Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study*

Adam Albright Bruce Hayes

Department of Linguistics

UCLA

November 2001

* Adam Albright and Bruce Hayes, Department of Linguistics, University of Califonia, Los Angeles. This research was supported by NSF grant BCS-9910686 and by an NSF Graduate Fellowship award to Adam Albright. Correspondence should be addressed to both authors, Dept. of Linguistics, UCLA, Los Angeles, CA 90095-1543; aalbrigh@ucla.edu, bhayes@humnet.ucla.edu.

Abstract

Are morphological patterns learned in the form of rules? Some models deny this entirely, attributing all morphological processes to analogical mechanisms. The dual mechanism model (Pinker & Prince, 1988) posits that speakers do internalize rules, but that these rules are few and cover only regular processes; the remaining patterns are attributed to analogy. We argue here for a third approach: a model that uses multiple stochastic rules and no analogy. This model employs inductive learning to discover multiple rules with different phonological contexts. These rules are assigned reliability scores according to their performance in the existing lexicon.

We evaluated a machine implemented version of our model using data from two "wug" test experiments on English past tenses. We found that participant ratings of novel pasts depended on the phonological shape of the stem. This held true for irregulars (*spling-splung* better than *glip-glup*), and, surprisingly, also for regulars (*blafe-blafed* better than *chake-chaked*). The ratings generally followed the statistical patterns of the English lexicon. For example, all verbs ending in voiceless fricatives are regular, and participants gave especially high ratings for regular pasts of wug verbs of this type, like *blafe*. These results are unexpected under a model that derives all regulars with a single rule, but they are predicted by our multiple-rule model.

We also argue against the hypothesis that all morphological processes are analogical. We implemented a version of Nosofsky's (1990) Generalized Context Model, which evaluates novel pasts based on their similarity to existing verbs. This analogical model underperformed our rule-based model in correlations to the wug test data. Moreover, it failed qualitatively in areas where rule-based and analogical treatments differ most saliently: it failed to locate patterns that require an abstract structural characterization, and it often favored implausible responses based on single, highly similar exemplars. We conclude that speakers extend morphological patterns based on abstract structural properties, of a kind appropriately described with multiple stochastic rules.

Rules vs. Analogy in English Past Tenses: A Computational/Experimental Study

1. Introduction: Rules in Regular and Irregular Morphology

Does language, as embodied in the mind/brain of the native speaker, employ rules? A major area in which this question has been debated has been inflectional morphology. Researchers in formal linguistic theory have commonly assumed that rules are the basis of all linguistic knowledge, including morphological knowledge. In contrast, many connectionists, dating from Rumelhart and McClelland (1986), have asserted that rules are an illusion suffered by the linguist, which vanishes under a more fine-grained, gradient approach to the data.

Out of this debate, there has also arisen a prominent compromise position: the *dual mechanism* approach advocated by Steven Pinker and his colleagues (Pinker & Prince, 1988, 1994; Pinker, 1999a; Clahsen, 1999). This approach adopts a limited set of rules to handle regular forms—in most cases just one, extremely general default rule—while irregular forms are handled not by rules but by an associative or analogical mechanism. According to this theory, rules are necessary for regulars, but they are inadequate to handle irregular forms because they do not explain the gradient similarity relations that characteristically hold between them (for example *cling-clung*, *fling-flung*, *dig-dug*, and so on).

The restriction of rules to regular processes has been a controversial feature of the dual mechanism approach. In a recent round of arguments (Clahsen, 1999, and responses), a number of critics have taken exception to this aspect of the model (Dressler, 1999; Indefrey, 1999; Wiese, 1999; Wunderlich, 1999). They note that traditional linguistic analyses frequently posit more than one rule per morphological process, and the rules posited often have a considerable amount of detail, in contrast to the extremely general rules often assumed by advocates of the dual mechanism approach.

The debate over the dual mechanism model forms the backdrop for our current study, because of the fundamental questions it involves: how many rules does a grammar contain? Which morphological phenomena are best described by rules, and which by analogy? The purpose of this paper is to argue for a model of morphology that employs many rules, including multiple rules for the same morphological process. We argue that this model makes predictions about morphological processes (both regular and irregular) that are more accurate than those of either the dual mechanism model or a purely analogical model.

Our strategy in testing the multiple-rule approach is inspired by a variety of previous efforts in this area. We begin by presenting a computationally implemented instantiation of our model; for purposes of comparison, we also describe an implemented analogical model, based on Nosofsky (1990) and Nakisa, Plunkett and Hahn (2001). Our use of implemented systems

follows a view brought to the debate by connectionists, namely, that simulations are the most stringent test of a model's predictions (Rumelhart & McClelland, 1986; MacWhinney & Leinbach, 1991; Daugherty & Seidenberg, 1994). We then present data from two new nonce-probe (*wug* test) experiments on English past tenses, allowing us to test directly, as Prasada and Pinker (1993) did, whether the models can generalize to new items in the same way as humans. Finally, we compare the performance of the rule-based and analogical models in capturing various aspects of the experimental data, under the view that comparing differences in how competing models perform on the same task can be a revealing diagnostic of larger conceptual problems (Ling & Marinov, 1993; Nakisa et al.).

2. Preliminaries

2.1 Rules and analogy

To begin, it will help to be explicit about what we mean by rules and analogy. The use of these terms varies a great deal, and the discussion that follows depends on having a clear interpretation of these concepts. This is especially crucial in light of Hahn and Chater's (1998) discussion of the overlap between rule-based and similarity-based models, and the difficulty of distinguishing them empirically.

Consider a simple example. In three wug testing experiments (Bybee & Moder, 1983; Prasada & Pinker, 1993; and the present study), participants have felt that *splung* [splʌŋ] is fairly acceptable as a past tense for *spling* [splɪŋ]. Plainly this is related to the fact that English has a number of existing verbs whose past tenses are formed in the same way: *swing*, *string*, *wring*, *sting*, *sling*, *fling*, and *cling*.[1] One possible account would be to say that *splung* is acceptable because *spling* is phonologically similar to many of the members of this set (cf. Nakisa et al., 2001, p. 201). In the present case, the similarity presumably involves ending with the sequence [ɪŋ], and perhaps also in containing a preceding liquid, *s*+consonant cluster, and so on. We will refer to any approach of this type, in which behavior on novel items is determined solely by their similarity to existing items, as analogical.

A rule-based approach, on the other hand, would involve generalizing over the data in some fashion, in order to locate a phonological context in which the [ɪ] → [ʌ] change is required, or at least appropriate. For example, it might discover an [ɪ] → [ʌ] rule restricted to the context of a final [ŋ], as in (1).

(1) $\text{ɪ} \rightarrow \text{ʌ} \; / \; \_\_\_ \; \text{ŋ} \;]_{[+past]}$

At first blush, the analogical and rule-based approaches seem to be different ways of saying the same thing—the context / ___ ŋ]$_{[+past]}$ in rule (1) forces the change to occur only in words that are similar to *fling*, *sting*, etc. But there is a critical difference. The rule-based approach requires

[1] The reader may have noticed that a number of English irregular verbs also form their past tenses by changing [ɪ] to [ʌ], but do not end in [ɪŋ]: *slink*, *stink*, *win*, *spin*, *dig*, and *stick*. The role of these verbs is discussed below in section 3.1.7.

that *fling*, *sting*, etc. be similar to *spling* in exactly the same way, namely by ending in /ɪŋ/. The structural description of the rule provides the necessary and sufficient conditions that a form must meet in order for the rule to apply. When similarity of a form to a set of model forms is based on a uniform structural description, as in (1), we will refer to this as **structured similarity**. A rule-based system can relate a set of forms only if they possess structured similarity, since rules are defined by their structural descriptions.

An analogical model, on the other hand, could allow each analogical form to be similar to *spling* in its own way. Thus, supposing hypothetically that English had verbs like *plip-plup* and *sliff-sluff*, then in a purely analogical model these verbs could gang up with *fling, sting*, etc. as analogical support for *spling-splung*, as shown in (2). When a form is similar in different ways to the various comparison forms, we will use the term **variegated similarity**.

(2)

Model form	s	p	l	ɪ	ŋ
fling-flung		f	l	ɪ	ŋ
sting-stung	s	t		ɪ	ŋ
"*plip*"-"*plup*"		p	l	ɪ	p
"*sliff*"-"*sluff*"	s		l	ɪ	f

There is nothing inherent in the analogical approach that prevents it from making use of variegated similarity. Therefore, analogical systems are potentially able to capture effects beyond the reach of structured similarity, and hence of rules. If we could find evidence that speakers form generalizations that rely on variegated similarity, then we would have good evidence that at least some of the morphological system is driven by analogy. In what follows, we attempt to search for such cases, and find that the evidence is less than compelling. We conclude that a model using "pure" analogy—i.e., pure enough to employ variegated similarity—is not restrictive enough as a model of morphology.

It is worth acknowledging at this point that conceptions of analogy are often more sophisticated than this, permitting analogy to zero in on particular aspects of the phonological structure of words (see section 6.3.1). However, when an analogical model is biased or restricted to pay attention to the same things that can be referred to in rules, it becomes difficult to distinguish the model empirically from a rule-based model. Therefore, following Hahn and Chater (1998), we have chosen to work with a formalization of pure analogy, which makes maximally distinct predictions by employing the full range of possible similarity relations.

2.2 Connectionism

In this light, we can explain why we have not included a connectionist simulation in this study. The problem is that a connectionist model is likely not to be a pure implementation of either rules or analogy. Certainly, connectionist models are commonly construed as being analogical. But it is quite possible for a network to mimic rules as well, by locating cases of structured similarity (Hanson & Burr, 1990; Dell, Reed, Adams, & Meyer, 2000). As Dell et al. note (p. 1357), "connectionist learning models are associated with flexibility in the specificity of what is learned. Some of the weight changes in the network can be characterized as the induction of 'rules' at various levels of generality." Thus, although it would certainly be interesting to

know whether connectionist networks could model our data, we find that for the purpose of evaluating the role of analogy in morphology, they are likely to yield inconclusive results.

It can be added that the results of a symbolic model are often easier to diagnose than those of a connectionist model (Clark & Karmiloff-Smith, 1993; Ling & Marinov, 1993; Hutchinson, 1994). In both models we employ, the basis on which the model rates a particular form in a particular way is always transparent.

2.3 Road Map

The remainder of this article is laid out as follows. First, we describe our models (the rule-based model, and its analogical counterpart). Second, we report the results of two wug test experiments we conducted on English past tenses. Next, we describe our attempts to model the wug test data, comparing in detail the performance of our two models. In the final section, we discuss some implications of our results.

3. Models

We believe that models of morphological learning must minimally possess three properties before they can be fully tested against human behavior. First, they should generate complete output forms for every word, rather than simply classifying them into coarse 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, models should be able to make multiple guesses for each word, because people, too, often entertain multiple possibilities. Lastly, models should assign well-formedness scores along a numerical scale to each output, rather than just giving a list of guesses. The scores permit comparison with comparable well-formedness scores assigned by people, and allow us to evaluate whether the models are able to capture gradient effects. Both our rule-based model and our analogical model satisfy these three criteria.

3.1 A Rule-Based Model

3.1.1 Finding Rules through Minimal Generalization

Our rule-based model builds on ideas from Pinker and Prince (1988, pp. 130-136). The basic principle is that rules can be gradually built up from the lexicon through a process of iterative generalization over pairs of forms. The starting point is to take each learning pair (here, a verb stem together with its past) and construe it as a rule; thus, for example, the stem-past pair *shine-shined*[2] [ʃaɪn]-[ʃaɪnd] is interpreted as "[ʃaɪn] becomes [ʃaɪnd]." Such rules can be factored into a **structural change** (here, addition of [d] in final position) and an invariant **context** (the part that is shared; here, the stem [ʃaɪn]), as in (3).

[2] *Shine* is a regular verb when transitive: *He shined his shoes*.

(3) ∅ → d / [ʃaɪn ___]$_{[+past]}$ = "Insert [d] after the stem [ʃaɪn] to form the past tense ."

We will refer to such one-form rules as "degenerate," because they express no generalization, but simply encode the data in a form that permits further generalization to take place.

Generalization is carried out by comparing rules with one another. Suppose that at some later time the algorithm encounters *consign-consigned,* spawning another degenerate rule:

(4) ∅ → d / [kənsaɪn ___]$_{[+past]}$

Since the structural change (∅ → d) in (4) is the same as the change in (3), it is possible to combine (3) and (4) to create a more general rule, as illustrated in (5).

	change		variable	shared features	shared segments	change location		
(5)a.								
b.	∅ →	d / [		ʃ	aɪn	___	]$_{[+past]}$	(*shine-shined*)
c.	∅ →	d / [	kən	s	aɪn	___	]$_{[+past]}$	(*consign-consigned*)
d.	∅ →	d / [	X	$\begin{bmatrix}+\text{strident}\\+\text{contin}\\-\text{voice}\end{bmatrix}$	aɪn	___	]$_{[+past]}$	(generalized rule)

The strategy here is to find the tightest rule that will cover both cases; hence we refer to the procedure as **minimal generalization**. In the present case, minimal generalization works like this: moving outward from the location of the change, any segments shared by the two degenerate rules (here, [aɪn]) are retained in the generalized rule. Where two segments differ, but can be grouped together as a more abstract category using phonological features, this is done to create a featural term. In our feature system, the sounds [ʃ] and [s] can be characterized as [+strident, +continuant, –voice]. Lastly, once featural generalization has been carried out for one segment, any further mismatches (here, [kən] mismatched to null) are resolved by adding a free variable ('X') to the generalized rule. When the change is medial, as in the [ɪ] → [æ] change of *sing-sang,* the search for shared material is carried out in parallel on both sides of the structural change. For a full description of the minimal generalization algorithm, see Albright and Hayes (1999).

3.1.2 Features

Phonological features play two roles in the minimal generalization algorithm. First, they permit it to achieve tighter and more accurate generalizations. For instance, the regular English past tense suffix has three phonetically distinct allomorphs: [-d] (as in *rubbed*), [-t] (as in *jumped*), and [-əd] (as in *voted* or *needed*). Of these, [-əd] attaches only to stems ending in [t] or [d]. When the algorithm compares the degenerate rules for *vote* and *need,* shown in (6a,b), it is crucial that it not immediately generalize all the remaining material to a free variable, as in (6c). If it did this, then [-əd] could be attached everywhere, yielding impossible forms like **jumpèd*

[dʒʌmpəd]. Instead, our implementation uses features to produce a much more conservatively generalized rule, namely (6d). The features [+coronal, +anterior, -nasal, -continuant] uniquely characterize the class [t, d]. Thus, the system will correctly attach [-əd] after only these sounds.

(6) a. $\varnothing \rightarrow$ əd / [vot ___]$_{[+past]}$

b. $\varnothing \rightarrow$ əd / [nid ___]$_{[+past]}$

c. $\varnothing \rightarrow$ əd / [X ___]$_{[+past]}$ (too general)

d. $\varnothing \rightarrow$ əd / [X $\begin{bmatrix} +\text{coronal} \\ +\text{anterior} \\ -\text{nasal} \\ -\text{continuant} \end{bmatrix}$ ___]$_{[+past]}$ (appropriately restricted)

Features also permit the system to generalize to segments it has never seen before. Pinker (1999a), adapting Halle (1978), gives a vivid example: an English speaker who can produce the velar fricative [x] will, in saying "Handel out-Bached ([baxt]) Bach," employ the [-t] allomorph of the regular past. This reflects the fact that [x] has the features of a voiceless consonant, but not those of an alveolar stop. Our rule-based model guesses [baxt] correctly, even if the input data does not contain [x], since the featural term permits it to discover contexts like "after voiceless segments."

3.1.3 Phonology

Outputs of morphological processes are often shaped by principles of phonological well-formedness. This is true of English past tenses, where the choice of regular allomorph is often guided by such principles (Pinker & Prince, 1988, pp. 101-108).

Our rule-based learner makes use of phonological principles to derive correct outputs. Suppose that the learning data include regular stems ending in [b], [g], and [n] (e.g. *rub-rubbed*, *sag-sagged*, *plan-planned*). The rule-based learner will invoke the featural generalization process to arrive at a rule that attaches [-d] to any stem ending in a sound that is [+voice, –continuant]. However, this class also includes [d], so that the generalized rule would predict incorrect forms like **needd* [nidd]. This incorrect prediction cannot be avoided by purely morphological means, because there is no combination of features that includes [b], [g], and [n] without also including [d].[3] Rather, the reason that the past tense of *need* is not [nidd] is phonological: *[dd] is not a possible final sequence in English.

Two different approaches to eliminating phonologically ill-formed outputs like *[nidd] have been proposed in the literature. In an approach using **phonological rules** (Bloomfield, 1939; Chomsky & Halle, 1968), the morphology is allowed to suffix [-d] to [nid], and the resulting *[dd] cluster is repaired by a phonological rule breaking up the illegal cluster with a schwa: /nid+d/ → [nidəd]. In **constraint-based** approaches (Prince & Smolensky, 1993; Bird, 1995),

[3] [b] and [g] are voiced stops, [n] is alveolar, and [d] is a voiced alveolar stop; hence any feature combination that includes [b], [g], and [n] will also include [d].

multiple candidate outputs compete (e.g. [nidd] and [nidəd]), and some of them are filtered by phonological constraints; thus a constraint like *[dd] eliminates [nidd].

Our rule-based learner is designed to accommodate either phonological rules or constraints. The various morphological rules it learns will generate candidate outputs [nidd] and [nidəd]. Armed with the knowledge that words cannot end in [dd], the learner can either filter out [nidd] (constraint-based approach) or discover a rule that converts /nidd/ to [nidəd] (rule-based approach). In either case, it is assumed that the phonologically illegal sequences are already known, prior to morphological learning.[4] In modeling our experimental data, we tried both approaches. It emerged that the constraint-based approach yielded slightly better results (and much better results for the analogical model discussed below), so we adopted it for purposes of the present study.

3.1.4 Iterative Generalization and Rule Evaluation

As morphological rules are learned, the first stages of generalization tend to produce rather arbitrary and idiosyncratic rules like (5d). However, when the process is iterated, comparing already generalized rules with other rules, increasingly general rules are discovered. Fairly quickly, rules emerge that are sufficiently general to cover all of the pairs in the learning set that share a particular change.

For English past tenses, the degree of generality that is attained depends on whether phonology is implemented by rules or constraints. When allowed to discover phonological rules (schwa insertion and voicing assimilation; Pinker & Prince, 1988, pp. 105-106), our procedure yields a completely general suffixation rule, which attaches [-d] to any stem (7a). If constraints are used, each of the three regular past tense allomorphs must be handled separately, as in (7b).

(7) a. $\varnothing \rightarrow$ d / [X ___]$_{[+past]}$

b. $\varnothing \rightarrow$ d / [X [+voice] ___]$_{[+past]}$

$\varnothing \rightarrow$ t / [X [–voice] ___]$_{[+past]}$

$\varnothing \rightarrow$ əd / [X $\begin{bmatrix} \text{+coronal} \\ \text{+anterior} \\ \text{–nasal} \\ \text{-continuant} \end{bmatrix}$ ___]$_{[+past]}$

Either way, in the process of arriving at these rules, the system also creates a large number of other, less general rules. What should be done with these rules? One option, advocated by Pinker and Prince (1988, 134), is that we should keep only those rules that are maximally *general* (as defined by number of forms they correctly derive); all other rules should be

[4] In making this assumption we rely on recent experimental work (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Friederici & Wessels, 1993; Jusczyk, Luce, & Charles-Luce, 1994; and for discussion Hayes, in press). This research indicates that infants learn a great deal about the legal sound sequences of their language by the age of ten months, probably well before they tackle morphological problems like past tenses.

discarded, having played their role. However, in modeling our experimental data, we have found that a more effective criterion is the *accuracy* of a rule in capturing the distribution of patterns in the language.

Our rule-based model assesses accuracy by collecting some simple statistics about how well rules perform in deriving the forms in the learning data. For example, (7a) $\varnothing \rightarrow$ [d] / [X ___]$_{[+past]}$, the most general rule for English pasts, is applicable everywhere; hence its **scope** (as we will call it) is equal to the size of the data set. For the learning data employed here (see section 3.3), this value is 4,253. If phonological rules are employed, this rule derives the correct output for all 4,034 regular forms; that is, it achieves 4,034 **hits**. To calculate an accuracy score for the rule, we divide hits by scope, obtaining a tentative score (which we call **raw confidence**) of .949. The rule [ɪ] $\rightarrow$ [ʌ] / {l, r} ___ ŋ, which covers past tenses like *sprung*, has a scope of 9 and 6 hits, hence a raw confidence of .667.

Generalizations can be trusted better when they are based on larger numbers of forms. Following Mikheev (1997), we use a lower confidence limit on raw confidence to penalize rules based on a small number of forms; thus, for instance, if the lower confidence limit (α) is 75%, a score of 5 correct outcomes out of 5 applicable cases is downgraded from 1.00 to an **adjusted confidence** of 0.825. A case of 1000 correct outcomes out of 1000 cases, however, is downgraded only from 1.000 to 0.999. The lower confidence limit is a parameter of the model, set by finding the value that best fits consultant intuitions; in modeling our experimental data, this value was α = .55. Generalizations can also be trusted better if the forms that instantiate them are uniformly distributed within the context they describe. For this purpose, we use upper confidence limits to penalize nonuniform distributions. For full discussion of how this works and why it is needed, see Albright and Hayes (2000). The value of the upper confidence limit was also set by fitting to experimental data, at α = .95.

3.1.5 Islands of Reliability

Intuitively, assessing the accuracy of rules in this way should allow us to locate the "correct" rules to describe the input data. In practice, however, the most accurate rules are rarely the ones that would traditionally be included in a grammar. Consider the following fact: every verb of English that ends in a voiceless fricative ([f, θ, s, ʃ]) is regular. (There are 352 such verbs in our learning dataset.) The minimal generalization algorithm, comparing forms like *missed* [mɪst], *wished* [wɪʃt], and *laughed* [læft], constructs a rule that covers just this subset of the regulars:

(8) $\varnothing \rightarrow$ t / [X $\begin{bmatrix} \text{–sonorant} \\ \text{+continuant} \\ \text{–voice} \end{bmatrix}$ ___]$_{[+past]}$ "Suffix [-t] to stems ending in voiceless fricatives."

The adjusted confidence of this rule is .998, which is higher than the general rules of (7).

The question at hand, therefore, is what is the status of highly accurate rules like (8) in the final grammar. The hypothesis we adopt and test here is that such rules are retained alongside more general context-free rules; that is, speakers know the contexts in which the regular change can be relied upon to a greater than average extent. We will refer to phonological contexts in

which a particular morphological change works especially well in the existing lexicon as **islands of reliability**. Naturally, islands of reliability occur for both regular and irregular changes.

It is in giving a grammatical status to islands of reliability that we most sharply part company with traditional linguistic analysis, which has (to our knowledge) generally contented itself with locating the single best formulation of a rule for any given pattern. Thus, the empirical evidence we present below concerning islands of reliability for regular forms bears on questions of linguistic theory itself, in addition to questions of morphological learning.[5]

3.1.6 Generating outputs

Giving rules probabilistic confidence values allows the rule-based model to generate multiple, competing outputs with numerical confidence values attached. When an input form is submitted to the model for wug testing, it is compared against all the rules in the grammar. Each rule that can apply does so, deriving a candidate output form. Naturally, in the usual cases, many rules will involve the same change and thus derive identical outputs. We assume that the candidate output is assigned the well-formedness scores of the best rule that derives it.

As an illustration of how the model works, here are the outcomes it derives for the wug verb *gleed*, along with their raw confidence values and the adjusted values.

Table 1: Past tenses for *gleed* derived by the rule-based learner

Output	Rule	Hits /Scope	Raw Conf.	Ad-justed Conf.	Hits/Failures
gleeded	$\varnothing \rightarrow$ əd / [X {d, t} ___]$_{[+past]}$	1146/1234	.929	.872	*want*, *need*, *start*, *wait*, *decide*, *etc.* / **get*, **find*, **put*, **set*, **stand*, etc.
gled	i $\rightarrow$ ɛ / [X {l, r} ___ d]$_{[+past]}$	6/7	.857	.706	*read*, *lead*, *bleed*, *breed*, *mislead*, *misread* / **plead*
glode	i $\rightarrow$ o / [X C ___ [+cons]]$_{[+past]}$	6/184	.033	.033	*speak*, *freeze*, *weave*, *interweave*, *bespeak* / **leak*, **teach*, **leave*, etc.
gleed	no change / [X {d, t} ___]$_{[+past]}$	29/1234	.024	.021	*put*, *shed*, *let*, *set*, *cut*, *hit*, *spread*, *beat*, *shut*, *hurt*, *cost*, *cast*, *burst*, *split*, etc. / **get*, **want*, **need*, etc.

[5] An alternative formulation of our claim is that there is just one rule for regulars, but it is annotated with a large set of contexts indicating where it can be applied with greater confidence. At least for present purposes, this differs from a multiple-regular-rule approach only in terms of economy of expression, and not empirically; so we will stick with multiple-rule terminology here.

3.1.7 Excursus: "Family Resemblance" and Prototypicality

Our rule-based treatment of *gleed* contrasts with a view held by Bybee and Slobin (1982) and by Pinker and his colleagues (Pinker & Prince, 1988, 1994; Prasada & Pinker, 1993; Pinker, 1999a,b). These scholars suggest that phenomena found in irregulars involving "prototypicality" or "family resemblance" imply that an adequate rule-based account is impossible. We quote Pinker (1999b):

> Just as we have a rule adding "ed" to form the regular past tense, we [could have] a suite of rules that generate irregular past tense forms by substituting vowels or consonants. For example, one rule changes "i" to "u" in verbs like "cling, clung"... A problem for this theory is the family resemblance among the verbs undergoing the rule, such as "string, strung", "sting, stung", "fling, flung", "cling, clung." How do you get the rule to apply to them?

Pinker goes on to suggest various possibilities. A rule like ɪ → ʌ / [X ___ Y]$_{[+past]}$ would be too general, because it lacks the phonological context that seems to be affiliated with the change. Thus Pinker notes that the verbs *fib*, *wish*, and *trip* are regular (cf. **fub*, **wush*, **trup*). On the other hand, a contextual rule like ɪ → ʌ / [X ___ ŋ]$_{[+past]}$ would be too specific, because there is a set of marginal forms that also change [ɪ] to [ʌ], but don't "quite" meet the crucial condition. For example, *stick-stuck* has a final velar consonant, but it is not nasal; while *spin-spun* has a final nasal consonant, but it is not velar. Pinker concludes that rules are fundamentally unable to capture irregular processes; instead, they must be derived by a mechanism that relies on prototypicality and family resemblance.[6]

We feel that this conclusion is premature, and that a rule-based approach can be adapted to account for prototypicality effects. First, we agree with dual mechanism theorists (and most of traditional linguistic theory) that irregulars are lexically listed; this is what prevents them from being regularized (Aronoff, 1976). Thus, we need not require that the rules for irregulars succeed in covering all forms perfectly. Rather, these rules characterize the (modest) productivity of the various irregular patterns, as seen in acquisition data and experimental work.

Second, we assume that grammars may contain multiple rules with the same structural change (e.g., [ɪ] → [ʌ]), but different confidence values. In our model, the cluster of [ɪ] → [ʌ] verbs gives rise to a cluster of rules, having varying degrees of generality. For example, the canonical forms *cling*, *fling*, and *sling* lead to a rule that characterizes them with considerable precision: ɪ → ʌ / [[–voice] l ___ ŋ]$_{[+past]}$. This rule works in 3/3 cases and yields a score of .718. But if *fub* were to be the past tense of *fib*, it would have to be generated by the more general rule ɪ → ʌ / [X C ___ [+voice, –continuant]]$_{[+past]}$. This rule has 11 hits (adding *win*, *swing*, *dig*, *spring*, *spin*, *sting*, *wring*, and *string*); but it also has a much larger scope (44), because it encompasses many forms like *bring*, *grin* and *rig*. As a result, the score for *fub* would

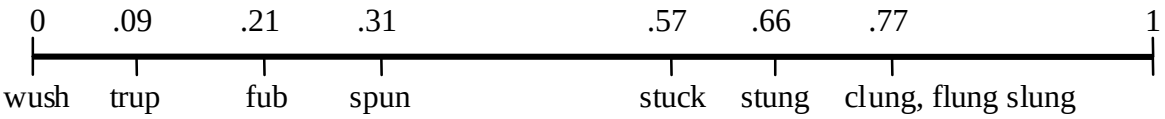
---

[6] Pinker also objects to ɪ → ʌ / [X ___ ŋ]$_{[+past]}$ because the forms *bring-brought* and *spring-sprang* would be exceptions to it. This strikes us as inconsistent with a position he adopts elsewhere, namely, that languages have rules for regular processes even when these rules suffer from exceptions. We see no reason why a stricter standard should necessarily be maintained for rules describing irregular processes.

be only .206. Similarly, to generate *trup* for *trip*, we would have to generalize even further, to ɪ → ʌ / [X C ___ [-continuant]]$_{[+past]}$. This rule has 12 hits (adding *stick-stuck*), but a much larger scope (110), so the score sinks to .092. Moreover, this rule encompasses all [ɪ] → [ʌ] verbs in the training set, and under the assumption of minimal generalization, no further [ɪ] → [ʌ] rules are created. *Wish* falls outside the scope of this rule (since its final segment is [+continuant]), and thus Pinker's hypothetical form *wush* would not be derived at all.

Intermediate cases like *stick* and *spin* are more complicated, since they make use of special left-side environments that improve their score slightly. However, their scores fit the general picture, which is summarized below:

(9)



Summing up, it is not at all clear to us that there is anything about the "family resemblance" phenomenon that makes it unamenable to treatment by multiple rules in a system of the sort we are proposing. One final note on this point: the graph in (9), despite superficial appearances, is *not* a metric of the similarity of *wish*, *trip*, etc. to the core verb set. The values are computed using the entire learning set, by assessing the effectiveness of rules. In other words, our rule-based model, unlike the analogical model discussed below, does not make direct use of similarity.

3.1.8 Summary of the Rule-Based Model

Our rule-based model locates a rich set of generalizations about morphological processes by proceeding bottom up from the lexicon. Generalization is minimal, in that it never proceeds beyond what is permitted by the learning data, within the format provided for rules. However, despite the tightness of minimal generalization, the model can also locate very general—even context-free—rules. This happens when a change occurs in a heterogeneous set of environments. The model provides gradient ratings of well formedness for each output, defined as the score of the best rule that derives it. The score for a rule is defined as the lower confidence limit of the ratio of its hits to its scope. The model can incorporate either rule-based phonology or filtrative phonology.

We turn now to our second learning model, which is designed to work very differently.

3.2 An Analogical Model

In developing a model that works purely on an analogical basis, we have adopted the Generalized Context Model (GCM) described in Nosofsky (1990). This model was proposed as a very general account of how similarity influences people's intuitive judgments; it is supported by a variety of data from domains outside language. Nakisa et al. (2001) have adapted the GCM to the analysis of English past tenses, and our own implementation follows their work in many respects.

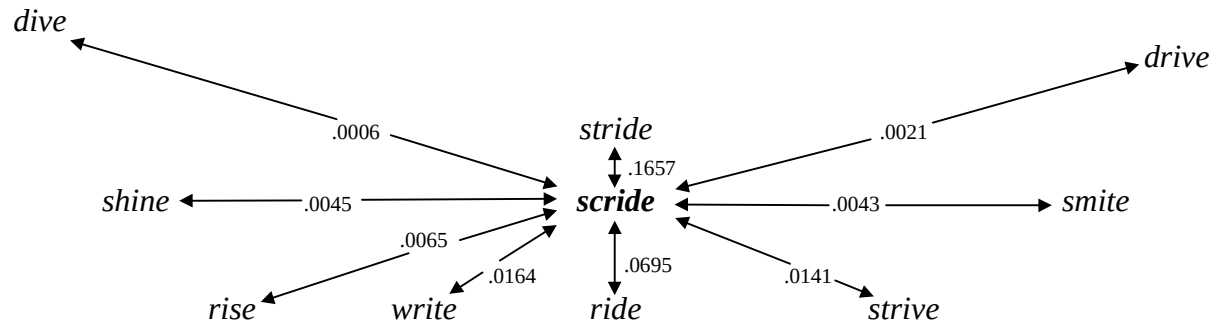
3.2.1 The Generative Front End

The GCM model does not generate, but only evaluates, candidates. To plug this gap, we augmented our implementation of the GCM with a generative front end, which is simply a clone of that portion of the rule-based learner that locates all the possible structural changes for past tense formation. This module creates candidates by applying all applicable structural changes freely. Thus, for a stem like *scride* [skraɪd], the module constructs (a) candidates with the three allomorphs for the regular past ([skraɪd**əd**], [skraɪd**d**], [skraɪd**t**]; (b) a no-change candidate ([skraɪd]; cf. *hit*); (c) a candidate changing /d/ to [t] ([skraɪt]; *bend*); and (d) candidates with the five vowel changes applicable to [aɪ]: *scrode* (*ride*), *scrid* (*hide*), *scroud* [skraʊd] (*find*), *scrud* (*strike*), and *scraud* [skrɔd] (*fight*). Of these, [skraɪdd] and [skraɪdt] are phonologically filtered; the remaining candidates are submitted to the core GCM algorithm for evaluation.

3.2.2 The Core of the Model

The intuitive idea behind the GCM can be given with a simple example. Suppose we wish to evaluate the well-formedness of the particular candidate *scrode* as the past tense of *scride*. The verbs in our learning database that share the structural change [aɪ] → [o] are *dive, drive, ride, rise, shine, smite, stride, strive,* and *write.* Assume that we possess a numerical measure of the similarity of *scride* to each member of this nine-verb set (this measure is described below). By adding the similarity values together, we obtain a measure of the similarity of *scride* to the [aɪ] → [o] class in general. This number will be larger: (a) the more verbs there are in the [aɪ] → [o] class; and (b) the more similar each of the [aɪ] → [o] verbs is to *scride*. It is intuitive, we think, that this number should correlate with the goodness of *scrode* as an output form. The scheme is illustrated below. The arrows are labeled with the actual similarity values used in our model.

(10) *Similarity of All* [aɪ] → [o] *Forms to* scride



To this basic idea, the GCM adds an important adjustment: we must compensate for how much *scride* resembles the verbs of English in general. This is done by summing the similarity of *scride* to all the verbs of the learning set, and dividing the total obtained in the previous paragraph by the result. Thus, the score that the model gives to *scrode* is:

(11) $$\frac{\textit{summed similarity of}\ \text{scride}\ \textit{to all members of the}\ [\text{aɪ}] \rightarrow [\text{o}]\ \textit{class}^{7}}{\textit{summed similarity of}\ \text{scride}\ \textit{to all verbs}} = \frac{.2837}{2.09} = \mathbf{.1358}$$

Well-formedness scores are calculated in similar fashion for all candidates, for all of the wug verbs under examination.

We turn next to the issue of how the similarity values are derived. We employ a theory of word similarity, based on a theory of segmental similarity.

3.2.3 Segmental Similarity

Our account of segmental similarity is taken from the work of Broe (1993). The similarity of two segments is defined as the ratio:

(12) $$\frac{\textit{number of shared natural classes}}{\textit{number of shared natural classes + number of non-shared natural classes}}$$

A natural class is a central notion in phonological theory; it is defined as a set of segments that share the same values for some particular set of phonological features. Thus, for instance, rounded vowels, voiced obstruents, and labial consonants are all natural classes. In defining natural classes, we used a fairly standard inventory of features, and calculated the ratio of shared to total natural classes using a computer program written for this purpose (Zuraw, 1999). For example, in our feature set [s] and [ʃ] ("sh") share 174 natural classes; and there are 32 natural classes to which one sound belongs but not the other; hence the similarity score is 174/(174 + 32) = .845 out of a possible 1.

The use of this metric of segmental similarity is justified, we think, by its strong track record in modeling speech errors and English phonotactics (Frisch, 1996), and in modeling the phonotactics of Arabic verb stems (Frisch, Broe, & Pierrehumbert, 1997).

3.2.4 Word Similarity

We convert segmental similarity (which, being a proportion, ranges from 0 to 1) to dissimilarity by subtracting from one; thus [s] and [ʃ] receive a dissimilarity score of .155. The dissimilarity of a pair of words can then be defined by pairing up their segments and summing the dissimilarity values of the paired segments. In addition, a specified numerical penalty is assessed for every segment (in either word) that is unpaired. The latter is a parameter of the model, established by fitting to data.

---

[7] In the full version of the model (Nosofsky, 1990; Nakisa, Plunkett & Hahn, 2001), there are additional factors. The model needs to know the frequency with which a given present stem takes a particular past; this value is 1 for almost all verbs, except for rare cases like *dive* ~ {*dived, dove*}. For simplicity, we assume 1 for all cases, and simply treat the two variants of *dive* as separate verbs. In addition, each output pattern is associated with a bias term; since we lack any principled means for assigning differences in bias terms, we follow Nakisa et al. in assuming a value of 1 everywhere.

In pairing the segments of two words with each other, we face the critical problem of **alignment**: normally, words are similar if there is a good correspondence of their segments. Thus, for example, *parade* and *prayed*, which seem quite similar, will yield a very good similarity score if they are aligned as in (13a), but they will yield a very poor score if they are left justified, as in (13b).

(13) a.	*parade*:	p	ə	r	e	d		b.	p	ə	r	e	d
	prayed:	p	null	r	e	d			p	r	e	d	null

Right justification would work better, but it cannot be the correct solution in general, since it would fail miserably for a pair like (say) *elephant*/*elephants*.

It appears that the best solution to this problem is to explore all possible alignments (see Kruskal, 1983), and pick the one that yields the lowest word dissimilarity score.[8] This procedure finds the correct alignments for *parade*/*prayed*, and appears in general to locate sets of forms that seem to us to be genuinely similar.

Suppose now that we are evaluating the well-formedness of past tenses for *scride*. We need to calculate the similarity of *scride* to all verbs in the learning set. For the particular verbs *shine* and *write*, here are the best alignments found, with the summed dissimilarities:

(14)	*shine*:	ʃ		null		null		a		ɪ		n			
	penalty:	.155	+	.4	+	.4	+	0	+	0	+	.667	=	1.622	
	scride:	s		k		r		a		ɪ		d			
	write:	null		null		r		a		ɪ		t			
	penalty:	.4	+	.4	+	0	+	0	+	0	+	.434	=	1.234	
	scride:	s		k		r		a		ɪ		d			

It can be noted that our procedure skips [kr] in matching *shine* to *scride*, but skips [sk] in matching *write* to *scride*. It thereby obtains the best available matches ([ʃ]-[s] and [r]-[r]) for segmental similarity. We conjecture that this is an improvement over the GCM model developed by Nakisa et al., which permits only rightward alignment, and thus would be forced to align the [ʃ] of *shine* with the less similar [r] of *scride*.

The penalty used here for matching segments to null, .4, strikes us intuitively as rather low. However, since it yielded the best fit to experimental data, it was adopted here. Larger values of the penalty yielded rather similar correlations, however.

When polysyllabic, words may be dissimilar in virtue of having a different stress pattern. We assess a penalty for stress matches, set to .6 by fitting to the experimental data.

[8] In our implementation, we follow Nakisa, Plunkett, and Hahn in assuming that vowels will not be aligned with consonants; presumably such alignments normally incur fairly extreme prosodic dissimilarity.

3.2.5 The Similarity/Dissimilarity Mapping

The last part of the model is the conversion of dissimilarity values to similarity. As it turns out, the way one does this conversion has crucial effects on the overall behavior of the model. We use the following equation (Nosofsky, 1990; Nakisa et al., 2001):

(15) $\eta_{ij} = e^{(-d_{ij}/s)^p}$ where

η_{ij} is the calculated similarity of two forms *i* and *j*
d_{ij} is the dissimilarity of *i* and *j*
e is the base for natural logarithms
s and *p* are parameters, fixed by fitting to the data

The parameter *s* in equation (15) turns out to have a fairly clear interpretation in practice: when *s* is low, the model tends to rely primarily on a small set of very similar forms in forming its judgments (i.e., all forms do make some positive contribution, but the formula renders the contribution of dissimilar forms very small). When *s* is large, the model is less sensitive to local similarity. As *s* approaches infinity, η_{ij} approaches 1 for all *i* and *j*, so the algorithm reduces to letting the learning data "take a vote." The effect of *p* is subtler and will not be reviewed here. The best-fit values for *s* and *p* turned out to be .4 and 1, respectively.

Applying the formula to the value $d_{shine,scride} = 1.622$, obtained above, we get .0045, which is the similarity value that appeared earlier in Fig. 1.

3.2.6 Generating Outputs

As an illustration of how the model works, Table 2 shows the outcomes it derives for *gleed*, along with their scores and the analog forms used in deriving each outcome.

Table 2: Past tenses for *gleed* derived by the analogical model

Output	Score	Analogs
gleeded	0.3063	*plead*, *glide*, *bleat*, *pleat*, *bead*, *greet*, *glut*, *need*, *grade*, *gloat*, and 955 others in our learning set
gled	0.0833	*bleed*, *lead*, *breed*, *read*, *feed*, *speed*, *meet*, *breast-feed*
gleed	0.0175	*bid*, *beat*, *slit*, *let*, *shed*, *knit*, *quit*, *split*, *fit*, *hit*, and 12 others
gleet	0.0028	*lend*, *build*, *bend*, *send*, *spend*
glade	0.0025	*eat*
glode	0.0017	*weave*, *freeze*, *steal*, *speak*
glud	0.0005	*sneak*

3.2.7 Summary of the Analogical Model

To sum up: in our GCM-based analogical model, the well-formedness of a wug output is the sum of its similarity to all forms undergoing the same change, divided by its summed similarity

to the whole data set. Similarity is calculated as a warped version (15) of dissimilarity, which is the minimized sum of segmental dissimilarity, dissimilarity to null, and stress dissimilarity.

We feel that a model of this sort satisfies a rigorous criterion for being "analogical". It plainly accesses variegated similarity, and (unless the data accidentally help it to do so) it utterly ignores the structured-similarity relations that are crucial to the functioning of our rule-based model.

3.3 Feeding the Models

We sought to feed both our rule-based and analogical models a diet of stem+past pairs that would resemble what had been encountered in the life experience of our experimental participants. We took our set of input forms from the English portion of the CELEX database, selecting all the verbs that had a frequency of 10 or greater. (Going into lower frequencies, we found many verbs that we anticipated would not be familiar to our experimental participants.) In addition, for verbs that show more than one past tense (like *dived*/*dove*), we included both (e.g. both *dive-dived* and *dive-dove*). The resulting corpus consisted of 4,253 stem/past tense pairs, of which 4,035 were regular and 218 were irregular.

Since our experimental participants were speakers of American English, we Americanized the British English pronunciations of the CELEX database, using a combination of translation rules and hand checking.

A current debate in the acquisition literature (Clahsen & Rothweiler, 1992; Marcus et al., 1995; Bybee, 1995) concerns whether prefixed forms of the same stem (e.g. *do*/*redo*/*outdo*) should be counted separately for purposes of learning. We prepared a version of our learning set from which all prefixed forms were removed, thus cutting its size down to 3,308 input pairs (3,170 regular, 138 irregular), and ran both learning models on both sets. [9] As it turned out, the rule-based model did slightly better on the full set, and the analogical model did slightly better on the edited set. The results below report the performance of each model on its own best learning set.

Another question in the theory of morphological learning concerns whether learning proceeds on the basis of types vs. tokens. In learning based on type frequency, all verbs in the learning set are given equal influence; in token-based learning, each verb is weighted by its frequency, e.g. in calculating hits or scope (rule-based model) or in counting the similar forms (analogical model). Bybee (1995, 2001) and Pierrehumbert (in press) have both argued that morphological patterns are extended on the basis of type frequency. Our results are consistent with this view, as both of our models match the experimental data somewhat better when they are run using types rather than tokens. The results reported below are based on type frequency.

---

[9] The notion "prefixed" is open to many possible definitions. Since our goal was to eliminate forms in which learners could easily recognize and remove the prefix, we used the following definition: a form was considered prefixed if it began with a known English prefix (or a noun or verb, in the case of compounds), the remaining stem occurred as a free stem, and the prefix was judged to contribute its "canonical" meaning to the word. Thus, for example, *disappear* and *disintegrate* were considered prefixed because the meaning of *dis-* is apparent in these words, but *disturb* and *display* were not.

3.4 Relating the Models to Data

One long-standing tradition in learning theory evaluates models by training them on part of the learning data, then testing them on the remainder. When we tested our models in this way, we found that both produced regular outputs as their first choice virtually 100% of the time.[10] We think that in a system as regular as English past tenses, this is probably the correct way for the models to behave. English speakers by and large favor irregular pasts only when they have memorized them as part of their lexicon. Occasionally they do prefer irregulars, and even innovate an irregular form like *dove* or *dwelt*. However, we think this is best attributed to the probabilistic nature of their grammars, which often gives an irregular form a status *almost* as good as the corresponding regular.

Ling and Marinov (1993, pp. 264-5) make a good case that testing against the learning corpus is not the right way to evaluate models in any event. The problem is that real speakers have the benefit of having memorized the irregulars, and the models do not; hence expecting the models to reproduce existing irregulars that they have never seen, simply by guessing, is unrealistic.

A better way to assess the models is to administer to them a wug test that has also been given to people. Here, we can be sure that models and people are on equal footing; both must use their capacity to generalize in order to decide how novel words should be treated, unaffected by factors like memory or frequency that would be involved with real verbs.

4. Experiments

We carried out two experiments on English past tenses, both of them modeled loosely on Prasada and Pinker (1993). In Experiment 1, participants were given a variety of wug verbs in the stem form, and volunteered past tense forms. In Experiment 2, in addition to volunteering past tense forms, participants also provided ratings of possible past tenses: the regular, and one or two possible irregular forms. Phonological well-formedness ratings of all of the wug stems were also collected in Experiment 1, in order to be able to factor out this potential confound in subsequent analyses.

For both experiments, wug verbs were presented and gathered exclusively in spoken form. This permitted us to avoid biasing the participants toward particular responses with the spelling

[10] We tested each system by randomly dividing the learning set in ten, and modeling each tenth using the remaining nine tenths as input data. The exact results depend on the parameter settings used. For the rule-based model, using parameter settings selected by fitting to the experimental data, 4192 of the 4199 forms were output as regular; 3 (*withstand*, *take*, *partake*) was retained as irregular, and 4 were irregularized (*clink-clunk*, *deride-derode*, *plead-pled*, *stake-stook*). For the analogical model, again using parameter settings selected to best model experimental data, 4198/4199 forms were output as regular; one (*stink*) was retained as irregular, and no forms were irregularized. Unlike the rule-based model, the analogical model occasionally used the wrong regular suffix, as in *bandièd* [ˈbændiəd] and *taxi't* [ˈtæksit]. Such errors occurred 1.2% of the time; we discuss them below in section 5.3.5.

of the wug verbs. Using spoken responses from the participants also avoided uncertainty about what they actually intended, since English orthography is notoriously ambiguous.

4.1 Stimuli

Our wug verbs were chosen to test a number of different hypotheses, and to this end were divided into what we will call a Core set and a Peripheral set.

4.1.1 The Core Set

The Core set was designed to test the following hypotheses:

(16) a. If a verb falls into an island of reliability for **irregular** pasts, will it receive higher ratings?

b. If a verb falls into an island of reliability for **regular** pasts, will it receive higher ratings?

(Recall that an island of reliability is a phonological context in which a particular morphological change works especially well in the existing lexicon; section 3.1.5). The questions in (16) are roughly the same as those asked by Prasada and Pinker (1993), substituting "falls into an island of reliability for" for "is phonologically close to". Prasada and Pinker's experiments were intended to show, we think, that the answer to question (16a) is "yes" (replicating Bybee & Moder, 1983), and to question (16b) is "no."

Prasada and Pinker designed their novel verbs using informal methods, such as finding verbs that rhymed with many regulars/irregulars, or changing just one phoneme vs. multiple phonemes to obtain greater distance. One problem with this approach is that it provides no quantitative control for how many existing rhymes a novel verb has, how similar they are, and so on. In addition, as Prasada and Pinker themselves note, this procedure introduces a confound: the only way for a novel verb to be dissimilar to all existing regulars is for it to be dissimilar to *all* English words. As a result, the verbs in Prasada and Pinker's "distant from existing regulars" condition were phonologically deviant as English words; e.g. *ploamph* and *smairg*.

In fact, such verbs did receive low participant ratings, which on the face of it suggests that regular processes *are* sensitive to islands of reliability (16b). However, as Prasada and Pinker point out, it is also possible that their participants disliked regular pasts like *ploamphed* and *smairged* because of their phonological deviance; i.e., *ploamphed* may have been a perfect past tense for *ploamph*, but received low ratings because of its phonologically deviant stem. Prasada and Pinker attempted to correct for this statistically by subtracting stem phonological well-formedness ratings from past tense ratings; when this is done, the difference between close similarity and distant similarity pseudo-regulars appears to vanish. However, such a result would surely be more persuasive if the confound had not been present in the first place. It seems fair to say that Prasada and Pinker's negative result for regulars is ambiguous and open to interpretation, because of the way in which novel verbs were created.

In an attempt to circumvent this problem in designing our own wug verbs, we used (a slightly earlier version of) our rule-based computational model as a tool for experimental design. We started by constructing a set of 2344 candidate wug forms, by concatenating combinations of

relatively common syllable onsets and relatively common syllable rhymes.[11] By starting with only phonologically "bland" candidate forms, we minimized the possibility that our past tense data would be influenced by phonological well-formedness. The entire list of potential wug forms was then submitted to our model, which generated and rated the regular past and several irregulars for each. We inspected this output, searching for forms to fill the four-way matrix in Table 3.

Table 3: Design of the Core set of wug stems

Stem occupies an island of reliability for both the regular output and at least one irregular output.	Stem occupies an island of reliability for the regular output only.
Stem occupies an island of reliability for at least one irregular output, but not for the regular output.	Stem occupies no island of reliability for either regular or irregular forms

Perhaps surprisingly, it was possible to fill all four cells of this matrix. The islands for regulars and irregulars form cross-classifying categories, and it is not the case that being in an island of reliability for regulars precludes being in an island for irregulars. For example, the novel stem *dize* [daɪz] meets the structural description for an [aɪ] → [o] rule that covers *rise*, *ride*, and *dive*, but it also meets the structural description for a very reliable regular rule suffixing [d] to stems that end in [z] (*suppose*, *realize*, *raise*, *cause*, and 211 others).

In filling the cells of the four-way matrix, we sought to find not just extreme cases, but rather a variety of "island strengths." This permitted a wider variety of islands to be included, and also facilitated correlation analysis by providing data that were closer to being normally distributed.

The wug verbs chosen for the four basic categories of the Core set are shown in Table 4. In the third and sixth columns, we include the irregular forms that were provided as options for participants to rate in Experiment 2 (normally just one, but occasionally two). They were devised partly by examining the outputs of the algorithmic learner, and partly by examining the forms volunteered in Experiment 1.

[11] We used here type frequency, since some highly unusual onsets (e.g. [ð]) occur in just a few very common words. In addition, we checked the preliminary list by hand, removing any forms where the onset-rhyme juncture struck us as phonologically unusual.

Table 4: Wug verbs (Core set)

Present stem	**Rated past**	**2nd rated past**	**Present stem**	**Rated past**	**2nd rated past**
a. Island of reliability for both regulars & irregulars			**b. Island of reliability for regulars only**[12]		
bize [baɪz]	*boze* [boz]		*blafe* [blef]	*bleft* [blɛft]	
dize [daɪz]	*doze* [doz]		*bredge* [brɛdʒ]	*broge* [brodʒ]	
drice [draɪs]	*droce* [dros]		*chool* [tʃul]	*chole* [tʃol]	
flidge [flɪdʒ]	*fludge* [flʌdʒ]		*dape* [dep]	*dapt* [dæpt]	
fro [fro]	*frew* [fru]		*gezz* [gɛz]	*gozz* [gaz]	
gare [ger]	*gore* [gor]		*nace* [nes]	*noce* [nos]	
glip [glɪp]	*glup* [glʌp]		*spack* [spæk]	*spuck* [spʌk]	
rife [raɪf]	*rofe* [rof]	*rif* [rɪf]	*stire* [staɪr]	*store* [stor]	
stin [stɪn]	*stan* [stæn]	*stun* [stʌn]	*tesh* [tɛʃ]	*tosh* [taʃ]	
stip [stɪp]	*stup* [stʌp]		*wiss* [wɪs]	*wus* [wʌs]	
c. Island of reliability for irregulars only			**d. Island of reliability for neither regs nor irregs**		
blig [blɪg]	*blug* [blʌg]		*gude* [gud]	*gude* [gud]	
chake [tʃek]	*chook* [tʃʊk]		*nold* [nold]	*neld* [nɛld]	*nold* [nold]
drit [drɪt]	*drit* [drɪt]	*drat* [dræt]	*nung* [nʌŋ]	*nang* [næŋ]	
fleep [flip]	*flept* [flɛpt]		*pank* [pæŋk]	*punk* [pʌŋk]	
gleed [glid]	*gled* [glɛd]	*gleed* [glid]	*preak* [prik]	*preck* [prɛk]	*proke* [prok]
glit [glɪt]	*glit* [glɪt]	*glat* [glæt]	*rask* [ræsk]	*rusk* [rʌsk]	
plim [plɪm]	*plum* [plʌm]	*plam* [plæm]	*shilk* [ʃɪlk]	*shalk* [ʃælk]	
queed [kwid]	*qued* [kwɛd]		*tark* [tark]	*tork* [tork]	
scride [skraɪd]	*scrode* [skrod]	*scrid* [skrɪd]	*trisk* [trɪsk]	*trusk* [trʌsk]	*trask* [træsk]
spling [splɪŋ]	*splung* [splʌŋ]	*splang* [splæŋ]	*tunk* [tʌŋk]	*tank* [tæŋk]	
teep [tip]	*tept* [tɛpt]				

4.1.2 The Peripheral Set

The Peripheral set of wug verbs was intended both to add to the diversity of forms, and also address some additional questions of interest. Eight verbs, listed in (17), were included that resembled existing verbs of the *burnt* class, in which [-t] is exceptionally attached to stems ending in /l/ or /n/. The real *burnt* verbs, which are not found in all dialects, include *burn*, *learn*, *dwell*, *smell*, *spell*, *spill*, and *spoil*. The reason for our particular interest in these verbs is described in Albright and Hayes (2000).

[12] Originally, this set includes two additional forms, *mip* [mɪp] and *slame* [slem]. These proved to be very often misperceived by the consultants (as [nɪp] and [slen]), so they were discarded from the analysis.

(17)	Present stem		Rated Past		2nd Rated Past	
	grell	[grɛl]	*grelt*	[grɛlt]		
	skell	[skɛl]	*skelt*	[skɛlt]		
	snell	[snɛl]	*snelt*	[snɛlt]	*snold*	[snold]
	scoil	[skɔɪl]	*scoilt*	[skɔɪlt]		
	squill	[skwɪl]	*squilt*	[skwɪlt]		
	murn	[mərn]	*murnt*	[mərnt]		
	shurn	[ʃərn]	*shurnt*	[ʃərnt]		
	lan	[læn]	*lant*	[lænt]		

The verbs in (18) were included because they are *not* supported by reasonable islands of reliability for any irregular form, but nevertheless closely resemble particular irregulars. We hoped to see if these verbs might give rise to effects that could be unambiguously interpreted as analogical.

(18)	Present stem		Rated Past		Real Model
	kive	[kɪv]	*kave*	[kev]	*give-gave*
	lum	[lʌm]	*lame*	[lem]	*come-come*
	pum	[pʌm]	*pame*	[pem]	*come-came*
	shee	[ʃi]	*shaw*	[ʃɔ]	*see-saw*
	zay	[ze]	*zed*	[zɛd]	*say-said*

The forms *chool-chole* and *nold-neld*, which were included for other reasons (Table 4), also served as potentially analogical cases, based on their similarity to *choose* and *hold*.

The remaining forms in (19) also relied on close similarity to a very few forms, rather than a rule-like pattern. *Shy'nt*, [13] *ry'nt*, and *gry'nt* were chosen because although they are phonetically similar, the closest existing verbs form their past tenses differently (*shone*/*wrote* vs. *ground*), so that they could serve as comparison test for individual-verb analogies.

(19)	Present stem		Rated Past		2nd Rated Past		Real Model
	chind	[tʃaɪnd]	*chound*	[tʃaʊnd]	*chind*	[tʃaɪnd]	*find-found*
	shy'nt	[ʃaɪnt]	*shoant*	[ʃont]	*shount*	[ʃaʊnt]	*shine-shone*
	gry'nt	[graɪnt]	*groant*	[gront]	*grount*	[graʊnt]	*grind-ground*
	ry'nt	[raɪnt]	*roant*	[ront]	*rount*	[raʊnt]	*write-wrote*
	flet	[flɛt]	*flet*	[flɛt]			*let-let*

4.2 Experiment 1 Procedure

Experiment 1 consisted of two parts: the first obtained baseline phonological well-formedness scores for each of the wug stems; the second elicited past tense forms of wug verbs in a production task. The experiment (as well as Experiment 2, below) was conducted in the sound booth of the UCLA Phonetics Laboratory. Twenty native speakers of American English

[13] This is our attempt to spell [ʃaɪnt], which rhymes with *pint*.

(predominantly UCLA undergraduates) were paid $10 for their participation in Experiment 1, which lasted one hour.

4.2.1 Phonological well-formedness ratings

In order to assess the possible confounding influence of phonological well-formedness on morphological intuitions, all of the wug stems were rated for phonological well-formedness in the first part of Experiment 1. For reasons discussed in section 4.1.1 above, the wug stems were all designed to be well-formed English words; thus, in addition to the 60 target wug forms, 30 additional ill-formed fillers were included as foils.

The wug stems and fillers were presented twice over headphones, first in isolation, and then in a simple frame sentence; e.g., "*Grell*. John likes to *grell*." Participants repeated the wug stem aloud ("*Grell.*"), in order to confirm that they had heard the novel word correctly, and then rated the naturalness of the stem on a scale from 1 (worst) to 7 (best):

(20) Scale for phonological well-formedness ratings

1	2	3	4	5	6	7
Completely bizarre, impossible as an English word			Not so good, but imaginable as an English word			Completely normal, would make a fine English word

Participants were instructed to rate novel words according to how natural, or English-like they sounded on first impression. If a rating was not entered within six seconds after the end of the audio stimulus, a message would appear indicated that time for that trial had expired. Five novel verbs were chosen as training stimuli: *bzarshk* [bzarʃk], *kip* [kɪp], *pint* [pɪnt], *plake* [plek], and *sfoond* [sfund].

Stimuli for both Experiments 1 and 2 were presented using Psyscope (Cohen, MacWhinney, Flatt, & Provost, 1993), and participants entered ratings directly using a specially modified keyboard. The phonological ratings portion of Experiment 1 took approximately 10 minutes to complete.

4.2.2 Volunteering wug pasts

In the second part of Experiment 1, participants volunteered past tense forms for all of the wug verbs listed in section 4.1, in an open response sentence completion task. The purpose of eliciting past tense forms in this way was twofold: first, we wished to compare the likelihood of volunteering novel past tense forms against the well-formedness ratings collected in Experiment 2. In addition, the volunteering portion of Experiment 1 aided in designing the ratings portion of Experiment 2, since we wanted to be sure to include all of the forms that participants were likely to volunteer.

For the sentence completion task, each wug verb was embedded in a frame dialog consisting of four sentences. In the first two sentences, the verb occurred in its stem form. In the third sentence, the verb appeared in a context that would require a present participle, and in the fourth, it appeared in a context requiring the past tense. Participants heard the first two sentences over

headphones. The frame dialogs were displayed simultaneously on a computer monitor, but with blanks in place of the wug verbs. Participants were instructed to read sentences 3 and 4 aloud, filling in the blanks with appropriately inflected forms of the given wug verbs.

(21)	**Screen:**	**Headphone input:**
Sentence 1	I dream that one day I'll be able to ___.	"I dream that one day I'll be able to *rife*."
Sentence 2	The chance to ___ would be very exciting.	"The chance to *rife* would be very exciting."
	Screen:	**Participant reads:**
Sentence 3	I think I'd really enjoy ___.	"I think I'd really enjoy [*response*]."
Sentence 4	My friend Sam ___ once, and he loved it.	"My friend Sam [*response*] once, and he loved it."

For example, for the dialog in (21), participants were expected to supply *rifing* for sentence 3, and *rifed*, *rofe*, or some other past tense form for sentence 4. The full set of dialogs used is given in Appendix A.

Responses for sentences 3 and 4 were recorded and transcribed by two listeners with phonetic training. Sentence 3 required participants to attach *-ing*, which is a completely regular morphological operation of English. Thus it could be used as a check to confirm that participants had heard and internalized the wug verb correctly. If either listener recorded something other than the expected *–ing* form for sentence 3, then the past tense response in sentence 4 was discarded for that trial. A total of 62 out of 1160 trials were discarded for this reason.

For the volunteering portion of Experiment 1, there was a training period of 5 verbs. In the first training dialog, we used a real English verb (*leap*), so participants could get used to the task; the remaining four training verbs were made-up verbs. Participants were instructed to complete the dialogs using whatever form of the made-up verb seemed most natural to them; they were also reminded that there were no right or wrong answers, and that we were merely interested in their opinion about how they would use the made-up verbs.

Each participant completed 60 frame dialogs, one for each of the wug verbs. The order of the wug verbs was randomized on a subject-by-subject basis. A total of 15 different frames were used, meaning that participants saw each frame 4 times, with a different wug verb each time. Each wug verb was embedded in three different frame dialogs, which were varied between subjects. In this way, no particular wug verb was seen in the exact same frame dialog by all participants, minimizing the chance that responses for a particular wug verb would be biased by some unintentional semantic influence of a particular frame dialog.

The volunteering portion of Experiment 1 lasted approximately half an hour, with a rest period halfway through.

4.3 Experiment 2 Procedure

The format of Experiment 2 was the same as the volunteering portion of Experiment 1, except that in addition to volunteering past tense forms, participants also provided acceptability ratings of various possible forms. Twenty-one participants, none of whom had participated in Experiment 1, were paid $10 each for their participation, which lasted approximately 45 minutes.

Wug stems were once again presented auditorily, using the same frame dialogs as in Experiment 1. Participants heard two sentences containing the wug verb in its stem form, and had to read two sentences aloud, providing correctly inflected present participle and past tense forms. This volunteering component was included because otherwise, with only auditory presentation of wug verbs and a purely passive ratings task, it would have been difficult to ensure that participants had heard and internalized the wug verbs correctly. By requiring participants to repeat the wug verbs in subsequent sentences, we were able to use these responses to check that they were paying attention, and rating the intended verbs. As in Experiment 1, participant responses were transcribed by two listeners, and if either transcriber recorded an unexpected/incorrect participial form, all subsequent data from that trial was excluded (119 out of 1,416 trials altogether).

After participants had completed the fill-in-the-blank portion of the dialog, they then heard an abbreviated version of the dialog, with either a regular or irregular past tense form provided for them to rate. Upon rating this form, they heard the voice repeat the mini-dialog once again, this time with the opposite past tense form to rate. The purpose of repeating the mini-dialog each time was to encourage participants to consider the goodness of novel pasts *in relation to the given wug stem*. The full protocol is shown in (22).

(22) *Frame dialog for ratings task*

Sentence 1:	[voice]	"I dream that one day I'll be able to *rife*."
Sentence 2:	[voice]	"The chance to *rife* would be very exciting."
Sentence 3:	[participant]	"I think I'd really enjoy _____."
Sentence 4:	[participant]	"My friend Sam _____ once, and he loved it."
Sentence 5:	[voice]	"I dream that one day I'll be able to *rife*. My friend Sam *rifed* once, and he loved it."
	(*participant rates*)	
Sentence 6:	[voice]	"I dream that one day I'll be able to *rife*. My friend Sam *rofe* once, and he loved it."
	(*participant rates*)	

Participants were instructed to rate each past tense option according to how natural it sounded *as the past tense of the verb*, on a scale of 1 (worst) to 7 (best):

(23) Scale for past tense acceptability ratings

1	2	3	4	5	6	7
completely bizarre, impossible as the past tense of the verb			not so good, but imaginable as the past tense of the verb			completely normal, would make a fine past tense of the verb

Unlike the phonological well-formedness rating task of Experiment 1, for past tense ratings there was no time limit to enter a rating.

Each participant rated each possible past tense form for all wug verbs; for most wug verbs, there were only two possible past tense forms provided (the regular and one irregular), but for eleven wug verbs, two different irregulars were provided (see section 4.1). The order of items to rate (regular first vs. irregular first) varied from item to item, but was counterbalanced in such a way that each form was rated in each position an equal number of times, and each participant saw an equal number of regulars first and irregulars first. As before, the order of wug items was randomized on a subject-by-subject basis, and the carrier frame for each wug verb was varied between subjects.

The training period for Experiment 2 consisted of four items. The first was designed to introduce participants to the idea of comparing multiple past tense forms: *frink* [frɪŋk], with past tenses *frank* [fræŋk], *frunk* [frʌŋk], and *fret* [frɛt]. When participants heard the form *fret*, they were reminded that sometimes a form could sound quite ordinary as an English past tense, but could nonetheless be an implausible way to form the past tense *of the nonsense verb in question* (in this case, *frink*). The remaining three training items were *pint* [pɪnt], past *punt* or *pinted*; *kip*, past *kap* [kæp] or *kipped*; and *prack*, past *pruck* or *pracked*.

4.4 Coding the Results

4.4.1 Correcting for phonological well-formedness

Recall from section 4.1.1 that any past tense wug experiment faces a potential confound: forms may receive lower ratings either because they are bad *as past tenses*, or because they are *phonologically deviant*; only the first of these is of interest here. We attempted to minimize the effect of phonological deviance by choosing wug verbs that were phonologically very bland. As it turned out, the phonological ratings data largely confirmed our hope that phonological well-formedness would have little effect on the past tense ratings. The average phonological rating for our wug verbs was 4.68 (s.d. = 1.62, n = 58), whereas the average rating for our ill-formed foils (rated phonologically, but not included in the wug tests) was 2.97 (s.d. = 1.46, n = 29). More important, the phonological ratings data were poorly correlated with the participants' ratings of past tense forms: $r(58) = .006$.[14] Thus, it seems that our scheme for avoiding major phonological ill-formedness effects was successful.

Nevertheless, as an added precaution, we also used the phonological well-formedness ratings gathered in Experiment 1 to try to correct for phonological effects in the ratings data. This correction was carried out as follows: first linear regressions were performed, trying to predict the regular and irregular past tense ratings of Experiment 2 using the phonological well-formedness ratings from Experiment 1. The residuals of this regression were then rescaled, so

[14] The comparable value for Prasada and Pinker's (1993) study was $r = .214$. We have reason to believe that the greater role of phonological well-formedness in Prasada and Pinker's study is due to the inclusion of strange forms and not to more accurate phonological well-formedness ratings: among the forms that overlapped in the two studies, the correlation for phonological ratings was $r(13) = .867$.

that they had the same means and standard deviations as the Experiment 2 ratings. The result was a set of ratings on the same scale as the original past tense ratings, but with all of the variance that could have been caused by the influence of phonological well-formedness removed. All analyses of Experiment 2 ratings were carried out both on the raw ratings and on these "adjusted" ratings (corrected for phonological well-formedness), with very similar results obtained either way; we report here the results using adjusted ratings.

4.4.2 Production probability

In discussing volunteered forms, we will use the statistic of **production probability**, following Prasada and Pinker (1993). The production probability of a form is defined as the number of experimental participants who volunteered it, divided by the total number of valid responses.

5. Results

The data collected in Experiments 1 and 2 are summarized in Appendixes B and C.

5.1 Preliminaries

5.1.1 The Preference for Regulars

Our results show plainly that English speakers prefer regular past tenses; regulars received a mean rating of 5.75, whereas irregulars received a mean of 4.22. Participants also volunteered regulars far more often: summing over both Experiment 1 and Experiment 2, 81.5% of all volunteered forms were regular. This replicates the results of earlier wug testing studies.

Although participants almost always prefer regular pasts, the magnitude of this preference can be influenced by the experimental design (cf. Prasada & Pinker, 1993, 27fn.). We found a large difference between the production probability for irregulars in Experiment 1 (8.7%) vs. Experiment 2 (18.5%). This is almost certainly due to a difference in the task. In Experiment 2, the participants alternated between volunteering and rating. The irregular forms presented for rating constituted an implicit invitation, we think, to offer irregular forms in the volunteering task, an invitation which most of the participants took up. In terms of our models, we would characterize the behavior of Experiment 2 participants as making use of the second and third choices that the models provide.

The global preference for regulars has an implication for evaluating our models: it is probably unilluminating to evaluate them by calculating *overall* correlations of their predictions against participant data, combining both regular and irregular data in the same analysis. The reason is that any model that rates all regulars above all irregulars could get a fairly high

correlation, without capturing any of the more subtle item-by-item differences.[15] Instead, we have calculated correlations for regulars and irregulars separately.

5.1.2 Ratings Data vs. Volunteered Forms

We find that the production probabilities for volunteered forms correlate reasonably well with ratings data. The correlation of the ratings data with the production probabilities is $r = .837$ (.929 among regulars; .690 among irregulars). Breaking this down between Experiment 1 (pure volunteering) and Experiment 2 (volunteering interspersed with rating), the correlations are: Expt. 1, $r = .788$ (regulars .814, irregulars .515); Expt. 2, $r = .865$ (regulars .902, irregulars .685). For the Experiment 2 forms, this is perhaps not too surprising, since participants might naturally wish to justify their volunteered form in the ratings that immediately followed. However, there is no such confound for Experiment 1, which was administered to a different group of participants. We conclude that the validation of ratings data by volunteering data was reasonably successful.

5.2 Results I: Islands of Reliability for Regulars and Irregulars

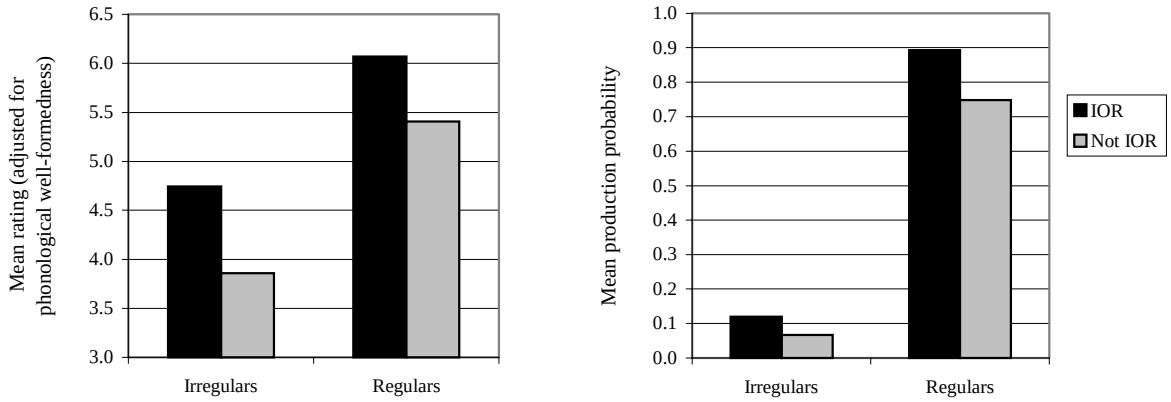
The first set of results addresses a prediction made by the dual mechanism model of morphology. Under a widely adopted interpretation, this model claims that all regular past tenses are derived by the same rule, and thus they should not differ in their acceptability. In contrast, irregulars are derived in the model by an associative network, and should differ significantly in their ratings, depending on their similarity to existing irregulars. Our Core set of wug verbs (section 4.1.1) was designed to test this prediction; it included wug verbs falling either within or outside the islands of reliability for both regulars and irregulars.

5.2.1 Results

Figs. 1a and 1b show the effect of islands of reliability for ratings data and volunteered forms, respectively. The first two columns of each figure show that for irregulars, wug pasts were rated higher, and were volunteered more often, when they occupied an island of reliability. This result is strongly reminiscent of the earlier findings of Bybee and Moder (1982) and of Prasada and Pinker (1993), although it is based on island of reliability effects as defined above rather than on neighborhood similarity or prototypicality. The rightmost two columns Figs. 1a and 1b show that a comparable effect was observed for regular pasts. For both ratings and volunteered production probabilities, two-way ANOVAs revealed highly significant main effects of past type (regulars higher than irregulars; ratings $F(1, 78) = 94.22$, $p < .0001$, production probabilities $F(1, 78) = 758.38$, $p < .0001$) and islandhood (islands of reliability higher than non-islands; ratings $F(1, 78) = 27.23$, $p < .0001$, production probabilities $F(1, 78) = 14.05$, $p < .001$), with no significant interaction. Thus, we find that both regulars and irregulars are susceptible to island of reliability effects, to an equal extent.

[15] For the ratings data for Experiment 2, the overall correlation with regulars and irregulars combined is: rule-based model, $r = .806$; analogical model .780. A model that guesses 1 for regulars and 0 for irregulars would achieve a correlation of .693.

Fig. 1: Effect of Islands of Reliability (IOR) on Irregulars and Regulars



(a) IOR Effect on Ratings (adjusted) (b) IOR Effect on Production Probabilities

Since the existence of island of reliability effects for regulars is one of our central claims, and since it is so much at odds with the findings of Prasada and Pinker, it deserves closer scrutiny.

First, we can point out that the effect cannot be due to differences of phonological well-formedness (the explanation Prasada and Pinker 1993 give for a comparable pattern in their own data), since we saw earlier that (a) the wug forms used in the present study were rated as quite bland; (b) the phonological well-formedness ratings correlated very poorly with past tense ratings; and (c) any small effects that were present were corrected for by fitting to residuals rather than the raw data.

A more sensitive test of the validity of this result is to examine not just the difference in means between IOR and non-IOR test items, but the actual correlation of the participant ratings to the predicted ratings of the rule-based model. This is in fact a better test of the gradient nature of the effect, since the wug verbs were selected to sample the whole range of reliability for irregular and regular past tense formation, rather than occupying just the four "corners" of the set of possibilities.

As (24) shows, the predictions of our rule-based model (see Appendix B for all values) for both regular and irregular past tense ratings in the Core data are positively correlated with the participants' ratings (again, with phonological well-formedness factored out). For completeness, we also include the predictions of our analogical model, which also show a positive correlation.

(24) *Correlations of participant ratings to the predictions of two models: Core verbs* (*n = 41*)

	Rule-based model	Analogical model
regulars	$r = .745$, $p < .0001$	$r = .448$, $p < .01$
irregulars	$r = .570$, $p < .0001$	$r = .488$, $p < .001$

The same is true for the production probability of volunteered forms, also adjusted for phonological well-formedness:

(25) *Correlations of production probabilities to predictions of two models: Core verbs* (*n = 41*)

	Rule-based model	Analogical model
regulars	$r = .695$, $p < .0001$	$r = .481$, $p < .001$
irregulars	$r = .333$, $p < .05$	$r = .517$, $p < .0001$

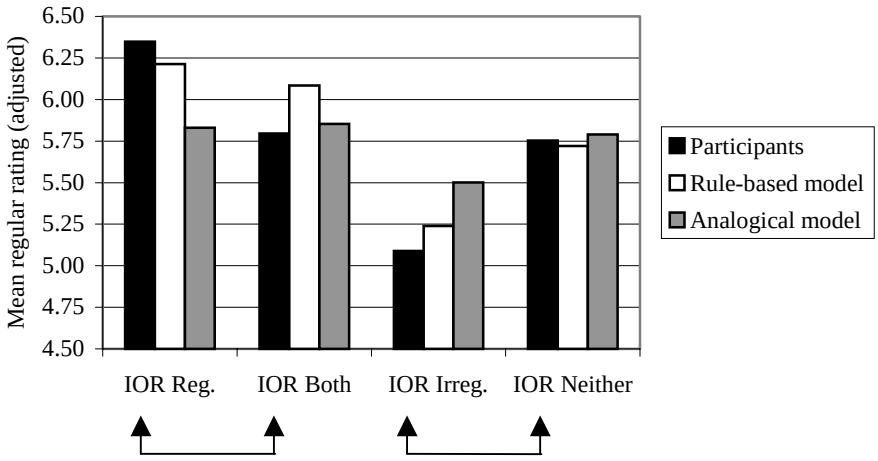
The upshot is that we find no evidence that island of reliability effects are weaker for novel regulars than for novel irregulars; for both, we observe item-by-item differences in ratings and production probabilities, which correspond to differences predicted by our models.

5.2.2 Trade-Off Behavior

As noted above, our participants rated both the regular and one or more irregular forms for each wug verb. We adopted this approach under this view that it would elicit more carefully considered responses from our consultants. However, it may increase the chance that consultants' ratings of regular forms might be influenced by their opinions about the corresponding irregular forms, and vice versa.

Fig. 2a gives the average regular ratings for all four categories in our Core data set (island of reliability for regulars, irregulars, both, and neither), along with the predictions of both of our models, rescaled to have the same mean and standard deviation as the participant ratings.

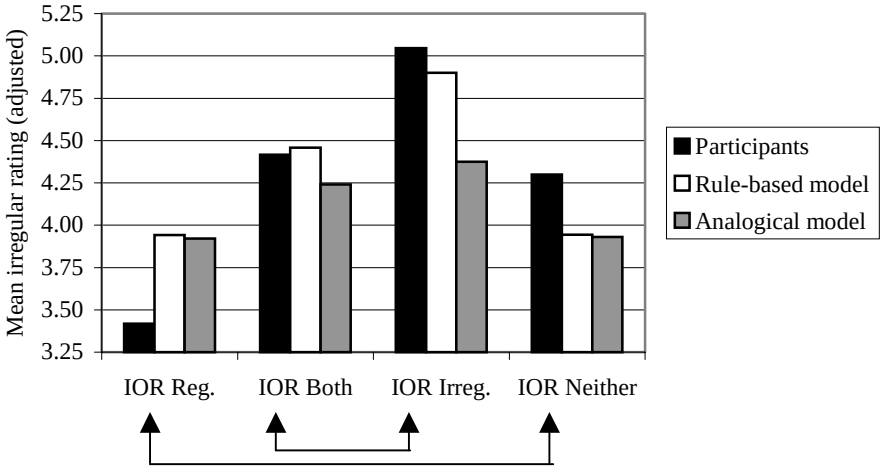
Fig. 2a: Mean ratings of regulars within four categories of islandhood



In general, participants rated regulars higher if they fell into islands of reliability (first and second column groups). However, if the regulars did not have to compete with favored irregulars, the rating was higher (see comparisons marked with arrows). In part, this appears to be simply the result of our choice of wug verbs, since our rule-based model predicts an effect in this direction. But the effect among the consultants is stronger, supporting the trade-off hypothesis.

Surprisingly, the very same effects are also found among the irregulars, as Fig. 2b shows. That is, all else being equal, irregulars are rated lower when they must compete with good regulars. Note that for ratings, unlike production probabilities, this is not a logical necessity: it is theoretically possible that a regular and its competing irregular could both receive high ratings.

Fig. 2b: Mean ratings of irregulars within four categories of islandhood



Combined, these effects produce a correlations of -.786 between the (phonologically adjusted) ratings of regulars and their corresponding irregulars (the better rated, if there were two).

The trade-off effect seen in the consultants' rating should not necessarily be assumed to be entirely an effect of the experimental situation, because our learning models predict a trade-off effect on their own: r = -.465 for the rule-based learner and -.277 for the analogical learner. In the analogical learner, this occurs because forms compete for their share of the same denominator (11). The rule-based learner does not inevitably predict trade-offs (else it would not have been possible to construct a four-way experimental design); nevertheless, it is easier for the model to locate islands of reliability for regulars in phonological territory that is also free of irregulars. Thus, at least part of the trade-off effect exhibited by the participants is predicted by the models.

The trade-off effects that occur are stronger than the models predict, however, and therefore we must consider whether the original conclusion—that there are island of reliability effects for both regulars and irregulars—is valid, or simply an artifact of this confound. To test this, we carried out partial correlations, with the goal of testing whether any island of reliability effects remain once trade-off effects are taken into account. We first carried out a multiple regression, using the factors of (a) phonological well-formedness and (b) the participants' ratings for competing irregulars (in the case of regulars) and competing regulars (in the case of irregulars), to try to predict past tense ratings. We then examined the correlation of the learning models with the remaining residuals. Our findings are shown in (26).

(26) *Correlations (partialing out phonological well-formedness and trade-off effects) of participant ratings to the predictions of two models: Core verbs* (n = 41)

	Rule-based model	Analogical model
regulars	r = .589, p < .0001	r = .258, p = .10
irregulars	r = .497, p < .0001	r = .343, p < .05

For the crucial case, the effect of islands of reliability on regulars, for at least the rule-based model, there remains a correlation of .589, which is highly significant. For the opposite case (irregular judgments potentially affected by trade-offs with competing regulars), the partial correlation is also still highly significant.

The upshot is that, although trade-off effects exist, there remains a correlation between the predictions of the rule-based model and the participants' ratings even when the influence of the competing past tense form has been removed completely. In conclusion, we find that speakers' intuitions about novel past tense forms are sensitive to the phonological shape of the stem. The fact that this is true for both regulars and irregulars is incompatible with a strict interpretation of the dual mechanism model (Prasada & Pinker 1993) in which the only mechanism for deriving regulars is a single default rule.

5.3 Results II: Rules vs. Analogy

Given this result, we must ask what mechanisms are responsible for the effect of phonological form on past tense ratings, both regular and irregular. We consider two possibilities: (a) a system with a large set of detailed rules, each annotated for its reliability; (b) a purely analogical system, which inflects novel forms according to their resemblance to existing verbs. We assess these two possibilities by comparing the predictions of our two computational implementations of them (section 3) against our experimental data.

Here, we will use our full data set, including both the Core and Peripheral forms (section 4.1). The Peripheral data included many forms that were explicitly chosen to assess analogical effects (for example, by being similar to just one or several high frequency model forms), and thus provide a more comprehensive test of the two models.

5.3.1 Correlations

In section 5.2.1, we saw that the rule-based model achieved higher correlations to participant ratings data for the Core forms than the analogical model did. For the full data set, it can be seen that the analogical model actually has a slight edge among the irregulars, but the rule-based model considerably outperforms it on regulars.

(27) *Correlations of participant ratings to the predictions of two models: All verbs*

	Rule-based model	Analogical model
regulars (n = 58)	$r = .714$, $p < .0001$	$r = .512$, $p < .0001$
irregulars (n = 75)	$r = .480$, $p < .0001$	$r = .496$, $p < .0001$

The comparative correlation values are informative as a rough guide, but the crucial aspect of the analysis is to determine why the models behaved as they did. To do this, it is useful to examine the behavior of the models on individual forms, attempting to diagnose what features of the models lead them to accurate or inaccurate predictions. This task is the object of the next few sections.

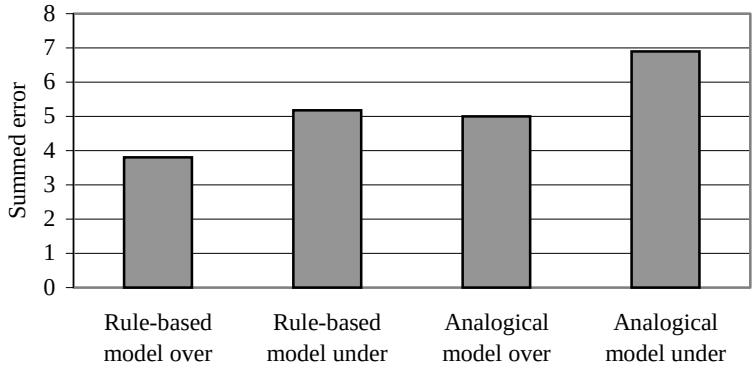
5.3.2 Failure of the Analogical Model to Locate Islands of Reliability

One way to diagnose the models' behavior is to compare their *relative* errors. The assumption we make is that in our experiment, a number of unplanned factors must have influenced the participants' ratings. Because of confounding factors and experimental error, neither model could hope to be exactly correct for all forms, so we believe it is instructive to compare which model was closer. This allows us to diagnose whether one of the models is systematically over- or underrating certain classes of test items.

To this end, we computed the absolute size of the errors made by each model, using the participant ratings adjusted for phonological well-formedness, and the predictions of the models rescaled to have the same means and standard deviations as the participant ratings. For each rated verb, we determined which model was closer, and then subtracted the error of the more accurate model from that of the less accurate model. Finally, these values were sorted according to whether the less accurate model was underestimating or overestimating the observed participant ratings. This yielded a four-way classification, with the categories Rule-Based Model Under, Rule-Based Model Over, Analogical Model Under, and Analogical Model Over.

In order to get a rough sense of the locus of errors for each of the models, we can simply sum the total error in each of these four categories. For the regulars, the result is shown in Figure 3.

Figure 3: Summed relative error of the two models for regulars



We see that when the analogical model is less accurate than the rule-based model, it tends to be because it is underestimating the goodness of forms. This tendency can be understood if we examine the particular verbs on which the analogical model made its greatest errors. Without exception, these are verbs that fall into excellent islands of reliability discovered by the rule-based learner. In Table 5, we list the twelve verbs on which the analogical model made its most serious underestimations. The predictions of both models are included, along with an informal description of the island of reliability used by the rule-based model in making its predictions, and the statistics that show how well this rule performs in the lexicon (i.e., the learning data).

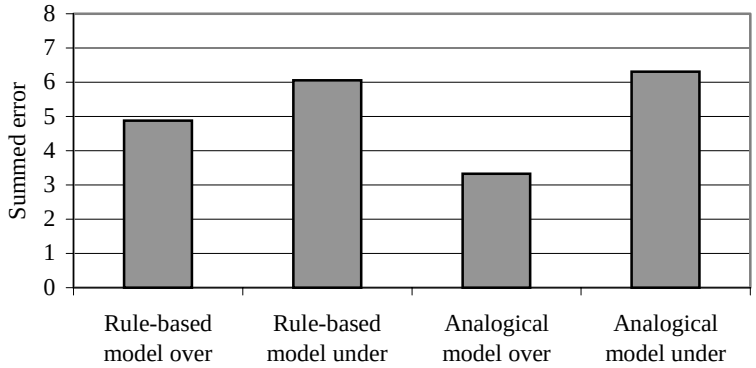
Table 5: Islands of reliability for regular pasts

Past form	Participant rating (adjusted)	Predicted rating: Rule-based model	Predicted rating: Analogical model	Relative error	Island of reliability used by rule-based model	Hits/ Scope
blafed	6.67	6.22	5.15	1.06	/ voiceless fric.___	352/352
driced	6.52	6.22	5.51	0.71	/ voiceless fric.___	352/352
naced	6.51	6.22	5.57	0.65	/ voiceless fric.___	352/352
teshed	6.23	6.22	5.59	0.63	/ voiceless fric.___	352/352
wissed	6.28	6.22	5.68	0.54	/ voiceless fric.___	352/352
flidged	6.41	6.16	5.46	0.70	/ [dʒ, ʒ] ___	110/110
bredged	6.60	6.16	5.85	0.32	/ [dʒ, ʒ] ___	110/110
daped	6.14	6.14	5.56	0.57	/ [V, –high] p ___	83/83
shilked	5.82	5.97	5.17	0.49	/ [C, +coronal] k ___	31/31
tarked	6.24	5.97	5.66	0.31	/ [C, +coronal] k ___	31/31
spacked	6.13	6.01	5.79	0.22	/ [V, +low, –round] k ___	37/37
bligged	5.95	5.66	5.45	0.21	/ g ___	41/42

The analogical model cannot mimic the rule-based model in finding these islands, because the similarity relations it computes are global, depending on the entire phonological material of a word, rather than being structured, based on possessing the crucial segments in just the right place. It is true that when we examine the similar existing verbs that determined the analogical model's behavior, we find that they do tend to fall into the relevant island of reliability more often than not. However, this mere tendency is apparently not strong enough for the analogical model to achieve the performance level reached by the rule-based model.

Figure 4 give the analogous results for irregulars:

Figure 4: Summed relative error of the two models for irregulars



As before, the main problem for the analogical model is underestimation. Inspection of the individual forms shows that the explanation is the same as for regulars: the problem lies in the inability of the analogical model to locate a good island of reliability (for example, *dize-doze*, which falls into the relatively good *rise*/*ride*/*dive* island). The rule-based model also conspicuously overrates some forms; however, these turn out to have an independent explanation which we discuss below in section 5.3.4. The rule-based model's aggregate overestimation (first column) results primarily from *blig-blug* and *drice-droce*; we have no explanation for why consultants disfavored these forms.

5.3.3 Single-Model Analogies

An analogical model predicts that judgments about novel forms could be based largely or entirely on a single existing form. For example, *shee* is extremely similar to the existing verb *see*, which is the only verb of English that undergoes the change [i] → [ɔ]. This resemblance alone leads our analogical model to predict a reasonably high score for the output *shaw*, and similarity for parallel cases. The rule-based model, in contrast, abstracts its structural descriptions from multiple forms; hence extreme similarity to any one learning datum cannot by itself lead to high well-formedness scores. Does the ability of the analogical model to extend a pattern based on a single form allow it to capture aspects of the participant data that the rule-based model cannot?

To obtain data on single-form analogy, we located all of the volunteered forms which employed a change found in only one existing verb. Recall that we had included several wug stems to test this explicitly (*zay*, *shee*, *pum*, *lum*, *kive*, *nold*, and *chool*[16]); among these, the only apparent cases of single-form analogy were 2 instances of *kave*, 1 of *chole*, and 4 of *neld*. Among the remaining verbs we found 37 candidates for single-form analogies.[17] However, inspecting this list, we found reason to believe that they are unconvincing as cases of single-form analogy: they are all quite distant from their alleged model forms, and moreover they virtually all fit product-oriented generalizations, a pattern discussed below in section 6.2.2.[18]

As a further test for single-form analogy, we inspected the data for the rhyming triplet *gry'nt*, *ry'nt*, and *shy'nt*. We anticipated that the participants might base their responses on the closest available analogical verbs (*grind-ground*, *write-wrote*, *shine-shone*). In the volunteered data, this did not occur; participants volunteered [o] more often for all three verbs, including

---

[16] In searching for single-form analogies, the issue arises of whether prefixed forms of the same stem should be counted separately; e.g. for *nold-neld* the possible models include *withhold* and *behold* as well as *hold*. If they are considered as separate verbs (see fn. 9), then *neld* would not count as a single form analogy.

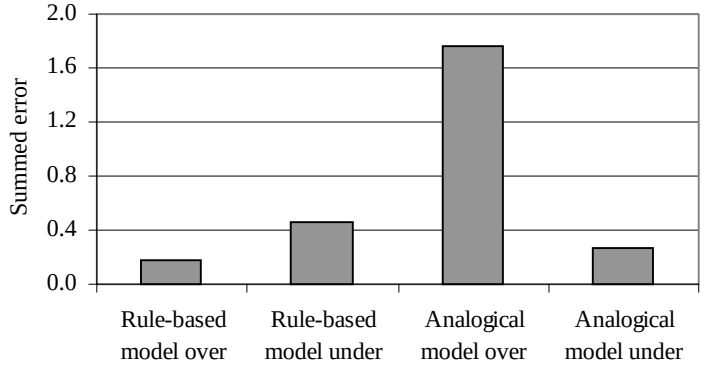
[17] There were: 16 forms using the [aɪ]-[ʌ] pattern of *strike-struck* (7 *shy'nt-shunt*, 4 *ry'nt-runt*, 2 *chind-chund*, 2 *gry'nt-grunt*, 1 *scride-scrud*); 7 with [æ]-[ʌ], like *hang-hung* (3 *spack-spuck*, 3 *pank-punk*, 1 *rask-rusk*); 7 with [ʌ]-[æ], like *run-ran* (2 *pum-pame*, 2 *nung-nang*, 2 *tunk-tank*, 1 *lum-lam*); 3 with [u]-[o], like *choose-chose* (all *gude-gode*); 2 with [i]-[ʌ], like *sneak-snuck* (1 *preak-pruck*, 1 *fleep-flup*); 1 with [i]-[ɛd], like *flee-fled* (*shee-shed*); and 1 with [ɪ]-[e], like *give-gave* (*plim-plame*).

[18] The single apparent exception is *plim-plame*. Many Americans speak a dialect in which words like *sang*, *rang*, *drank*, and *shrank* have the vowel [e] rather than [æ], so it is possible that *plame* employs the [ɪ]-[e] change.

gry'nt: the numbers were *groant* 2, *grount* 1; *roant* 8, *rount* 1; *shoant* 3, *shount* 1. In the ratings data, [aʊ] was indeed preferred for *grint* and [o] for *ry'nt* and *shy'nt*, but the effect was weak (*groant* 3.92, *grount* 4.27; *roant* 3.95, *rount* 3.36; *shoant* 3.58, *shount* 3.14). We conclude that this subset of the data at best supports a modest effect of single-form analogy.

A more systematic test of single-form analogy can be made by examining the behavior of our analogical learning model. We collected all of the wug forms in which the contribution of a single existing form accounted for at least two thirds of the analogical model's total predicted score for that form. These were as follows (contribution of the most similar existing verb given as a percentage): *zay-zed* (*say*; 100%), *lum-lame* (*come*; 100%), *pum-pame* (*come*; 100%), *kive-kave* (*give*; 100%), *nold-neld* (*hold*; 100%), *shee-shaw* (*see*; 100%), *chool-chole* (*choose*; 100%), *nung-nang* (*run*; 100%), *tunk-tank* (*run*; 100%), *pank-punk* (*hang*; 100%), *rask-rusk* (*hang*; 100%), *spack-spuck* (*hang*; 100%), *gezz-gozz* (*get*; 96%), *gry'nt-grount* (*grind*; 92%), *snell-snold* (*sell*; 91%), *preak-proke* (*speak*; 83%), *shy'nt-shont* (*shine*; 79%), *flidge-fludge* (*fling*; 77%), *scride-scrid* (*slide*; 69%), *ry'nt-ront* (*write*; 68%), *tesh-tosh* (*tread*; 68%). We then repeated the procedure described above for Fig. 3 for this subset of the test items, sorting the models' errors into four categories and summing the total error in each category. The result is shown in Figure 5.

Figure 5: Summed relative error of the two models: single-form analogies



It appears that the analogical model's ability to base predictions on a single model largely harms, rather than helps, its performance.

We do note that there were three cases in which participants rated single-form analogies higher than the rule-based model predicted (second column of Fig. 5): *kave* (rule-based model under by .32), *neld* (.15), and *zed* (.09). However, the magnitude of these errors is relatively small, and their aggregate effect is greatly outweighed by the cases in which the analogical model is led astray by single-form analogies.

Summing up, the evidence for single-form analogy in our data appears to consist of a handful of volunteered forms (*kave*, *pame*, *neld*), a slight preference in the ratings for *grount* over *groant*, and better performance by the analogical model on *kave*, *neld*, and *zed*. Against this, there is the fact that the use of single-form analogy seriously impairs the overall accuracy of our analogical model's predictions. It appears that our participants may have used analogy

sporadically, but not in any systematic fashion. Certainly, people are able to manipulate single-form analogies at a conscious level—after all, *zay-zed* makes sense to us in a way that *zay-blif* does not. But we think our data do not support the claim that single-form analogy plays a central role in the morphological system.

5.3.4 Underestimation of burnt*-class Forms by the Rule-Based Model*

By far the largest and most systematic error made by the rule-based model was in its underestimation of the goodness of novel *burnt*-class verbs (*murnt*, *skelt*, etc.; see (17)). For these test items, the mean predicted rating of the rule-based model was 4.12, whereas the mean adjusted rating by participants in Experiment 2 was 5.02. For this set of words, the analogical model was much closer to the experimentally obtained value, with a mean predicted rating of 5.19. This kind of error accounts for 78% of the "Rule-based model under" error column in Figure 4 above.

The underestimation of *burnt* forms may reflect a defect in our rule-based model; however, there is another possibility. Although in general the results of our two experiments were similar (see section 5.1.2), in the case of *burnt*-class forms, there was a large difference. In Experiment 2, participants volunteered a fair number of *burnt* forms (20, out of 366 valid responses for sonorant-final wug verbs). However, in Experiment 1, only 1 such form (*murnt*) was volunteered, out of 301 valid responses. We conjecture that this large difference resulted from our having presented forms of the *burnt* type to the participants in Experiment 2 (which included a ratings task), but not in Experiment 1. It appears that exposing participants to actual models of the *burnt* type may have led them to volunteer *burnt* forms more often and to rate them higher.

A possible explanation may be seen in the study of Quirk (1970), who examined *burnt* verbs in British and American English. Quirk found that Americans seldom use *burnt* forms, but they are highly aware of their existence in other dialects. It seems possible that the Experiment 2 participants, who heard *burnt* forms, produced them at a higher rate than they ordinarily would, as a marker of what they perceived as a prestige register. If this is correct, then we may take the results of Experiment 1, untrammeled by this effect, as a better characterization of the status of *burnt* forms in the natural, spontaneous speech of our participants.

We may also point out that the high ratings that the analogical model assigned to novel *burnt* forms is not necessarily to be construed as a virtue of that model; in fact, we will argue in the next section that it results from the model's inability to learn the correct allomorphic distribution of [-t], [-d], and [-əd].

5.3.5 A Role for Variegated Similarity?

As discussed in section 2.1, an important difference between the rule-based model and the analogical model is that the latter can make use of what we have termed variegated similarity in constructing the analogical set for the behavior of novel words. In this section, we consider whether this capacity is necessary: does the analogical model outperform the rule-based model in cases that rely on variegated similarity?

The fact that the analogical model *can* make use of variegated similarity does not guarantee that it actually did so. However, when we inspected its outputs, we found that the model forms which played the greatest role in determining the outcome characteristically were similar to the base form in variegated ways. For example, the top five model forms that contributed to the analogical model's score for the past tense form *scoiled* are shown in (28). The shaded boxes, which show the places where models diverge from *scoil*, cover all of the territory of the word except the final [l].

(28) Variegated similarity among the most influential analogs for *scoiled*

Analog	s	k	ɔɪ	l	Similarity	Contribution
soil	s		ɔɪ	l	0.264	11.6%
coil		k	ɔɪ	l	0.264	11.6%
spoil	s	p	ɔɪ	l	0.137	6.0%
scowl	s	k	aʊ	l	0.106	4.7%
scale	s	k	e	l	0.080	3.5%

Other forms work similarly, though the amount of variegation, as we have informally assessed it, varies somewhat.

Given that the analogical model does make use of variegated similarity, is this helpful in modeling human intuitions? If so, we would expect to find numerous cases in which the rule-based model underestimated participant ratings, because it could not find support from batches of existing verbs with variegated similarity, and a paucity of cases in which the analogical model overestimated. Our data are uninformative in this respect. Among regulars, the total error in these two categories is about equal (see Figure 3). Among irregulars, both models err, but largely for reasons we have already located: the rule-based model underrates *burnt* forms, and the analogical model overrates forms based on a single form. The residue in both cases is small and rather symmetrical (rule-based model underestimations: .618, analogical model overestimations .837).

Although the error comparisons are uninformative, there is another way of assessing the role of variegated similarity, namely, the behavior of the analogical model in predicting the distribution of the three allomorphs of the regular past tense suffix. It does not suffice simply to predict correctly that a verb will be regular; rather, an adequate model must predict which of the three regular suffix allomorphs ([-d], [-t], and [-əd]) will be used.

The analogical model approaches this task by trying all three suffixes, assigning its predicted score to each. Then, some of these outputs get phonologically filtered (section 3.1.3). In particular, filtering will block any output in which [-d] is added to a stem ending in a voiceless consonant, [-t] is added to a stem ending in a voiced obstruent, or either [-d] or [-t] are added to a stem ending in [t] or [d]. However, filtration cannot account for the full distribution of the past tense allomorphs. The allomorph [-t] is incorrect, but phonologically legal, after any voiced sonorant (cf.: *plant*, *heart*, *vote*), and [-əd] is legal everywhere (*lurid*, *wicked*, *fluid*).

Locating the final consonant to determine the correct ending is a canonical case where structured similarity is required: the past tense allomorph depends solely on the final segment of the stem, and more particularly on just a few of its features. The analogical model, however, is inherently unable to focus on these crucial structural elements. Instead, it gets distracted by variegated similarity, and makes wrong guesses. For instance, when constructing a past tense for the existing verb *render*, the analogical model guesses **renderèd* [rɛndərəd], based largely on the following analogical set (the 10 most similar forms): *rend*, *end*, *rent*, *vend*, *raid*, *fend*, *mend*, *tend*, *round*, and *dread*. These stems bear an accidental similarity to *render*, which (in this case) suffices to outweigh the influence of legitimate model forms like *surrender*.

The participants in our experiment misattached [-əd] precisely once, in the volunteered form *bliggèd* [blɪgəd]. This may be compared to the 936 responses in which the correct [-d] was attached to stems ending in a non-alveolar voiced segments. We conjecture that the basis for [blɪgəd] may have been archaic forms of English (e.g. *banishèd*), encountered in music and poetry, or perhaps it was merely a speech error.

The analogical model also invoked variegated similarity to overgeneralize the allomorph [-t]. For instance, for the existing verb *whisper*, it guessed *whispert* [wɪspərt], using as its basis forms like the following: *whip*, *wish*, *whisk*, *wince*, *quip*, *lisp*, *swish*, *rip*, *work*, and *miss*.

In a sense, these wrong guesses are only the tip of the iceberg: even where the analogical model's first choice is correct, it usually gives relatively high scores to rival outputs containing the wrong past tense allomorph. For instance, the model assigns to *lan* [læn] the past tense *lannèd* [lænəd] with a (reasonably good) score of .147, despite the fact that *lan* does not end with a /t/ or a /d/. As before, the reason is that *lan* is rather similar—in variegated ways—to a number of existing verbs that do end with /t/ or /d/: *land*, *plant*, *slant*, *add*, *sand*, *last*, *rant*, *hand*, *pant*, and *chant*.

The rule-based model avoids outputting the incorrect allomorph. For instance, it does not generate **renderèd* or **lannèd*, because the principle of minimal generalization leads it never even to consider attaching [-əd] other than after an alveolar stop. It also gives *lant* a very low score, reflecting its status as an irregular. More generally, the model correctly reproduces the canonical distribution of the three regular past tense allomorphs: [-əd] only after alveolar stops, [-t] only after voiceless segments other than [t], and [-d] elsewhere.

From this perspective, we can now address the question of why the analogical model guessed fairly high scores for verbs of the *burnt* class (section 5.3.4): the effect was due to model forms ending in voiceless segments. Consider, for example, the most similar analogs for the novel past tense form *squilt*: *squelched*, *spilt*, *squeaked*, *swished*, *switched*, *skipped*, *quipped*, *scalped*, *spelt*, *kissed*. Only two of these are actually irregular, but the diverse nature of the final consonants is irrelevant. The analogical model predicts that *squilt* should sound relatively good because its onset is similar to that of many regular verbs that end in voiceless consonants. This prediction strikes us as extremely counterintuitive.

We conclude that there is little evidence that morphology makes crucial use of variegated similarity for either regulars or irregulars; moreover, variegated similarity leads to catastrophic results in predicting the distribution of the allomorphs of the regular past.

6. Discussion

6.1 Summary

We summarize here our main results. By employing our rule-based learner, we found that the English lexicon contains islands of reliability for regular past tenses. Experimental evidence shows that speakers are also aware of these islands; our participants showed a marked preference for the regular outcome for past tenses that fall within these islands. This is true both for the ratings data and for the volunteered forms. The preference cannot be due to greater phonological well-formedness for such verbs, since our experiments fully controlled for this confound. Moreover, the preference cannot be attributed to a trade-off effect from rival irregulars, as the preference remains when this trade-off is partialed out in a correlation analysis. Our data thus are counterevidence to the strict interpretation of the dual mechanism model (Prasada & Pinker, 1993): when speakers form or evaluate novel regular past tenses, they do not rely solely on a single, context-free rule.

Given this result, we sought to determine whether the mechanism used by speakers in forming past tenses is best described by multiple rules, as our own model supposes, or rather by a form of analogy. Our adaptation of the GCM model was intended to clarify this comparison by embodying the analogical approach in its purest possible form, namely a model that relies on variegated rather than structured similarity. Comparing the performance of the rule-based and analogical models, we found that the analogical model underperformed the rule-based model in correlations to the experimental data. More important, this underperformance can be attributed to essential characteristics of the analogical model; specifically:

- The analogical model systematically underrated regular forms falling within islands of reliability, because its reliance on variegated similarity made it impossible for it to locate these islands.

- The analogical model systematically overrated forms on the basis of similarity to particular individual verbs. This is an error type that is avoided in the rule-based model, which pays attention only to generalizations based on multiple verbs.

- The analogical model made drastic errors in distributing the three allomorphs of the past tense suffix. This was again the result of its relying on variegated rather than structured similarity, which prevented it from locating the crucial structural conditions determining the allomorphic distribution.

From this we infer that a purely analogical model of the GCM type is not promising as an account of how morphological systems in human languages work; and that a multiple-rule approach is currently the more plausible account.

6.2 How our models could be improved

6.2.1 Phonological Theory

Our models employ a theory of phonology in which words are represented as simple sequences of feature bundles (Chomsky & Halle, 1968). Contemporary phonological theory posits that representations include various other formal properties such as moras, syllables, feet, and tiers (Kenstowicz, 1994; Goldsmith, 1995). We believe that incorporating these developments could improve the performance of our models. To give a simple example, Pinker and Prince (1988, p. 114) observe that all English verbs with polysyllabic roots are regular. This is a huge island of reliability, but our model cannot access it, since it lacks any concept of syllables. It is likely that this lack would have been very noticeable if our wug test had included any polysyllabic forms. More generally, many concepts of contemporary phonological theory appear to us to be crucial to improving the performance of the learner, particularly as it is extended to languages with more complex morphophonological systems than English.

6.2.2 Product-Oriented Generalizations

Both of our models are source-oriented, in that past tense formation is described as a morphological operation performed on an input stem (suffix [-əd], change [ɪ] to [ʌ], no-change, etc.) An important insight by Bybee and her colleagues (Bybee & Moder, 1982; Bybee & Slobin, 1983) is that speakers form generalizations not just about the relation between inputs and outputs, but also about the outputs themselves. Examples of such *product-oriented* generalizations about English past tenses might include statements such as "past tense forms should end in an alveolar stop," "past tense forms should contain the vowel [ʌ]," and so on.

Past research has shown that speakers do seem to be guided by product-oriented generalizations when inflecting novel forms, and this is true in our data as well. For example, we found 9 cases in which participants changed [ɪ] to [o], even though no real English verb forms its past tense in this way. The basis of these responses seems to be that English has quite a few verbs (20 in our full learning set) that form their past tense by changing the vowel to [o], although the vowel that gets changed is [aɪ] (*ride-rode*), [e] (*break-broke*), [i] (*speak-spoke*), or [u] (*choose-chose*), and never [ɪ]. We also found many "no-model" changes for the vowels [ʌ], [æ], and [ɛ], all of which occur frequently in existing irregular pasts. Moreover, we think that the putative cases of single-form analogy discussed above in 5.3.3 (e.g. *gude-gode, shy'nt-shunt*) are more likely product-oriented formations, since they, too, favor the output vowels [ʌ], [æ], and [o]. Altogether, about 22% of the volunteered irregulars were formed with vowel changes attested in at most one real verb, and thus could not be accounted for in our input-oriented model.

There are two ways that product-oriented responses might be accommodated in an input-based model. The first possibility would be to allow generalization across multiple structural changes, instead of restricting generalization to occur within a given change. Thus, comparison of the changes [ɪ] → [ʌ], [aɪ] → [ʌ], etc. could yield rules of the type "form the past by changing the stem vowel to [ʌ]." Another possibility is that product-oriented effects could be handled by surface constraints in the phonology, as suggested by Russell (1999) or MacBride (2001).

Plainly, more work is needed on how product-oriented phenomena should be modeled and learned.

6.2.3 Treatment of [t]- and [d]-Final Stems

We found that for wug verbs ending in an alveolar stop ([t]/[d]), our consultants were surprisingly reluctant to attach the expected [-əd] suffix, preferring instead no-change pasts for these verbs. Overall, no-change past tense forms were volunteered 28% of the time for stems ending in alveolar stops (18.4% for [t], 37.7% for [d]), compared to only 0.5% of the time for stems ending in other segments. The same pattern emerged from the ratings data; we offered 7 irregular past tense forms of the no-change type (*flet*, *glit*, *drit*, *chind*, *nold*, *gude*, and *gleed*), and they received generally high ratings (mean 5.38, all others 4.10). The extreme case was *chind* [tʃaɪnd] ("John likes to *chind*, yesterday he *chind* for three hours"), which received a mean adjusted rating of 6.01, versus 3.96 and 3.92 from the rule-based and analogical models, respectively.

A related fact is that for same set of stems, the *regular* outcome (like *glit-glitted*) was disfavored, with a mean adjusted rating of 4.77, compared to 6.01 for regular pasts of non-alveolar final stems. This difference was predicted by both of our models, but not to such a great extent (5.28 vs. 5.88 for the rule-based model, 5.33 vs. 5.86 for the analogical model).

The preference for no-change in wug stems ending in [t] and [d] has been observed before in children, but apparently not among adults (Berko, 1958; Bybee & Slobin, 1982; Marcus et al., 1992). Bybee and Slobin suggest that it is the result of a product-oriented generalization, in which the goal is to create a form that ends in an alveolar stop, and stems that already do so do not require any further additional change. In a similar vein, MacWhinney (1978) proposes an "affix-checking" filter that prevents suffixation on words that are already suffixed; in the case of alveolar-final stems, the filter is fooled by the final alveolar stop, and further suffixation is blocked. It is possible that the alveolar-final stem effect could be captured by allowing product-oriented generalizations. Alternatively, it could be handled by morphological or phonological constraints against repeated material (Yip, 1998; Plag, 1998). We leave it as a goal for future research to test which of these proposals is best able to account for the effect.

6.2.4 The tendency to output existing words

Our experiments found one other pattern that has also been seen in earlier work: when asked to volunteer past tenses, participants often produce existing words. Of the 443 irregular volunteered forms (both experiments), 106 were real words; of these, 22 were real past tense verbs (most of them were regular pasts for other verbs). Expressed in percentage terms, these figures are (respectively) 4.4% and 0.9% of all valid responses, and 23.9% and 5.0% of all valid irregular responses.

This effect was considerably weaker than the effect found by Bybee and Slobin (1982), in which existing past tenses dominated the participants' responses. The high number of such responses in Bybee and Slobin's data was probably due to the fact that their participants were required to volunteer their forms under time pressure. Our own participants, given as much time

as they wanted to reflect, were able to synthesize novel forms, rather than using whatever lay at hand in the mental lexicon.

The ability to favor existing words is beyond the capacity of our models; they address only the task of synthesizing novel verbs. However, to model the productions with greater accuracy, it would be necessary to integrate our models into larger scale processing models, in which the grammatical mechanism of morphological synthesis competes with the lexicon in determining the output, perhaps in the way proposed by Baayen et al. (1997).

6.3 Implications for Other Models

6.3.1 Analogical models

In this paper, we have pointed out serious problems with a model intended to represent a purely analogical theory of morphology: because it has no access to structured similarity, it fails to find islands of reliability, exaggerates the influence of individual forms, and grossly overgenerates. A better analogical model would need to avoid these problems.

In point of fact, many current analogical models are already restricted, in varying degrees, to structured similarity. They typically impose structure on their inputs in advance, by aligning them with a template or reducing them to a limited set of preselected variables. MacWhinney and Leinbach (1991), for example, fitted input verbs into templates of CCCVVCCC syllables, and fed their model both a left-aligned representation of the full verb and a right-aligned representation of just the final rhyme. Similarly, in Eddington's (2000) recent analysis of English past tenses using the Analogical Model of Language (Skousen, 1989), verbs were coded using a predefined set of variables, including the final phoneme, an indication of whether the final syllable was stressed, and a right-aligned representation of the last two syllables.

Our results show that zeroing in on a restricted set of the relevant structural properties of words is more than just an implementational convenience; rather, it is a crucial part of how speakers learn morphology. Furthermore, the relevant structural properties may vary considerably from language to language: English past tenses require knowing whether the final phoneme is an alveolar stop, while Korean case suffixes require knowing if it is a consonant or a vowel (Martin, 1992), Dutch plurals depend on the stress pattern of the root (Booij, 1998; van der Hulst & Kooij, 1998), and so on. Hence, we believe that models that rely on templatifying inputs or decomposing them into preselected variables are incomplete without a cross-linguistically valid mechanism that guides templatification or variable selection.

We also find in inspecting the literature that analogical models have generally not been tested for the problems of single-form analogy and overgeneration (e.g. **renderèd*, **whispert*). We hope that this paper extends the empirical domain for measuring the success of future English past tense models.

6.3.2 The Dual Mechanism Model

Our finding of island of reliability effects for regulars appears to contradict what is by now a massive body of research driven by the dual mechanism theory of morphology. It is important to remember, however, that the dual mechanism literature makes two distinct claims. The first is

that some morphologically complex words are stored while others are derived on-line (the "words and rules" hypothesis; Pinker, 1999a). Our rule-based model is compatible with the idea that existing irregular forms are lexically stored—in fact, it depends on it, since the grammar it learns for English would prefer the regular outcome in virtually all cases. Our model is also compatible with lexical storage of regular forms (Schreuder et al., 1999; Sereno et al., 1999; and others), but it would not require it, as regulars could be produced by the grammar as well. This model is intended solely as a model of morphological productivity, and not as a model of how existing words are stored and produced.

The second claim of the dual mechanism theory is that the mechanism for deriving words is simple, and contains rules for regular patterns only. We argue here against this second claim. The fact that phonological form influenced participant ratings of novel regulars shows that the morphological system must contain specific information about the applicability of patterns in different phonological contexts. Moreover, our comparison of the rule-based and analogical models shows that irregular patterns, too, are extended in a way that seems to be appropriately described by rules. Our model represents the view that grammar should capture all of the generalizations it can about the existing lexicon, not just the largest or most productive ones. In order to do this, it employs multiple rules, describing all morphological patterns.

As we see it, the rule-based model advocated here has three main advantages over current versions of the dual mechanism model. First, it can capture the experimentally observed item-by-item differences among regulars. Second, it uses the same grammar to derive both regular and irregular processes, and thus does not rely on some unspecified control mechanism to explain what participants are doing when they rate novel regulars vs. novel irregulars. Finally, it includes an inductive learning algorithm, ensuring that the proposed adult grammar could be learned from input data.

6.4 General Conclusion

We feel that our results support a view of morphology that integrates elements from sharply divergent intellectual traditions.

With connectionist researchers, we share the view that inductive learning of detailed generalizations plays a major role in language. In particular, although learners of English could get by with only a single default rule for regulars, it appears they go beyond this: they learn a set of detailed environments that differentiate the degrees of confidence for the regular outcome.

On the other hand, we share with the mainstream tradition of formal linguistic theory the view that linguistic knowledge is best characterized by rules:

- Because they contain variables, rules permit correct outputs to be derived even for unusual input forms that lack neighbors (the central argument made by Pinker and Prince, 1988).

- Rules can form very tight systems that avoid overgeneration (**renderèd*, **whispert*).

- Rules limit themselves to structured similarity, and cannot access variegated similarity. Our tentative conclusion from our experimental results is that this limitation is correct, or very close to being so.

- Because they are based on formation of generalizations, rules avoid single-form analogies, which appear to have a marginal (perhaps metalinguistic) status in human productions.

In other words, our opinion of rules is perhaps even higher than traditional formal linguistics has held: when they are discovered by an inductive learning algorithm, rules are the appropriate means of expressing both macro- *and* micro-generalizations.

Appendix A: Frame Dialogs for Experiments 1 and 2

1. I dream that one day I'll be able to ___. The chance to ___ would be very exciting. I think I'd really enjoy ___. My friend Sam ___ once and he loved it.

2. Every day I like to ___. I usually ___ first thing in the morning. After I'm finished ___, I'm ready to start the day. This morning I ___ as usual.

3. Why won't John ___? I want him to ___. He's only tolerable when he's ___. Like yesterday he ___, and he was fine

4. You should ___. It's always worthwhile to ___. I'm ___, and you should too. Last week you ___, why won't you now?

5. Egbert loves to ___ all the time. It seems he was born to ___. ___ is what he's good at. Last week he ___ six days out of seven.

6. Nobody wants to ___ these days. I don't know why people don't ___. This country has a long tradition of ___. Years ago everyone ___ and life was much better.

7. I don't want to ___. I try not to ___. I know that ___ isn't good for you. I know that because I ___ a lot when I was a teenager.

8. Fred couldn't ___. All he wanted was to ___. Finally, he succeeded in ___. After he ___ his mind was at ease.

9. It's difficult to ___. I've always found it hard to ___. Some people say ___ is easy. But I ___ last week and I can tell you it's not easy.

10. Everyone wants to ___. Magazines are telling me to ___. It seems like ___ is "in". But I ___ last week, and I don't see what all the fuss is about.

11. Jane refuses to ___. She's too frightened to ___. I don't know why she's so scared of ___. Everyone else ___ last year.

12. When I was a kid I used to ___. My father taught me to ___. I used to enjoy ___ very much. But I probably ___ one too many times.

13. Next week we're going to ___. We've been waiting ages for the chance to ___. We're reading a book about ___. My friend ___ once before, but this will be my first time.

14. Nick tries to ___ every day. He gets off work early so he can ___. He says that ___ holds his life together. But I ___ once, and it was nothing special.

15. In the 80s everyone used to ___. If you didn't ___, you weren't living. You could say ___ was the national pastime. Of course I ___ along with everyone else.

Appendix B: Phonological Ratings and Past Tense Scores for Regulars

Table B1: Core verbs

Stem	Stem Rating	Category	Regular Past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Regular Mean Rating	Regular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
1. *bize*	4.57	IOR both	*bized*	0.778	0.571	0.667	5.30	5.32	6.06	5.87
2. *dize*	4.62	IOR both	*dized*	0.889	0.762	0.821	5.42	5.42	6.06	5.95
3. *drice*	3.86	IOR both	*driced*	1.000	0.913	0.953	6.26	6.52	6.22	5.51
4. *flidge*	4.05	IOR both	*flidged*	0.947	0.783	0.857	6.21	6.41	6.16	5.46
5. *fro*	5.84	IOR both	*froed*	0.950	0.833	0.886	5.83	5.50	5.40	6.16
6. *gare*	5.24	IOR both	*gared*	1.000	0.955	0.976	6.57	6.44	6.02	6.27
7. *glip*	4.95	IOR both	*glipped*	1.000	0.857	0.925	5.95	5.88	6.07	5.80
8. *rife*	5.61	IOR both	*rifed*	0.950	0.762	0.854	5.95	5.69	6.22	5.07
9. *stin*	5.40	IOR both	*stinned*	0.900	0.522	0.698	5.30	5.08	5.83	6.02
10. *stip*	5.45	IOR both	*stipped*	1.000	0.708	0.841	5.92	5.70	6.07	5.88
11. *blafe*	3.57	IOR regular	*blafed*	1.000	0.818	0.892	6.32	6.67	6.22	5.15
12. *bredge*	3.86	IOR regular	*bredged*	0.950	0.905	0.927	6.33	6.60	6.16	5.85
13. *chool*	3.76	IOR regular	*chooled*	1.000	0.957	0.977	6.13	6.41	6.12	6.38
14. *dape*	5.14	IOR regular	*daped*	1.000	0.957	0.976	6.25	6.14	6.14	5.56
15. *gezz*	4.19	IOR regular	*gezzed*	1.000	0.955	0.976	6.61	6.79	6.06	5.89
16. *nace*	5.00	IOR regular	*naced*	1.000	1.000	1.000	6.57	6.50	6.22	5.57
17. *spack*	5.05	IOR regular	*spacked*	1.000	0.739	0.860	6.22	6.13	6.01	5.79
18. *stire*	5.62	IOR regular	*stired*	1.000	0.818	0.902	6.00	5.74	6.02	6.29
19. *tesh*	4.71	IOR regular	*teshed*	1.000	0.870	0.925	6.22	6.23	6.22	5.59
20. *whiss*	5.76	IOR regular	*whissed*	0.950	0.952	0.951	6.57	6.28	6.22	5.68
21. *blig*	3.71	IOR irregular	*bligged*	0.941	0.652	0.775	5.67	5.95	5.66	5.44
22. *chake*	5.33	IOR irregular	*chaked*	0.950	0.818	0.881	5.74	5.55	4.77	5.65
23. *drit*	4.30	IOR irregular	*dritted*	0.842	0.591	0.707	4.96	5.04	5.43	5.29
24. *fleep*	4.24	IOR irregular	*fleeped*	1.000	0.478	0.721	5.00	5.10	5.69	5.56
25. *gleed*	5.29	IOR irregular	*gleeded*	0.684	0.455	0.561	4.22	3.98	4.36	5.15
26. *glit*	5.25	IOR irregular	*glitted*	0.778	0.542	0.643	5.00	4.80	5.43	5.37
27. *plim*	4.43	IOR irregular	*plimmed*	0.950	0.682	0.810	6.13	6.22	5.74	5.96
28. *queed*	3.81	IOR irregular	*queeded*	0.700	0.364	0.524	4.65	4.86	4.36	5.10
29. *scride*	4.05	IOR irregular	*scrided*	0.556	0.292	0.405	4.17	4.30	4.58	4.89
30. *spling*	4.56	IOR irregular	*splinged*	0.667	0.368	0.514	4.36	4.34	5.14	5.35

Table B1: Core verbs (cont.)

Stem	Stem Rating	Category	Regular Past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Regular Mean Rating	Regular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
32. *gude*	4.25	IOR neither	*guded*	0.625	0.500	0.556	4.90	4.99	6.07	5.26
33. *nold*	4.10	IOR neither	*nolded*	0.833	0.273	0.525	4.64	4.76	4.78	5.54
34. *nung*	3.21	IOR neither	*nunged*	0.933	0.737	0.824	5.37	5.78	5.14	5.97
35. *pank*	5.62	IOR neither	*panked*	1.000	0.810	0.900	6.30	6.05	5.62	5.92
36. *preak*	4.90	IOR neither	*preaked*	0.900	0.792	0.841	5.83	5.77	5.37	5.80
37. *rask*	5.30	IOR neither	*rasked*	1.000	0.870	0.930	6.42	6.26	5.97	6.11
38. *shilk*	4.60	IOR neither	*shilked*	1.000	0.950	0.975	5.79	5.82	5.97	5.17
39. *tark*	5.10	IOR neither	*tarked*	1.000	0.870	0.930	6.33	6.24	5.97	5.66
40. *trisk*	5.14	IOR neither	*trisked*	1.000	0.789	0.897	6.29	6.17	5.97	6.05
41. *tunk*	4.65	IOR neither	*tunked*	1.000	0.826	0.907	5.67	5.67	5.62	5.80

Table B2: Peripheral verbs

Stem	Stem Rating	Category	Regular Past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Regular Mean Rating	Regular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
1. *grell*	4.52	Burnt	*grelled*	1.000	0.810	0.902	5.86	5.91	5.86	6.17
2. *murn*	5.43	Burnt	*murned*	0.947	0.957	0.952	6.57	6.38	6.02	6.29
3. *scoil*	3.84	Burnt	*scoiled*	0.944	0.947	0.946	6.45	6.72	5.93	6.51
4. *shurn*	5.00	Burnt	*shurned*	0.947	0.857	0.900	6.57	6.50	6.02	6.31
5. *skell*	5.05	Burnt	*skelled*	0.944	0.682	0.800	6.05	5.95	5.86	6.24
6. *snell*	5.38	Burnt	*snelled*	0.947	0.826	0.881	6.17	5.99	5.86	6.18
7. *squill*	4.67	Burnt	*squilled*	0.895	0.810	0.850	5.92	5.93	5.86	6.30
8. *kive*	4.38	Single form analogy	*kived*	0.950	0.864	0.905	6.00	6.10	5.58	5.87
9. *lum*	4.81	Single form analogy	*lummed*	1.000	0.826	0.905	6.35	6.33	5.74	6.14
10. *pum*	4.81	Single form analogy	*pummed*	1.000	0.826	0.902	6.17	6.15	5.74	6.06
11. *shee*	5.95	Single form analogy	*sheed*	1.000	0.875	0.929	6.17	5.81	5.94	5.89
12. *zay*	4.14	Single form analogy	*zayed*	0.950	1.000	0.977	6.39	6.57	6.13	5.99
13. *chind*	4.62	Few form analogy	*chinded*	0.235	0.368	0.306	3.89	3.84	4.36	5.41
14. *flet*	4.24	Few form analogy	*fletted*	0.889	0.368	0.622	4.50	4.58	5.43	5.39
15. *gry'nt*	5.16	Few form analogy	*gry'nted*	0.842	0.545	0.683	5.26	5.10	6.20	5.59
16. *ry'nt*	3.00	Few form analogy	*ry'nted*	0.778	0.250	0.476	5.00	5.46	6.20	5.51
17. *shy'nt*	3.52	Few form analogy	*shy'nted*	0.800	0.364	0.571	5.17	5.49	6.20	5.49

Appendix C: Phonological Ratings and Past Tense Scores for Irregulars

Table C1: Core verbs

Stem	Stem Rating	Category	Irregular past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Irregular Mean Rating	Irregular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
1. *bize*	4.57	IOR both	*boze*	0.056	0.381	0.231	4.57	4.55	4.11	4.04
2. *dize*	4.62	IOR both	*doze*	0.111	0.190	0.154	5.04	5.04	4.73	4.18
3. *drice*	3.86	IOR both	*droce*	0.000	0.087	0.047	4.48	4.31	5.15	4.28
4. *flidge*	4.05	IOR both	*fludge*	0.000	0.043	0.024	4.88	4.76	4.22	4.10
5. *fro*	5.84	IOR both	*frew*	0.050	0.125	0.091	4.33	4.57	4.97	4.38
6. *gare*	5.24	IOR both	*gore*	0.000	0.000	0.000	3.39	3.49	4.30	4.23
7. *glip*	4.95	IOR both	*glup*	0.000	0.048	0.025	3.45	3.50	4.02	3.97
8. *rife*	5.61	IOR both	*rofe*	0.000	0.190	0.098	4.14	4.33	4.61	4.35
9. *stin*	5.40	IOR both	*stun*	0.100	0.261	0.186	4.78	4.94	4.34	4.63
10. *stip*	5.45	IOR both	*stup*	0.000	0.083	0.045	4.50	4.66	4.15	4.26
11. *blafe*	3.57	IOR regular	*bleft*	0.000	0.045	0.027	4.09	3.86	3.94	3.85
12. *bredge*	3.86	IOR regular	*broge*	0.050	0.048	0.049	3.43	3.25	3.94	3.85
13. *chool*	3.76	IOR regular	*chole*	0.000	0.043	0.023	3.71	3.51	3.94	4.05
14. *dape*	5.14	IOR regular	*dapt*	0.000	0.000	0.000	4.00	4.09	3.94	3.85
15. *gezz*	4.19	IOR regular	*gozz*	0.000	0.000	0.000	2.52	2.40	3.94	3.95
16. *nace*	5.00	IOR regular	*noce*	0.000	0.000	0.000	2.91	2.96	4.00	3.89
17. *spack*	5.05	IOR regular	*spuck*	0.000	0.130	0.070	3.96	4.03	3.94	3.85
18. *stire*	5.62	IOR regular	*store*	0.000	0.091	0.049	3.22	3.40	3.94	4.03
19. *tesh*	4.71	IOR regular	*tosh*	0.000	0.000	0.000	3.13	3.12	3.94	3.88
20. *whiss*	5.76	IOR regular	*wus*	0.000	0.048	0.024	3.35	3.56	3.94	3.99
21. *blig*	3.71	IOR irregular	*blug*	0.000	0.130	0.075	4.17	3.97	5.19	4.08
22. *chake*	5.33	IOR irregular	*chook*	0.000	0.000	0.000	5.04	5.19	5.13	4.17
23. *drit*	4.30	IOR irregular	*drit*	0.053	0.091	0.073	5.13	5.07	4.62	4.11
24. *drit*	4.30	IOR irregular	*drat*	0.000	0.182	0.098	3.65	3.57	4.06	3.99
25. *fleep*	4.24	IOR irregular	*flept*	0.000	0.435	0.233	6.09	6.02	5.15	4.40
26. *gleed*	5.29	IOR irregular	*gled*	0.158	0.318	0.244	6.00	6.15	5.07	4.53
27. *gleed*	5.29	IOR irregular	*gleed*	0.105	0.227	0.171	4.09	4.21	3.94	3.99
28. *glit*	5.25	IOR irregular	*glit*	0.167	0.125	0.143	5.21	5.34	4.89	4.32
29. *glit*	5.25	IOR irregular	*glat*	0.000	0.167	0.095	3.75	3.86	4.06	3.91
30. *plim*	4.43	IOR irregular	*plum*	0.000	0.136	0.071	4.17	4.12	4.52	4.10
31. *plim*	4.43	IOR irregular	*plam*	0.000	0.045	0.024	3.57	3.51	4.21	3.92
32. *queed*	3.81	IOR irregular	*qued*	0.100	0.318	0.214	5.35	5.19	4.43	4.09

Table C1: Core verbs (cont.)

Stem	Stem Rating	Category	Irregular past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Irregular Mean Rating	Irregular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
33. *rife*	5.61	IOR irregular	*riff*	0.000	0.000	0.000	3.24	3.42	3.94	3.90
34. *scride*	4.05	IOR irregular	*scrode*	0.111	0.250	0.190	4.39	4.26	4.98	4.73
35. *scride*	4.05	IOR irregular	*scrid*	0.000	0.042	0.024	3.57	3.43	4.12	3.95
36. *spling*	4.56	IOR irregular	*splung*	0.222	0.421	0.324	5.45	5.45	5.19	5.42
37. *spling*	4.56	IOR irregular	*splang*	0.056	0.158	0.108	4.50	4.48	4.36	4.54
38. *stin*	5.40	IOR irregular	*stan*	0.000	0.000	0.000	2.74	2.87	4.27	4.03
39. *teep*	4.95	IOR irregular	*tept*	0.000	0.087	0.047	4.70	4.76	4.73	4.20
40. *gude*	4.25	IOR neither	*gude*	0.375	0.300	0.333	5.55	5.48	3.96	3.99
41. *nold*	4.10	IOR neither	*nold*	0.167	0.500	0.350	6.05	5.95	3.96	3.91
42. *nold*	4.10	IOR neither	*neld*	0.000	0.182	0.100	5.14	5.03	3.94	4.10
43. *nung*	3.21	IOR neither	*nang*	0.000	0.105	0.059	4.32	4.02	3.94	3.89
44. *pank*	5.62	IOR neither	*punk*	0.000	0.143	0.075	4.00	4.19	3.94	3.89
45. *preak*	4.90	IOR neither	*proke*	0.100	0.167	0.136	3.92	3.96	3.98	3.93
46. *preak*	4.90	IOR neither	*preck*	0.000	0.000	0.000	3.54	3.58	3.94	3.98
47. *rask*	5.30	IOR neither	*rusk*	0.000	0.043	0.023	4.08	4.21	3.94	3.85
48. *shilk*	4.60	IOR neither	*shalk*	0.000	0.000	0.000	3.67	3.64	3.94	4.13
49. *tark*	5.10	IOR neither	*tork*	0.000	0.043	0.023	3.71	3.79	3.94	3.85
50. *trisk*	5.14	IOR neither	*trask*	0.000	0.105	0.051	3.76	3.85	3.94	4.01
51. *trisk*	5.14	IOR neither	*trusk*	0.000	0.053	0.026	3.62	3.71	3.94	3.94
52. *tunk*	4.65	IOR neither	*tank*	0.000	0.087	0.047	3.92	3.91	3.94	3.86

Table C2: Peripheral verbs

Stem	Stem Rating	Category	Irregular past	Exp. 1 Production Probability	Exp. 2 Production Probability	Overall Production Probability	Irregular Mean Rating	Irregular Mean Rating Residuals	Rule-based Model Predicted	Analogical Model Predicted
1. *grell*	4.52	Burnt	*grelt*	0.000	0.143	0.073	4.90	4.88	4.06	5.39
2. *murn*	5.43	Burnt	*murnt*	0.053	0.043	0.048	4.74	4.90	4.60	5.23
3. *scoil*	3.84	Burnt	*scoilt*	0.000	0.000	0.000	5.15	4.99	4.01	4.91
4. *shurn*	5.00	Burnt	*shurnt*	0.000	0.000	0.000	4.22	4.28	3.95	5.11
5. *skell*	5.05	Burnt	*skelt*	0.000	0.227	0.125	5.32	5.41	4.06	5.23
6. *snell*	5.38	Burnt	*snelt*	0.000	0.130	0.071	5.30	5.46	4.06	5.25
7. *squill*	4.67	Burnt	*squilt*	0.000	0.048	0.025	5.21	5.22	4.06	5.20
8. *kive*	4.38	Single form analogy	*kave*	0.000	0.091	0.048	4.41	4.35	3.94	4.12
9. *lum*	4.81	Single form analogy	*lame*	0.000	0.000	0.000	2.87	2.88	3.94	3.93
10. *pum*	4.81	Single form analogy	*pame*	0.000	0.000	0.000	2.71	2.71	3.94	4.01
11. *shee*	5.95	Single form analogy	*shaw*	0.000	0.000	0.000	3.25	3.50	3.94	4.18
12. *zay*	4.14	Single form analogy	*zed*	0.000	0.000	0.000	4.39	4.28	3.94	4.05
13. *chind*	4.62	Few form analogy	*chound*	0.647	0.368	0.500	6.00	6.01	3.96	3.92
14. *chind*	4.62	Few form analogy	*chound*	0.000	0.000	0.000	4.05	4.04	4.83	4.45
15. *flet*	4.24	Few form analogy	*flet*	0.111	0.474	0.297	5.65	5.58	5.22	4.30
16. *gry'nt*	5.16	Few form analogy	*groant*	0.000	0.091	0.049	4.17	4.27	3.94	4.07
17. *gry'nt*	5.16	Few form analogy	*grount*	0.053	0.000	0.024	3.83	3.92	3.94	4.67
18. *ry'nt*	3.00	Few form analogy	*roant*	0.056	0.292	0.190	4.29	3.95	3.94	4.32
19. *ry'nt*	3.00	Few form analogy	*rount*	0.000	0.042	0.024	3.71	3.36	3.94	4.13
20. *shy'nt*	3.52	Few form analogy	*shoant*	0.000	0.136	0.071	3.83	3.58	3.94	4.23
21. *shy'nt*	3.52	Few form analogy	*shount*	0.000	0.045	0.024	3.39	3.14	3.94	4.02
22. *snell*	5.38	Few form analogy	*snold*	0.000	0.000	0.000	2.83	2.95	3.94	4.17

References

Albright, A. & Hayes, B. (1999). An automated learner for phonology and morphology. Unpublished manuscript, University of California, Los Angeles. Retrieved November 26, 2001 from http://www.linguistics.ucla.edu/people/hayes/learning/learner.pdf

Albright, A. & Hayes, B. (2000). Distributional encroachment and its consequences for morphological learning. *UCLA Working Papers in Linguistics, 4*, 179–190.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, R. H., Dijkstra, T., & Schreuder R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37*, 94–117.

Baroni, M. (2000). *Distributional Cues in Morpheme Discovery: A Computational Model and Empirical Evidence*. Unpublished doctoral dissertation, University of California, Los Angeles.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.

Bird, S. (1995). *Computational Phonology - A Constraint-Based Approach*. Cambridge University Press.

Bloomfield, L. (1939). Menomini morphophonemics. *Travaux du cercle linguistique de Prague, 8*, 105–115.

Booij, G. (1998). Phonological output constraints in morphology. In W. Kehrein & R. Wiese (Eds.), *Phonology and Morphology of the Germanic Languages*. Tübingen: Niemeyer.

Broe, M. (1993). *Specification theory: the treatment of redundancy in generative phonology*. Unpublished doctoral dissertation, University of Edinburgh.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425–455.

Bybee, J. (2001). *Phonology and Language Use*. Cambridge University Press.

Bybee, J. & Moder, C. L. (1983). Morphological classes as natural categories. *Language, 59*, 251–270.

Bybee, J. & Slobin, D. (1982). Rules and schemas in the development and use of the English past tense. *Language, 58*, 265–289.

Chater, N. & Hahn, U. (1998). Rules and similarity: Distinct? Exhaustive? Empirically distinguishable? *Cognition, 65*, 197–230.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. New York: Harper and Row.

Clark, A. & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language, 8*, 487–519.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences, 22*, 991–1013.

Clahsen, H. & Rothweiler, M. (1992). Inflectional rules in children's grammars: Evidence from the development of participles in German. In G. Booij & J. van Marle (Eds.) *Yearbook of Morphology 1992*. Dordrecht: Kluwer.

Cohen, J.D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25*, 257–271.

Daugherty, K., & Seidenberg, M. (1994). Beyond rules and exceptions: A connectionist modeling approach to inflectional morphology. In S. Lima, R. Corrigan, & G. Iverson (Eds.), *The Reality of Linguistic Rules*. Amsterdam: John Benjamins.

Dell, G., Reed, K., Adams, D., & Meyer A. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1355–1367.

Dressler, W. U. (1999). Why collapse morphological concepts? *Behavioral and Brain Sciences, 22*, 1021.

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua, 110*, 281–298.

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

Frisch, S. (1996). *Similarity and Frequency in Phonology*. Unpublished doctoral dissertation, Northwestern University.

Frisch, S., Broe, M., & Pierrehumbert, J. (1997). *Similarity and phonotactics in Arabic*. Manuscript submitted for publication. Retrieved November 14, 2001, from http://roa.rutgers.edu/files/223-1097/roa-223-frisch-2.pdf

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27*, 153–198.

Halle, M. (1978). Knowledge unlearned and untaught: what speakers know about the sounds of their language. 294-303. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

Hayes, B. (in press). Phonological acquisition in Optimality Theory: The early stages. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Fixing Priorities: Constraints in Phonological Acquisition*. Cambridge University Press.

Hanson, S. J. & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences, 13*, 471–489.

van der Hulst, H. & Kooij, J. G. (1998). Prosodic choices and the Dutch nominal plural. In W. Kehrein & R. Wiese (Eds.), *Phonology and Morphology of the Germanic Languages*. Tübingen: Niemeyer.

Hutchinson, A. (1994). *Algorithmic Learning*. Oxford: Clarendon Press.

Indefrey, P. (1999). Some problems with the lexical status of nondefault inflection. *Behavioral and Brain Sciences, 22*, 1025.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y. & 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. & 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 & J. B. Kruskal (Eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley Publishing Company.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. *Cognition, 49*, 235–290.

MacBride, A. (2001, April). An approach to alternations in the Berber verbal stem. Paper presented at the 6th Southwestern Workshop on Optimality Theory, University of Southern California.

MacWhinney, B. (1978). The Acquisition of Morphophonology. *Monographs of the Society for Research* in *Child Development, 43*(1-2).

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*, 121-157.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology, 29*, 186–256.

Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in Language Acquisition. *Monographs of the Society for Research in Child Development, 57*(4, Serial No. 228).

Mikheev, A (1997). Automatic rule induction for unknown-word guessing. *Computational Linguistics, 23*, 405–423.

Nakisa, R. C., Plunkett, K. & Hahn, U. (2001). A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder & J. Murre (Eds.), *Models of Language Acquisition: Inductive and Deductive Approaches*. Cambridge, MA: MIT Press.

Nosofsky, R. M. (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393–418.

Pierrehumbert, J. (in press) Stochastic phonology. *GLOT 5*(6), 1–13. Retrieved November 14, 2001, from http://www.ling.nwu.edu/~jbp/GLOT.pdf

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

Pinker, S. (1999b, Oct. 29). Regular habits. *Times Literary Supplement*. Retrieved November 14, 2001, from http://www.mit.edu/~pinker/tls.html

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

Pinker, S., & Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. Lima, R. Corrigan, & G. Iverson (Eds.), *The Reality of Linguistic Rules*. Amsterdam: John Benjamins.

Plag, I. (1998). Morphological haplology in a constraint-based morpho-phonology. In W. Kehrein & R. Wiese (Eds.), *Phonology and Morphology of the Germanic Languages*. Tübingen: Niemeyer.

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

Prince, A. & Smolensky, P. (1993). *Optimality Theory: Constraint Interaction in Generative Grammar* (Tech. Rep. No. 2). Rutgers University, Center for Cognitive Science.

Quirk, R. (1970). Aspect and variant inflection in English verbs. *Language*, *46*(2), 300-311.

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

Russell, K. (1999). MOT: Sketch of an OT approach to morphology. Unpublished manuscript, University of Manitoba. Retrieved November 14, 2001 from http://roa.rutgers.edu/files/352-1099/roa-352-russell-3.pdf

Schreuder, R., de Jong, N., Krott, A., & Baayen, H. (1999). Rules and rote: Beyond the linguistic either-or fallacy. *Behavioral and Brain Sciences*, *22*, 1038–1039.

Sereno, J., Zwitserlood, P., & Jongman, A. (1999). Entries and operations: The great divide and the pitfalls of form frequency. *Behavioral and Brain Sciences*, *22*, 1039.

Skousen, R. (1989). *Analogical Modeling of Language*. Dordrecht: Kluwer Academic Publishers.

Wiese, R. (1999). On default rules and other rules. *Behavioral and Brain Sciences*, *22*, 1043–1044.

Wunderlich, D. (1999). German noun plurals reconsidered. *Behavioral and Brain Sciences*, *22*, 1044–1045.

Yip, M. (1998). Identity avoidance in phonology and morphology. In S. G. Lapointe, D. K. Brentari, & P. M. Farrell (Eds.), *Morphology and its Relation to Phonology and Syntax*. Stanford, CA: CSLI Publications.

Zuraw, K. (1999). Similarity.exe [Computer software]. Los Angeles: Department of Linguistics, University of Southern California.