

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Dissociations Between Regularities and Irregularities in Language Processing:
Computational Demonstrations Without Separable Processing Components

Permalink

<https://escholarship.org/uc/item/0794z85p>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26)

ISSN

1069-7977

Authors

Sibley, Daragh E.
Kello, Christopher T.

Publication Date

2004

Peer reviewed

Dissociations Between Regularities and Irregularities in Language Processing: Computational Demonstrations Without Separable Processing Components

Daragh E. Sibley (dsibley@gmu.edu)

Department of Psychology, George Mason University
Fairfax, VA 22030-4444 USA

Christopher T. Kello (ckello@gmu.edu)

Department of Psychology, George Mason University
Fairfax, VA 22030-4444 USA

Abstract

Two models are presented that compute a quasi-regular mapping. One was based on localist representations of items in the quasi-regular domain, the other was based on distributed representations. In each model, a control parameter termed *input gain* was modulated over the one and only level of representation that mapped inputs to outputs. Input gain caused both 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, each model exhibited a double dissociation 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.

Introduction

The quasi-regular nature of language has played a central role in theories of language processing in the mind and brain. On the one hand, language processes must be able to handle novel inputs, e.g., skilled readers can give reasonable pronunciations and conjugations to verbs that they have never encountered before. These abilities demonstrate how language usage can be generative on the basis of regularities. On the other hand, irregular items often exist for which the regularities do not apply. Thus, language processes must be able to override the regularities, when appropriate, with knowledge that is applicable to only a few items, or even to just one. How are language processes structured to handle both regularities, and the exceptions to those regularities?

One answer to this question is that any given quasi-regular domain is processed by two complementary routes. A *regularity-based* route is specialized to capture the regularities that span across linguistic items in the domain, and an *item-based* route is specialized to capture knowledge that is specific to items in the domain. For instance, in the words-and-rules theory (Pinker, 1999), rules are used to process regular inflectional morphologies (e.g., WALK-WALKED), and a lexicon is used to process irregular inflections (e.g., GO-WENT). In the dual-route cascaded (DRC) theory of word reading (Coltheart, Curtis, Atkins, & Haller, 93; Coltheart et al., 2001), a set of grapheme-to-phoneme correspondence rules is used to capture

regularities between the spellings and sounds of words, and a system of lexical knowledge serves to override the rules when necessary (e.g., PINT does not rhyme with MINT).

Alternatively, *single-route* theories have been proposed in which the mechanisms and representations for handling regularities and irregularities are inseparable. For instance, Rumelhart and McClelland (1986) proposed a theory in which a single route of processing was used to generate the past tense of both regular and irregular verbs (also see, e.g., Joanisse & Seidenberg, 1999). Kello and Plaut (2003) proposed a theory of word reading in which the mapping from spelling to sound is mediated by a single level of learned representations (also see Plaut & Gonnerman, 2000).

A wide variety of evidence has been brought to bear on dual-route and single-route theories of language processing (for reviews, see Coltheart et al., 2001; McClelland & Patterson, 2002; Pinker, 1999; Pinker & Ullman, 2002; Plaut, McClelland, Seidenberg, & Patterson, 1996). Much of this evidence speaks to one or another particularity of a given theory. Every piece of evidence contributes to the overall debate, but here we focus on one kind of evidence that is relevant to all theories in question: dissociations between regularity-based and item-based processing.

Double dissociations have been observed in language processing, and some have been interpreted as evidence for separable regularity-based and item-based components of the language system. In the area of inflectional morphology, Ullman and his colleagues (Ullman et al., 1997) reported evidence for a dissociation between the past tense formation of regular and irregular verbs in English. They 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 found the opposite pattern for Parkinson's patients and aphasics with anterior lesions. Marslen-Wilson and Tyler (1997; 1998) found a similar dissociation in a priming paradigm with language-impaired patients.

In the area of word reading, deficits found in surface and phonological dyslexia have been interpreted analogously to those found in posterior versus anterior aphasics. For instance, Berhmann and Bub (1992) reported on a surface dyslexic patient MP for whom the ability to read exception words (particularly of low frequency) was greatly impaired,

whereas the ability to read both regular words and nonwords was mostly intact. By contrast, Funnell (1983) reported on a phonological dyslexic patient WB for whom the ability to read nonwords (even simple CVC nonwords) was greatly impaired, whereas the ability to read both easy and difficult words was mostly intact.

The impairments of these and other patients have a straightforward explanation in terms of separable item-based and regularity-based processing components. The deficits in Alzheimer's patients, posterior aphasics, and surface dyslexics all reflect damage to an item-based component of processing (e.g., a lexicon) that is responsible for irregular items (not necessarily the same component across types of deficits). The deficits in Parkinson's patients, anterior aphasics, and phonological dyslexics all reflect damage to a regularity-based component of processing (e.g., rules) that is responsible for novel items.

These double dissociations appear to challenge single-route theories because item-based and regularity-based processes are not separable in single-route theories. Proponents of single-route theories have responded to this evidence in a number of ways. In some cases, methodologies or interpretations of data have been called into question (e.g., McClelland & Patterson, 2002). In other cases, the data have been explained in terms of dissociations between semantic and phonological components of processing, rather than item-based and regularity-based components (e.g., Joanisse & Seidenberg, 1999). The research to date has left open the question of whether dissociations between the processing of novel and irregular items can be explained without reference to an architectural dichotomy in the language system.

Current Work

The primary aim of the current study was to demonstrate how a dissociation between item-based and regularity-based processing can occur in a single-route architecture without any manipulation of separable processing components, i.e., without reference to separable semantic and phonological contributions to processing. The basic idea is that a single component of processing can shift between two qualitatively different "modes" of processing as a function of one control parameter. Specifically, we present two different kinds of connectionist models that possess a control parameter termed *input gain*. We show that, in both types of models, input gain can cause a shift in processing between an item-based mode and a regularity-based mode. Furthermore, we show how this shift can give rise to a double dissociation in performance on irregular versus novel inputs.

The models were built to process an abstract, quasi-regular mapping. Properties of the mapping were analogous to basic properties of quasi-regularity in language domains. However, items did not correspond to any particular words in a particular language domain. The mapping was created primarily to facilitate analysis of the models, rather than to simulate a particular language phenomenon such as the past

tense formation in English. Therefore, the models are intended and reported only as proofs-of-concept.

The first model used a single level of localist nodes to map input patterns onto output patterns. Each node represented one item in the training corpus, and the activation of each node was a function of the similarity between the item it represented, and the current input to the model. Thus, this model could be considered as *analogy-based* because both known and novel inputs were explicitly processed in terms of the similarity of their input patterns to that of all items in the corpus (see Albright & Hayes, 2003; Nakisa, Plunkett, & Hahn, 2000).

The second model used a *distributed* level of representation to map input patterns onto output patterns. Hidden representations were learned via backpropagation (Rumelhart, Hinton, & Williams, 1985), and each hidden unit contributed to the processing of many, if not all, items in the training corpus. Representations learned through backpropagation tend to map similar inputs onto similar outputs (Rumelhart et al., 1995). Thus, as in the analogy model, the distributed model processed both known and novel inputs in terms of their similarity to items in the corpus. But unlike the analogy model, hidden representations were shaped by similarities among both input and output patterns in the corpus, as well as the relationships between inputs and outputs.

In both models, input gain is a multiplicative scaling parameter on the net inputs to units, be they localist nodes or hidden units. The current simulation results show that the modulation of input gain at testing caused similar effects in both models. At low levels of input gain, both models failed to map irregular items to their appropriate outputs, but succeeded in mapping regular items and novel inputs. At high levels of input gain, both models succeeded at mapping both regular and irregular items, but performed poorly with novel inputs.

The reason why input gain caused this double dissociation was different for each model. In the analogy model, input gain modulated the intensity of competition for activation among localist nodes. Low levels of competition caused outputs to be based on the summed contributions from many partially activated nodes. Regularities across nodes were extracted in these summations to the point of overriding any exceptions to the regularities. By contrast, high levels of competition caused a winner-take-all mode of processing in which a known input correctly activated its corresponding node, whereas a novel input incorrectly activated a node corresponding to a similar, known item.

In the distributed model, input gain modulated the sharpness of a sigmoidal activation function. Low levels of input gain caused hidden units to operate mostly in their linear range, thereby emphasizing the componential (i.e., regular) relationships that were learned between inputs and outputs. High levels of input gain caused hidden units to operate mostly in their asymptotic range, thereby emphasizing the conjunctive relationships that were learned between inputs and outputs (for a discussion of

componential and conjunctive coding, see O'Reilly, 2001). Componential relationships supported only the processing of regular and novel items, whereas conjunctive relationships supported only the processing of known items.

Simulation Methods

Input and Output Representations were constructed from a 12 dimensional binary space. Out of $2^{12} = 4096$ possible input patterns, one fourth (1024) were chosen at random to constitute the corpus of items. Each chosen input pattern was associated with one output pattern. Output patterns were created in two steps. First, each input pattern was copied to its corresponding output pattern (i.e., the identity mapping. Note, however, that the results apply to all linearly separable mappings). Second, the bit value of each dimension, for each output pattern, was flipped with a 5% probability. Thus, the identity mapping was a regularity, and flipped values were exceptions to that regularity. This procedure resulted in 563 fully regular items (no flipped bits), and 461 irregular items with one to four flipped bits per item. The 3072 remaining patterns served as novel items during testing.

For the analogy model, there were 12 input units corresponding to the 12 input dimensions, and dimension values were coded as activations of ± 1 on the inputs. For the distributed model, there were 24 input units, half of which coded the 12 dimension values as activations of 0 or 1. The other half were activated as flipped values of the first half, i.e., $1-x$, where x was each of the first 12 activations. The $x|1-x$ coding scheme was used because the distributed model was trained via backpropagation (this scheme was not necessary in the analogy model because it was not trained; see next two sections). In backpropagation, no learning will occur on a unit's sending weights when the activation value of that unit is zero. Therefore, the $x|1-x$ coding scheme ensured that weight derivatives were generated for every input dimension, on every training episode.

For both models, there were 12 output units corresponding to the 12 output dimensions, and dimension values were coded as targets of 0 or 1 on the outputs.

Analogy Model Architecture. In the analogy model, input units were fully connected to 1024 "logogen" units. Each logogen represented one item in the corpus, and the weights on incoming connections from input units were set according to each logogen's input pattern, i.e., +1 weights for positive input dimensions, and -1 weights for negative dimensions. Each logogen projected outgoing connections to all 12 output units, and the weights on outgoing connections were set according to each logogen's output pattern (as for incoming connections).

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

$$a_j = e^{\gamma I_j} / \sum_i e^{\gamma I_i},$$

where I was the net input to a unit, 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 logogens. Each output unit was then calculated as the sigmoid of the dot product between the logogen vector and its incoming weight vector. Noise was included to break perfect ties between very small (e.g., two or three) numbers of activated logogens. Such ties occurred more often at high levels of input gain.

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. However, results were very similar over a range of hidden unit numbers. Hidden units were calculated with the hyperbolic tangent function,

$$a_j = \tanh(\gamma I_j),$$

which is analogous to the logistic, except it has asymptotes at ± 1 instead of 0 and 1. 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 analogy model) during both training and testing. Output units were calculated as in the analogy model.

Connection weights were initialized to random values in the range ± 0.1 , and weights were learned by gradient descent,

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

where w_{ij} was the connection weight from unit j to i , η was the learning rate (fixed at 0.001), and E was cross-entropy error (Rumelhart et al., 1995). Weight changes were made each time after weight derivatives had been accumulated over all 1024 items in the corpus. Weight derivatives were calculated for each 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 backpropagated to generate the weight derivatives. Weight updates were repeated until every output unit was within 0.1 of its target for every item in the training corpus. This criterion was reached after 3000 passes through the corpus.

Testing Procedure. For both models, performance was assessed on each test item by setting the input units to the item's input pattern, and then determining whether the activation of each output unit was within 0.5 of its target (which was either 0 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 3072 novel items were set according to each item's input pattern, i.e., the identity mapping.

To dissociate item-based and regularity-based processing, input gain was varied as a single control parameter over the logogen units in the analogy model and over the hidden units in the distributed model. The reported levels of input gain were between 0.5 and 3 for the analogy model, and 0.333 and 3 for the distributed model. These ranges were chosen to show asymptotic performance at the lower and upper ends, i.e., the patterns of behavior did not change substantially beyond these ranges.

Simulation Results

Mean accuracies for the analogy model are graphed in Figure 1 as a function of input gain and item type (regular, irregular, or novel). The same are graphed for the distributed model in Figure 2.

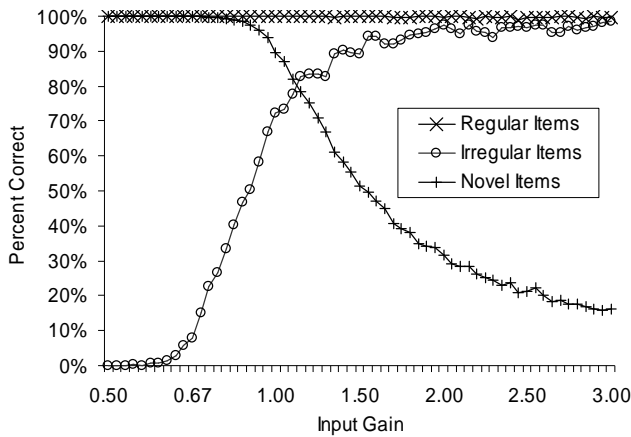


Figure 1: Mean accuracies for the analogy model

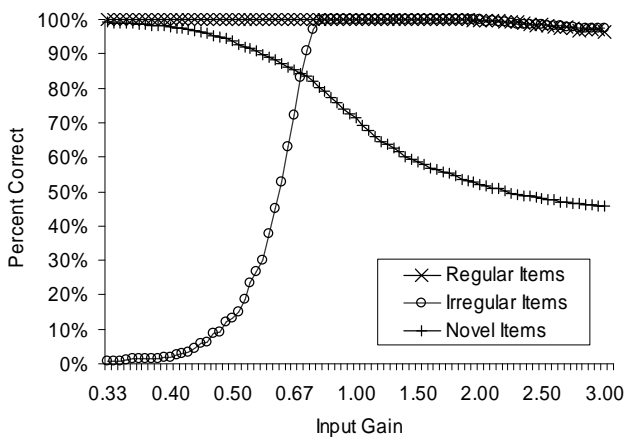


Figure 2: Mean accuracies for the distributed model

Figures 1 and 2 show that both models exhibited a clear dissociation in performance on irregular items compared with novel 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 can be

considered as a *regularization* error because, for the quasi-regular domain constructed here, the identity mapping is the regular mapping.

At high levels of input gain, performance on all items in the corpus was near perfect in both models. By contrast, mean accuracies for the novel items dropped to as low as 16% for the analogy model, and 46% for the distributed model. Of all the analogy model’s erroneous responses to novel items at the highest level of input gain, 97% were output patterns that corresponded to output patterns in the training corpus. These responses can be considered as *lexicalization* errors because they are responses for other items in the model’s “lexicon”. The same analysis of errors made by the distributed model showed only 27% lexicalization errors (where the chance rate was 25%).

These results show that the manipulation of input gain as a single control parameter, over a single level of representation, caused a clear double dissociation in both models. To better understand the similarities and differences in processing between these models, three visualizations of the input-output mappings for each model are shown in Figure 3.

In each visualization, all 4096 points in the 12 dimensional input space are arranged on a grid such that all adjacent vertices differ by only one bit. To illustrate, near the lower left-hand corner of each plot is the vertex where all 12 input dimensions are negative. The next vertex up and the next vertex to the right each have one positive input dimension, and so on. Each grid “wraps around” such that vertices on the left edge are adjacent to the corresponding vertices on the right edge, and likewise for the top and bottom edges. Thus, the 2D space of each grid represents a portion of the similarity structure in the 12D input space. In addition, 10 evenly spaced points are interpolated in each space between each pair of vertices. Given that each side has 64 vertices ($64^2 = 4096$), there are $640^2 = 409,600$ points of the input space represented in each plot.

At each point, a gray scale value is plotted that represents the summed activation of four output units for the corresponding input pattern. The same four output units (chosen arbitrarily) are shown at all points in all plots. The gray scale values are calculated such that, the darker the point, the closer the outputs were to 0.5. Conversely, whiter points indicate where the outputs were at their asymptotes (0 or 1). Thus, the dark borders in each plot represent the decision boundaries in each model, that is, where one or more of the four outputs crossed the middle point between asymptotes as a function of change in the input space.

Plots are shown for each model, at three different levels of input gain: the low end (0.5 in the analogy model and 0.333 in the distributed model; top row), the high end (3 in both models; bottom row), and the point at which accuracies for irregular items and novel items are equal (1.1 in the analogy model and 0.8 in the distributed model; middle row). Overall differences in plot densities for the analogy model, compared with plot densities for the distributed model, were due to differences in the polarity of the output

units: outputs in the distributed model tended to be closer to 0 or 1, i.e., values that corresponded to white points on the plots.

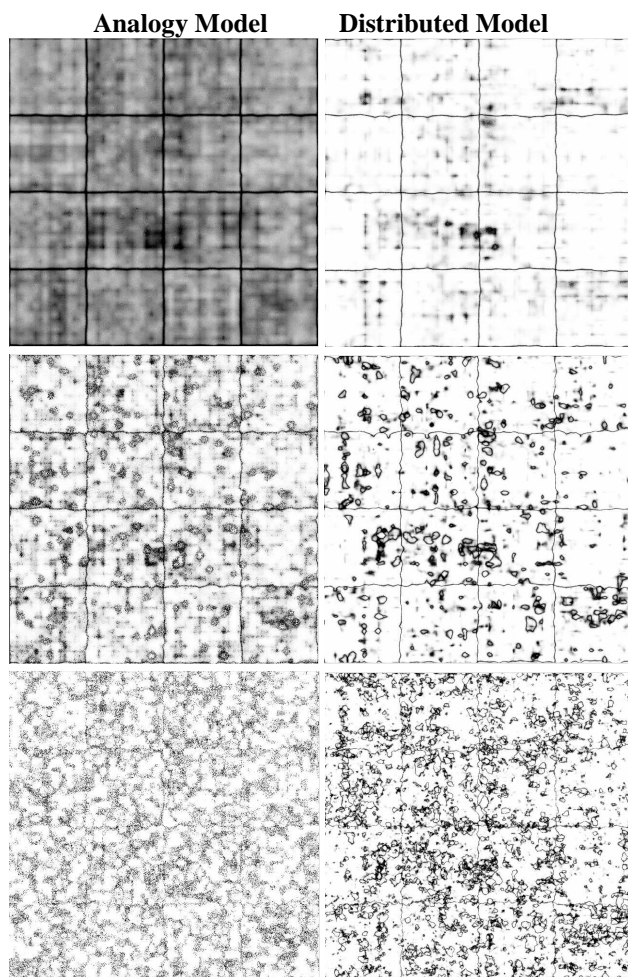


Figure 3. Visualizations for each model at low (top), medium (middle), and high (bottom) levels of input gain

The grid patterns seen in the top two plots of Figure 3 show that both models processed the identity mapping at the low end of input gain. In fact, if all 12 outputs had been represented, each plot would show a 64 by 64 grid pattern, where the grid lines fall exactly between the vertices. Thus, the grid reflects the finding that, at low input gain, the identity mapping was generalized to all inputs, including those for novel and irregular items. The grid is a depiction of regularity-based processing in each model because the identity mapping was the regularity in our quasi-regular domain.

The middle two plots 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 “warping” of the identity mapping that was necessary to process the irregular items.

Moreover, given that mean accuracies were about 80% for novel items as well, one can infer that these distortions and pockets were mostly isolated to the irregular items. These plots show that a balance was struck at moderate levels of input gain between item-based and regularity-based processing.

The bottom two plots 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 analogy model. Recall that, at the high end of input gain, 97% of the errors for novel items were lexicalizations. What this means is that the pockets show where known inputs were mapped correctly, and where novel items were mapped incorrectly to similar known items. These “item pockets” are a depiction of item-based processing in the analogy 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%). It appears that the distortions needed for accurate mappings of irregular items had “spread out” 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 spread out over untrained (novel) regions of the space more than they did over trained (known) regions. It is this selective spread of decision boundaries that indicates item-based processing at the high end of input gain.

Conclusions

The current simulations provide a new demonstration of how double dissociations can occur without separable processing components (see also Devlin & Gonnerman, 1998; Juola, 2000). 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.

It is important to acknowledge that the current work only opens the door to an alternative to the rules/lexicon and phonology/semantics explanations of double dissociations. It is unclear whether input gain would provide a satisfying account of specific empirical results. For instance, input gain would not appear to handle dissociations in which all regular items, both novel and known, are impaired (Marslen-Wilson & Tyler, 1997, 1998; Ullman et al., 1997). Also, the current simulations did not include subregularities or variations in the frequency of items. These factors have been simulated successfully (Kello, Sibley, & Plaut, submitted), but only as demonstrations. Subregularities allowed for model errors that were more like patient errors, but further work is necessary to test the simulated errors.

The current simulations also raise a number of larger questions, such as: Are there any testable differences between the analogy and distributed models presented here? Do these simulation results have implications for current

theories of word reading and inflectional morphology? Are the reported models consistent with the localization of regularity-based and item-based processing in the brain, to the extent that evidence exists for such localization? What might be the neural bases of input gain? These and other questions await further research.

Acknowledgments

This work was funded in part by NIH Grant MH55628, and NSF Grant 0239595. The computational simulations were run using the Lens network simulator (version 2.6), written by Doug Rohde (<http://tedlab.mit.edu/~dr/Lens>). We thank David Plaut for his input on precursors to this work.

References

- Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, *90*, 119-161.
- Behrmann, M., & Bub, D. (1992). Surface dyslexia and dysgraphia: Dual routes, single lexicon. *Cognitive Neuropsychology*, *9*, 209-251.
- 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.
- Funnell, E. (1983). Phonological processes in reading: New evidence from acquired dyslexia. *British Journal of Psychology*, *74*, 159-180.
- Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading, and dyslexia: Insights from connectionist models. *Psychological Review*, *163*, 491-528.
- 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.
- 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. (2003). Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language*, *48*, 207-232.
- Kello, C.T., Sibley, D.E., & Plaut, D.C. (submitted). Dissociations in performance on novel versus irregular items: Single-route demonstrations with input gain in localist and distributed models. *Manuscript under review*.
- Marslen-Wilson, W.D. & Tyler, L.K. (1997). 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 Science*, *2*, 428-436.
- McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences*, *6*, 465-472.
- Nakisa, R., Plunkett, K. & Hahn, U. (2000). Single- and dual-route models of inflectional morphology. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches*. New York: Oxford University Press.
- O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199-1241.
- 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.
- Plaut, D. C. & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*, 445-485.
- Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. New York: Basic Books.
- Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456-463.
- Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In C. Yves & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures, and applications* (pp 1-34). Hillsdale, NJ: Lawrence Erlbaum.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533-536.
- Rumelhart, D. E., McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart, McClelland, J. L., and The PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216-271). Cambridge, MA: MIT Press.
- Ullman, M. T., Corkin, S., Coppola, M., Hickok, G., & et al. (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.