Cognitive Science 29 (2005) 627–654

Dissociations in Performance on Novel Versus Irregular Items: Single-Route Demonstrations With Input Gain in Localist and Distributed Models

Christopher T. Kello[a], Daragh E. Sibley[a], David C. Plaut[b]

[a]*Department of Psychology, George Mason University*
[b]*Department of Psychology, Carnegie Mellon University*

Received 6 May 2004; received in revised form 4 October 2004; accepted 3 November 2004

Abstract

Four pairs of connectionist simulations are presented in which quasi-regular mappings are computed using localist and distributed representations. In each simulation, a control parameter termed *input gain* was modulated over the only level of representation that mapped inputs to outputs. Input gain caused both localist and distributed models to shift between regularity-based and item-based modes of processing. Performance on irregular items was selectively impaired in the regularity-based modes, whereas performance on novel items was selectively impaired in the item-based modes. Thus, the models exhibited double dissociations without separable processing components. These results are discussed in the context of analogous dissociations found in language domains such as word reading and inflectional morphology.

Keywords: Connectionist models; Localist and distributed representations; Double dissociations; Word reading; Inflectional morphology; Dyslexia; Input gain; Control parameters

1. Introduction

Many domains of cognition have *quasi-regular* structures in their representations (Plaut, McClelland, Seidenberg, & Patterson, 1996). The structuring of natural categories such as "game" and "bird" are well-known examples in that they are somewhat defined by regularities, yet the existence of exceptions to those regularities cannot be denied (Wittgenstein, 1953). Quasi regularity can also be found in domains such as problem solving (Sloman, 1996), reasoning (Anderson, Fincham, & Douglass, 1997), and skill acquisition (Medin & Ross, 1989). It has played a particularly strong role in research on language processing (Pinker, 1999).

Requests for reprints should be sent to Christopher T. Kello, Department of Psychology, George Mason University, Fairfax, VA 22030. E-mail: ckello@gmu.edu

Two well-known examples of quasi regularity in the English language are the relation between spelling and sound, and the past-tense formation of verbs. Every grapheme has a tendency to correspond to one given sound (e.g., *S* is usually /s/), but each tendency has its exceptions (e.g., SURE). Some verbs have typical past-tense formations (e.g., STAY–STAYED), but others do not (e.g., SAY–SAID). Other examples include pluralization in English (Haskell, MacDonald, & Seidenberg, 2003) and Hebrew (Berent, Pinker, & Shimron, 2002), and past-participle formation in German (Beretta, Carr, Huang, & Cao, 2003).

Quasi regularity has driven many language researchers to propose that there are two separate systems of processing, one to handle regularities and another to handle exceptions. This separation has drawn support from a number of findings, but the strongest evidence has come from selective deficits in the processing of regularities versus exceptions. These selective and complimentary deficits constitute *double dissociations,* which are often thought to arise from separable processing components (but see Plaut, 1995; Shallice, 1988; Van Orden, Pennington, & Stone, 2001).

In this study, we challenge the widespread assumption that selective deficits in the processing of regularities and exceptions entail a corresponding division in the language system. Two types of connectionist models, one using *localist* codes and the other using *distributed* codes, are presented in which a quasi-regular mapping from inputs to outputs is computed. A parameter termed *input gain* is shown to transition both types of model between *regularity-based* or *item-based* modes of processing (Kello, 2003). At high levels of input gain, performance on novel items was selectively impaired. At low levels of input gain, performance on exception items was selectively impaired. Neither the localist nor distributed models contained an architectural division between regularity-based and item-based processing, or any other architectural division that could have contributed to the simulated double dissociation.

The simulations did not account for any particular set of empirical results, nor were they meant to. Instead, they demonstrate a novel and general way that double dissociations can occur without separate system components. The simulations also provide the groundwork for future research to determine whether dissociations between regularity-based and item-based processing in brain and behavior can emerge from aberrant changes in control parameters.

1.1. Dual-route theories

A regularity-based process is governed by the regularities that span across items in a given linguistic domain. An item-based process is governed by information that is specific to individual items in the domain. Dual-route theories are designed to leverage the complementary strengths of regularity-based and item-based processes. In the domain of word reading, the most prominent dual-route theory has been implemented as the dual-route cascaded (DRC) model (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). The DRC model contains a system of grapheme-to-phoneme correspondence rules that capture regularities between spellings and sounds of words, and a system of lexical knowledge to capture word-specific information (the lexical system is composed of both semantic and lexical representations). Words are processed by running these two systems in parallel and combining their outputs at an integration stage. The model has been built with a vocabulary of over

7,500 monosyllabic words in English, and it has been applied to a wide range of results from naming and lexical-decision experiments.

In the domain of inflectional morphology, Pinker (1999) argued that a set of rules exists to generate the inflected forms of words by means of combining stems and affixes, and a separate lexicon exists to store irregularly inflected forms that are not handled properly by the rules (see also Clahsen, 1999). This *words-and-rules* theory has been applied primarily in the domain of English past-tense formation, but it has also been applied to tense formation in other languages, as well as to pluralization. The theory has not been made explicit in a computational model, but the rules have been associated with neural circuits implicated in procedural processing, and the lexicon has been associated with neural circuits implicated in declarative processing (Pinker & Ullman, 2002; Ullman, 2001).

The DRC model and the words-and-rules theory are similar in that they both propose a set of symbolic rules to capture the regularities in a quasi-regular domain, and a lexicon to handle exceptions to those regularities. By contrast, a multiple-route theory was outlined by Seidenberg and McClelland (1989; hereafter referred to as SM89). They proposed that word reading can be theorized in terms of activation patterns that span semantic, orthographic, and phonological representations and extend into other mediating and modulating levels of representation (e.g., representations of context). These patterns are learned and processed by a common set of connectionist mechanisms. Thus, the theoretical components are distinguished by the kinds of information that they represent, rather than the kinds of processing mechanisms that subserve them.

The SM89 theory (Seidenberg & McClelland, 1989) proposed two routes or pathways that contribute to pronouncing written words. One is the relatively direct route from orthography to phonology, computed via hidden units, and the other is an indirect route mediated by semantics. The indirect, *semantic* route is primarily item based because it is shaped by semantic knowledge that is specific to individual words and because the semantic structures of words are mostly unrelated to their orthographic and phonological forms (at least at the level of the morpheme). The direct, *phonological* route is primarily regularity-based because it is shaped by the systematic, sublexical regularities that exist between the orthographic and phonological forms of words in a language such as English.

The semantic and phonological routes in the SM89 theory (Seidenberg & McClelland, 1989) bear some resemblance to the words and rules of Pinker's (1999) theory and to the rule and lexical routes of the DRC model. That said, there are some important differences. Most relevant to our work, the connectionist basis of the SM89 theory means that gradations of regularity can be represented in either processing route, both in terms of scale (e.g., regularities at the level of the letter, grapheme, or larger groups of letters) and consistency (e.g., regularities that hold for most or all words, or only for some smaller subset of words). By contrast, the rules proposed in the DRC model and in Pinker's theory were designed to capture only a single level of regularity (McClelland & Patterson, 2002b). Gradations of regularity provide some of the motivation for single-route alternatives to dual-route theories, as discussed next.

1.2. Single-route theories

The dual-route approach to language processing is appealing for the reasons already discussed (among others), but it has its disadvantages as well. Perhaps the most basic of these dis-

advantages is that quasi-regular domains are rarely characterized by the simple dichotomy of regularities and exceptions. Instead, studies have shown that quasi-regular domains often contain gradations of regularity, from fully systematic (regular) to fully idiosyncratic (irregular) forms (e.g., Bybee, 2001).

The relation between spelling and sound in English is a prime example. Each vowel grapheme has a vowel sound that it corresponds to most often, but for many of these graphemes, there are multiple exceptions with varying degrees of irregularity. For instance, the grapheme OU corresponds most often to the diphthong /aU/ (as in OUT and LOUD). However, it also corresponds to the reduced schwa in some derivational suffixes (as in RIGHTEOUS and CONSCIOUS), and the vowel /U/ in a handful of other cases (as in GHOUL, SOUP, GROUP, and THROUGH). Still other correspondences are even more exceptional (as in ROUGH, TOUGH, SOUL, THOUGH, THOUGHT, and OUGHT).

Graded regularities, as in the OU example just given, are suggestive of a language system designed to capture the full spectrum of relations that might exist in a given domain. Some have argued that the dichotomy proposed in dual-route accounts such as the DRC model (Coltheart et al., 1993, 2001) and the words-and-rules theory (Pinker, 1999) is too discrete given the graded nature of quasi regularity (Rumelhart, Hinton, & McClelland, 1986). These theories are forced to treat at least some graded regularities as completely idiosyncratic, which prohibits the theories from capturing their graded structure. This criticism is less applicable to the SM89 theory (Seidenberg & McClelland, 1989) because, as mentioned earlier, connectionist representations are well suited for capturing gradations of regularity. Nonetheless, one must ask whether the unified design of a single-route architecture is more apt for capturing the spectrum of relations in a quasi-regular domain.

Some single-route theories have handled graded regularities by means of similarity-based processing. In localist single-route theories, linguistic items are stored as individual elements or nodes, with item features linked to each node. Linguistic inputs are processed on the basis of their featural similarity to stored items. Regularities are captured in the consistency of featural mappings among stored items. A regularity is strong when many items share a given featural mapping, and weaker when fewer items share the mapping. An irregularity occurs when the featural mapping for one item contrasts with a featural mapping that is shared by other, similar items. Localist single-route theories that employ similarity-based processing have been proposed in the domain of word reading (Glushko, 1979; Kay & Marcel, 1981; Morton, 1969; Taraban & McClelland, 1987), as well as inflectional morphology (Skousen, 1989).

Similarity-based processing has also been employed by distributed single-route theories. Rumelhart et al. (1986) proposed that a single set of learned, distributed associations could capture the sound patterning between present- and past-tense verb formations in English. Although their specific implementation was roundly criticized (Lachter & Bever, 1988; Pinker & Prince, 1988), their work has played a central role in the ongoing debate between connectionist and symbolic accounts of language processing (McClelland & Patterson, 2002a; Pinker & Ullman, 2002). Joanisse and Seidenberg (1999) contributed a connectionist model of past-tense formation to this debate (hereafter referred to as the JS99 model), and we use their model here to demonstrate the distributed approach because it is particularly relevant to our work.

In the JS99 model (Joanisse & Seidenberg, 1999), processes of speech comprehension and production were abstracted as phonological inputs and outputs, respectively. Comprehension

was linked to production via one internal level of representation. This internal level was also linked to a level of localist representation that served as a proxy for semantics. Internal representations consisted of patterns of activation distributed across 100 hidden units, and these patterns were learned via the back-propagation of error that was generated on the output units. Error came from four language tasks given to the model: speech production (mapping from semantics to phonological outputs), speech comprehension (mapping from phonological inputs to semantics), speech imitation (mapping phonological inputs to phonological outputs), and past-tense formation (mapping present-tense phonological inputs to past-tense phonological outputs). The last task forced internal representations to capture the quasi-regular relation between present and past-tense formations.

The JS99 model (Joanisse & Seidenberg, 1999) was an implementation of a single-route theory in that regular, irregular, and novel verb forms were all mapped through a single level of representation. This single-route model was based on the following principles. Input and output units were designed to represent phonological components of the present- and past-tense forms of verbs, respectively. The internal representations captured any consistent relations between the input and output units by virtue of the way that distributed representations are learned via back-propagation. Given that regularities in English past-tense formations are carried by the phonological components of words (e.g., verbs ending in /-t/ and /-d/ usually take the /-ld/ suffix to form their past tense), the internal representations were driven to capture these regularities. The same representations also had to capture the exceptions to those regularities that come from irregular past-tense forms (e.g., BE, GO, HAVE).

Irregular forms were processed by learning to associate certain *conjunctions* of features on the input units with irregular patterns on the output units. For instance, it is the conjunction of the letters *R* and *U* with the ending letter *N* that indicates the irregular past tense RAN instead of RUNNED. By contrast, regular forms were processed by *componential* relations that were learned between inputs and outputs. For instance, the ending letter *N* is related to the ending sound /-d/ for the regular past tense. The hidden representations were able to process both componential and conjunctive relations by virtue of nonlinearities in the hidden unit activation function (see O'Reilly, 2001). This property of the JS99 model (Joanisse & Seidenberg, 1999) played an important role in our work, and it shall be revisited later.

1.3. Double dissociations between regularity-based and item-based processing

A critical source of support for dual-route theories comes from observed double dissociations in regularity-based and item-based processing among brain-damaged patients. In the domain of word reading, this corresponds to the distinction between phonological and surface dyslexia. For instance, Funnell (1983) reported on a phonological dyslexic patient WB for whom the ability to read nonwords (even simple consonant–vowel–consonant nonwords) was greatly impaired, whereas the ability to read both easy and difficult words was mostly intact. By contrast, Behrmann and Bub (1992) reported on a surface dyslexic patient MP for whom the ability to read irregular words (particularly of low frequency) was greatly impaired, whereas the ability to read both regular words and nonwords was mostly intact. The deficits of patients WB and MP (as well as those of other patients) have been simulated in the DRC model

(Coltheart et al., 2001) by damaging the regularity-based and item-based components, respectively.

Analogous dissociations have been argued to arise in the domain of English inflectional morphology. Ullman and his colleagues (Ullman, Corkin, Coppola, Hickok, & et al., 1997) found that Alzheimer's patients, as well as aphasics with posterior lesions, were poor at generating the past tense of verbs with irregular inflections, but relatively normal with regular inflections. They reported the opposite pattern for Parkinson's patients and aphasics with anterior lesions. Marslen-Wilson and Tyler (1997, 1998) reported a similar dissociation through a priming paradigm. For two patients with acquired language deficits, target responses were primed by the regular past tense of the target verb (e.g., WALKED–WALK), but not by the irregular past tense (e.g., FOUND–FIND). Two other patients exhibited the opposite pattern of priming.

These dissociations in word reading and past-tense formation are challenging for single-route theories because it is not obvious how regularity-based and item-based processes could be dissociated in a system that does not have them separated. In some cases, proponents of single-route theories have argued that the dissociations are not as clear and prevalent as some believe (Bird, Ralph, Seidenberg, McClelland, & Patterson, 2003; Patterson & Hodges, 1992; Patterson, Lambon Ralph, Hodges, & McClelland, 2001; Patterson & Marcel, 1992). In other cases, they have argued that a division between phonological and semantic processes can account for the dissociations (Joanisse & Seidenberg, 1999). Phonological processes are argued to support regularity-based processing, and semantic processes are argued to support item-based processing. On this basis, damage to phonological or semantic processes would impair regularity-based or item-based processing, respectively.

The connection between phonological and regularity-based processing has been drawn because phonological representations must be componential to support regularity-based processing (Patterson & Marcel, 1992). If this componential structure is compromised when phonological processes are damaged, then phonological deficits would lead to deficits in regularity-based processes. The connection between semantic and item-based processing has been drawn because irregular forms require support from semantic representations to counteract the influence of learned regularities (Patterson & Hodges, 1992). If this support is compromised when semantic processes are damaged, then semantic deficits would lead to deficits in item-based processes.

The phonological–semantic division predicts that phonological and regularity-based deficits should coincide, as should semantic and item-based deficits. A number of findings have been reported that are consistent with these predictions (for overviews, see Bird et al., 2003; Patterson & Lambon Ralph, 1999). These findings are inconsistent with dual-route theories, because these theories posit semantic and phonological components of processing that are separate from regularity-based and item-based components (but for criticisms see Coltheart, 1996; Coltheart et al., 2001).

1.4. A single-process account of surface and phonological dyslexia

The phonological–semantic division may account for some instances of selective deficits in regularity-based or item-based processing, but it may not account for all of them (see Coltheart

et al., 2001). Motivated by this uncertainty, Kello (2003) analyzed a single-route model of word reading to investigate whether surface and phonological dyslexia could be simulated without damage to separable processing components such as semantics and phonology (see Figure 1). In that model, there was a bidirectional mapping via hidden units between semantics and phonology, much like in the JS99 model (Joanisse & Seidenberg, 1999). Orthography was mapped via an "integrated" pathway directly onto the hidden units between semantics and phonology. Orthography was *not* mapped separately onto semantics and phonology, as it is in the SM89 theory (Seidenberg & McClelland, 1989).

The main theoretical principle behind this architecture is the following: The representations that mediate semantics and phonology, learned during spoken language acquisition, should provide an apt interface between the written and spoken language systems (Kello & Plaut, 2003). In reading, one must access semantic and phonological information about words, information that is learned primarily through spoken language. This information can be accessed via one integrated pathway of processing from orthography into the "junction" between semantics and phonology, rather than separate pathways into semantics and phonology.

The basic characteristics of surface and phonological dyslexia were simulated in this model by manipulating the input-gain parameter. Input gain is a multiplicative scalar on the net inputs to connectionist processing units. When unit activations are computed as a sigmoid of their net inputs (as they commonly are), input gain modulates the *sensitivity* of a unit's activation to its net input: At lower levels of input gain, larger magnitudes of net input are necessary to push a unit's output toward one of its two asymptotes (i.e., 0 or 1 for the logistic), and vice versa for higher levels of input gain. The idea of using input gain to simulate dyslexia was based on previous work in which input gain was used to simulate word-naming errors under time pressure (Kello & Plaut, 2003).

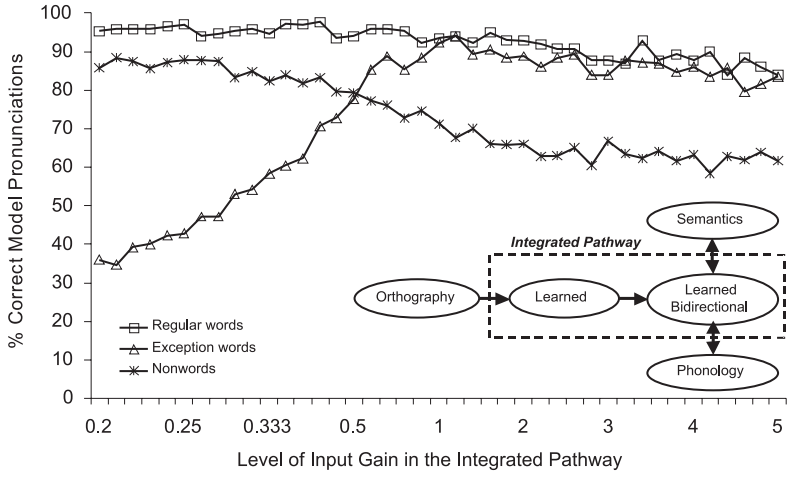

Fig. 1. Performance of the single-route model of word reading as a function of input gain and item type. The single route of processing from orthography into the spoken word mapping is outlined by the dashed rectangle. Input gain was manipulated on the units in these representations. Taken from Kello (2003).

Results from the dyslexia simulation showed that input gain can dissociate performance on irregular words from performance on nonwords. The model was trained at a normative input gain of 1 and was then tested under aberrantly low and high levels of input gain on the hidden units. Low levels caused a selective deficit in the naming of irregular words, whereas high levels caused a selective deficit in the naming of nonwords. Moreover, the same model was shown elsewhere (Kello & Plaut, 2003) to account for basic effects in skilled reading, such as the interaction of frequency and regularity in naming latencies.

2. Current simulations

Simulations with the single-route model of word reading demonstrate that input gain can dissociate regularity-based and item-based processing, even without separate regularity-based and item-based components of processing. However, the complexity and detail of that model makes it prohibitively difficult to understand how input gain had its dissociating effect. For instance, were distributed representations and recurrent connections necessary to cause the dissociation? Did word frequency and quasi regularity influence the dissociation?

In this work, simple connectionist models were built to elucidate the dissociating effect of input gain. The models computed simple quasi-regular mappings that were similar in design to those used in previous simulations of language phenomena, but the mappings were abstracted away from any particular domain. Two different types of models were investigated.

In the localist model, inputs were mapped onto outputs via a set of localist units that competed with each other for activation. Input gain modulated the amount of competition among the mediating units. The representations and processing mechanisms used in the localist model were similar in many ways to those used in other models of word reading (e.g., as in the lexical nonsemantic route in Coltheart et al., 2001) and inflectional morphology (Albright & Hayes, 2003; Hahn & Nakisa, 2000).

In the distributed model, by contrast, each input–output pairing activated multiple units, and each unit participated in multiple pairings. The patterns of activation across a set of hidden units were learned via back-propagation to mediate the quasi-regular mapping. Activations of the hidden units were computed as a sigmoid of their net inputs, and input gain was used to modulate the sensitivity of these units to changes in its net input. The representations and processing mechanisms used in the distributed model were similar in many ways to those used in other distributed models of word reading (Plaut et al., 1996; Seidenberg & McClelland, 1989) and inflectional morphology (Joanisse & Seidenberg, 1999; Rumelhart et al., 1986).

Four pairs of simulations are reported. The first two pairs of simulations were basic explorations of the effect of input gain on the localist and distributed model types. The second two pairs of simulations were designed to more closely mimic two basic aspects of quasi regularity in language. The four pairs of simulations, when taken together, provide a thorough analysis of how input gain can dissociate regularity-based and item-based processing in quasi-regular domains. They also reveal some of the similarities and differences of using input gain as a mechanism of dissociation in models with localist versus distributed coding schemes.

2.1. Simulations 1a and 1b

In Simulations 1a and 1b, feed-forward versions of the localist and distributed models were built to compute a simple quasi-regular mapping in which there was one regularity and a number of items that deviated from the regularity. Input gain was varied from low to high levels, and performance on regular, irregular, and novel inputs was assessed.

2.1.1. Input and output representations

The input space and the target output space each consisted of 12 binary dimensions. Out of $2^{12} = 4{,}096$ possible points in the 12-dimensional input space, one fourth (1,024) were chosen at random to constitute the corpus of "known" input patterns. The 3,072 remaining input patterns served as novel items during testing. Each known input pattern was associated with one target output pattern. Target patterns were created as follows:

Each target dimension was assigned to one of the input dimensions, thereby created a one-to-one relation from inputs to targets. For each known item, its targets were first set directly to their corresponding inputs; this constituted the identity mapping. The identity mapping was then "distorted" with some probability. In particular, each target bit value was flipped with a .05 probability. Thus, the identity mapping was the only regularity in this simple quasi-regular mapping, and flipped values were exceptions to the regularity. This procedure resulted in 563 fully regular items (no flipped values), and 461 irregular items with one to four flipped values per item. This proportion of irregular items is somewhat higher than what one might find in real quasi-regular domains, but it ensured that there were sufficient numbers of items of each type for the analyses.

For both the distributed model and the localist model, target values were either +1 or –1. For the localist model, input values were also either +1 or –1. For the distributed model, ± input values were not used because input vectors that point in the opposite direction are problematic to learn via back-propagation. Instead, input values were either 0 or 1. The problem with 0/1 coding is that input values of 0 do not have a direct effect on processing, whereas all input values had direct effects on processing in the localist model. To more closely mimic the localist coding scheme, two input values were used in the distributed model to code each input dimension. One coded the input dimension directly as x, and the other coded its opposite, $1 - x$. This coding scheme ensured that each input dimension had a direct effect on the processing of every item. It is important to note that the input representations coded the same 12 input dimensions in both the localist and distributed models, that is, no information was added by the $x|1 - x$ coding scheme.

The input and output representations captured the essential properties of quasi regularity as it is implemented in most connectionist models of language processes, such as word reading and inflectional morphology. Specifically, each input unit had a mostly systematic relation with one output unit, much like the way that each orthographic unit would have a mostly systematic relation with at least one phonological unit in a model of word reading (e.g., a unit for the letter *P* in the initial position would have a mostly systematic relation with a unit for the phoneme /p/ in the first position). Moreover, these relations were never entirely systematic, much like the case in real quasi-regular domains.

2.1.2. Localist model architecture

In the localist model, the 12 input units were fully connected to 1,024 "lexical" units (similar to logogens as proposed by Morton, 1969). Each lexical unit represented one item in the corpus, and the weights on incoming connections from input units were set according to each unit's input pattern. Learning was not necessary, given that the representations were predetermined by the corpus itself. Weights were set according to the input and output features of each given item. This meant that lexical units would be activated based on their similarity to a given input pattern.

Specifically, each connection going into a given lexical unit was weighted either +1 or –1 in accordance with the input pattern for the corresponding item. For instance, if a lexical unit was assigned to an input pattern for which all 12 units were set to +1, then all 12 incoming connections were set to +1. Each lexical unit also had 12 outgoing connections, one into each of the 12 output units. The outgoing connections were set using the same procedure as for the incoming connections: An output pattern was associated with each input pattern, and the outgoing connections from each lexical unit were weighted +1 or –1 in accordance with the output pattern for the corresponding item.

To process a given item, input units were first set to the item's input pattern. Lexical unit activations were then calculated with the normalized exponential function (see Nosofsky, 1990),

$$a_j = e^{\gamma \varepsilon I_j} \Big/ \sum_i e^{\gamma \varepsilon I_i} \tag{1}$$

where I_j was the net input to unit *j,* calculated as the dot product between the input vector and the incoming weight vector, γ was input gain, ε was noise sampled evenly in the range ± 0.1, and *i* spanned all lexical units. The denominator normalized the sum of all lexical unit activations to 1, and input gain modulated the distribution of activation (i.e., competition) among these units. At relatively low input gain, activations were more evenly distributed because they were on the shallow part of the exponential function. At high input gain, activations were garnered by lexical units with the largest net inputs because they were on the steeper part of the exponential. Noise was included to break perfect ties between very small (e.g., two or three) numbers of activated units. Such ties occurred more often at high levels of input gain. Output units were calculated as the hyperbolic tangent of the dot product between the vector of lexical unit activations, and its incoming weight vector (the hyperbolic tangent is a sigmoidal function with asymptotes at +1 and –1; see next section).

2.1.3. Distributed model architecture

In the distributed model, the input units were fully connected to 200 hidden units, and the hidden units were fully connected to the output units. The number of hidden units was determined through pilot testing to be about 50 units more than the minimum needed to learn the mapping. A range of numbers of hidden units was tested, but these analyses are not reported because the results were very similar across hidden unit numbers. The activation of each hidden unit was computed with the hyperbolic tangent function,

$$a_j = \tanh(\gamma \varepsilon I_j), \tag{2}$$

which is analogous to the logistic, except it has asymptotes at +1 and –1 instead of +1 and 0. Input gain altered the sigmoidal shape of the function. Low values of input gain flattened the function and made it more linear. High values sharpened it to more closely mimic a step function.

Input gain (γ) was fixed at 1 during training and varied during testing (see next section). Noise (ε) was fixed at 0.1 (as in the localist model) during both training and testing. Output unit activations were computed as in the localist model. Connection weights were initialized to random values in the range [+0.1, –0.1), which is generally thought be a good default range such that initial weight derivatives are not too large or too small. Weight derivatives were computed through the back-propagation of error generated on the output units. Specifically,

$$\Delta w_{ij} = \eta\left(\partial E/\partial w_{ij}\right)$$

where w_{ij} was the connection weight from unit j to i, η was the learning rate (fixed at 0.001), and E was the cross-entropy error (Rumelhart et al., 1995). Weight derivatives were accumulated over each pass through all 1,024 known items, and weights were updated at the end of each pass. Weight derivatives were calculated for each known item as follows: Input units were set to the item's input pattern, activation was propagated forward through the network, an error signal was calculated from the difference between actual and target outputs, and the error signal was back-propagated to generate the weight derivatives. Weight updates were repeated until every output unit was within 0.1 of its target for all 1,024 known items. This criterion was chosen to ensure highly accurate performance. It was reached after 3,000 passes through the known items.

2.1.4. Testing procedure

For both models, performance was assessed on each item by setting the input units to the item's input pattern and then determining whether the activation of each output unit was within 1 of its target (which was either +1 or –1). Model outputs were correct only when the activations of all 12 output units were within range. Targets for items in the corpus were set according to each item's output pattern. Targets for the 3,072 novel items were set according to each item's input pattern, that is, the identity mapping, which was the one and only regularity in the quasi-regular mapping that we constructed.

Input gain was varied as a single control parameter over the lexical units in the localist model, and over the hidden units in the distributed model. Input gain was not manipulated over the output units because it would have no qualitative effect on performance (i.e., it would not change the signs of output activations). The reported levels of input gain are between 0.5 and 3 for the localist model and 0.33 and 3 for the distributed model. These ranges were chosen to show asymptotic performance at the lower and upper ends of input gain for each model. Levels of input gain outside these ranges are not reported because the patterns of behavior did not change substantially beyond these ranges.

2.1.5. Model performance

Mean accuracies for the localist model and the distributed model are graphed in Figure 2 as a function of input gain and item type (regular, irregular, or novel). The graphs show that both models exhibited a clear dissociation in performance on irregular items compared with novel

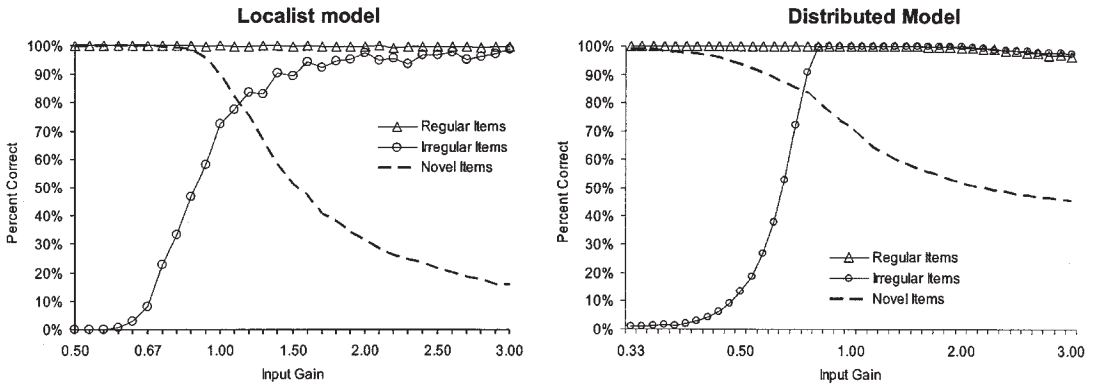

Fig. 2. Performance of the localist and distributed models in Simulations 1a and 1b as a function of input gain and item type.

items. At low levels of input gain, generalization of the identity mapping to novel inputs was essentially perfect, as was performance on regular items. By contrast, performance on irregular items dropped to 0%, at which point all inputs resulted in the identity mapping. For irregular items, application of the identity mapping was a regularization error because the identity mapping was the regularity.

At high levels of input gain, performance on all known items was near perfect in both models. By contrast, mean accuracies for the novel items dropped to as low as 16% for the localist model, and 46% for the distributed model. Of all the localist model's erroneous responses to novel items at the highest level of input gain, 97% were output patterns that corresponded to items in the training corpus. These responses were lexicalization errors because they are responses for known items in the model's "lexicon." The same analysis of errors made by the distributed model showed only 27% lexicalization errors, which was not much higher than the chance rate of 22.7%. The chance rate was calculated simply as the percentage of points in the output space that were assigned as targets; the percentage was less than 25% because some points in the output space served as targets for more than one input pattern, due to the creation of irregular items.

2.1.6. Model visualizations

To further investigate how input gain had its effects on performance, a method was devised to visualize the structure of a model's input–output mapping at a given level of input gain. In this method, the space of all possible input vectors is laid out on a square, and a model's *decision boundaries* are drawn onto the square. Decision boundaries occur wherever a model changes its output qualitatively (i.e., one or more output units change sign). Decision boundaries illuminate the processing of a model by virtue of depicting the arrangement of the output space with respect to the input space.

Six visualizations are shown in Figure 3. The left column shows three visualizations for the localist model, and the right column shows three visualizations for the distributed model. The top row shows visualizations at the low end of input gain (0.5 in the localist model and 0.333 in the distributed model; top row), the bottom row shows visualizations at the high end of input

gain (3 in both models), and the middle row shows visualizations for the point at which accuracies for irregular items and novel items were equal (1.1 in the localist model and 0.8 in the distributed model). The information to be extracted from these visualizations is explained next, and the details of how these visualizations were created are given in an Appendix.

The gray-scale values represent closeness to a decision boundary, with a black dot meaning that the output vector at that point on the square was "teetering" on a decision boundary. The overall differences in plot densities for the localist model, compared with plot densities for the

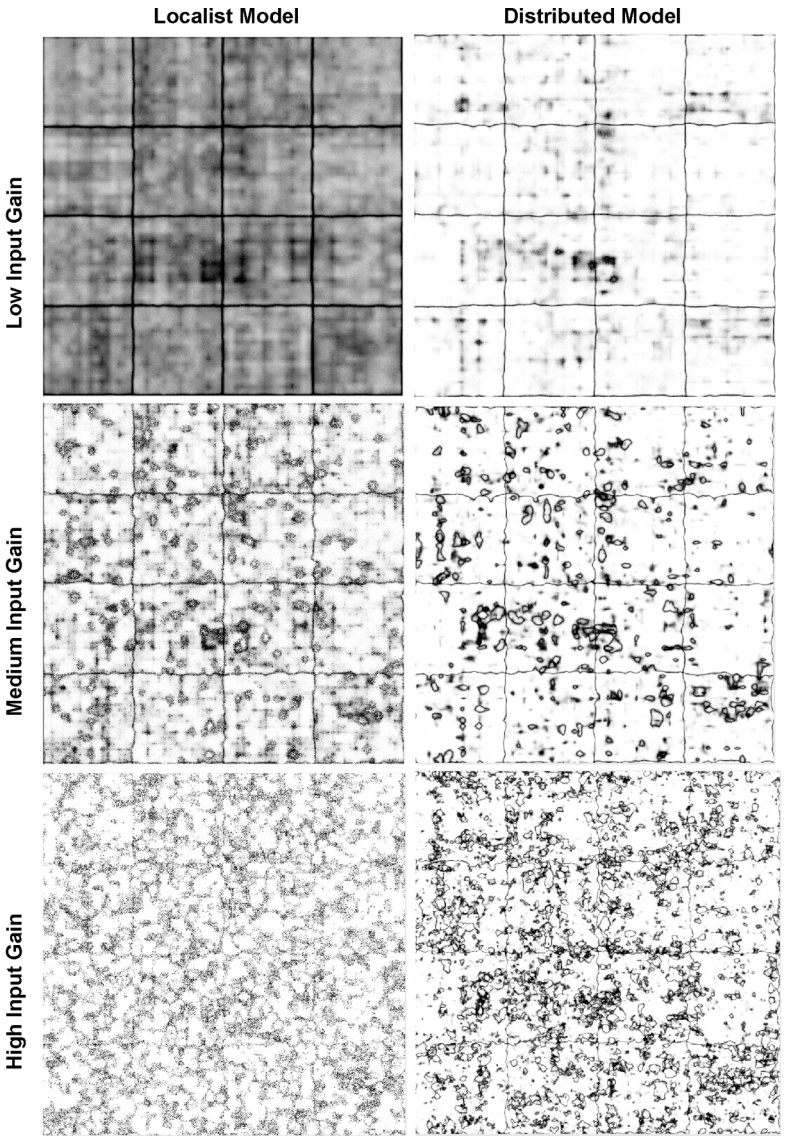

Fig. 3. Visualizations of the input–output mapping for the localist and distributed models at three different levels of input gain.

distributed model, were due to arbitrary differences in the overall magnitudes of output unit activations. The relevant information is in the patterning of darker and lighter areas.

In this method of visualization, the identity mapping shows up as a 4×4 grid, and any deviations from computing the identity mapping show up as distortions of the grid, or decision boundaries in addition to the grid. The grid patterns seen in the top two visualizations of Figure 3 show that both models processed the identity mapping at the low end of input gain. The fact that the grid comprises the only clear decision boundaries in these visualizations means that all irregular items were given the identity mapping, that is, regularized.

The middle two visualizations show that the grid pattern became distorted for both models at moderate levels of input gain, and "pockets" of decision boundaries began to appear. Given that mean accuracies were about 80% for irregular items at these levels of input gain, one can infer that the distortions and pockets reflect the "elaboration" of the identity mapping that was necessary to process the irregular items. Moreover, given that mean accuracies were about 80% for novel items and 100% for regular items at these levels of input gain, one can infer that the distortions and pockets were mostly isolated to the irregular items. The visualizations also show that placements and shapes of the decision boundaries were similar between the two models, again reflecting their similar patterns of performance.

The bottom two visualizations show that, for each model, the grid pattern was mostly replaced by pockets of decision boundaries at the high end of input gain. These pockets have a fairly simple interpretation for the localist model. Recall that, at the high end of input gain, 97% of the localist model errors for novel items were lexicalizations. Thus, these "item pockets" depict the boundaries established to process items in the training corpus. A lexicalization occurs when a novel input pattern is mapped into an item pocket. Item pockets are a clear depiction of item-based processing in the localist model.

In the distributed model, the pockets cannot be readily interpreted as item pockets because a substantial number of novel items were mapped correctly at the high end of input gain (46%), and the proportion of lexicalization errors for novel items was not much above chance (27% compared with 22.7% chance rate). It appears that the distortions needed for accurate mappings of irregular items had widened beyond their normal scope at high levels of input gain. Because the mapping of regular items is mostly correct at the high end of input gain, one can infer that the decision boundaries tended not to spread into regions of the space occupied by known items. It is this selective spread of decision boundaries that indicates item-based processing at the high end of input gain.

One can also see that the visualizations differed somewhat between the models: The pockets were smaller, more regular in shape, and more evenly distributed in the visualization for the localist model. This difference in item pockets reflects the ability of input gain to shift the localist model into a purely item-based mode of processing. By contrast, input gain in the distributed model did not cause such an absolute deficit in performance with novel items.

2.1.7. Discussion of simulations 1a and 1b

The results of Simulations 1a and 1b demonstrate that a double dissociation between regularity-based and item-based processing can emerge as a function of one control parameter in a single-route system. The same multiplicative scaling parameter, input gain, gave rise to the observed dissociation in the localist model as well as the distributed model.

This double dissociation is analogous to the one between irregular words and nonwords that was observed in the single-route model of word reading reported by Kello (2003). However, the localist model presented here used competitive localist units, whereas the word-reading model used hidden units trained via back-propagation. Also, the models presented here had feed-forward connections only, and their unit activations were computed instantaneously. By contrast, the word-reading model had recurrent connections and activation dynamics. The implementation of the word-reading model had left it unclear whether recurrence and dynamics played a key role in the dissociating effect of input gain. Simulations 1a and 1b show clearly that neither recurrence nor dynamics are necessary for this dissociating effect.

2.2. Simulations 2a and 2b

In Simulations 2a and 2b, recurrence and dynamics were introduced into the localist and distributed models to examine their effect on the function of input gain. The purpose of these simulations is to determine how recurrence and dynamics affect the dissociation observed in the first two simulations. It is important to confirm that, as suggested by the single-route model of word reading reported by Kello (2003), the dissociation is robust enough to hold in the context of recurrence and dynamics.

2.2.1. Method

The representations and model architectures were identical to those used in Simulations 1a and 1b, except that a new set of 1,024 known items was chosen randomly from the 12-dimensional input space. Recurrent connections were added to the model architectures as follows:

For the localist model, recurrent connections were made by taking each feed-forward connection from the lexical units to the output units and copying it in the reverse direction. To illustrate, if the connection weight from lexical unit A to output unit B was +1, then the connection weight from output unit B to lexical unit A was also +1. For the distributed model, recurrent connection weights were initialized with random values in the range [–0.5,+0.5), whereas the range for feed-forward connection weights was [–0.1,+0.1), as in Simulation 1b. The larger range was used to ensure that, after training, the recurrent connections would make a contribution to the net inputs of hidden units that was comparable to that of the feed-forward connections. Connection weights were learned with a version of back-propagation adapted for continuous recurrent networks (Pearlmutter, 1995), with targets being applied for only the last 12 ticks of processing. The learning rate was set to 0.0001, and 4,000 epochs of training were required to reach criterion.

Time was added to the models by integrating the net inputs of the lexical, hidden, and output units over a sequence of 18 ticks of processing for each given input pattern (input units were simply set according to the given input pattern for all 18 ticks). Specifically, the change in a unit's net input from one tick to the next was

$$\Delta I_j^{[t]} = \Delta t \left(\sum w_{ij} a_i^{[t-1]} - I_j^{[t-1]} \right),$$

where Δt was an integration constant fixed at 0.166, w_{ij} was the connection weight from unit i to unit j, $a_j^{[t-1]}$ was the activation of unit i on the previous tick, and $I_j^{[t-1]}$ was the net input to unit j on the previous tick (net inputs were initialized to zero on the first tick).

The testing procedure was the same as the one used in Simulations 1a and 1b, except that a criterion had to be set to determine the tick at which a model's output would be assessed. Outputs were tracked over the 18 ticks of processing and were assessed on the tick at which activations of the 12 output units changed less than 20% from their activations on the previous tick. This settling criterion normalized for the absolute magnitude of activations, which was different between the two types of models. However, pilot work not reported here showed that the kind of stopping criterion used did not affect results.

2.2.2. *Model performance*

Mean accuracies for the localist model and the distributed model are graphed in Figure 4 as a function of input gain and item type. The graphs show that, as in the first pair of simulations, both models exhibited a double dissociation in performance on irregular items compared with novel items as a function of input gain. Low levels of input gain had the same effect in the recurrent models as in the feed-forward models: Performance on regular and novel items was at ceiling, whereas performance on irregular items fell to 0%. High levels of input gain impaired performance on novel items more than on regular items as was found in the feed-forward models. However, there were also clear differences between the distributed and localist models, and between the feed-forward and recurrent models.

In the recurrent distributed model, manipulating input gain moved performance through four qualitatively different regions. From low to high input gain, the regions were as follows: selective deficit on irregular items, accurate performance on all items, selective deficit on novel items, and deficit on all items. This last region was absent in the performance profile of the recurrent localist model. It appears that recurrent connections in the distributed model had the effect of amplifying the distortions caused by high levels of input gain (the distortions were seen in the visualizations of Simulations 1a and 1b). These distortions were amplified to the point of impairing performance on all items.

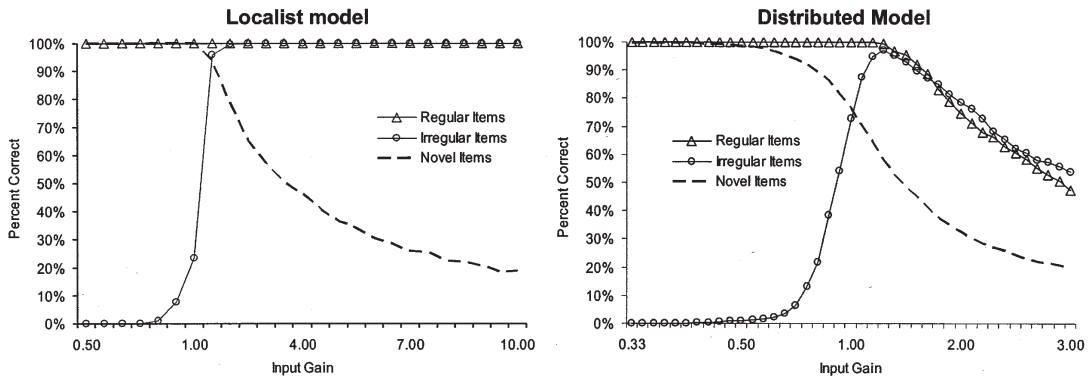

Fig. 4. Performance of the localist and distributed models in Simulations 2a and 2b as a function of input gain and item type.

Although recurrent connections in the localist model did not amplify any distortions, they did have a more quantitative effect on performance. In particular, much higher levels of input gain were needed to impair performance on novel items compared with Simulation 1a: The accuracy plots for the feed-forward and recurrent localist models are comparable, but the maximum level of input gain shown in the feed-forward plot is 3, whereas it is 10 in the recurrent plot. Thus, adding recurrent connections to the localist model weakened the deficit in performance on novel items at high levels of input gain. The reason for this difference is that, in the early ticks of processing, net inputs to the lexical units were relatively small in magnitude, regardless of the setting of input gain. These small net inputs meant that low levels of input gain were mimicked in the early ticks of processing. As a result, the output units were driven by the identity mapping at first, and in turn, the recurrent connections reinforced the identity mapping in the early ticks of processing. To the extent that this reinforcement overpowered the competition among lexical units caused by high levels of input gain, novel items were correctly generalized to the identity mapping.

2.3. Simulations 3a and 3b

In the first two pairs of simulations, the basic dissociating effect of input gain was established in feed-forward and recurrent versions of the localist and distributed models. In Simulations 3a and 3b, the dissociating effect of input gain was further tested in a quasi-regular mapping that more closely represented the quasi regularity that is often found in natural language domains.

Linguistic items typically vary in their frequency of occurrence in the language, and irregularities are more likely to occur in the more frequent items. For instance, a disproportionately large number of the most common verbs in the English language have irregular past-tense formations, such as IS, GO, and HAVE. Also, a disproportionately large number of the most common words in the English language contain irregular spelling–sound correspondences, such as THE, SAID, and HAVE. Frequency and its relation with irregularity were not considered in the first pair of simulations for purposes of clarity and simplicity, but it would be useful to know whether these factors change the pattern of simulation results.

2.3.1. Method

The representations and model architectures were identical to those used in Simulations 1a and 1b, except that a new set of 1,024 known items was chosen randomly from the 12-dimensional input space. The known items were randomly assigned a rank r from 1 to 1,024, and the frequency of each item was set according to $f = r^{-0.5}$. The target output pattern for each item was first set according to its input pattern, and then the sign of each target value was flipped with probability $p = .82r^{-0.5}$. The result of this formula was that the more frequent items were more likely to be irregular, and more likely to be more irregular (i.e., have more flipped values), compared with the less frequent items. The multiplicative constant of .82 was determined via a method of approximation to ensure that there was a .05 probability on average of flipping each target value across the set of known items, as in Simulations 1a and 1b.

For the distributed model, frequencies were used to scale the error generated by each presentation of each item during training, and 56,000 passes through the training corpus were required to reach criterion. For the localist model, frequency values were multiplied by 6 and

added to the net input of each respective item. The multiple of 6 was set to give the most frequent item a 50% advantage over the least frequent item in terms of baseline net input. The frequency parameters for both models were determined on the basis of pilot work such that the effect of frequency on processing was neither too strong nor too weak.

2.3.2. Model performance

To simplify the presentation of results, known items were divided into categories of high frequency and low frequency. The high-frequency category consisted of the 256 most frequent items, and the low-frequency category consisted of the 256 least frequent words. Mean accuracies for the localist model and the distributed model are graphed in Figure 5 as a function of input gain and item type (high-frequency regular, high-frequency irregular, low-frequency regular, low-frequency irregular, or novel). The graphs show that, as in the first pair of simulations, both models exhibited a clear double dissociation in performance on irregular items compared with novel items as a function of input gain. Moreover, performance on high-frequency items was generally more accurate than performance on low-frequency items in both models. Also, both models exhibited a characteristic interaction between frequency and regularity: Averaged across levels of input gain, there was generally a larger effect of regularity on low-frequency items compared with high-frequency items.

Although the simulation results were mostly the same for the two types of models, there were three differences that should be noted. First, as in Simulations 1a and 1b, the deficit in performance for novel items at high levels of input gain was more severe in the localist model. This difference was addressed in the discussion of results from the first pair of simulations. Second, performance on known items was worse overall for the localist model at high levels of input gain. This difference is not informative because it would disappear if a free parameter were changed, for example, if the level of noise was decreased. We did not adjust any free parameters because we wanted to fix them at reasonable default values across all of the simulations.

The third and final difference was that the interaction between frequency and regularity was maintained across most levels of input gain in the localist model, whereas performance was at ceiling at moderately high levels of input gain in the distributed model. This difference is also not informative because adding more noise to the distributed model "recovered" the interac-

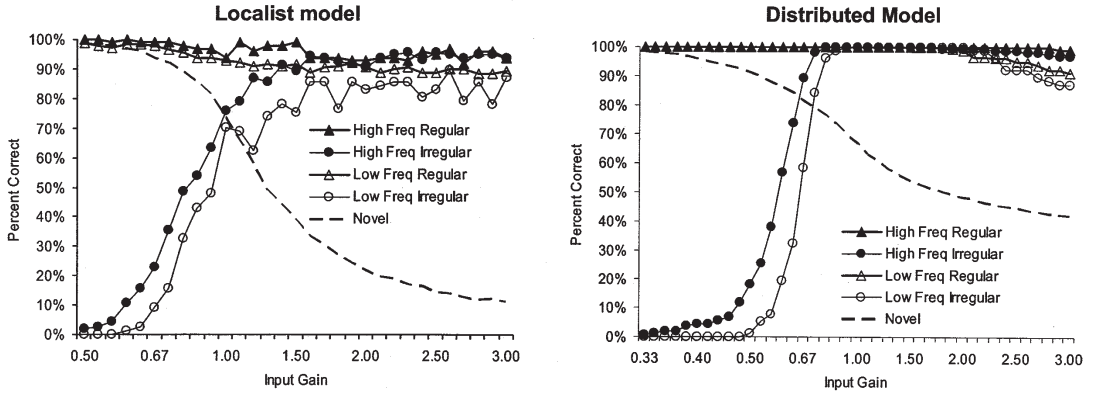

Fig. 5. Performance of the localist and distributed models in Simulations 3a and 3b as a function of input gain and item type.

tion: When the range of noise was increased to ±1.0, mean accuracies for high-frequency items and low-frequency regular items were all in the range of 83% to 87%, whereas the mean accuracy of low-frequency irregular items was 72%.

2.4. Simulations 4a and 4b

Simulations 4a and 4b were designed to test the effect of input gain in the context of another common, if not defining, characteristic of quasi-regular domains. Specifically, quasi-regular domains often include a range of regularities that vary in their scope. For instance, the -/d/ past-tense suffix in English is wide in scope because it applies to most verbs that end in a vowel, for example, PLAY, KEY, SIGH, SHOW, CHEW, and FLAW; the -/d/ suffix also applies to most verbs ending in a voiced, nonalveolar consonant. However, there is a small group of vowel-ending verbs (all starting with an orthographic consonant cluster) that have -EW as their past-tense formations: SLAY, FLY, BLOW, GROW, KNOW, THROW, and DRAW. These words form a subregularity in the quasi-regular domain of past-tense formation in English. To illustrate the same point in the domain of spelling–sound correspondences, consider that the letter *I* is usually pronounced as /I/ when a monosyllabic word does not end in *E* (e.g., HIT, BRIM, and WIND as a noun, but not DINE). However, for most monosyllabic words ending in -IND, the letter *I* is pronounced as /aI/: BIND, BLIND, FIND, GRIND, HIND, KIND, MIND, RIND, and WIND as a verb. These words form a subregularity in the quasi-regular domain of spelling–sound correspondences in English.

There were no subregularities in the quasi-regular mapping used in the first three pairs of simulations; the identity mapping was the only regularity, and dimension values were randomly flipped to create a scattering of exceptions to that regularity. For Simulations 4a and 4b, a new quasi-regular mapping was generated in which there was one large-scale regularity (the identity mapping) and two smaller scale subregularities. For items that followed the *flip* subregularity, four target values were set opposite to the identity mapping. For items that followed the *shift* subregularity, four target values were set on the basis of input dimensions *other* than the ones they normally corresponded to.

The flip and shift subregularities were designed to approximate the way that subregularities might be coded in connectionist models of language processing. For instance, a model of past-tense formation in English might have a unit corresponding to the letter *T* at the end of a verb, and this unit would tend to activate a unit corresponding to the -/Id/ past-tense suffix (e.g., CHAT–CHATTED, DOT–DOTTED, etc.). However, the -/Id/ would often need to be deactivated when the vowel letter is *E* or *I,* as in HIT, FIT, SPIT, QUIT, BET, LET, SET, and WET. The flip subregularity captures the essence of this kind of input–output relation.

As another example, a model of word reading in English might have units corresponding to the vowel letters such as *A* and *I,* and these units would tend to activate units corresponding to the sounds /{/ and /I/. However, when a monosyllabic word ends with the letter *E,* those same vowel letters will tend to activate a different set of units, that is, those corresponding to long vowels such as /el/ and /aI/. The shift subregularity captured the essence of this kind of input–output relation.

2.4.1. Method

The representations and model architectures were identical to those used in Simulations 1a and 1b, except that a new set of 1,024 known items was chosen randomly from the

12-dimensional input space. As in the previous simulations, target patterns were created by first copying each input pattern to its respective target pattern, that is, the identity mapping. Exceptions to the identity mapping were set according to three procedures, one for the flip subregularity, one for the shift subregularity, and one for random irregularities.

Three input dimensions were chosen up front to determine the "flip region" of input space. The flip subregularity was applied to items for which these three input dimensions were all positive. The flip procedure was to change the sign of four output dimensions, also chosen up front. Three other input dimensions were chosen up front to determine the shift region of input space. The shift subregularity was applied to items for which these three input dimensions were all positive. The shift procedure was to set four output dimensions according to four input dimensions other than those dictated by the identity mapping. Finally, random irregularities were also applied with some probability, as in the previous simulations. The probability was set such that the proportion of random irregularities was equated across simulations.

Connection weights in the localist model were set using the same procedure as in Simulations 1a and 2a, and the connection weights in the distributed model were learned to criterion after 3,500 passes through the training corpus.

2.4.2. Model performance

Known items were classified as either regular, irregular, flip subregular, or shift subregular. Novel items were classified as either regular, flip subregular, or shift subregular. Mean accuracies for the localist model and the distributed model are graphed in Figure 6 as a function of input gain for known items of each class and for the regular novel items. The results for regular items (both known and novel) and irregular items (i.e., those with random exceptions) replicated the results found in Simulations 1a and 1b: performance on irregular items versus regular novel items dissociated as a function of input gain. Also, the deficit in performance on irregular items at low levels of input gain was comparable between the two models, whereas the deficit in performance on regular novel items at high levels of input gain was more severe in the localist model.

For known items classified as flip or shift subregular, performance was comparable between the two models. At low levels of input gain, both types of items were incorrectly assigned the

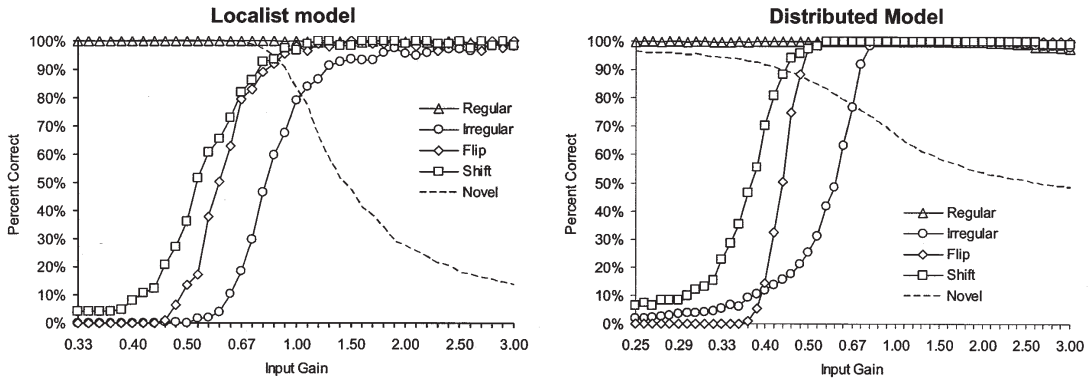

Fig. 6. Performance of the localist and distributed models in Simulations 4a and 4b as a function of input gain and item type (not including novel items in the subregular regions).

identity mapping, that is, the large-scale regularity overrode the smaller scale subregularities. As input gain was increased, accuracies improved for the shift subregular items first, followed by the flip subregular items. Moreover, accuracies for both shift and flip subregular items improved before accuracies for irregular items improved. Responses to subregular items were more resistant to regularization because of the support that each subregular item drew from its subregular neighbors. Responses to the shift items were more resistant than flip items because it turned out that shift subregularity was more similar to the identity mapping than the flip subregularity.

For novel items classified as flip and shift subregular, it was informative to examine performance in terms of the outputs that conformed to the identity mapping (regularizations), versus the respective subregular mapping (subregularizations). The rates of these different types of outputs are graphed in Figure 7 as a function of input gain. At low levels of input gain, both models produced nothing but regularizations for flip and shift subregular items. This result is consistent with the effect of low levels of input gain for all item types in all the simulations reported herein.

As input gain increased to higher levels, regularizations disappeared, whereas subregularizations became more prominent. For the localist model, the rate of subregularizations peaked at about 35% for flip subregular items, and at about 50% for shift subregular items. For the distributed model, the rate of subregularizations peaked at about 95% for both flip and shift subregular items. The rate of subregularizations was lower in the localist model because it was more prone to making lexicalization errors. The rate of subregularizations was high because subregularities were defined by conjunctions of input dimensions, and high levels of input gain emphasized conjunctive processing.

In summary, Simulations 4a and 4b showed that the dissociating effect of input gain was essentially maintained when the models computed a quasi-regular mapping with subregularities. The subregular mappings were processed somewhat differently in the two models, and these differences provided a more complete picture of the effect that input gain has on competitive localist representations versus learned distributed representations.

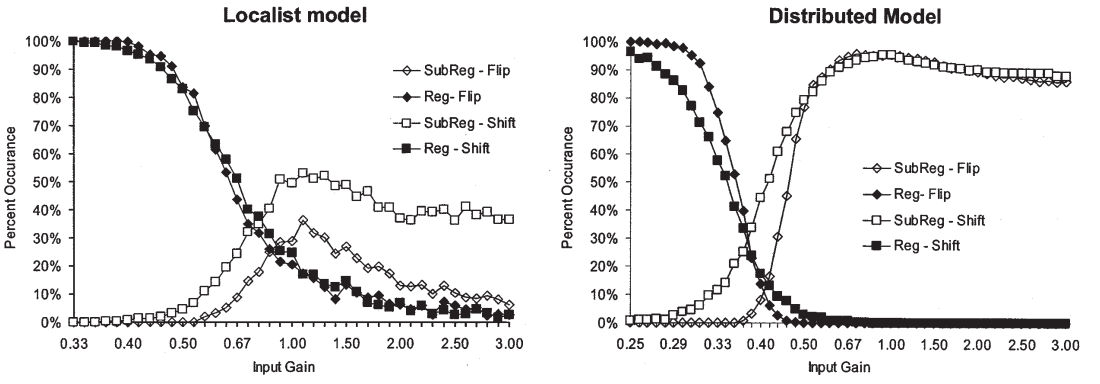

Fig. 7. Percentage of occurrence of regularization and subregularization outputs as a function of input gain, shown separately for novel items classified as flip subregular and shift subregular. SubReg-Flip = occurrence of subregularizations for novel flip items, Reg-Flip = occurrence of regularizations for novel flip items, SubReg-Shift = occurrence of subregularizations for novel shift items, Reg-Shift = occurrence of regularizations for novel shift items.

3. General discussion

The meaning of double dissociations in language processing, and neuropsychology in general, is an ongoing topic of debate (see Dunn & Kirsner, 2003, and accompanying Discussion Forum). Some researchers have questioned the assumptions that underlie the use of double dissociations to infer processing components (Van Orden et al., 2001). Others have demonstrated how double dissociations can emerge without separable processing components. Plaut (1995) simulated a dissociation in the reading of abstract versus concrete nouns by damaging two different parts of a nonmodular model of word reading. Devlin, Gonnerman, Andersen, & Seidenberg (1998) simulated a dissociation in the processing of living versus nonliving things as a function of degree of impairment in a nonmodular model of semantic processing. Juola (2000) simulated a dissociation in the past-tense formation of regular versus irregular verbs as chance occurrences of nonselective lesions in a nonmodular model of verb inflection. Kinder and Shanks (2003) simulated a dissociation in priming versus recognition performance by altering two different parameters in a nonmodular model of memory.

Our simulations contribute a new demonstration to this list. Performance on novel versus irregular stimuli was dissociated by shifting between regularity-based and item-based modes of processing. Unlike previous demonstrations, these modes existed at the ends of a continuum created by one control parameter, namely, input gain. Moreover, input gain had analogous effects in both localist and distributed architectures, even though its computational function in the two types of architectures is quite different.

The reason why input gain had analogous effects is that similarity among input patterns played a similar role in both types of models. In the localist model, the role of similarity is transparent: The activation of a lexical unit was a monotonic function of similarity between an input pattern and the known item it represented. The role of similarity in the distributed model is more indirect, but equally important: Similar input patterns tended to activate similar hidden unit patterns, as is generally true when mapping via distributed representations.

For both types of models, input gain affected the *scope* of similarity. Scope is defined by the relation between similarity and influence on processing. A wide scope means that the processing of a given input pattern is influenced by a wide range of known items, some more and some less similar to the input pattern. A narrow scope means that processing is influenced primarily by items that are very similar to the input pattern. For both types of models, lowering the levels of input gain widened the scope of similarity, whereas raising the levels of input gain narrowed it. The means by which input gain had this effect was different between the two model types.

In the localist model, input gain had a direct effect on the scope of similarity. Lower levels reduced competition among lexical units, which allowed them to be partially activated in response to input patterns that were only vaguely similar to the items they represented, that is, a wide scope of similarity. Higher levels meant that activation was given only to the few items most similar to a given input pattern, that is, a narrow scope of similarity.

In the distributed model, input gain affected the scope of similarity by determining the region of activation space that hidden units operated in. Lower levels pushed hidden units to operate in the center of activation space. This encouraged a linear, componential mapping between inputs and outputs. Through learning, the center of activation space captured the regular relations between inputs and outputs that hold over wide ranges of known items. Higher levels

of input gain pushed hidden units to operate in the corners of activation space. This encouraged a nonlinear, conjunctive mapping between inputs and outputs. Through learning, the corners of activation space captured item-specific relations between inputs and outputs.

We do not go further into the computational details, because they are unnecessary for understanding how input gain caused the selective deficits in processing irregular items versus novel items. At low levels of input gain, irregular items were selectively impaired because idiosyncratic input–output relations were "averaged out" by regular input–output relations. This averaging occurred because processing was based on the aggregate effects of many items similar to the target. By the same token, averaging supported the processing of novel items. At high levels of input gain, novel items were selectively impaired because generalization cannot reliably be based on the input–output relations contained in a few items similar to the novel input pattern. By the same token, a restriction in the scope of similarity is advantageous for the processing of irregular items.

In summary, our simulations show how selective deficits in item-based processing versus regularity-based processing can arise from changes in the scope of similarity-based processing. Such changes can be effected by a single control parameter in a system without separable processing components. We are not arguing that input gain, or control parameters more generally, provide better accounts of double dissociations compared with other component-based accounts. It is quite possible that there are many ways to selectively impair regularity-based and item-based processing in real cognitive systems and that each account has a subset of cases for which it is best suited.

3.1. Neural bases of input gain

To further investigate the possibility that at least some empirically observed dissociations are caused by aberrant changes in control parameters such as input gain, it would be helpful to formulate the neural bases of those control parameters. It turns out that some well-established neural mechanisms have effects analogous to those of input gain. With respect to the competition-modulation function of input gain in the localist model, lateral inhibition is a widespread mechanism of neural processing that could implement competition among cognitive representations. It has been hypothesized that the strength of lateral inhibition can be modulated (e.g., Winder, 1999), analogous to the function of input gain in the localist models. There are also plausible candidates for a neural mechanism of sensitivity-modulation, analogous to the function of input gain in the distributed models. Most directly, it is well-established that some neuromodulators can change the sensitivity of neuronal firing rates in response to excitatory and inhibitory inputs (see Fellous & Linster, 1998).

Whatever the neural mechanism of input gain might be, brain damage would need to disrupt its function to account for selective deficits in regularity-based versus item-based processing. One possibility is that brain damage can disrupt the function of an appropriate neuromodulator in one or more neural systems. Another possibility is that brain damage can affect a modulatory system that is responsible for controlling the scope of similarity-based processing in another system. Another possibility is that brain damage can disrupt the balance of constraints on language processing, such as those from semantics and phonology. Such a disruption would be hypothesized to alter the level of competition (localist), or the sensitivity to inputs (distributed), in the affected neural systems (see also McNellis & Blumstein, 2001). These are clearly

speculations for now, but they show how it is at least plausible that brain damage could, in effect, cause a chronic increase or decrease in input gain.

3.2. The division of phonological and semantic processing

The idea that phonological and semantic constraints might play a role in certain dissociations is not a new one. There is substantial evidence that these constraints are subserved by separate neural systems (for a review and evidence from brain imaging, see McDermott, Petersen, Watson, & Ojemann, 2003). In line with this evidence, Patterson and her colleagues have argued that phonological processes are more important for regularity-based processing, and semantic processes are more important for item-based processing (as discussed earlier and summarized in Patterson & Lambon Ralph, 1999).

Input gain and the phonological–semantic hypothesis are not mutually exclusive. Indeed, reported cases in which both regular and novel items are dissociated from irregular items (e.g., Marslen-Wilson & Tyler, 1997; Ullman et al., 1997) may be particularly difficult to account for based solely on a manipulation of input gain, which did not affect known regular items in these analyses. Further empirical and computational work is needed to investigate this issue.

3.3. Measurements of cortical activity during language processing

Our simulations show how neuropsychological dissociations do not necessitate separate regularity-based and item-based processing components, but dissociations in cortical activity must also be considered. A number of imaging and electrophysiological studies (mostly using word-reading tasks) have suggested that there are at least two brain regions that play complementary roles in language processing (for a review, see Pugh et al., 2001). A temporoparietal region has been shown to respond more strongly to pseudowords and phonological tasks compared with words and other tasks. To complement, an occipitoventral region has been shown to respond more strongly for words, and equally for phonological tasks. Also, the former has slower response times, on average, compared with the latter.

The temporoparietal and occipitoventral regions have characteristics associated with regularity-based versus item-based processes, respectively. The measurements of cortical activity are consistent with dual-route theories, but they do not rule out other possible interpretations. One possibility is that the regions support phonological and semantic processes, as suggested by the studies discussed in the previous section. Another possibility is that the faster system supports the bulk of language processing in skilled performance (as supported by the data), including both regularity-based and item-based processing, and the slower system serves as a modulator or controller for the faster system. It is also possible that regularity-based and item-based processes exist only in the interactions between the two systems. Thus, the data are consistent with the idea that these two processes are components of a modular system, but the data are also consistent with two modes of processing in a nonmodular system.

3.4. Input gain and rate of processing

Input gain in models of language processing was first introduced by Kello and Plaut (2000) as a mechanism of control over rate of processing, rather than level of competition, or conjunc-

tive versus componential modes of processing. When processing has a time course (as in Simulations 2a and 2b), high levels of input gain can cause units to reach their asymptotes in fewer time steps (Kello & Plaut, 2003). This function of input gain raises the question of whether high and low levels of input gain always induce fast and slow rates of processing. If so, the input gain hypothesis would make the strange, and probably incorrect, prediction that the rate of processing becomes fast when regularity-based processing is impaired, and slow when item-based processing is impaired.

The input gain hypothesis does not, in fact, make this prediction because input gain has at least two different functions, depending on how it is manipulated. When input gain is manipulated uniformly across all levels of processing, it can alter the rate of processing, as shown in previous models (Kello & Plaut, 2003; Kello, Plaut, & MacWhinney, 2000). Input gain did not, by contrast, alter the rate of processing in this way when it was manipulated selectively over the representations mediating inputs and outputs, as in our simulations. Instead, the rates of processing were fastest at moderate levels of input gain, and somewhat slower at the extremes, as measured by the time it took for the output units to settle (these analyses were not reported because the effects were small and irrelevant). Thus, it appears that input gain is a multipurpose control mechanism in connectionist models, which raises the question of whether such a mechanism might exist in real neural and cognitive systems.

3.5. Concluding remarks

The tension between dual-route and single-route theories can be found in many areas of research in the cognitive sciences. The simulations reported here demonstrate how a single-route theory might account for dissociations that are traditionally seen as evidence for dual-route theories. Dual-route and single-route theories will ultimately stand or fall on the basis of research aimed at specific empirical issues. The input gain hypothesis may, in time, be proved or disproved by the results of such research efforts. Whatever their outcome, these efforts are advanced by work that helps to lay out all of the hypotheses that are consistent with the data.

Acknowledgments

This research was funded by NIH Grant MH55628 and NSF Grant 0239595.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition, 90,* 119–161.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* 932–945.

Behrmann, M., & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, single lexicon. *Cognitive Neuropsychology, 9,* 209–251.

Berent, I., Pinker, S., & Shimron, J. (2002). The nature of regularity and irregularity: Evidence from Hebrew nominal inflection. *Journal of Psycholinguistic Research, 31,* 459–502.

Beretta, A., Carr, T. H., Huang, J., & Cao, Y. (2003). The brain is not single-minded about inflectional morphology: A response to the commentaries. *Brain & Language, 85,* 531–534.

Bird, H., Ralph, M. A. L., Seidenberg, M. S., McClelland, J. L., & Patterson, K. (2003). Deficits in phonology and past-tense morphology: What's the connection? *Journal of Memory and Language, 48,* 502–526.

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

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral & Brain Sciences, 22,* 991–1060.

Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychology, 13,* 749–762.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience, 10,* 77–94.

Dunn, J. C., & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39,* 1–7.

Fellous, J.-M., & Linster, C. (1998). Computational models of neuromodulation. *Neural Computation, 10,* 771–805.

Funnell, E. (1983). Phonological processes in reading—New evidence from acquired dyslexia. *British Journal of Psychology, 74,* 159–180.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception & Performance, 5,* 674–691.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology, 41,* 313–360.

Haskell, T. R., MacDonald, M. C., & Seidenberg, M. S. (2003). Language learning and innateness: Some implications of compounds research. *Cognitive Psychology, 47,* 119–163.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 7592–7597.

Juola, P. (2000). Double dissociations and neurophysiological expectations. *Brain & Cognition, 43,* 257–262.

Kay, J., & Marcel, A. (1981). One process, not two, in reading aloud: Lexical analogies do the work of non-lexical rules. *Quarterly Journal of Experimental Psychology A, 33A,* 397–413.

Kello, C. T. (2003). The emergence of a double dissociation in the modulation of a single control parameter in a nonlinear dynamical system. *Cortex, 39,* 132–134.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo-naming task. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26,* 719–750.

Kello, C. T., & Plaut, D. C. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory & Language, 48,* 207–232.

Kello, C. T., Plaut, D. C., & MacWhinney, B. (2000). The task dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference in speech production. *Journal of Experimental Psychology: General, 129,* 340–360.

Kinder, A., & Shanks, D. R. (2003). Neuropsychological dissociations between priming and recognition: A single-system connectionist account. *Psychological Review, 110,* 728–744.

Lachter, J., & Bever, T. G. (1988). The relation between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. *Cognition, 28,* 195–247.

Marslen-Wilson, W. D., & Tyler, L. K. (1997, June 5). Dissociating types of mental computation. *Nature, 387,* 592–594.

Marslen-Wilson, W. D., & Tyler, L. K. (1998). Rules, representations, and the English past tense. *Trends in Cognitive Sciences, 2,* 428–435.

McClelland, J. L., & Patterson, K. (2002a). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences, 6,* 465–472.

McClelland, J. L., & Patterson, K. (2002b). "Words or rules" cannot exploit the regularity in exceptions: Reply to Pinker and Ullman. *Trends in Cognitive Sciences, 6,* 464–465.

McDermott, K. B., Petersen, S. E., Watson, J. M., & Ojemann, J. G. (2003). A procedure for identifying regions preferentially activated by attention to semantic and phonological relations using functional magnetic resonance imaging. *Neuropsychologia, 41,* 293–303.

McNellis, M. G., & Blumstein, S. E. (2001). Self-organizing dynamics of lexical access in normals and aphasics. *Journal of Cognitive Neuroscience, 13,* 151–170.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence: Vol. 5* (pp. 189–223). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

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

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13,* 1199–1241.

Patterson, K., & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia, 30,* 1025–1040.

Patterson, K., Lambon Ralph, M., Hodges, J., & McClelland, J. (2001). Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsychologia, 39,* 709–724.

Patterson, K., & Lambon Ralph, M. A. (1999). Selective disorders of reading? *Current Opinion in Neurobiology, 9,* 235–239.

Patterson, K., & Marcel, A. (1992). Phonological alexia or phonological alexia? In J. Alegria & D. Holender (Eds), *Analytic approaches to human cognition* (pp. 259–274). New York: North-Holland.

Pearlmutter, B. A. (1995). Gradient calculations for dynamic recurrent neural networks—A survey. *IEEE Transactions on Neural Networks, 6,* 1212–1228.

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

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., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6,* 456–463.

Plaut, D. C. (1995). Double dissociation without modularity—Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology, 17,* 291–321.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Pugh, K. R., Mencl, W., Jenner, A. R., Katz, L., Frost, S. J., Lee, J. R., et al. (2001). Neurobiological studies of reading and reading disability. *Journal of Communication Disorders, 34,* 479–492.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 45–76). Cambridge, MA: MIT Press.

Rumelhart, D. E., Durbin, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. Chauvin & D. E. Rumelhart (Eds), *Backpropagation: Theory, architecture, and applications* (pp. 1–34). Hillsdale, NJ: Lawrence Erlbaum.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge, England: Cambridge University Press.

Skousen, R. (1989). *Analogical modeling of language.* Dordrecht, The Netherlands: Kluwer Academic.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language, 26,* 608–631.

Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30,* 37–69.

Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. J., & Pinker, S. (1997). A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience, 9,* 266–276.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science, 25,* 111–172.

Winder, S. A. (1999). A model for biological winner-take-all neural competition employing inhibitory modulation of NMDA-mediated excitatory gain. *Neurocomputing, 26–27,* 587–592.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: Macmillan.

Appendix

To visualize decision boundaries, the 12-dimensional input space was "folded" on itself such that it could be plotted on a 2-dimensional grid. The grid was structured such that all points vertically or horizontally adjacent to each other were different by only one dimension. To illustrate, one corner of the grid was assigned the point in input space where all dimensions have the value –1. The next point from the corner moving vertically on the grid had one positive value, say, on Dimension A for the sake of illustration. The next point from the corner moving horizontally on the grid also had one positive value, say, on Dimension B. Given that points horizontally and vertically adjacent could be different on only one dimension, the point moving diagonally from the corner was not free to vary; in this example, it was forced to have positive values on Dimensions A and B only.

This scheme of laying out points in the input space was extended to fill a 64×64 grid, which covered all 4,096 points in the input space. The grid was only able to capture a portion of the similarity structure of the full 12-dimensional input space, due to the folding of 12 dimensions into 2. The limitation can be seen in the fact that, whereas each adjacent point differed by only one dimension, it was not true that all points differing by only one dimension were adjacent. This limitation did not interfere with the basic purpose of the visualizations, which was to give a general sense of how the models computed quasi-regular mappings.

In addition to the binary-valued points in the input space, 9 intermediate points were interpolated between each column and between each row of the grid. For instance, if Dimension A of the input space was set to –1 on the bottom row of the grid, and +1 on the next row up, then nine rows were interpolated in between, with the value of Dimension A ranging from –0.8 to +0.8 in steps of 0.2. The same was done to interpolate the columns. The grid was interpolated because most decision boundaries fell between the binary-valued points. At each point on the grid, a gray-scale value was plotted that reflected the combined values of four output dimensions, chosen arbitrarily. Only four values were combined because pilot work showed that the visualizations became overly saturated when more than four dimensions were combined.

Each gray-scale value at each point of the grid was determined by presenting the corresponding input pattern to the model under examination, at the level of input gain being plotted. The model produced an output pattern, and activations were read off the four chosen output units. The gray-scale value was set according to these activations: It was blackest when one or more of the activation values were in the middle of the sigmoid, that is, at a decision boundary. The gray-scale value was whitest when all four activation values were at asymptote, that is, maximally distant from a decision boundary. The change from black to white was linear between the middle of the sigmoid and its asymptotes.

The result of this plotting method is that dark outlines in the visualizations represent the shapes and locations of decision boundaries. The overall distribution of outlines depicts the overall character of the input–output mapping.