ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cognition 90 (2003) 119–161

COGNITION

www.elsevier.com/locate/COGNIT

Rules vs. analogy in English past tenses: a computational/experimental study

Adam Albright[a,*], Bruce Hayes[b,*]

[a]Department of Linguistics, University of California, Santa Cruz, Santa Cruz, CA 95064-1077, USA
[b]Department of Linguistics, University of California, Los Angeles, Los Angeles, CA 90095-1543, USA

Received 7 December 2001; revised 11 November 2002; accepted 30 June 2003

Abstract

Are morphological patterns learned in the form of rules? Some models deny this, attributing all morphology to analogical mechanisms. The dual mechanism model (Pinker, S., & Prince, A. (1998). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73–193) posits that speakers do internalize rules, but that these rules are few and cover only regular processes; the remaining patterns are attributed to analogy. This article advocates a third approach, which uses multiple stochastic rules and no analogy. We propose a model that employs inductive learning to discover multiple rules, and assigns them confidence scores based on their performance in the lexicon. Our model is supported over the two alternatives by new "wug test" data on English past tenses, which show that participant ratings of novel pasts depend on the phonological shape of the stem, both for irregulars and, surprisingly, also for regulars. The latter observation cannot be explained under the dual mechanism approach, which derives all regulars with a single rule. To evaluate the alternative hypothesis that all morphology is analogical, we implemented a purely analogical model, which evaluates novel pasts based solely on their similarity to existing verbs. Tested against experimental data, this analogical model also failed in key respects: it could not locate patterns that require abstract structural characterizations, and it 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 rules.

Keywords: Rules; Analogy; Similarity; Past tenses; Dual mechanism model

[*] Corresponding authors.
E-mail addresses: albright@ucsc.edu (A. Albright); bhayes@humnet.ucla.edu (B. Hayes).

0022-2860/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0010-0277(03)00146-X

1. Introduction: rules in regular and irregular morphology

What is the mental mechanism that underlies a native speaker's capacity to produce novel words and sentences? Researchers working within generative linguistics have commonly assumed that speakers acquire abstract knowledge about possible structures of their language and represent it mentally as rules. An alternative view, however, is that new forms are generated solely by analogy, and that the clean, categorical effects described by rules are an illusion which vanishes under a more fine-grained, gradient approach to the data (Bybee, 1985, 2001; Rumelhart & McClelland, 1986; Skousen, 1989).

The debate over rules and analogy has been most intense in the domain of inflectional morphology. In this area, a compromise position has emerged: the dual mechanism approach (see e.g. Clahsen, 1999; Pinker, 1999a; Pinker & Prince, 1988, 1994) adopts a limited set of rules to handle regular forms – in most cases just one, extremely general default rule – while employing an analogical mechanism to handle irregular forms. There are two motivating assumptions behind this approach: (1) that regular (default) processes are clean and categorical, while irregular processes exhibit gradience and are sensitive to similarity; and (2) that categorical processes are a diagnostic for rules, while gradient processes must be modeled only by analogy.

Our goal in this paper is to challenge both of these assumptions, and to argue instead for a model of morphology that makes use of multiple, stochastic rules. We present data from two new experiments on English past tense formation, showing that regular processes are no more clean and categorical than irregular processes. These results run contrary to a number of previous findings in the literature (e.g. Prasada & Pinker, 1993), and are incompatible with the claim that regular and irregular processes are handled by qualitatively different mechanisms. We then consider what the best account of these results might be. We contrast the predictions of a purely analogical model against those of a model that employs many rules, including multiple rules for the same morphological process, and that includes detailed probabilistic knowledge about the reliability of rules in different phonological environments. We find that in almost every respect, the rule-based model is a more accurate account of how novel words are inflected.

Our strategy in testing the multiple-rule approach is inspired by a variety of previous efforts in this area. We begin in Section 2 by presenting a computational implementation 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 (Daugherty & Seidenberg, 1994; MacWhinney & Leinbach, 1991; Rumelhart & McClelland, 1986). We then present data in Section 3 from two new nonce-probe (*wug* test; Berko, 1958) 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, in Section 4 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., 2001).

2. Models

2.1. Rules and analogy

To begin, we lay out what we consider the essential properties of a rule-based or analogical approach. The use of these terms varies a great deal, and the discussion that follows depends on having a clear interpretation of these concepts.

Consider a simple example. In three wug testing experiments (Bybee & Moder, 1983; Prasada & Pinker, 1993; and the present study), participants have found *splung* [splʌŋ] fairly acceptable as a past tense for *spling* [splɪŋ]. This is undoubtedly 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*. In an analogical approach, these words play a direct role in determining behavior on novel items: *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 apparently involves ending with the sequence [ɪŋ], and perhaps also in containing a preceding liquid, *s* + consonant cluster, and so on (Bybee & Moder, 1983).

Under a rule-based approach, on the other hand, the influence of existing words is mediated by rules that are generalized over the data in order to locate a phonological context in which the [ɪ] → [ʌ] change is required, or at least appropriate. For example, one might posit an [ɪ] → [ʌ] rule restricted to the context of a final [ŋ], as in (1).

(1) ɪ → ʌ / ___ ŋ]$_{[+\text{past}]}$

At first blush, the analogical and rule-based approaches seem to be different ways of saying the same thing – the context / ___ ŋ]$_{[+\text{past}]}$ 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 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.

In contrast, there is nothing inherent in an analogical approach that requires similarity to be structured; each analogical form could be similar to *spling* in its own way. Thus, if English (hypothetically) had verbs like *plip-plup* and *sliff-sluff*, in a purely analogical model these verbs could gang up with *fling*, *sting*, etc. as 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

Since analogical approaches rely on a more general – possibly variegated – notion of similarity, they are potentially able to capture effects beyond the reach of structured similarity, and hence of rules. If we could find evidence that speakers are influenced by variegated similarity, then we would have good reason to think 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, in a way that is tailored to the task at hand. We are certainly not claiming that *all* analogical models are susceptible to the same failings that we find in the model presented here. However, when an analogical model is biased or restricted to pay attention to the same things that could be referred to in the corresponding rules, it becomes difficult to distinguish the model empirically from a rule-based model (Chater & Hahn, 1998). Our interest is in testing the claim of Pinker and others that some morphological processes cannot be adequately described without the full formal power of analogy (i.e. beyond what can be captured by rules). Thus, we adopt here a more powerful, if more naïve, model of analogy, which makes maximally distinct predictions by employing the full range of possible similarity relations.

2.2. *Criteria for models*

Our modeling work takes place in the context of a flourishing research program in algorithmic learning of morphology and phonology. Some models that take on similar tasks to our own include connectionist models (Daugherty & Seidenberg, 1994; MacWhinney & Leinbach, 1991; Nakisa et al., 2001; Plunkett & Juola, 1999; Plunkett & Marchman, 1993; Rumelhart & McClelland, 1986; Westermann, 1997), symbolic analogical models such as the Tilburg Memory-Based Learner (TiMBL; Daelemans, Zavrel, van der Sloot, & van den Bosch, 2002), Analogical Modeling of Language (AML; Eddington, 2000; Skousen, 1989), the Generalized Context Model (Nakisa et al., 2001; Nosofsky, 1990), and the decision-tree-based model of Ling and Marinov (1993).

In comparing the range of currently available theories and models, we found that they generally did not possess all the features needed to fully evaluate their predictions and performance. Thus, it is useful to start with a list of the minimum basic properties we think are necessary to provide a testable model of the generative capabilities of native speakers.

First, a model should be fully explicit, to the point of being machine implemented. It is true that important work in this area has been carried out at the conceptual level (for example, Bybee, 1985; Pinker & Prince, 1988), but an implemented model has the advantage that it can be compared precisely with experimental data.

Second, even implemented models differ in explicitness: some models do not actually generate outputs, but merely classify the input forms into broad categories such as "regular", "irregular", or "vowel change". As we will see below, the use of such broad categories is perilous, because it can conceal grave defects in a model. For this reason, a model must fully specify its intended outputs.

Third, where appropriate, models should generate multiple outputs for any given input, and they should rate each output on a well-formedness scale. Ambivalence between different choices, with gradient preferences, is characteristic of human judgments in morphology, including the experimental data we report below.

Fourth, models should be able to discover the crucial phonological generalizations on their own, without human assistance. This means that models should not require that the analyst select in advance a particular group of phonological properties for the model to attend to.[1] Models that satisfy this criterion are more realistic, and also produce clearer comparative results, since their performance does not depend on the ability of the analyst in picking out the right learning variables in advance.

Finally, for present purposes, we need a pair of models that embody a maximally clear distinction between rules and analogy, following the criterion of structured vs. variegated similarity laid out in the previous section. From this point of view, a number of existing models could be described as hybrid rule-analogy models. While such models are well worth exploring on their own merits,[2] they are less helpful in exploring the theoretical predictions of analogical vs. rule-based approaches.

Below, we describe two implemented models that satisfy all of the above criteria.

2.3. A rule-based model

2.3.1. 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 iterative generalization over pairs of forms. The starting point is to take each learning pair (here, a verb stem and its past) and construe it as a word-specific rule; thus, for example, the pair *shine-shined*[3] [ʃaɪn]-[ʃaɪnd] is interpreted as "[ʃaɪn] becomes

[1] Some examples: Plunkett and Juola (1999) fitted input verbs (all monosyllabic) into templates of the form CCCVVCCC. They used right alignment, so that final consonants were always placed in the final C slot (whereas initial consonants would be placed in any of the first three slots, depending on the initial cluster length). In Eddington's (2000) analysis of English past tenses using AML and TiMBL, verbs were coded with a predefined set of variables that included the final phoneme, an indication of whether the final syllable was stressed, and a right-aligned representation of the last two syllables. In both cases, the choice was highly apt for learning English past tenses – but would not have been if some quite different morphological process such as prefixation had happened to be present in the learning data.

In contrast, the actual input data to children consist of whole words, composed of dozens or even hundreds of (frequently correlated) feature values. Furthermore, phonological environments are often formed from conjunctions of two or more features (e.g. [-əd] is selected when the final segment is both alveolar and a stop), and different features are relevant for different classes (cf. [-t] when the final segment is voiceless). Recent work in the area of instance-based learning has made headway on the task of finding the relevant features from among a larger set (see Daelemans et al., 2002; Howe & Cardie, 1997; Wettschereck, Aha, & Mohri, 1997; Zavrel & Daelemans, 1997); however, we are not aware of any feature-selection technique that would allow the learner, on the basis of limited data, to isolate all the different combinations of features that we find to be relevant below.

[2] To encourage such exploration, we have posted our learning sets, features, and experimental data on the Internet (http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/).

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

[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).

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

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

(4) $\varnothing \rightarrow$ d / [kənsaɪn ___]$_{[+\text{past}]}$

Since the structural change ($\varnothing \rightarrow$ 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).

(5)	a.	Change	Variable	Shared features	Shared segments	Change location		
	b.	$\varnothing \rightarrow$ d / [		ʃ	aɪn	___	]$_{[+\text{past}]}$	(*shine-shined*)
	c.	$\varnothing \rightarrow$ d / [	kən	s	aɪn	___	]$_{[+\text{past}]}$	(*consign-consigned*)
	d.	$\varnothing \rightarrow$ d / [	X	$\begin{bmatrix}+\text{strident}\\+\text{continuant}\\-\text{voice}\end{bmatrix}$	aɪn	___	]$_{[+\text{past}]}$	(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**. Moving outward from the location of the change, any segments shared by the two specific rules (here, [aɪn]) are retained in the generalized rule. Where two segments differ, but can be grouped together using phonological features, this is done to create a featural term; here, [ʃ] and [s] reduce to their common features [+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 minimal generalization, see Albright and Hayes (2002).

2.3.2. Features

Phonological features permit minimal generalization 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 word-specific 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, then [-əd] could be attached everywhere, yielding impossible forms like **jumpèd* [dʒʌmpəd]. Instead, our implementation uses features to generalize more conservatively, as in (6d). The featural expression in (6d) uniquely characterizes the class [t, d]. Thus, the system will correctly attach [-əd] after only these sounds.

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

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

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

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

Features also permit the system to generalize to segments it has never seen before. Pinker (1999a), adapting Halle (1978), gives the following 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 generates [baxt] correctly, even if the learning data do not contain [x], since the featural term permits it to discover contexts like "after voiceless segments".

2.3.3. Phonology

Outputs of morphological processes are often shaped by principles of phonological well-formedness. In English, such principles guide the choice of the regular allomorph of the past tense suffix (Pinker & Prince, 1988, pp. 101–108). Our rule-based model 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 model will employ featural generalization 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 restricting the morphological rule, because there is no combination of features that includes [b], [g], and [n] without also including [d].[4] 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. One approach uses **phonological rules** (Bloomfield, 1939; Chomsky & Halle, 1968), allowing the morphology to suffix [-d] to [nid], and then repairing the resulting *[dd] cluster by a phonological rule that inserts schwa: /nid + d/ → [nidəd]. Alternatively, in a **constraint-based** approach (Bird, 1995; Prince & Smolensky, 1993), multiple candidate outputs are produced (e.g. [nidd] and [nidəd]), and some of them are filtered by phonological constraints; thus, a constraint like *[dd] eliminates [nidd].

Our rule-based model can 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 model 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

[4] [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].

known, prior to morphological learning.[5] In modeling our experimental data, we have found that the constraint-based approach yielded slightly better results (and much better results for the analogical model discussed below), so we adopt it for purposes of the present study.

2.3.4. Iterative generalization and rule evaluation

The first stages of generalization tend to produce rather arbitrary and idiosyncratic rules like (5d). However, when the process is iterated, 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. $\varnothing \rightarrow \text{d} / [\text{X} \_\_\_]_{[+\text{past}]}$

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

$\varnothing \rightarrow \text{t} / [\text{X} [-\text{voice}] \_\_\_]_{[+\text{past}]}$

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

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, p. 124), is to keep only those rules that are maximally *general* (as defined by the number of forms they correctly derive). Here, however, we will adopt a different strategy, rewarding instead those generalizations which are maximally *accurate*. In doing this we follow earlier work, both in inductive rule-learning (Michalski, 1983; Mikheev, 1997) and in instance-based learning (Daelemans et al., 2002; Skousen, 1989, 2001).

To assess accuracy, our rule-based model collects some simple statistics about how well the rules perform in deriving the forms in the learning data. For example, (7a) $\varnothing \rightarrow [\text{d}] / [\text{X} \_\_\_]_{[+\text{past}]}$, the most general rule for English pasts, is applicable to all verbs; 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 2.5), this value is 4253. If phonological rules are employed, this rule derives the correct output for all 4034 regular forms; that is, it achieves 4034 **hits**. To calculate an accuracy score for the rule, we divide hits by scope, obtaining a tentative score (which we call **raw confidence**) of 0.949. The rule [ɪ] → [ʌ] / {l, r} ___ ŋ, which covers past tenses like *sprung*, has a scope of 9 and 6 hits, yielding a raw confidence of 0.667.

Generalizations can be trusted better when they are based on more data. Following Mikheev (1997), we use lower confidence limit statistics on the raw confidence ratio to

[5] This appears to be a realistic assumption; for literature review and discussion see Hayes (in press).

penalize rules based on a small number of forms. For instance, if the lower confidence limit (α) is 75%, a score of 5/5 is downgraded from 1.00 to an **adjusted confidence** of 0.825. A score of 1000/1000, however, is downgraded only from 1.000 to 0.999. The lower confidence limit is a parameter of the model; in modeling our experimental data, the best fit was achieved by setting its value to 0.55. 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 non-uniform distributions, following Bayardo, Agrawal, and Gunopulos (1999); for discussion of how this works and why it is needed, see Albright and Hayes (2002). The value of the upper confidence limit was also set by fitting to experimental data, at $\alpha = 0.95$.

2.3.5. Islands of reliability

The goal of assessing the accuracy of rules is 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 data set.) 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)

$$\varnothing \rightarrow t \,/\, [X \begin{bmatrix} -\text{sonorant} \\ +\text{continuant} \\ -\text{voice} \end{bmatrix} \underline{\quad\quad}]_{[+\text{past}]}$$

"Suffix [-t] to stems ending in voiceless fricatives."

The adjusted confidence of this rule is 0.998, 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** (Albright, 2002). Islands of reliability are found 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.[6]

2.3.6. Generating outputs

Probabilistic confidence values allow the rule-based model to generate multiple, competing outputs with numerical confidence values. When an input form is submitted to the model for wug testing, it is compared against all the rules in the grammar. Each rule

[6] 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.

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

Output	Rule	Hits/Scope	Raw confidence	Adjusted confidence	Hits/Failures
gleeded	∅ → əd / [X {d, t} ___]$_{[+\text{past}]}$	1146/1234	0.929	0.872	*want*, *need*, *start*, *wait*, *decide*, *etc.* / **get*, **find*, **put*, **set*, **stand*, etc.
gled	i → ɛ / [X {l, r} ___ d]$_{[+\text{past}]}$	6/7	0.857	0.793	*read*, *lead*, *bleed*, *breed*, *mislead*, *misread* / **plead*
glode	i → o / [X C ___ [+cons]]$_{[+\text{past}]}$	6/184	0.033	0.033	*speak*, *freeze*, *weave*, *interweave*, *bespeak* / **leak*, **teach*, **leave*, etc.
gleed	No change / [X {d, t} ___]$_{[+\text{past}]}$	29/1234	0.024	0.014	*shed*, *spread*, *put*, *let*, *set*, *cut*, *hit*, *beat*, *shut*, *hurt*, *cost*, *cast*, *burst*, *split*, etc. / **get*, **want*, **need*, etc.

that can apply does so, deriving a candidate output form. In many cases, there will be numerous rules that involve the same change and derive identical outputs. We assume that the candidate output is assigned the well-formedness score of the best rule that derives it.

As an illustration of how the model works, Table 1 shows the outcomes it derives for the wug verb *gleed*, along with their raw confidence values and the adjusted values.

2.3.7. 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, 1999a,b; Pinker & Prince, 1988, 1994; Prasada & Pinker, 1993). These scholars argue that rules are fundamentally inadequate for describing irregular patterns, because they characteristically involve "prototypicality" or "family resemblance" effects. We quote Pinker (1999b):

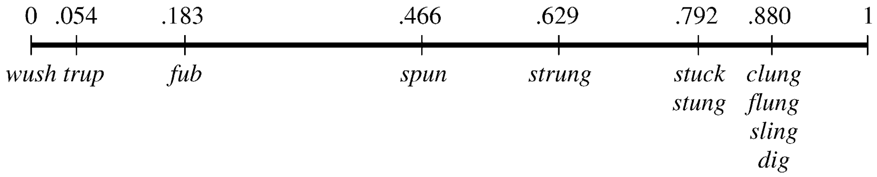
> Just as we have a rule adding "ed" to form the regular past tense, we [could have] a suite of rules that generate irregular past tense forms by substituting vowels or consonants. For example, one rule changes "i" to "u" in verbs like "cling, clung"… A problem for this theory is the family resemblance 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]$_{[+\text{past}]}$ 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]$_{[+\text{past}]}$ 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; *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.[7]

We feel that this conclusion is premature, and holds only of more traditional rule-based accounts (e.g. Halle & Mohanan, 1985; Hoard & Sloat, 1973). An approach based on stochastic rules, such as the one advocated here, can easily be adapted to account for prototypicality effects. First, we agree with dual mechanism theorists (and most of traditional linguistic theory; cf. Aronoff, 1976) that irregulars are lexically listed; this is what prevents them from being regularized. 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 central forms *cling*, *fling*, *sling*, and *dig* lead to a rule ("replace [ɪ] with [ʌ] between a voiced dental consonant and a final voiced velar consonant") that characterizes them with considerable precision; it works in 4/4 cases and yields a score of 0.880. But if *fub* were to be the past tense of *fib*, it would have to be produced by a more general but less accurate rule ("replace [ɪ] with [ʌ] between any consonant and any final voiced consonant"). This rule has 11 hits (adding *win*, *swing*, *spring*, *spin*, *sting*, *wring*, and *string*); but it also has a much larger scope (45), because it encompasses many forms like *bring*, *grin* and *rig*. As a result, the score for *fub* would be only 0.183. *Trup* requires an even more general and less accurate rule, with 12 hits (adding in *stick*) and 98 misses, for a score of 0.054. Adding in the scores for the other verbs Pinker mentions, we obtain the values shown in (9).

(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.[8]

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

[7] Pinker also objects to ɪ → ʌ / [X ___ ŋ]$_{[+\text{past}]}$ 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 be maintained for rules describing irregular processes.

[8] Despite superficial appearances, the graph in (9) 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.

2.4. An analogical model

In developing a model that works purely on an analogical basis, we have adopted a version of the Generalized Context Model (GCM; Nosofsky, 1990). This model is intended to be a very general account of how similarity influences people's intuitive judgments, and has been used to capture 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 most respects.

2.4.1. The core of the model

The intuitive idea behind the GCM can be illustrated with a simple example. Suppose we wish to evaluate the likelihood of uttering *scrode* as the past tense of *scride*. To do this, we compare *scride* with all of the existing verbs that form their past tenses with the change [aɪ] → [o]. (In our learning set, these are *dive*, *drive*, *ride*, *rise*, *shine*, *smite*, *stride*, *strive*, and *write*.) We then assess the similarity of *scride* to each member of this set, following a procedure described below. By adding the similarity scores 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 in Fig. 1; arrows are labeled with the actual similarity values used in our model.

To this basic idea, the GCM adds an important adjustment: we must compensate for how much *scride* resembles 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.[9] Thus, the score that the model gives to *scrode* is:

(10) $$\frac{\textit{summed similarity of}\ \text{scride}\ \textit{to all members of the}\ [\text{aɪ}] \rightarrow [\text{o}]\ \textit{class}}{\textit{summed similarity of}\ \text{scride}\ \textit{to all verbs}} = \frac{0.3997}{3.72} = \mathbf{0.1079}$$

2.4.2. Calculating similarity

The similarity of two forms is calculated by first finding their optimal alignment; that is, the alignment that minimizes the string edit distance (Kruskal, 1999). In order to ensure that phonetically similar segments are preferentially aligned with one another, the substitution cost function is made sensitive to the relative similarity of the segments involved. In calculating the similarity of segments, we adopt the natural class based theory proposed by Broe (1993), which provides a good estimate of the relative similarity of individual pairs of segments (Frisch, 1996; Frisch, Broe, & Pierrehumbert, 1997).[10] In addition to penalizing mismatched segments, the model also assesses a cost for leaving segments unaligned, and for misaligning stressed syllables; the optimal

[9] Like Nakisa et al. (2001), we omitted bias terms for output patterns.

[10] Broe's theory derives a similarity score for each pair of segments; since what we need is a dissimilarity penalty, we subtract the similarity score from one.

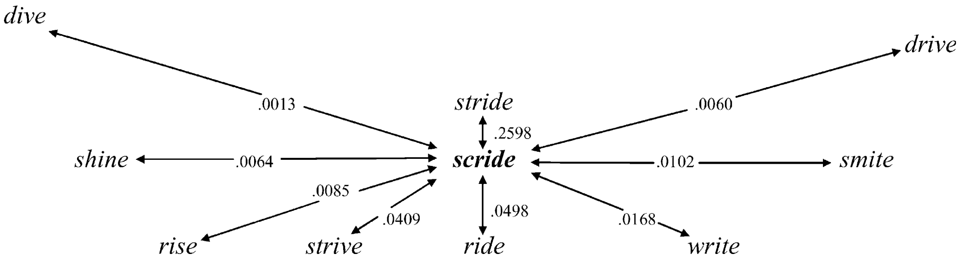


Fig. 1. Similarity of all [aɪ] → [o] forms to *scride*.

values for these penalties were established by fitting to the data; both turned out to be 0.6.

Taking all of these considerations together, the model finds the least costly alignment of the novel word to existing words; that is, the alignment that minimizes the sum of all penalties. For the case of *scride* and *shine*, this alignment is shown below:

(11)														
	shine:	ʃ		null		null		a		ɪ		n		
	penalty:	0.155	+	0.6	+	0.6	+	0	+	0	+	0.667	=	2.022
	scride:	s		k		r		a		ɪ		d		

The last step is to convert this overall dissimilarity value to overall similarity, using the following equation (Nakisa et al., 2001; Nosofsky, 1990):

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

where η_{ij} is the calculated similarity of two forms i and j, d_{ij} is the dissimilarity of i and j, and s and p are parameters, fixed by fitting to the data.

The parameter s has the following effect: when s is low, the model tends to rely primarily on a small set of very similar forms in forming its judgments. As s increases, the model becomes more sensitive to a broader range of forms. The effect of p is subtler and will not be reviewed here. The best-fit values for s and p turned out to be 0.4 and 1, respectively.

Applying (12) to the value $d_{shine,scride} = 2.022$, obtained in (11) above, we get 0.0064, which is the similarity value that appeared earlier in Fig. 1. Once we have the similarity of *scride* to all other forms in the learning set, we can use (12) to determine that *scrode* should receive a score of 0.1075 as the past tense of *scride*. All other candidates are evaluated in the same way.

2.4.3. Generating outputs

As just described, the model does not generate, but only evaluates candidates. To plug this gap, we augmented our GCM implementation with a generative front end, which simply locates all the possible structural changes for past tense formation, and creates candidates by applying all applicable structural changes freely. Thus, for a stem like *scride* [skraɪd], the module constructs candidates with the three regular past allomorphs ([skraɪd**əd**], [skraɪd**d**], [skraɪd**t**]; plus candidates that follow all applicable irregular patterns: *scrode* (discussed above), *scride* (cf. *shed/shed*), *scrite* (*bend*), *scrid* (*hide*), *scroud* [skraʊd] (*find*), *scrud* (*strike*), and *scraud* [skrɔd] (*fight*). Of these, [skraɪdd] and

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*, *bottle-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*

[skraɪdt] are phonologically filtered; the remaining candidates are submitted to the core GCM algorithm for evaluation, as described above.

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.

To conclude, we feel that a model of this sort satisfies a rigorous criterion for being "analogical", as it straightforwardly embodies the principle that similar forms influence one another. The model moreover satisfies the criteria laid out in Section 2.1: it is fully susceptible to the influence of variegated similarity, and (unless the data accidentally help it to do so) it utterly ignores the structured-similarity relations that are crucial to our rule-based model.

2.5. Feeding the models

We sought to feed both our rule-based and analogical models a diet of stem/past tense pairs that would resemble what had been encountered by our experimental participants. We took our set of input forms from the English portion of the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995), selecting all the verbs that had a lemma frequency of 10 or greater. In addition, for verbs that show more than one past tense (like *dived*/*dove*), we included both as separate entries (e.g. both *dive-dived* and *dive-dove*). The resulting corpus consisted of 4253 stem/past tense pairs, 4035 regular and 218 irregular. Verb forms were listed in a phonemic transcription reflecting American English pronunciation.

A current debate in the acquisition literature (Bybee, 1995; Clahsen & Rothweiler, 1992; Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 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 3308 input pairs (3170 regular, 138 irregular), and ran both learning models on both sets. 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 reflect 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 scope and hits (rule-based model) or in counting the similar forms (analogical model). Bybee (1995, 2001) and Pierrehumbert (2001) 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.

2.6. 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.[11] 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 *snuck*. 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.

A similar point is made by Ling and Marinov (1993, pp. 264–265), who argue that testing against the learning corpus is not the right way to evaluate models. The problem is that real speakers have the benefit of having memorized the irregulars, and the models do not; hence it is unrealistic to expect the models to reproduce existing irregulars that they have never seen, simply by guessing. 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.

3. Experiments

To this end, we carried out two experiments on English past tenses, 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, both regular and irregular. Phonological well-formedness ratings of all of the wug stems

[11] We tested each system by randomly dividing the learning set in ten, and modeling each tenth using the remaining nine tenths as input data (Weiss & Kulikowski, 1991). For the rule-based model, 4192 of the 4199 forms were output as regular; three (*withstand*, *take*, *partake*) were retained as irregular, and four were irregularized (*clink-clunk*, *deride-derode*, *plead-pled*, *stake-stook*). For the analogical model, 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* ['bændiəd] and *taxi*'t ['tæksit]. Such errors occurred 1.2% of the time; we discuss them in Section 4.3.5.

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 of the wug verbs, and avoided uncertainty about the intended responses due to ambiguous spelling.

3.1. Stimuli

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

3.1.1. The Core set

The Core set was designed to test the following hypotheses:

(13) a. If a verb falls into an island of reliability for **irregular** pasts (e.g. *spling-splung*), will it receive higher ratings?
b. If a verb falls into an island of reliability for **regular** pasts (e.g. *blafe-blafed*), 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.) The questions in (13) 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 (13a) is "yes" (replicating Bybee & Moder, 1983), and the answer to question (13b) 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 (13b). 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 be a perfect past tense for *ploamph*, but receives low ratings because *ploamph* itself is odd. Prasada and Pinker attempted to correct for this by subtracting stem phonological well-formedness ratings from past tense ratings; when this is done, the similarity effects for regulars appear 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.

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.

In an attempt to circumvent this problem in designing our own wug verbs, we used our rule-based model as a tool for experimental design. We constructed a set of 2344 candidate wug forms, by concatenating combinations of relatively common syllable onsets and syllable rhymes. By starting with phonologically "bland" initial candidates, 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 the 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.

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.

Examples of the wug verbs chosen for the four basic categories of the Core set are given in (14). For each verb, we include the irregular forms that were provided as options for participants to rate in Experiment 2 (normally just one, occasionally two). For the complete set of Core verbs, see Appendix A.

(14) a. Island of reliability for both regulars and irregulars
dize [daɪz] (*doze* [doz]); ***fro*** [fro] (*frew* [fru]); ***rife*** [raɪf] (*rofe* [rof], *riff* [rɪf])

b. Island of reliability for regulars only[12]
bredge [brɛdʒ] (*broge* [brodʒ]); ***gezz*** [gɛz] (*gozz* [gaz]); ***nace*** [nes] (*noce* [nos])

c. Island of reliability for irregulars only
fleep [flip] (*flept* [flɛpt]); ***gleed*** [glid] (*gled* [glɛd], *gleed*); ***spling*** [splɪŋ] (*splung* [splʌŋ], *splang* [splæŋ])

d. Island of reliability for neither regulars nor irregulars
gude [gud] (*gude*); ***nung*** [nʌŋ] (*nang* [næŋ]); ***preak*** [prik] (*preck* [prɛk], *proke* [prok])

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

3.1.2. 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 (15), were included that resembled existing verbs of the *burnt* class, in which [-t] is exceptionally suffixed 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 (2002).

(15) Pseudo-*burnt* verbs
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 (16) 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.

(16) Potential single-form analogy forms
kive [kɪv] (*kave* [kev]), cf. *give-gave*; ***lum*** [lʌm] (*lame* [lem]), cf. *come-came*; ***pum*** [pʌm] (*pame* [pem]), cf. *come-came*; ***shee*** [ʃi] (*shaw* [ʃɔ]), cf. *see-saw*; ***zay*** [ze] (*zed* [zɛd]), cf. *say-said*

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

The remaining forms in (17) 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 they could serve as a comparison test for individual-verb analogies.

(17) Other potentially analogical forms
chind [tʃaɪnd] (*chound* [tʃaʊnd], *chind* [tʃaɪnd]), cf. *find-found*; ***shy'nt*** [ʃaɪnt] (*shoant* [ʃont], *shount* [ʃaʊnt]), cf. *shine-shone*; ***gry'nt*** [graɪnt] (*groant* [gront], *grount* [graʊnt]), cf. *grind-ground*; ***ry'nt*** [raɪnt] (*roant* [ront], *rount* [raʊnt]), cf. *write-wrote*; ***flet*** [flɛt] (*flet*), cf. *let-let*

3.2. Participants

All of the experimental participants were native speakers of American English, primarily UCLA undergraduates. They were paid $10 for their participation, which took between 45 minutes and 1 hour, and took place in the sound booth of the UCLA Phonetics Laboratory.

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

3.3. Experiment 1 procedure

Experiment 1 consisted of two parts: an initial pretest to obtain baseline phonological well-formedness scores, and then the main production task to elicit past tense forms. There were 20 participants.

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 a pretest. For reasons discussed in Section 3.1.1 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.

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 ("completely bizarre, impossible as an English word") to 7 ("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. Stimuli for both Experiments 1 and 2 were presented using PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993), and participants entered ratings using a specially modified keyboard. For the ratings task there was a training period of five novel verbs.

After the pretest, participants volunteered past tense forms for all of the wug verbs listed in Section 3.1, in an open response sentence completion task. For the sentence completion task, each wug verb was embedded in a frame dialog consisting of four sentences:

(18)	**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."

Participants heard the first two sentences over headphones, but on the screen saw 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; thus (for (18)), *rifing* for sentence 3, and *rifed*, *rofe*, or some other past tense form for sentence 4.[14]

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. If either listener recorded something other

[14] The full set of frame dialogs may be downloaded from http://www.linguistics.ucla.edu/people/hayes/rulesvsanalogy/FrameSentences.htm.

than the expected *-ing* form for sentence 3 (as occurred in 62/1160 trials), then the past tense response in sentence 4 was discarded for that trial.

For the volunteering portion of Experiment 1, there was a training period of five 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 wug verb. The order was randomized on a subject-by-subject basis. 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.

3.4. 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. There were 24 participants, none of whom had participated in Experiment 1.

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 helped to ensure that participants had heard and internalized the wug verbs correctly. 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 mini-dialog repeated, this time with the opposite past tense form to rate. The purpose of the mini-dialog was to encourage participants to consider the goodness of novel pasts *in relation to the given wug stem.* The full protocol is shown in (19).

(19) *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):

(20) 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

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 11 wug verbs, two different irregulars were provided (see Section 3.1). 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 items was randomized on a subject-by-subject basis, and the frame for each 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*, with past tenses *frank*, *frunk*, and *fret*. 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] (*punt*, *pinted*), *kip* (*kap* [kæp], *kipped*), and *prack* (*pruck*, *pracked*).

3.5. Coding the results

3.5.1. Correcting for phonological well-formedness

Recall from Section 3.1.1 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 ($\mathrm{SD} = 1.62$, $n = 58$), whereas the average rating for our ill-formed foils (rated phonologically, but not included in the wug tests) was 2.97 ($\mathrm{SD} = 1.46$, $n = 29$). More important, the phonological ratings data were poorly correlated with the participants' ratings of past tense forms: $r(58) = 0.006$.[15] Thus, it seems that our scheme for avoiding major phonological ill-formedness effects was successful.

Nevertheless, as an added precaution, we used the phonological well-formedness ratings gathered in Experiment 1 to correct for any phonological effects in the ratings data. 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

[15] The comparable value for Prasada and Pinker (1993) was $r = 0.214$. The greater role of phonological well-formedness in Prasada and Pinker's study was probably 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 $r(13) = 0.867$.

Experiment 1. The residuals of this regression were then rescaled, so 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 extremely similar results obtained either way; we report here the results using adjusted ratings.

3.5.2. 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.

4. Results

The data collected in Experiments 1 and 2 are summarized in Appendix A.

4.1. Preliminaries

4.1.1. The preference for regulars

The participants preferred 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 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, participants alternated between volunteering and rating. The irregular forms presented for rating were apparently an implicit invitation to offer irregular forms in the volunteering task. In terms of our models, we would characterize the behavior of Experiment 2 participants as making more frequent 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.[16] Instead, we have calculated correlations for regulars and irregulars separately.

[16] For the ratings data for Experiment 2, the overall correlation with regulars and irregulars combined is: rule-based model, $r = 0.806$; analogical model, $r = 0.780$. A model that guesses 1 for regulars and 0 for irregulars would achieve a correlation of $r = 0.693$.

4.1.2. Ratings data vs. volunteered forms

The production probabilities for volunteered forms correlate reasonably well with ratings data: $r = 0.837$ (0.929 among regulars; 0.690 among irregulars). Breaking this down between Experiment 1 (pure volunteering) and Experiment 2 (volunteering interspersed with rating), the correlations are: Experiment 1, $r = 0.788$ (regulars 0.814, irregulars 0.515); Experiment 2, $r = 0.865$ (regulars 0.902, irregulars 0.685). For the Experiment 2 forms, the correlation is unsurprising, 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.

4.2. 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. As it is generally interpreted, 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 3.1.1) was designed to test this prediction; it included wug verbs falling either within or outside the islands of reliability for both regulars and irregulars.

4.2.1. Results

Figs. 2a and 2b 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

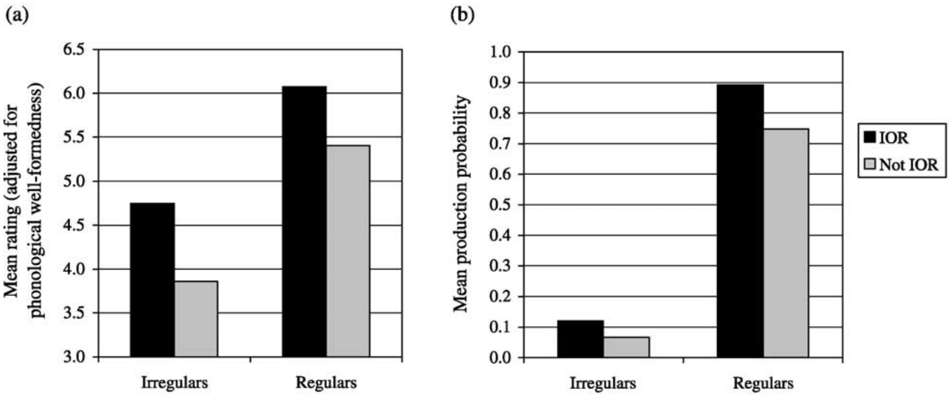


Fig. 2. Effect of islands of reliability (IOR) for irregulars and regulars. (a) IOR effect on ratings (adjusted). (b) IOR effect on production probabilities.

findings of Bybee and Moder (1983) 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 in Figs. 2a and 2b show a more surprising result, namely that island of reliability effects were also 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 $F(1, 78) = 94.22$, $P < 0.0001$, production probabilities $F(1, 78) = 758.38$, $P < 0.0001$) and islandhood (islands of reliability higher than non-islands; ratings $F(1, 78) = 27.23$, $P < 0.0001$, production probabilities $F(1, 78) = 14.05$, $P < 0.001$), with no significant interaction. Thus, we find that both regulars and irregulars are susceptible to island of reliability effects, to an equal extent.

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 (1993), 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 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 acceptable, (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 this result is to examine not just the difference in means between island of reliability and non-island of reliability 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 (21) shows, both the ratings and production probabilities are positively correlated with the predictions of the two models, to varying degrees (see Appendix A for all values). Crucially, positive correlations are seen not just for irregular pasts, but for regulars as well.

(21) Correlations (r) of participant responses to model predictions: Core verbs ($n = 41$)

	Rule-based model		Analogical model	
	Ratings	Production probabilities	Ratings	Production probabilities
Regulars	0.745 ($P < 0.0001$)	0.678 ($P < 0.0001$)	0.448 ($P < 0.01$)	0.446 ($P < 0.01$)
Irregulars	0.570 ($P < 0.0001$)	0.333 ($P < 0.05$)	0.488 ($P < 0.001$)	0.517 ($P < 0.0001$)

In summary, we find no evidence that island of reliability effects are weaker for novel regulars than for novel irregulars.

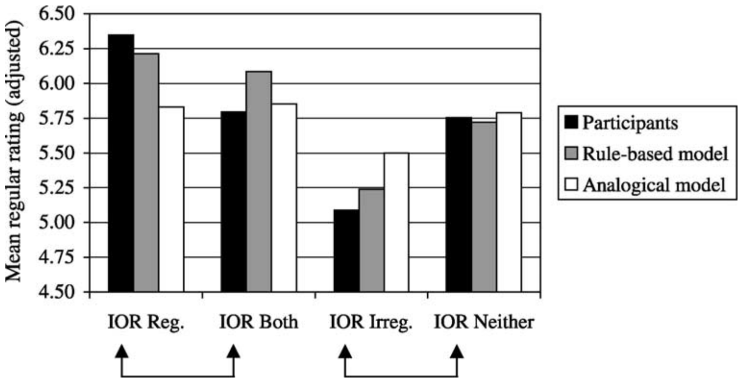


Fig. 3. Mean ratings of regulars within four categories of islandhood.

4.2.2. *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 the 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. 3 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.

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), supporting the trade-off hypothesis.

Surprisingly, the same effects are also found among the irregulars, as Fig. 4 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.

In part, this trade-off effect appears to be simply the result of our choice of wug verbs, since it is also predicted to a certain extent by our models.[17] However, the effect among the consultants is stronger, lending partial support to the trade-off hypothesis.

Given that the responses for irregulars can influence those for regulars and vice versa, 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. In other words, are item-by-item differences in regular ratings really just caused by

[17] In the analogical model, trade-offs inevitably occur because forms compete for their share of the same denominator (10). The rule-based learner does not necessarily 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.

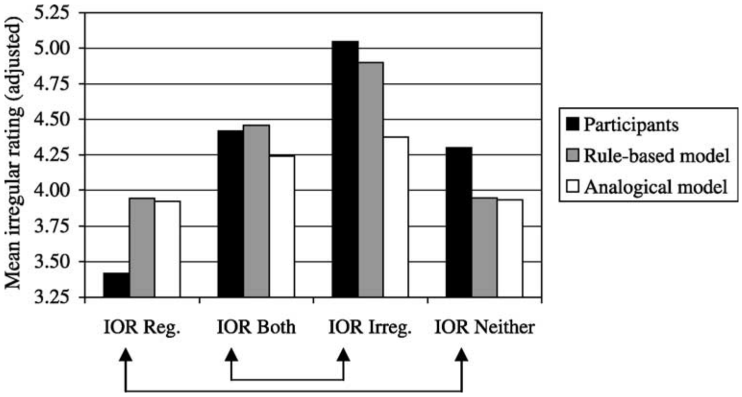


Fig. 4. Mean ratings of irregulars within four categories of islandhood.

competition from irregulars? 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 (22).

(22) 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	$r = 0.589$, $P < 0.0001$	$r = 0.322$, $P < 0.05$
Irregulars	$r = 0.497$, $P < 0.0001$	$r = 0.316$, $P < 0.05$

For the crucial case, the effect of islands of reliability on regulars, for at least the rule-based model, there remains a highly significant correlation of 0.589. 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, ratings of regulars and irregulars are to a large extent independent of one another, and 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 is removed.

In conclusion, we find that speakers' intuitions about novel past tense forms are sensitive to the phonological content 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.

4.3. 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 the two possibilities described earlier: (a) a system with a large set of detailed rules, each annotated for its reliability (Section 2.3); (b) a purely analogical system (Section 2.4), which inflects novel forms according to their resemblance to existing verbs. We assess these two possibilities by comparing their predictions against our experimental data.

Here, we will use our full data set, including both the Core and Peripheral forms (Section 3.1). 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.

4.3.1. Correlations

In Section 4.2.1, we saw that the rule-based model achieved higher correlations to participant ratings data for the Core forms than the analogical model did. The same is true for the full data set, particularly among regulars.

(23) Correlations of participant ratings to the predictions of two models: all verbs

	Rule-based model	Analogical model
Regulars ($n = 58$)	$r = 0.714$, $P < 0.0001$	$r = 0.545$, $P < 0.0001$
Irregulars ($n = 75$)	$r = 0.480$, $P < 0.0001$	$r = 0.471$, $P < 0.0001$

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.

4.3.2. Failure of the analogical model to locate islands of reliability

One way to diagnose the models' behavior is to compare their *relative* errors, attempting to diagnose whether one of the models is systematically over- or underrating certain classes of test items. We calculated relative error by first computing the absolute size of the errors made by each model. This was done by rescaling the predictions of each model to have the same mean and standard deviation as the participant ratings (adjusted for phonological well-formedness), and then calculating the difference between the models and the participant ratings. For each verb, we determined which model was closer, 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 determine the locus of errors for each of the models, we summed the total error in each of these four categories. For the regulars, the result is shown in Fig. 5. 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 model. In Table 4, we list

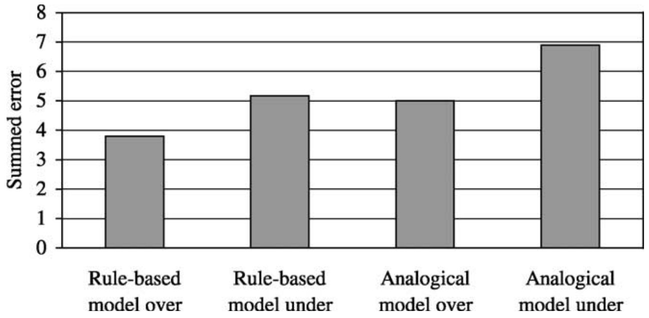


Fig. 5. Summed relative error of the two models for regulars.

the 12 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).

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. Fig. 6 give the analogous results for irregulars.

Table 4
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	/ $\begin{bmatrix} \text{V} \\ -\text{high} \end{bmatrix}$ p ___	83/83
shilked	5.82	5.97	5.17	0.49	/ $\begin{bmatrix} \text{C} \\ +\text{coronal} \end{bmatrix}$ k ___	31/31
tarked	6.24	5.97	5.66	0.31	/ $\begin{bmatrix} \text{C} \\ +\text{coronal} \end{bmatrix}$ k ___	31/31
spacked	6.13	6.01	5.79	0.22	/ $\begin{bmatrix} \text{V} \\ +\text{low} \\ -\text{round} \end{bmatrix}$ k ___	37/37
bligged	5.95	5.66	5.45	0.21	/ g ___	41/42

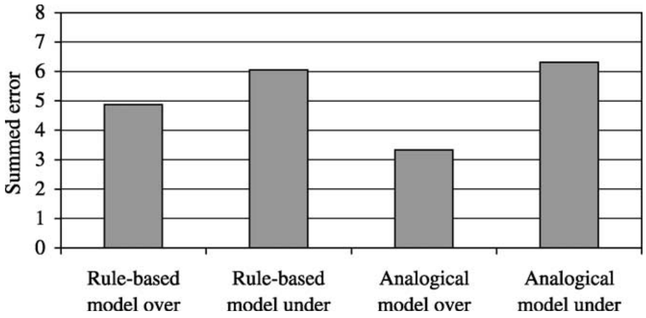


Fig. 6. Summed relative error of the two models for irregulars.

As Fig. 5 shows, 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 analogical model is unable to locate good islands 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, discussed below in Section 4.3.4. 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.

4.3.3. 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 similarly 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*); among these, the only apparent cases of single-form analogy were two instances of *kave*, two of *pame*, one of *chole*, and four of *neld*. Among the remaining verbs we found 35 candidates for single-form analogies.[18] However, inspecting this list, we found them to be unconvincing as cases of single-form analogy: they are all quite distant from their alleged model forms,

[18] 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*); 5 with [ʌ]-[æ], like *run-ran* (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*).

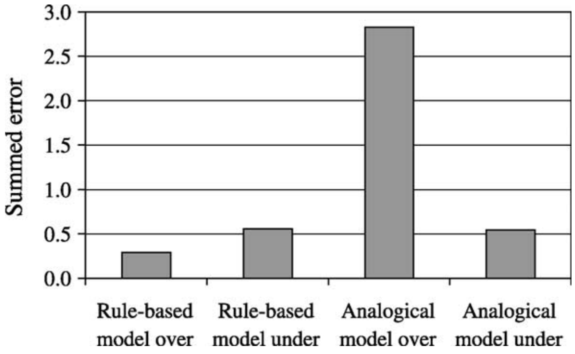


Fig. 7. Summed relative error of the two models: single-form analogies.

and moreover they virtually all fit product-oriented generalizations, a pattern discussed below in Section 5.2.[19]

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 *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 *gry'nt* 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. For instance, for *zay-zed*, *say* contributed 100% of the outcome; for *preak-proke*, *speak* contributed 83%, and so on. There were 21 such verbs. For this set of verbs we repeated the procedure described above under Figs. 5 and 6, sorting the models' errors into four categories and summing the total error in each category. The result is shown in Fig. 7. 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. 7): *kave* (rule-based model under by 0.32), *neld* (0.15), and *zed* (0.09). 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.

[19] 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 an attested [ɪ]-[e] change.

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.

4.3.4. 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 (15)). For these 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 Fig. 6 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 4.1.2), 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 one such form (*murnt*) was volunteered out of 301 valid responses. We conjecture that this large difference resulted from the fact that Experiment 2 included a ratings task, in which participants were presented with forms of the *burnt* type. It appears that participants were unlikely to think of *burnt* forms on their own, but once they had been suggested, they were volunteered more often, and rated 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 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].

4.3.5. A role for variegated similarity?

Unlike our rule-based model, our analogical model can make use of what we have termed variegated similarity (Section 2.1) 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 were often 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 (24). The shaded boxes, which show the places where models diverge from *scoil*, cover all of the territory of the word except the final [l].

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

Analog	s	k	ɔɪ	l	Similarity	Contribution
spoil	s	p	ɔɪ	l	0.225	6.0%
soil	s		ɔɪ	l	0.223	6.0%
coil		k	ɔɪ	l	0.223	6.0%
scowl	s	k	aʊ	l	0.185	4.9%
scale	s	k	e	l	0.151	4.0%

Other forms work similarly, though the amount of variegation 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 Fig. 5). 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: 0.618, analogical model overestimations: 0.837).

Although the error comparisons are uninformative, there turns out to be a much clearer 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. As noted above (Section 2.2), 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 2.3.3). 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, in particular on just a few of its features. Our 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, for the existing verb *render*, the analogical model guesses **renderèd* [rɛndərəd], based largely on the following analogical set (the ten most similar forms): *rend*, *end*, *rent*, *vend*, *raid*, *fend*, *mend*, *tend*, *round*, and *dread*. These stems bear an irrelevant similarity to *render*, which (in this case) suffices to outweigh the influence of legitimate model forms like *surrender*. The analogical model also invoked variegated similarity to overgeneralize the allomorph [-t], e.g. *whispert* [wɪspərt], based on forms like *whip*, *wish*, *whisk*, *wince*, *quip*, *lisp*, *swish*, *rip*, *work*, and *miss*.

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 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.

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 0.132, 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*, etc.).

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 (Section 2.3.2). 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.

We can now explain why the analogical model guessed fairly high scores for verbs of the *burnt* class (Section 4.3.4): 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.

The analogical model we employ here is especially susceptible to the effects of variegated similarity, because it does not assign any structure to the data. However, it appears that other, more sophisticated analogical models also have serious problems in selecting the correct past tense allomorph. Derwing and Skousen (1994), in their application of the AML model to past tenses, carefully preselected the ten variables to which their model was to attend, thus giving it a priori help in learning the right suffix allomorphs. They found that despite this head start, their model was unable to predict the suffix allomorphs correctly, and their diagnosis was the same as ours: variegated similarity (p. 213). Similarly, the connectionist simulation of Plunkett and Juola (1999), which performed impressively in reproducing its training set, nevertheless had severe problems in picking the right suffix allomorph for novel verbs;[20] variegated similarity again seems the most likely culprit.

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 poor results in predicting the distribution of the allomorphs of the regular past.

5. Discussion

5.1. Summary

We summarize here our main results. By employing our rule-based model, 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 verbs 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, since it 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 having access to the full power of the analogical approach, in particular, the ability to invoke variegated 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, as follows. (a) The analogical model systematically underrated regular forms falling within islands of reliability, because its reliance on variegated similarity prevented it from locating these islands. (b) The analogical model made drastic errors in distributing the three allomorphs of the past tense suffix, again because of its reliance on variegated similarity.

[20] In wug testing, their model picked the wrong allomorph at a rate somewhere between ten and twenty percent (pp. 479, 483). Moreover, during training, the number of suffix errors is highly correlated with the number of verbs newly encountered; from their Table 3, we compute a value of $r = 0.908$. It would appear, then, that Plunkett and Juola's model could not reliably pick the right suffix allomorph unless it had been fully trained on the verb in question.

(c) The analogical model systematically overrated forms on the basis of similarity to individual verbs.

From this we infer that analogy, in its most basic form, is too powerful a mechanism to account for how morphological systems in human languages work, and that a multiple-rule approach is a more accurate model of how speakers create novel forms.

5.2. 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, 2001; Bybee & Moder, 1983; Bybee & Slobin, 1982) 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 nine 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 involving 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 Section 4.3.3 (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 Myers (1999), Russell (1999), MacBride (2001), and Burzio (2002).

5.3. Implications for 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 guided 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 (a claim for which evidence is accumulating[21]), 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.[22]

The second claim of the dual mechanism theory is that the grammar is simple, and contains only extremely general rules for regular patterns. We argue here against this second claim. We are by no means the first to contest this assertion; many critics have taken exception to this aspect of the model, noting that morphological processes frequently involve multiple, competing subgeneralizations (Bybee, 2001; Dressler, 1999; Indefrey, 1999; Wiese, 1999; Wunderlich, 1999). Furthermore, our finding of island of reliability effects in English mirrors a similar finding for Italian by Albright (2002). The current study brings two new findings to the debate: (1) that even an extremely regular process like English past tense formation can display subtle contextual effects that cannot be modeled with a single, general rule; and (2) that the productivity of irregular subgeneralizations is best modeled by stochastic rules, rather than by analogy. Taken together, these findings support a model in which learners posit rules that attempt to capture generalizations about all morphological processes, not just the largest or most productive ones.

5.4. 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 specific environments that differentiate the degrees of confidence for the regular outcome.

[21] Recent work in this area includes Schreuder, de Jong, Krott, and Baayen (1999), Sereno, Zwitserlood, and Jongman (1999), Hare, Ford, and Marslen-Wilson (2001) and Baayen, Schreuder, de Jong, and Krott (in press). For surveys of earlier work, see also Bybee (2001, pp. 111–113).

[22] For this reason, we do not address the neurolinguistic or pathological evidence that has been brought to bear on the dual mechanism model, except to note that many of these results would be compatible with a multiple rule based model. For example, Ullman et al. (1997) found that Alzheimer's patients perform poorly on irregulars (60% correct for the five most anomic patients), but better on regulars (89%) and on wug verbs (84%). With Ullman et al., we attribute this to the fact that the grammar of English (whether it consists of one rule or many) virtually always prefers regulars. Thus, regulars can be derived by the grammar, but irregulars must normally be retrieved from memory, which Alzheimer's patients fail to do consistently. Conversely, Parkinson's patients, posited to have impaired rules, did well on irregulars (88% for the five most hypokinetic patients), as we would expect if memory is unimpaired. Moreover, they retained a modest capacity for producing regulars (80%), which is to be expected if regulars are often memorized. They did poorly on wug verbs (65%), where rule application is the only possibility.

On the other hand, we share with the mainstream tradition of generative linguistics 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 & 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 generative linguistics has held: when they are discovered by an inductive learning algorithm, rules are the appropriate means of expressing both macro- *and* micro-generalizations.

Acknowledgements

This research was supported by NSF grant BCS-9910686 and by an NSF Graduate Fellowship award to Adam Albright. We would like to thank Marco Baroni, James Myers, Colin Wilson, and two *Cognition* reviewers for helpful comments on an earlier version of this article. We also gratefully acknowledge the help of Stephen Wilson in carrying out the experiments reported here, as well as all of our experimental participants.

Appendix A. Phonological ratings, past tense scores, and model predictions

Table A1[23]
Core verbs

	Stem	Stem rating	Past	Experiment 1 production probability	Experiment 2 production probability	Overall production probability	Mean rating	Adjusted mean rating	Rule-based model predicted	Analogical model predicted
Island of reliability for both regulars and irregulars:										
1.	*bize*	4.57	*bized*	0.778	0.571	0.667	5.30	5.32	6.06	5.87
			boze	0.056	0.381	0.231	4.57	4.55	4.11	4.04
2.	*dize*	4.62	*dized*	0.889	0.762	0.821	5.42	5.42	6.06	5.95
			doze	0.111	0.190	0.154	5.04	5.04	4.73	4.18
3.	*drice*	3.86	*driced*	1.000	0.913	0.953	6.26	6.52	6.22	5.51
			droce	0.000	0.087	0.047	4.48	4.31	5.15	4.28

(*continued on next page*)

Table A1[23] (*continued*)

	Stem	Stem rating	Past	Experiment 1 production probability	Experiment 2 production probability	Overall production probability	Mean rating	Adjusted mean rating	Rule-based model predicted	Analogical model predicted
4.	*flidge*	4.05	*flidged*	0.947	0.783	0.857	6.21	6.41	6.16	5.46
			fludge	0.000	0.043	0.024	4.88	4.76	4.22	4.10
5.	*fro*	5.84	*froed*	0.950	0.833	0.886	5.83	5.50	5.40	6.16
			frew	0.050	0.125	0.091	4.33	4.57	4.97	4.38
6.	*gare*	5.24	*gared*	1.000	0.955	0.976	6.57	6.44	6.02	6.27
			gore	0.000	0.000	0.000	3.39	3.49	4.30	4.23
7.	*glip*	4.95	*glipped*	1.000	0.857	0.925	5.95	5.88	6.07	5.80
			glup	0.000	0.048	0.025	3.45	3.50	4.02	3.97
8.	*rife*	5.61	*rifed*	0.950	0.762	0.854	5.95	5.69	6.22	5.07
			rofe	0.000	0.190	0.098	4.14	4.33	4.61	4.35
			riff	0.000	0.000	0.000	3.24	3.42	3.94	3.90
9.	*stin*	5.40	*stinned*	0.900	0.522	0.698	5.30	5.08	5.83	6.02
			stun	0.100	0.261	0.186	4.78	4.94	4.34	4.63
			stan	0.000	0.000	0.000	2.74	2.87	4.27	4.03
10.	*stip*	5.45	*stipped*	1.000	0.708	0.841	5.92	5.70	6.07	5.88
			stup	0.000	0.083	0.045	4.50	4.66	4.15	4.26
Island of reliability for regulars only:										
11.	*blafe*	3.57	*blafed*	1.000	0.818	0.892	6.32	6.67	6.22	5.15
			bleft	0.000	0.045	0.027	4.09	3.86	3.94	3.85
12.	*bredge*	3.86	*bredged*	0.950	0.905	0.927	6.33	6.60	6.16	5.85
			broge	0.050	0.048	0.049	3.43	3.25	3.94	3.85
13.	*chool*	3.76	*chooled*	1.000	0.957	0.977	6.13	6.41	6.12	6.38
			chole	0.000	0.043	0.023	3.71	3.51	3.94	4.05
14.	*dape*	5.14	*daped*	1.000	0.957	0.976	6.25	6.14	6.14	5.56
			dapt	0.000	0.000	0.000	4.00	4.09	3.94	3.85
15.	*gezz*	4.19	*gezzed*	1.000	0.955	0.976	6.61	6.79	6.06	5.89
			gozz	0.000	0.000	0.000	2.52	2.40	3.94	3.95
16.	*nace*	5.00	*naced*	1.000	1.000	1.000	6.57	6.50	6.22	5.57
			noce	0.000	0.000	0.000	2.91	2.96	4.00	3.89
17.	*spack*	5.05	*spacked*	1.000	0.739	0.860	6.22	6.13	6.01	5.79
			spuck	0.000	0.130	0.070	3.96	4.03	3.94	3.85
18.	*stire*	5.62	*stired*	1.000	0.818	0.902	6.00	5.74	6.02	6.29
			store	0.000	0.091	0.049	3.22	3.40	3.94	4.03
19.	*tesh*	4.71	*teshed*	1.000	0.870	0.925	6.22	6.23	6.22	5.59
			tosh	0.000	0.000	0.000	3.13	3.12	3.94	3.88
20.	*wiss*	5.76	*wissed*	0.950	0.952	0.951	6.57	6.28	6.22	5.68
			wus	0.000	0.048	0.024	3.35	3.56	3.94	3.99
Island of reliability for irregulars only:										
21.	*blig*	3.71	*bligged*	0.941	0.652	0.775	5.67	5.95	5.66	5.44
			blug	0.000	0.130	0.075	4.17	3.97	5.19	4.08
22.	*chake*	5.33	*chaked*	0.950	0.818	0.881	5.74	5.55	4.77	5.65
			chook	0.000	0.000	0.000	5.04	5.19	5.13	4.17
23.	*drit*	4.30	*dritted*	0.842	0.591	0.707	4.96	5.04	5.43	5.29
			drit	0.053	0.091	0.073	5.13	5.07	4.62	4.11
			drat	0.000	0.182	0.098	3.65	3.57	4.06	3.99
24.	*fleep*	4.24	*fleeped*	1.000	0.478	0.721	5.00	5.10	5.69	5.56
			flept	0.000	0.435	0.233	6.09	6.02	5.15	4.40

(*continued on next page*)

Table A1[23] (*continued*)

	Stem	Stem rating	Past	Experiment 1 production probability	Experiment 2 production probability	Overall production probability	Mean rating	Adjusted mean rating	Rule-based model predicted	Analogical model predicted
25.	*gleed*	5.29	*gleeded*	0.684	0.455	0.561	4.22	3.98	4.36	5.15
			gled	0.158	0.318	0.244	6.00	6.15	5.07	4.53
			gleed	0.105	0.227	0.171	4.09	4.21	3.94	3.99
26.	*glit*	5.25	*glitted*	0.778	0.542	0.643	5.00	4.80	5.43	5.37
			glit	0.167	0.125	0.143	5.21	5.34	4.89	4.32
			glat	0.000	0.167	0.095	3.75	3.86	4.06	3.91
27.	*plim*	4.43	*plimmed*	0.950	0.682	0.810	6.13	6.22	5.74	5.96
			plum	0.000	0.136	0.071	4.17	4.12	4.52	4.10
			plam	0.000	0.045	0.024	3.57	3.51	4.21	3.92
28.	*queed*	3.81	*queeded*	0.700	0.364	0.524	4.65	4.86	4.36	5.10
			qued	0.100	0.318	0.214	5.35	5.19	4.43	4.09
29.	*scride*	4.05	*scrided*	0.556	0.292	0.405	4.17	4.30	4.58	4.89
			scrode	0.111	0.250	0.190	4.39	4.26	4.98	4.73
			scrid	0.000	0.042	0.024	3.57	3.43	4.12	3.95
30.	*spling*	4.56	*splinged*	0.667	0.368	0.514	4.36	4.34	5.14	5.35
			splung	0.222	0.421	0.324	5.45	5.45	5.19	5.42
			splang	0.056	0.158	0.108	4.50	4.48	4.36	4.54
Island of reliability for neither regulars nor irregulars:										
31.	*gude*	4.25	*guded*	0.625	0.500	0.556	4.90	4.99	6.07	5.26
			gude	0.375	0.300	0.333	5.55	5.48	3.96	3.99
32.	*nold*	4.10	*nolded*	0.833	0.273	0.525	4.64	4.76	4.78	5.54
			nold	0.167	0.500	0.350	6.05	5.95	3.96	3.91
			neld	0.000	0.182	0.100	5.14	5.03	3.94	4.10
33.	*nung*	3.21	*nunged*	0.933	0.737	0.824	5.37	5.78	5.14	5.97
			nang	0.000	0.105	0.059	4.32	4.02	3.94	3.89
34.	*pank*	5.62	*panked*	1.000	0.810	0.900	6.30	6.05	5.62	5.92
			punk	0.000	0.143	0.075	4.00	4.19	3.94	3.89
35.	*preak*	4.90	*preaked*	0.900	0.792	0.841	5.83	5.77	5.37	5.80
			proke	0.100	0.167	0.136	3.92	3.96	3.98	3.93
			preck	0.000	0.000	0.000	3.54	3.58	3.94	3.98
36.	*rask*	5.30	*rasked*	1.000	0.870	0.930	6.42	6.26	5.97	6.11
			rusk	0.000	0.043	0.023	4.08	4.21	3.94	3.85
37.	*shilk*	4.60	*shilked*	1.000	0.950	0.975	5.79	5.82	5.97	5.17
			shalk	0.000	0.000	0.000	3.67	3.64	3.94	4.13
38.	*tark*	5.10	*tarked*	1.000	0.870	0.930	6.33	6.24	5.97	5.66
			tork	0.000	0.043	0.023	3.71	3.79	3.94	3.85
39.	*teep*	4.95	*teeped*	1.000	0.783	0.884	5.91	5.84	5.70	5.61
			tept	0.000	0.087	0.047	4.70	4.76	4.73	4.20
40.	*trisk*	5.14	*trisked*	1.000	0.789	0.897	6.29	6.17	5.97	6.05
			trask	0.000	0.105	0.051	3.76	3.85	3.94	4.01
			trusk	0.000	0.053	0.026	3.62	3.71	3.94	3.94
41.	*tunk*	4.65	*tunked*	1.000	0.826	0.907	5.67	5.67	5.62	5.80
			tank	0.000	0.087	0.047	3.92	3.91	3.94	3.86

[23] International Phonetic Alphabet (IPA) transcriptions for verbs whose spelling could be ambiguous: *broge* [brodʒ], *chook* [tʃʊk], *drat* [dræt], *glat* [glæt], *gozz* [gaz], *gude* [gud], *nang* [næŋ], *rask* [ræsk], *shalk* [ʃælk], *stan* [stæn], *stup* [stʌp], *trask* [træsk], *wiss* [wɪs], *wus* [wʌs].

Table A2
Peripheral verbs

	Stem	Stem rating	Past	Experiment 1 production probability	Experiment 2 production probability	Overall production probability	Mean rating	Adjusted mean rating	Rule-based model predicted	Analogical model predicted
1.	*grell*	4.52	*grelled*	1.000	0.810	0.902	5.86	5.91	5.86	6.17
			grelt	0.000	0.143	0.073	4.90	4.88	4.06	5.39
2.	*murn*	5.43	*murned*	0.947	0.957	0.952	6.57	6.38	6.02	6.29
			murnt	0.053	0.043	0.048	4.74	4.90	4.60	5.23
3.	*scoil*	3.84	*scoiled*	0.944	0.947	0.946	6.45	6.72	5.93	6.51
			scoilt	0.000	0.000	0.000	5.15	4.99	4.01	4.91
4.	*shurn*	5.00	*shurned*	0.947	0.857	0.900	6.57	6.50	6.02	6.31
			shurnt	0.000	0.000	0.000	4.22	4.28	3.95	5.11
5.	*skell*	5.05	*skelled*	0.944	0.682	0.800	6.05	5.95	5.86	6.24
			skelt	0.000	0.227	0.125	5.32	5.41	4.06	5.23
6.	*snell*	5.38	*snelled*	0.947	0.826	0.881	6.17	5.99	5.86	6.18
			snelt	0.000	0.130	0.071	5.30	5.46	4.06	5.25
			snold	0.000	0.000	0.000	2.83	2.95	3.94	4.17
7.	*squill*	4.67	*squilled*	0.895	0.810	0.850	5.92	5.93	5.86	6.30
			squilt	0.000	0.048	0.025	5.21	5.22	4.06	5.20
8.	*kive*	4.38	*kived*	0.950	0.864	0.905	6.00	6.10	5.58	5.87
			kave	0.000	0.091	0.048	4.41	4.35	3.94	4.12
9.	*lum*	4.81	*lummed*	1.000	0.826	0.905	6.35	6.33	5.74	6.14
			lame	0.000	0.000	0.000	2.87	2.88	3.94	3.93
10.	*pum*	4.81	*pummed*	1.000	0.826	0.902	6.17	6.15	5.74	6.06
			pame	0.000	0.000	0.000	2.71	2.71	3.94	4.01
11.	*shee*	5.95	*sheed*	1.000	0.875	0.929	6.17	5.81	5.94	5.89
			shaw	0.000	0.000	0.000	3.25	3.50	3.94	4.18
12.	*zay*	4.14	*zayed*	0.950	1.000	0.977	6.39	6.57	6.13	5.99
			zed	0.000	0.000	0.000	4.39	4.28	3.94	4.05
13.	*chind*	4.62	*chinded*	0.235	0.368	0.306	3.89	3.84	4.36	5.41
			chound	0.000	0.000	0.000	4.05	4.04	4.83	4.45
14.	*flet*	4.24	*fletted*	0.889	0.368	0.622	4.50	4.58	5.43	5.39
			flet	0.111	0.474	0.297	5.65	5.58	5.22	4.30
15.	*gry'nt*	5.16	*gry'nted*	0.842	0.545	0.683	5.26	5.10	6.20	5.59
			groant	0.000	0.091	0.049	4.17	4.27	3.94	4.07
			grount	0.053	0.000	0.024	3.83	3.92	3.94	4.67
16.	*ry'nt*	3.00	*ry'nted*	0.778	0.250	0.476	5.00	5.46	6.20	5.51
			roant	0.056	0.292	0.190	4.29	3.95	3.94	4.32
			rount	0.000	0.042	0.024	3.71	3.36	3.94	4.13
17.	*shy'nt*	3.52	*shy'nted*	0.800	0.364	0.571	5.17	5.49	6.20	5.49
			shoant	0.000	0.136	0.071	3.83	3.58	3.94	4.23
			shount	0.000	0.045	0.024	3.39	3.14	3.94	4.02

References

Albright, A. (2002). Islands of reliability for regular morphology: evidence from Italian. *Language*, *78*, 684–709.

Albright, A., & Hayes, B. (2002). Modeling English past tense intuitions with Minimal Generalization. In M. Maxwell (Ed.), *Proceedings of the 6th meeting of the ACL Special Interest Group in Computational Phonology*, New Brunswick, NJ; Association for Computational Linguistics.

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

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2)*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Baayen, H., Schreuder, R., de Jong, N., & Krott, A. (in press). Dutch inflection: the rules that prove the exception. In S. Nooteboom, F. Weerman, & F. Wijnen (Eds.), *Storage and computation in the language faculty*. Dordrecht: Kluwer.

Bayardo, R., Agrawal, R., & Gunopulos, D. (1999). Constraint-based rule mining in large, dense databases (pp. 188–197). *Proceedings of the 15th International Conference on Data Engineering*, Los Alamitos, CA: IEEE Computer Society Press.

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

Bird, S. (1995). *Computational phonology – a constraint-based approach*. Cambridge: Cambridge University Press.

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

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

Burzio, L. (2002). Missing players: phonology and the past-tense debate. *Lingua*, *112*, 157–199.

Bybee, J. (1985). *Morphology: a study of the relation between form and meaning*. Amsterdam: John Benjamins.

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

Bybee, J. (2001). *Phonology and language use*. Cambridge: 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.

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.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2002). *TiMBL: Tilburg Memory-Based Learner, version 4.3 reference guide*. Tilburg: ILK.

Daugherty, K., & Seidenberg, M. (1994). Beyond rules and exceptions: a connectionist modeling approach to inflectional morphology. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules*. Amsterdam: John Benjamins, pp. 353–388.

Derwing, B., & Skousen, R. (1994). Productivity and the English past tense: testing Skousen's analogy model. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 193–218). Amsterdam: John Benjamins.

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.

Frisch, S. (1996). *Similarity and frequency in phonology*. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

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

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

Halle, M., & Mohanan, K. P. (1985). Segmental phonology of modern English. *Linguistic Inquiry*, *16*, 57–116.

Hare, M., Ford, M., & Marslen-Wilson, W. (2001). Ambiguity and frequency effects in regular inflection. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

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: Cambridge University Press.

Hoard, J., & Sloat, C. (1973). English irregular verbs. *Language*, *49*, 107–120.

Howe, N., & Cardie, C. (1997). Examining locally varying weights for nearest neighbor algorithms. In D. Leake, & E. Plaza (Eds.), *Case-based reasoning research and development: Second International Conference on Case-Based Reasoning* (pp. 455–466). Heidelberg: Springer Verlag.

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

Kruskal, J. B. (1999). 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* (2nd ed.). Reading, MA: Addison-Wesley.

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, Los Angeles, CA.

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.

Michalski, R. (1983). A theory and methodology of inductive learning. *Artificial Intelligence*, *20*, 111–161.

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

Myers, J. (1999). Lexical phonology and the lexicon. Retrieved November 20, 2001 from http://roa.rutgers.edu/files/330-0699/roa-330-Myers-3.pdf.

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. (2001). Stochastic phonology. *GLOT*, *5*(6), 1–13.

Pinker, S. (1999a). *Words and rules: the ingredients of language*. New York: Basic Books.

Pinker, S. (1999b). Regular habits. *Times Literary Supplement*, Oct. 29, 1999.

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. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules*. Amsterdam: John Benjamins.

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*, *23*, 463–490.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21–69.

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). New Brunswick, NJ: 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.), (*2*) (pp. 216–271). *Parallel distributed processing: explorations in the microstructure of cognition*, Cambridge, MA: MIT Press.

Russell, K. (1999). *MOT: sketch of an OT approach to morphology*. Unpublished manuscript, University of Manitoba, Winnipeg. 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.

Skousen, R. (2001). *Analogy and Structure*. Dordrecht: Kluwer Academic.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, M., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: lexicon a part of declarative memory, grammar processed by procedural system. *Journal of Cognitive Neuroscience*, *9*, 266–276.

Weiss, S., & Kulikowski, C. (1991). *Computer systems that learn*. San Mateo, CA: Morgan Kaufmann.

Westermann, G. (1997). A constructivist neural network learns the past tense of English verbs. *Proceedings of the GALA '97 conference on language acquisition*, Edinburgh: HCRC.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, *11*, 273–314.

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.

Zavrel, J., & Daelemans, W. (1997). Memory-based learning: using similarity for smoothing (pp. 436–443). *Proceedings of the 35th annual meeting of the ACL*, New Brunswick, NJ: ACL.