnment as a solution to the distributional encroachment problem

The problem we are faced with is to let the model identify cases of distributional encroachment as such, and not be fooled into grouping *burnt* and *laughed* together under the same [-t] generalization. Intuitively, the problem with the rule attaching [-t] after any consonant is that it is internally heterogeneous; it consists of one very consistent subset of cases (the obstruent-final stems) and one fundamentally different case (*burnt*). We can characterize internal heterogeneity more precisely if we compare the scope and hits of the "correct" rule (after obstruents) and the "spurious" rule (after any consonant):

(8)	*Rule*	*Hits*		*Scope*
	∅ → t / [–sonorant]__#	300	/	360
	∅ → t / [–syllabic]__#	301	/	511

We see that the rule adding [-t] after any consonant gains just one hit, but adds a significant number of exceptions (150).

Formalizing this intuition, we propose a refinement of the way that confidence is calculated, in order to diagnose when a subpart of a generalization is doing most of the work of the larger generalization. When we consider the confidence of a context $\mathbb{C}$ associated with a change A → B, we must consider every other context $\mathbb{C}'$ associated with A → B, checking to see whether $\mathbb{C}'$ covers a subset of the cases that $\mathbb{C}$ covers. In the present case, when we assess the confidence of adding [-t] after any consonant, we would check all of the other rules adding [-t], including the one that adds [-t] after obstruents. For each $\mathbb{C}'$ that covers a subset of $\mathbb{C}$, we must ask whether the rule $A \rightarrow B / \mathbb{C}'$ is actually "doing most of the work" of the larger rule $A \rightarrow B / \mathbb{C}$.

To find out if the smaller rule is doing most of the work, we calculate how well the larger rule ($\mathbb{C}$) performs outside the area covered by the smaller rule ($\mathbb{C}'$). The reliability of the residue area ($\mathbb{C} - \mathbb{C}'$) is calculated as follows:

(9)
$$\text{Reliability}(\mathbb{C} - \mathbb{C}') = \frac{\text{hits}(\mathbb{C}) - \text{hits}(\mathbb{C}')}{\text{scope}(\mathbb{C}) - \text{scope}(\mathbb{C}')}$$

From the reliability of this residue area ($\mathbb{C} - \mathbb{C}'$), we can then calculate its confidence, using confidence limit statistics in a way similar to that described above in section 3.2. However, there is a crucial difference: when we are assessing whether a rule explains enough cases to be trustable, we are interested in the denseness of cases within the generalization. But when we are assessing whether a rule offers an improvement over a subpart, we are interested in the sparseness of cases in the residue outside of the subpart. Therefore, when calculating the confidence of the residue, we must use the *upper* confidence limit rather than the lower confidence limit.

If the upper confidence limit of the reliability of the residue ($\mathbb{C} - \mathbb{C}'$) is lower than the lower confidence limit of the reliability of the larger context ($\mathbb{C}$), then we can infer that the smaller rule ($A \rightarrow B / \mathbb{C}'$) is doing most of the work of the larger rule ($A \rightarrow B / \mathbb{C}$). Therefore, we penalize the larger rule by replacing its confidence value (Lower confidence($\mathbb{C}$)) with the confidence value of the residue (Upper confidence($\mathbb{C} - \mathbb{C}'$)). We call this penalty *impugnment*, because the validity of the larger rule is being called into question by the smaller rule. Impugnment is carried out for all contexts of all rules.

This impugnment algorithm is similar to the pruning algorithm proposed by Anthony and Frisch (1997). However, their algorithm requires that the smaller rule cover at least as many positive cases (hits) as the larger rule. In this case, the larger rule does cover one more case than the smaller rule (the form *burnt*), so it would not be eligible for pruning under their system. Impugnment is also similar to the pruning strategies based on "minimum improvement" or "lift" (e.g., Bayardo, Agrawal and Gunopulos 1999), but in this case, we are considering the improvement of a more general (less specified) context, rather than a more specific one, and the criterion of improvement is built in rather than user-specified.

3.8 The status of impugnment

We find that in general, impugnment suffices to relegate forms of the *burnt* class to the status of minor irregular classes, and thus saves the model from serious overgeneralization. Since distributional encroachment appears to be common in languages (Albright and Hayes 1999), we feel that impugnment or some other algorithm of equivalent effect is crucial for accurate morphological learning.

This said, we must add a somewhat puzzling postscript. In the experiment described below, we found that speakers gave forms like *prant* surprisingly high ratings. As a result, we found that we could achieve the closest match in modeling the experimental data by turning impugnment off. We feel that the high ratings for *prant* forms most likely were an artifact, reflecting the sociolinguistic status of *burnt* pasts (they are most often encountered by Americans as literary forms and may be felt to be prestigious). The upshot is that at present the empirical necessity of impugnment remains to be demonstrated.

4 Testing the Model

4.1 Training

Before a model can be tested, it must be trained on a representative learning set. For our studies of English past tenses, we used a corpus of 4253 verbs, consisting of all the verbs that had a frequency of 10 or greater in the English portion of the CELEX database (Burnage 1991). We trained our model to predict the past tense form from the present stem.

The model, implemented in Java,[7] accomplished its task fairly rapidly, learning the English past tense pattern in about 20 minutes on a 450 MHz PC. Most of this learning time was spent expanding and refining the more detailed rules; the broad generalizations governing the system were in place after only a few dozen words had been examined.

4.2 Corpus testing

As a first test of our model's performance, we divided the training data randomly into ten parts, and used the model to predict past tenses for each part based on the remaining nine tenths. For virtually every verb, the first choice of our model was the regular past tense, in its phonologically correct form: [-t], [-d], or [-əd], depending on the last segment of the stem. We consider this preference to be appropriate, given that English past tenses are on the whole a highly regular system; human speakers output irregulars only because they have memorized them.

4.3 Testing on novel forms

In our opinion, the most important criterion for a model like ours is the ability to deal with novel, made-up stems in the same way that people do. Novel stems access the native speaker's generative ability, abstracting away from whatever behavior results from memorization of existing verbs.

To begin, we have found that the model correctly inflects unusual words like Prasada and Pinker's (1993) *ploamph* and *smairg*; i.e. as [plomft] and [smergd]. The model can do this because it learns highly general rules that encompass these unusual items. Moreover, when confronted with the non-English sound [x] in *to out-Bach* [aʊtbax], our model correctly predicts [aʊtbaxt̲]. The model is able to do this because it can generalize using features, and thus can learn a rule that covers [x] based on phonetically similar segments like [f] and [k].

On a more systematic level, we have explored the behavior of the model with a carefully chosen

[7] Source code available at http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/

set of made-up verbs, which were rated both by our model and by groups of native speakers. We carried out two experiments, which are described in detail in Albright and Hayes (2001).

In our first experiment, we asked participants to complete a sentence by using the past tense of a made-up verb that had been modeled in previous sentences. For example, participants filled in the blank in the frame "The chance to *rife* would be very exciting. My friend Sam ___ once, and he loved it." Typically, they would volunteer *rifed*, or occasionally *rofe* or some other irregular form. In the second experiment, participants were given a number of choices, and rated each on a scale from 1 (worst) to 7 (best).

In selecting verbs to use in the experiments, we tried to find a set of verbs for which our learning model would make a wide range of different predictions. We began with a constructed corpus of phonologically-ordinary monosyllables (i.e. combinations of common onsets and rhymes), and used the model to predict past tenses for each. Based on these predictions, we selected four kinds of verbs, which according to the model:

I. should sound especially good as regular, but not as irregular
II. should sound especially good as (some kind of) irregular, but not as regular
III. should sound good both as regular and as some kind of irregular
IV. should not sound especially good either as regular or as any kind of irregular

Here are examples of all four categories. I. *Blafe* is expected to sound particularly good as a regular (because it falls within the scope of the high confidence voiceless-fricative rule), but not as an irregular. II. *Spling* is expected to sound especially good as an irregular (*splung*), because it fits a high-reliability [ɪ] → [ʌ] rule, but it is not predicted to be especially good as a regular. III. *Bize* is predicted to sound good as both a regular and an irregular, since it falls into a highly reliable context for regulars (final fricatives) and also falls into a highly reliable context for the [aɪ] → [o] change (before a coronal obstruent). IV. *Gude* is not covered by any especially reliable rules for either regulars or irregulars. The full set of verbs is given in the Appendix.

When we tested these four categories of made-up verbs, we found that our participants gave them ratings that corresponded fairly closely to the predictions of our model. Not only did participants strongly prefer regulars, as we would expect, but there was also a good match of model to data within the categories I-IV defined above. The following graphs show this for both regulars and irregulars (all responses are rescaled to the same vertical axis):

Figure 1. Mean ratings for regulars

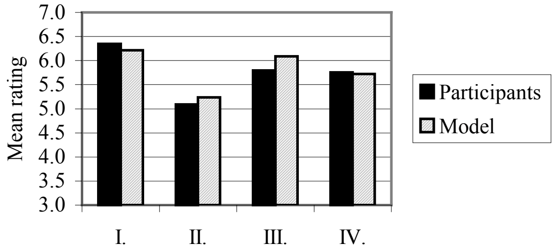
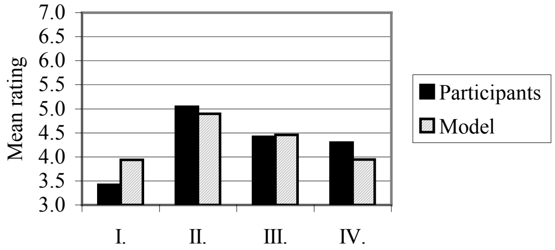


Figure 2. Mean ratings for irregulars



The graphs show data for participant ratings; similar results were obtained when we counted how frequently the various past tenses were volunteered when the participants were asked to fill in the blanks themselves.

As a more stringent test, we can examine not just mean values, but word-by-word predictions. A measure of this is the correlation between the model's predictions and the experimental results. The correlations are carried out separately for regulars and irregulars, since an overall correlation only establishes that the model knows that it should rate regulars highly.

(10) Ratings Data ($n = 41$)

regulars	$r = .745, p < .0001$
irregulars	$r = .570, p < .0001$

(11) Volunteered Data (% volunteered, $n = 41$)

regulars	$r = .695, p < .0001$
irregulars	$r = .333, p < .05$

In summary, our experiments validate a number of the model's predictions. First, participants prefer regulars over irregulars. Second, their intuitions are gradient, ranging continuously over the scale. Third, participants favor only those irregular forms that fall within a context characteristic of existing irregular verbs, like *-ing* ~ *-ung*. Finally, and most surprisingly, the participants followed the predictions of our model in favoring regular forms that can be derived by rules with high reliability.[8]

We conclude that our model captures a number of subtle but important patterns in the preferences of human speakers for past tense formation of novel verbs. Some of these preferences (e.g., the special preference for voiceless-fricative regulars) are not predicted by traditional linguistic analyses. We have obtained similar results in other languages (Albright 1999; to appear) We suspect that there may be many generalizations in morphology that are apprehended by native speakers but have been missed by traditional analysis. The use of machine learning may be useful in detecting such generalizations.

5 How the model could be improved

5.1 Phonological representations

Our model uses a very simple kind of phonological representation, from Chomsky and Halle (1968), and a very simple schema for rules ((5)). While this works well in systems that involve only local phonological generalizations, more complex systems are likely to require better representations if the correct generalizations are to be discovered. For example, the notion "closest vowel" is needed to characterize vowel harmony (e.g. Hungarian *könyv-nAk* → *könyv-nek* 'book-dative'). Our model cannot ignore the consonants that intervene between vowels, so it would not do well in learning this kind of rule. Our model also lacks any notion of syllables or syllable weight. Thus it could not learn the generalization that all polysyllabic English verb stems are regular (Pinker and Prince 1988); nor could it learn the distribution of the Latin abstract noun suffixes [-ia] and [-ieːs], which depends on the weight of the stem-final syllable ([graː.ti.a] 'favor', [kle.men.tia] 'mercy' vs. [maː.te.ri.eːs] 'matter'; Mester 1994). Lastly, the model lacks any notion of foot structure. Thus, it could not learn the distribution of the Yidin^y locative suffixes [-la] and [-ː] (prelengthening), which is arranged so that the output will have an even number of syllables, that is, an integral number of disyllabic stress feet ([ˈgabu][ˈd^{y}ula] 'clay-loc.' vs. [ŋuˈnaŋ][gaˈraː] 'whale-loc'; Dixon 1977). Phonological theory provides some of the means to solve these problems: theories of long-distance rules (e.g. Archangeli and Pulleyblank 1987), of syllable weight (McCarthy 1979), and of foot structure (Hayes 1982). We anticipate that incorporating such mechanisms would permit these phenomena to be learned by our system.

At the same time, however, we must consider the possibility that introducing new structures may expand the hypothesis space so much that it cannot be searched effectively by minimal generalization. Thus, where there are alternative phonological theories available, they should be assessed for whether they permit the right generalizations to be found without excessively expanding search time. It may also be possible to cut back on search time by using better algorithms for searching the hypothesis space.

5.2 Multiple changes

A number of morphological processes involve multiple changes, as in the German past participle *geschleppt* 'dragged', derived from *schlepp-* using both prefixation and suffixation. Our model (specifically, our method for detecting affixes) cannot characterize such cases as involving two simple changes, and would treat the relation as arbitrary. Two methods that might help here would be (a) to use some form of string-edit distance (Kruskal 1983), weighted by phonetic similarity, to determine that *-schlepp-* is the string shared by the two forms; (b) to adopt some method of morpheme discovery (e.g. Baroni 2000; Goldsmith 2001; Neuvel, to appear; Schone and Jurafsky 2001; Baroni et al. 2002) and use its results to favor rules that prefix *ge-* and suffix *-t*.

[8] Statistical testing reported in Albright and Hayes (2001) indicates that the effect on regulars cannot be attributed (entirely) to a "trade-off" effect with irregulars; i.e. *splinged* does not sound bad just because *splung* sounds good. In fact, the observable tradeoff effects are equally strong in both directions: some irregular forms sound worse because they also fall into a strong context for regulars.

Summarizing, we anticipate that improvements in the model could result from better phonological representations, better methods of search, and more sophisticated forms of string matching.

Appendix: Made-Up Verbs Used in the Experiments

I. Expected to be especially good as regular

blafe [blef], *bredge* [brɛdʒ], *chool* [tʃul], *dape* [dep], *gezz* [gɛz], *nace* [nes], *spack* [spæk], *stire* [staɪr], *tesh* [tɛʃ], *wiss* [wɪs]

II. Expected to be especially good as irregular

blig [blɪg], *chake* [tʃek], *drit* [drɪt], *fleep* [flip], *gleed* [glid], *glit* [glɪt], *plim* [plɪm], *queed* [kwid], *scride* [skraɪd], *spling* [splɪŋ], *teep* [tip]

III. Expected to be good both as regular and as irregular

bize [baɪz], *dize* [daɪz], *drice* [draɪs], *flidge* [flɪdʒ], *fro* [fro], *gare* [ger], *glip* [glɪp], *rife* [raɪf], *stin* [stɪn], *stip* [stɪp]

IV. Not expected to be especially good either as regular or as irregular

gude [gud], *nold* [nold], *nung* [nʌŋ], *pank* [pæŋk], *preak* [prik], *rask* [ræsk], *shilk* [ʃɪlk], *tark* [tark], *trisk* [trɪsk], *tunk* [tʌŋk],

References

Albright, Adam. 1999. Phonological subregularities in inflectional classes: Evidence from Italian. In Matthew Gordon, ed., *UCLA Working Papers in Linguistics, Vol. 1* (*Papers in Phonology 2*), pp. 1-47.

Albright, Adam. To appear. The productivity of infixation in Lakhota. To appear in Pamela Munro, ed., *UCLA Working Papers in Linguistics* (*Studies in Lakhota*).

Albright, Adam and Bruce Hayes 1999. Distributional encroachment and its consequences for phonological learning. *UCLA Working Papers in Linguistics* 4 (*Papers in Phonology* 4), 179-190.

Albright, Adam and Bruce Hayes. 2001. Rules vs. analogy in English past tenses: A computational/experimental study. Ms., Department of Linguistics, UCLA. http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/

Archangeli, Diana and Douglas Pulleyblank. 1987. Maximal and minimal rules: effects of tier scansion. In Joyce McDonough and Bernadette Plunkett, eds., *Proceedings of the North Eastern Linguistic Society* 17, Graduate Linguistics Student Association, University of Massachusetts, Amherst, pp. 16-35.

Baroni, Marco. 2000. Distributional cues in morpheme discovery: A computational model and empirical evidence. Ph.D. dissertation, UCLA. http://www.ai.univie.ac.at/~marco/

Baroni Marco, Johannes Matiasek, and Trost Harald. 2002. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. To appear in *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002*. http://www.ai. univie.ac.at/~marco/

Bayardo, R., R. Agrawal and D. Gunopulos. 1999. Constraint-based rule mining in large, dense databases. In ICDE-99. http://citeseer.nj.nec.com/bayardo99constraintbased.html

Burnage, G. 1991. *CELEX - A Guide for Users*. Nijmegen: Centre for Lexical Information, University of Nijmegen.

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

Daugherty, Kim G. and Mark S. Seidenberg. 1994. Beyond rules and exceptions: a connectionist approach to inflectional morphology. In S. D. Lima, R. L. Corrigan and G. K. Iverson, eds., *The Reality of Linguistic Rules.* Amsterdam: J. Benjamins.

Dixon, Robert M. W. 1977. *A Grammar of Yidin^y*. Cambridge: Cambridge University Press.

Dzeroski, Saso and Tomaz Erjavec. 1997. Learning Slovene declensions with FOIDL. In W. Daelemans, A. Van den Bosch, and A. Weijters, eds., *Workshop Notes of the ECML/Mlnet Workshop on Empirical Learning of Natural Language Processing Tasks*, Prague, pp. 49-60.

Eddington, David. 2000. Analogy and the dual-route model of morphology. *Lingua* 110:281–298.

Friederici, A. D. and Wessels, J. E. 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics* 54:287–295.

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27:153-198.

Halle, Morris. 1978. Knowledge unlearned and untaught: what speakers know about the sounds of their language. In Morris Halle, Joan Bresnan, and George Miller, eds., *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press, pp. 294-303.

Hayes, Bruce. 1982. Metrical structure as the organizing principle of Yidin^y phonology. In H. van der Hulst and N. Smith, eds., *The Structure of Phonological Representations, Part I.* Dordrecht: Foris Publications, pp. 97-110.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y. and Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* 32:402–420.

Jusczyk, P. W., Luce, P. A. and Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33:630–645.

Kruskal, J. B. 1983. An overview of sequence comparison. In D. Sankoff and J. B. Kruskal, eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley Publishing Company.

Ling, Charles X. and Marin Marinov. 1993. Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. *Cognition* 49:235–290.

McCarthy, John. 1979. On stress and syllabification. *Linguistic Inquiry* 10:443-465.

MacWhinney, Brian and Jared Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40:121-157.

Mester, Armin. 1994. The quantitative trochee in Latin. *Natural Language and Linguistic Theory* 12: 1-61.

Mikheev, Andrei. 1997. Automatic rule induction for unknown-word guessing. *Computational Linguistics* 23:405–423.

Mooney, Raymond J. and Mary Elaine Califf. 1995. Learning the past tense of English verbs using inductive logic programming. In Stefan Wermter, Ellen Riloff, and Gabriele Scheler, eds., *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*. Springer Verlag, pp. 370-384

Nakisa, Ramin. C., Kim Plunkett, and Ulrike Hahn. 2001. A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder and J. Murre, eds., *Models of Language Acquisition: Inductive and Deductive Approaches*. Cambridge, MA: MIT Press.

Neuvel, Sylvain. To appear. Whole word morphologizer. To appear in *Brain and Language*.

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

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.

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

Rumelhart, David. E. and McClelland, Jay L. 1986. On learning the past tenses of English verbs. In D. E. Rumelhart, J. L. McClelland, and The PDP Research Group, eds., *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Vol. 2*. Cambridge, MA: MIT Press, pp. 216-271.

Schone, Patrick and Daniel Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. *Proceedings of the North American chapter of the Association for Computational Linguistics* (*NAACL-2001*). http://www.colorado.edu/linguistics/jurafsky/pubs.html.

Simon, Anthony and Alan Frisch. 2001. Cautious induction in inductive logic programming. In *ILP97, The Seventh International Workshop on Inductive Logic Programming*, pp. 45-60.