Journal of Memory and Language, to appear.
Revised April 19, 2000, corrected June 13, 2000.

Testing Assumptions in Computational Theories of Aphasia

Wheeler Ruml, Alfonso Caramazza,
Jennifer R. Shelton, and Doriana Chialant
Harvard University

Abstract

We present the performances of thirteen aphasic patients on a picture-naming task, and attempt to model these data using computer simulations. We systematically manipulate the assumptions underlying several interactive, two-step, spreading-activation models, including the proposals of Dell, Schwartz, Martin, Saffran, and Gagnon (1997), Foygel and Dell (1999), and Rapp and Goldrick (in press). Using a numerical regression procedure and multiple views of each model's possible output, we find that peripheral pragmatic assumptions play a role equal to that of theoretically more central model components. None of the models we consider can account for all of the patients, leading us to conclude that one or more of the assumptions underlying each model is flawed. We argue that there are strong limitations on the conclusions that can legitimately be drawn from such simulation studies, but that close analysis of individual patients can allow sound testing of potentially more accurate models.

Keywords: computational modeling, aphasia, lexical access, computational neuropsychology.

(Caution: This is a preprint.
Please check any quotations against the published version.)

Wheeler Ruml, Division of Engineering and Applied Sciences, Harvard University; Alfonso Caramazza, Jennifer R. Shelton, and Doriana Chialant, Cognitive Neuropsychology Laboratory, Harvard University.

We would like to thank Gary Dell, Randi Martin and two anonymous reviewers for their many helpful comments, Nadine Martin for help in scoring some patient responses, Brenda Rapp and Matthew Goldrick for providing detailed information regarding their model's lexicon, Angelos Kottas for running some preliminary experiments, and Michele Miozzo and the Harvard Cognitive Neuropsychology Laboratory for many stimulating discussions regarding this research. Support was provided in part by the National Science Foundation under grants CDA-94-01024 and IRI-9618848, and by the National Institutes of Health under grant NS-22201.

Please address correspondence concerning this article to Wheeler Ruml, `ruml@eecs.harvard.edu`, Maxwell Dworkin Laboratory, Harvard University, 33 Oxford Street, Cambridge, MA 02138.

The promise of computational models of human language processing is widely recognized. Not only does the act of constructing a simulation force one to specify one's theory precisely, but the resulting model can be quantitatively tested against empirical data. Furthermore, the ease of simulation allows one to experiment with models that deviate from normal behavior, and thereby form theories about the interactions between brain damage and language processing. Data from aphasic patients can then be used to test the adequacy of the combined model of normal processing and damage in aphasia. One could even imagine using simulation results to provide insight into the breakdown occurring in specific patients. Examples of recent computational investigations of low-level language processing include the word reading models of Plaut, McClelland, Seidenberg, and Patterson (1996) and Shallice, Glasspool, and Houghton (1995) and the word production models of Levelt, Roelofs, and Meyer (1999) and Dell, Schwartz, Martin, Saffran, and Gagnon (1997).

In practice, a computational model is often constructed with the aim of testing claims about one or two specific theoretical issues, such as the role of interaction between levels of representation during word production. But the collection of theoretical principles that one wishes to put to empirical test does not usually describe a complete mechanism suitable for simulation. Details beyond the scope of the theory at stake must be filled in, such as the exact semantic relations between words in the model. And details supposedly within the purview of the theory must often be left out for the sake of reducing computation time, such as the full inventory of a typical human lexicon. In this paper, we systematically examine the role played by these seemingly minor assumptions by evaluating three closely related models of word production. We present data from thirteen aphasic patients on a picture naming task, and attempt to account for their performance using each of the three models. By manipulating both the minor assumptions of the models, as well as those that are usually interpreted as corresponding to important theoretical claims, we also generate a range of models along the spectrum spanned by the original three.

We use as a starting point the model proposed by Dell et al. (1997), which assumes that, in fluent aphasics, brain damage affects all parts of the lexical access system equally. Ruml and Caramazza (2000) showed that this model involving global damage has difficulty accounting for some of Dell et al.'s patient data, and we will see that it cannot match two of the thirteen patients we present here. Given this mismatch, we construct two variations of the theory which assume more localized damage. One of these is similar to the recent proposal of Foygel and Dell (1999). However, we will see that this model too has difficulty accounting for patient data. This suggests that some of the model's more basic assumptions require modification. Guided by proposals of Rapp and Goldrick (in press), we systematically construct five new models of aphasic naming, and test the ability of each one to fit our patient data. Although a hybrid model in between the proposals of Foygel and Dell and Rapp and Goldrick seems promising, none of the models we consider can account for all of the patients. Our investigation highlights the dramatic effect that theoretically peripheral assumptions can have on model performance, and we conclude that great care must be taken when attempting to draw theoretical inferences from simulation studies.

A Model of Global Damage

We begin with the model of aphasic naming proposed by Dell et al. (1997). As we noted, models of aphasia consist of a full model of normal processing with additional supple-

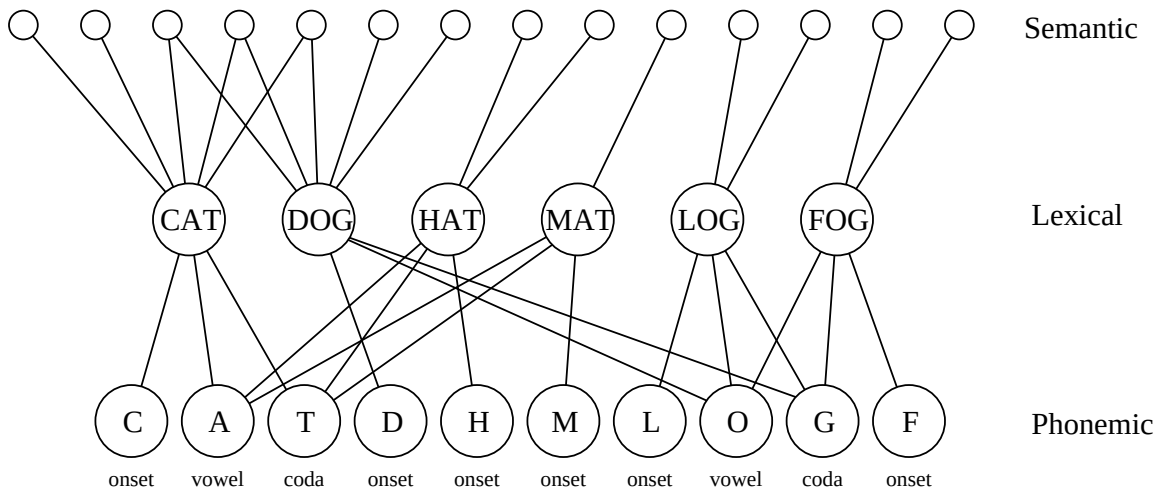

Figure 1. The structure of part of the core model.

mentary assumptions regarding the reaction of the system to brain damage. A clear division between the two is necessary in order for any of the insights gained through simulation of aphasic behavior to help constrain theories of normal processing. We will discuss Dell et al.'s model of normal processing in some detail, since it will serve as a core for most of the other theories we consider in this paper. The other theories differ mainly in their approach to modeling damage to the normal system.

The Model of Normal Performance

The core model of lexical access is based on notions taken from the theory of Gary Dell (1986). The undamaged system is postulated to consist of three levels of interacting representations: semantic, lexical, and phonological. Each level interacts only with the adjacent levels. The model is connectionist in style, and its structure is indicated in Figure 1. The representations in each level are represented by *activation values*, which are fractional numbers. These values are updated during processing according to the activation levels of neighboring representations, the decay of the original activation, and the influence of random noise. More precisely, if $a_t(m)$ represents the activation level at time t of a representation m which interacts with a set of neighbors N, and $R(x)$ represents a random sample drawn from the normal distribution with mean zero and standard deviation x, and *decay*, *connection*, *intrinsic*, and *activation* are parameters of the model, then

$$a_{t+1}(m) = old + incoming + noise$$

where

$$\begin{aligned} old = &\quad (1 - decay) \times a_t(m) \\ incoming = &\quad \textstyle\sum_{n \in N} \max(0, (connection \times a_t(n))) \\ noise = &\quad R(intrinsic) + (R(activation) \times a_t(m)). \end{aligned}$$

Negative levels of activation can exist at nodes, due to noise, but negative values do not influence neighboring nodes. Since the sum of all the activation in the network is not

Table 1: Parameters in the theory of normal naming performance and their default values.

Parameter	Description	Value
Nodes	the number of nodes in the semantic, lexical, and phonological layers	57, 6, and 10
Connectivity	the nodes each node connects to	see Figure 1 and text
Connection strength	the coefficient by which a given node's neighbors' activation levels are multiplied during spreading	0.1
Decay rate	the coefficient by which a given node's activation is multiplied during spreading	0.5
Semantic jolt	the activation level to which semantic nodes are set to represent the model's input	10
Lexical jolt	activation level to which the selected lexical node is set	100
Spreading steps	the number of time steps for which activation is spread through the network before lexical selection or phoneme selection takes place	8
Intrinsic noise	standard deviation of the distribution of activation-independent noise	0.01
Activation noise	standard deviation of the distribution of noise that is proportional to a node's activation level	0.16

Note: the four noise and jolt parameters can be scaled together without changing the behavior of the network, and so represent three rather than four true parameters.

constant, activation is best viewed as an attribute of each representation, rather than as something which is routed through the network like a fluid flow. The activation levels of representations in each layer are affected by representations in all adjacent layers because each connection is bidirectional: if node a can send activation to node b, then b can also send activation to a.

Lexical access is simulated in the model by setting the activation level of the semantic representations associated with the target word to a predefined constant value. The activation levels of all representations in the network are then updated for eight time steps. After this time, the activation level of the most active lexical representation is raised to a predefined high level, corresponding to the notion of lexical selection. Activations are further propagated for another eight time steps, after which the most active onset, vowel, and coda phonemes are chosen as the output of the model. In this way, the model incorporates the idea of two separate selection steps, one lexical, the next phonological. A summary of the parameters of the model is given in Table 1.

Because of the contribution of noise to the activation levels, the selected lexical rep-

Table 2: The lexicon of Dell et al.'s naming model.

Network 1		Network 2	
Word	Relation to Target	Word	Relation to Target
cat	target	cat	target
dog	semantic	dog	semantic
mat	formal	mat	formal
hat	formal	rat	mixed
log	unrelated	log	unrelated
fog	unrelated	fog	unrelated

resentation does not necessarily correspond to the target word, and even if it does, the phonological representations that are eventually selected are not necessarily those associated with the selected lexical node. By simulating many complete trials, one can accumulate an estimate of the probabilities of various kinds of responses. Dell et al. (1997) categorize errors as semantic (related to the target in meaning), formal (phonologically related to the target), mixed (both semantic and formal), unrelated (but a true word), and nonword (or gibberish). The distribution of the model's responses can be compared to the distribution measured from a human experimental participant to determine whether the model represents a mechanism sufficient to summarize human behavior.

Dell et al. followed a principled method for constructing the network and assigning the connectivity pattern among the representations. To reduce the computational burden of simulation, they chose a lexical layer of six words, and distinguished a particular word as always playing the role of the target. (This amounts to an assumption that all targets lie in typical semantic and phonological neighborhoods.) All words have a simple consonant–vowel–consonant structure. The phonological layer follows immediately, being determined by the lexical nodes and English pronunciation. To construct the semantic layer, Dell et al. stipulate that all lexical representations are associated with ten semantic features, and that semantically related words share three of their features.

To provide some assurance that this small network captures some of the relevant properties of English, Dell et al. measure its *error opportunities*, that is, the probability that a response of each possible type would result from selecting a random phoneme in the model at each position. This represents the model's behavior when it is generating random phonologically-legal strings of phonemes.[1] They compare this distribution to an estimate of random errors among English-speaking aphasic patients. In order to approximate the small frequency of mixed errors, Dell et al. actually set up two separate network simulations, only one of which contains a possible mixed error. The lexicons of the two networks are shown in Table 2. Only one out of every ten trials uses the network with the mixed error, while the other network is used the remainder of the time. This dual-network approximation of a typical lexicon allows the possibility of an occasional mixed error. Although they do not statistically test the similarity, Dell et al. are able to obtain error opportunities for their simulation that are in general accord with their estimates for English. They also show that

[1] Phonological legality is assured due to the simple CVC structure of the entire lexicon.

the model can generate an error distribution similar to that of control participants in a picture-naming task.

Assumptions Regarding Damage

Since the model's error distribution depends on its parameters, changing the parameters can cause the model to emit errors following a different distribution. Dell et al. (1997) propose what they call the *globality assumption*: that the damage to the lexical access system in aphasia can be modeled as changes to their simulation's *connection* and *decay* parameters. By changing these parameters throughout the network, this model of brain damage embodies the claim that the variety of patterns of errors observed in aphasic patients can be explained by uniform damage to all parts of the system simultaneously. Dell et al. propose that fluent aphasic patients can be modeled by finding, for each patient, values for *connection* and *decay* that cause the model to match that patient's error distribution. Although Dell et al. also propose a related model of single-word repetition performance, in this paper we will focus on error patterns during picture naming.

Patient Data

To allow the testing of Dell et al.'s theory using additional data beyond that presented in their original paper, we gathered data from thirteen fluent aphasic patients. They were recruited from area clinics and hospitals and agreed to participate in a research study. Each patient was compensated $10/hour for participation. Patients with dysarthria of speech or production deficits resulting in unintelligible jargon were excluded from participation. All the patients included in this study suffered a left cerebrovascular accident. Patient E.A. has been previously reported by Shelton and Weinrich (1997) and patient I.O.C. by Shelton, Fouch, and Caramazza (1998). The group included three different subjects with the initials J.R.

The patients were given a screening battery (Harvard Cognitive Neuropsychology Laboratory Screening Battery, unpublished) designed to assess a wide variety of language skills.[2] Background information for each patient is provided in Table 3, along with the percentage of correct responses on tests measuring auditory comprehension, phrase repetition, and object naming. These tests involve the use of simple stimuli and are very similar to tests found on clinical assessment tools such as the Boston Diagnostic Aphasia Exam (Goodglass & Kaplan, 1983).

Each patient was administered the Philadelphia Naming Test (P.N.T.) (Roach, Schwartz, Martin, Grewal, & Brecher, 1996) in a similar manner. Pictures were presented one at a time and patients were asked to name them and were encouraged to provide a response (i.e., they were encouraged to respond, even incorrectly, rather than say "I don't know"). Patients were given ten practice pictures prior to the test trials. For each patient, naming data were tape-recorded and scored for accuracy and error type following the session. The number of sessions a patient took to complete the 175 test pictures for naming varied from one session to four sessions. Patients who had a great deal of difficulty naming

[2] E.A.'s scores are from sections of the Boston Diagnostic Aphasia Exam. In previous work, E.A. has been described as nonfluent, based on the number of words he produces in conversational speech and discourse production tasks. However, he does not have any articulatory problems and he rarely makes distortion errors, so he would be classified as fluent according to Dell et al.'s criteria.

Table 3: Patient background information and correctness on certain language tasks.

Patient	Age†	Education†	Post Onset†	Auditory Comp.	Phrase Rep.	Object Naming
A.B.	73	12	3	1.0	1.0	1.0
E.M.	65	> 16	2.5	1.0	.80	1.0
T.H.	63	13	17	1.0	–	–
R.C.	62	12	4	1.0	.61	.92
L.S.	52	16	3	1.0	1.0	.90
J.R.3	59	16	3	1.0	.92	.62
L.T.	53	16	8 mo.	1.0	–	.90
M.M.	72	16	2	1.0	–	.70
J.R.2	43	16	5	1.0	–	.77
P.C.	51	15	11 mo.	.92	.40	.80
J.R.1	50	12	3	–	–	–
E.A.‡	65	15	18	.81	.25	.47
I.O.C.	55	14	1.5	1.0	–	.10

†in years. ‡from the B.D.A.E. – not available.

tended to become frustrated after approximately forty to fifty trials and we discontinued testing for that day when their frustration became apparent.

The scoring procedures for naming were identical to those described by Dell et al. (1997).[3] An answer was considered correct only if it matched the pictured item perfectly (e.g., 'ape' was not considered correct for 'monkey'). Errors were scored according to the criteria established by Dell et al. and could be semantic (e.g., bowl $\rightarrow$ cup), formal (e.g., cane $\rightarrow$ cab), mixed, both semantically and phonologically related to the target (e.g., foot $\rightarrow$ finger), unrelated (e.g., clown $\rightarrow$ house), or nonword (e.g., pyramid $\rightarrow$ kuramids). 'Other' errors included descriptions or definitions of the item (e.g., tractor $\rightarrow$ farmers use it), no response errors, visual errors, naming parts of objects (e.g., nurse $\rightarrow$ a particular dress that girls wear) or any other error type that did not fit in the 5 categories described above.

Patient Performance.

Table 4 summarizes each patient's responses on the P.N.T. (The data are presented in percentage terms in later tables.) Inspection of the data reveals a wide range in performance, from near perfect (A.B.) to quite poor (I.O.C.). The large number of 'other' errors for I.O.C., J.R.1, and E.A. result from these patients' tendencies to provide descriptions of items that they could not name. Almost all the 'other' errors for these patients were descriptions. In fact, the majority of 'other' errors for all patients involved descriptions of the items they could not name (e.g., hose $\rightarrow$ a thing you use to spray water with).

There are two obvious differences between the behavior of our patients and those reported by Dell et al. First, several of our patients did not make any formal and/or nonword errors. The patient with the highest frequency of nonwords was R.C., with 7 of 169 codable responses (4%). Second, there are several patients (T.H., J.R.3, M.M.)

[3]Nadine Martin kindly helped score several difficult responses.

Table 4: Picture naming performance of the patients on the Philadelphia Naming Test.

Patient	Naming response						
	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Other
A.B.	167	3	0	1	0	2	2
E.M.	165	2	0	0	0	4	4
T.H.	161	10	0	3	0	0	1
R.C.	148	10	0	3	1	7	6
L.S.	147	14	0	2	0	1	11
J.R.3	146	17	0	7	1	0	4
L.T.	144	11	3	4	0	2	11
M.M.	136	18	0	4	0	0	17
J.R.2	117	13	1	5	0	5	34
P.C.	101	19	8	8	22	4	13
J.R.1	79	6	1	2	0	0	87
E.A.	59	39	6	14	19	2	36
I.O.C.	29	6	0	1	0	0	139

Table 5: The mixed errors of T.H., J.R.3, and M.M.

Patient	Target		Response
T.H.	plant	→	flowers
	microscope	→	magnifying glass
	ruler	→	tape measure
J.R.3	plant	→	flowers
	vest	→	suit
	crown	→	queen
	boot	→	foot
	scale	→	weigh
	ruler	→	measure
	garage	→	car
M.M.	skull	→	skeleton
	ghost	→	goblin
	pear	→	peach
	garage	→	carport

Table 6: Picture naming performance of the patients, scored according to Mitchum et al. (1990).

Patient	Naming response							
	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Misc.	No Resp.
A.B.	167	4	0	0	0	2	2	0
E.M.	165	4	0	0	0	2	3	1
T.H.	161	13	0	0	0	0	1	0
R.C.	148	13	0	0	1	7	3	3
L.S.	147	16	0	0	0	1	9	2
J.R.3	146	24	0	0	1	0	4	0
L.T.	144	14	1	1	2	2	9	2
M.M.	136	22	0	0	0	0	17	0
J.R.2	117	18	0	1	0	5	29	5
P.C.	101	28	0	0	29	4	13	0
J.R.1	79	7	0	1	1	0	86	1
E.A.	59	49	0	4	25	2	30	6
I.O.C.	29	6	0	1	0	0	136	3

who made mainly semantic errors, with some mixed errors (which bear a semantic and phonological relationship to the target), no formal or nonword errors, and few unrelated errors. The error pattern cannot be attributed to level of correctness since T.H. performed quite well (making only 14 errors), whereas M.M. performs much worse (making 39 errors). For M.M. and J.R.3, all the 'other' errors were fairly detailed circumlocutions, and T.H.'s one 'other' error was naming of a part of the picture (bride $\rightarrow$ veil). Although these patients did commit some 'mixed' errors, a closer examination of these errors reveals how little phonological overlap many of them shared. Table 5 presents the mixed errors from these three patients and demonstrates that the target and error shared little phonological overlap. The errors are classified as mixed errors according to Dell et al.'s scoring system, which measures phonological overlap as "...target and error started or ended with the same phoneme; had a phoneme in common at another corresponding syllable or word position, aligning words left to right; or had more than one phoneme in common in any position (excluding unstressed vowels)" (p. 809).

Other scoring systems define phonological similarity as having 50% overlap between target and response, and it is often the convention when scoring naming errors that phonological errors share a proportion of overlap greater than one phoneme. To assess the effect of this phonological criterion, we rescored the patient responses using the system of Mitchum, Ritgert, Sandson, and Berndt (1990). As shown in Table 6, T.H., J.R.3, and M.M. make few, if any, mixed errors under the stricter criterion. In particular, T.H. would have made almost only semantic errors. His three mixed errors share very little phonological overlap (plant $\rightarrow$ flowers, sharing only $|l|$ in the second position; microscope $\rightarrow$ magnifying glass, sharing only $|m|$ in the first position; ruler $\rightarrow$ tape measure, sharing only $|er|$ in the final position) and his single 'misc' error, as we mentioned, could be interpreted as semantic in nature (bride $\rightarrow$ veil). The same is true for both J.R.3 and M.M., although these patients

Table 7: Verb responses during picture naming.

Target		Response
rake	→	sweep
toilet	→	flush
bed	→	sleep
ear	→	hear
skis	→	skiing
letter	→	write
bench	→	sit down
scissors	→	cutting
pillow	→	sleep
cane	→	limping
grapes	→	eat
fork	→	eat

made a number of 'other' errors as well (especially M.M.).[4]

Another feature of the patients' responses is the surprisingly large number of times patients produced a verb in response to the noun targets (see also Berndt, Haendiges, Mitchum, & Sandson, 1997). This happened especially frequently with E.A. (11 verb responses in 175 targets), but was also noted in responses from J.R.3 (three verb responses) and J.R.1 (one verb response). Some examples of these errors are shown in Table 7. These verb responses are surprising because Dell et al.'s theory models brain damage as abnormal values for parameters within the lexical network, and syntactic processes are assumed to operate normally. In particular, selection of lexical nodes is assumed to respect syntactic class, and non-noun responses can arise only indirectly, via errors during phoneme selection. (Dell et al. (1997) used this assumption to derive predictions regarding overabundance of nouns among patients' phonological errors, but did not address how their theory might account for a dearth of nouns.) Of course, these responses from our patients are either semantic or mixed errors, and are likely have a semantic basis rather than a phonological basis. Although these verb responses seem to call into question the relevance of Dell et al.'s theory, such errors are still perfectly scorable, and we will disregard the grammatical class issue. There is another, perhaps more serious problem with the model's relation to the patient data.

Relating the Patients to the Model

Since Dell et al.'s model always produces three phonemes, its responses can always be scored, and it cannot simulate trials on which patients offer a description of the stimulus or decline to provide a response. This poses problems when evaluating the model's ability to account for the patient data, especially patients such as I.O.C. and J.R.1, for whom 79% and 50% of responses were in the 'other' category, respectively. Directly comparing the model's

[4]Many of the patients also completed repetition of the target items, including E.M., R.C., J.R.3, M.M., J.R.2, P.C., and E.A. All of the patients except R.C. repeated the items perfectly.

output to the patients' error patterns would cause it to fail any statistical test of fit, since it cannot account for events it cannot generate. If we assume that unscorable responses arise due to a mechanism in the lexical access system that is missing from the model, then we must find a way of determining which portion of a patient's performance, if any, the model should be expected to account for. (Patients with damage outside of the lexical access system would certainly be outside the scope of the model.) Without an explicit theory of the interaction of the hypothesized missing mechanism with the rest of the model, one might prefer to test the model using only those patients who make no unscorable responses, in the hope that the mechanism plays no role in those patients' behavior. Unfortunately, such patients are so rare that this would render the model useless. No such patients were among the twenty-three reported by Dell et al. or the thirteen presented here. So practicality demands that we somehow relate the model to error patterns containing 'other' responses.

Perhaps the simplest approach to dealing with the unscorable trials is to ignore them and test the model on its ability to match each patient's distribution over the remaining response categories. This assumes that the mechanism responsible for the 'other' responses is independent of the portions of lexical access that are implemented in Dell et al.'s model, in the specific sense that the distribution over the scorable response categories would be the same if the mechanism did not exist. (Of course, we are already assuming that the mechanism is not involved during the scorable trials themselves, since it is not part of the model.) The situation is as if there were a process that blocked a certain percentage of attempts at lexical access, on trials chosen randomly without regard for the response that would have been generated otherwise. By making this independence assumption, we can remove the unscorable trials without adjusting the remaining counts. (Alternatively, one could allow the model an extra free parameter which allowed it to perfectly match any patient's number of 'other' responses.)

While simple, the independence assumption may seem implausible. One might suspect that patients generate circumlocutions or fail to respond only in those cases in which they believe that they haven't correctly retrieved the desired word. If patients tend to be correct in their self-diagnosis, then their 'other' trials would have tended to be distributed over only the error response categories had the hypothesized mechanism not been present. The independence assumption will therefore ask the model to generate correct responses too frequently. However, the proper redistribution of counts seems difficult to determine, since we don't know whether patients are preferentially able to detect failures in situations that would otherwise lead to a certain type of error. Similarly, we don't know how providing the explicit instruction to provide a response, even if it may be incorrect, might influence the situations in which the mechanism comes into play. In the absence of these kinds of information, we will use the independence assumption and renormalize the patient data while ignoring unscorable responses.

To mitigate the effects of our ignorance, one might choose to disregard patients who make 'other' responses very frequently. If the truly correct reweighting of response frequencies were very different from that accomplished by the independence assumption, then the error patterns of patients who make many unscorable responses will be distorted. There is a chance that an incorrectly reweighted distribution will cause the model to fail to match a patient whose true distribution of errors (in the absence of the 'other' mechanism) would have been successfully fit. In view of this possibility, we will follow Dell et al. and seg-

regate those patients whose 'other' frequencies are greater than 15% (J.R.2, J.R.1, E.A., and I.O.C.). We will include them in our experiments, but present analyses both with and without them.

Evaluation of the Global Damage Model

To test the ability of Dell et al.'s model of global damage to account for the patients, we need to find, for each patient, those values of *connection* and *decay* that cause the model's error distribution to come as close as possible to that patient's. By 'close,' we will mean having the greatest chance of having been sampled from the same underlying multinomial distribution. This can be measured using the X^2 statistic. If P represents the distribution of a patient's responses over n possible categories, and M represents the distribution of the responses generated by the model, then

$$X^2(M,P) = \sum_{j \in \{M,P\}} \sum_{i=1}^{n} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

where

$$expected_{ij} = total_j \times \frac{observed_{iM} + observed_{iP}}{total}.$$

Its value can be compared to the distribution of χ^2 to derive a significance level.

We began this process by systematically varying the parameter values and running the simulation at each combination. We varied the value of *decay* from $\frac{1}{2}$ to 1, in increments of $\frac{1}{128}$, and the logarithm (base 10) of *connection* from -1 to -4, in increments of $\frac{1}{16}$ (corresponding to *connection* values from 0.1 to 0.0001). At each of the 4,225 combinations, we ran the simulation until the widest 95% confidence interval on any error type was 2%. This required from 200 to 9,600 trials, depending on the resulting distribution. Then, starting from the best combination we found, we used the numerical optimization algorithm of Ruml and Caramazza (2000) to fine-tune the fit. This automated regression procedure uses estimated confidence intervals to quickly identify promising parameter values. Although the stochastic nature of the simulation implies that it is impossible to guarantee that a particular fit is optimal, Ruml and Caramazza showed that their algorithm does at least as well as manual fitting of Dell et al.'s model.

The fits found by the algorithm are shown in Table 8. The table shows the error distribution of each patient, and underneath, the distribution produced by the simulation. The parameter settings used are shown underneath each patient's initials. The four patients who made many 'other' responses are grouped at the bottom of the table. The last two columns give the X^2 value of the fit and its significance, p, the probability that a fit with that X^2 value (or larger) would have occurred if the patient and model were following the same distribution.[5] We also present a second measure of fit, the root mean squared difference (RMSD). Given two probability distributions M and P over a set of n response types,

$$\text{RMSD}(M,P) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i - P_i)^2}.$$

[5] We assume five degrees of freedom. Since two parameters were estimated from the data, this is conservative and we may fail to reject some mismatches.

Table 8: Fits of the global damage model to the patients.

Patient and	Naming response						Fit		
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	p
A.B.	.97	.02	.00	.01	.00	.01			
conn .0138, *dec* .52	.96	.02	.00	.00	.00	.01	.004	1.6	.898
E.M.	.96	.01	.00	.00	.00	.02			
conn .0107, *dec* .50	.95	.03	.01	.00	.00	.02	.011	3.5	.618
T.H.	.93	.06	.00	.02	.00	.00			
conn .0931, *dec* .72	.93	.05	.00	.02	.00	.00	.006	1.2	.942
R.C.	.88	.06	.00	.02	.01	.04			
conn .0583, *dec* .71	.85	.08	.02	.02	.00	.03	.017	5.2	.395
L.S.	.90	.09	.00	.01	.00	.01			
conn .0806, *dec* .74	.91	.06	.00	.02	.00	.01	.012	2.6	.760
J.R.3	.85	.10	.00	.04	.01	.00			
conn .0937, *dec* .81	.84	.10	.01	.03	.00	.02	.013	8.4	.134
L.T.	.88	.07	.02	.02	.00	.01			
conn .0586, *dec* .70	.88	.07	.01	.02	.00	.02	.005	1.3	.939
M.M.	.86	.11	.00	.03	.00	.00			
conn .0898, *dec* .79	.86	.08	.01	.03	.00	.02	.014	5.3	.382
P.C.	.62	.12	.05	.05	.14	.02			
conn .0649, *dec* .77	.51	.12	.12	.03	.06	.17	.086	51.1	**.000**
J.R.2	.83	.09	.01	.04	.00	.04			
conn .0938, *dec* .81	.83	.10	.01	.03	.00	.03	.005	0.6	.987
J.R.1	.90	.07	.01	.02	.00	.00			
conn .0583, *dec* .70	.90	.06	.01	.02	.00	.01	.006	1.3	.932
E.A.	.42	.28	.04	.10	.14	.01			
conn .0898, *dec* .87	.46	.14	.14	.04	.06	.17	.104	77.8	**.000**
I.O.C.	.81	.17	.00	.03	.00	.00			
conn .0937, *dec* .81	.84	.10	.01	.03	.00	.02	.032	2.9	.711

Table 9: Summary of the performance of the global damage model.

Patient	VAF by category						Summary VAF			
Group	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Mean	Wtd.	$\sum X^2$	Failures
Ruml et al.	.83	.80	-1.5	.65	.59	-12	-1.8	.56	80	1/9 (2/13)
Dell et al.	.98	-.40	.91	.11	.55	.83	.49	.87	276	5/21
Combined	.97	.11	.85	.21	.56	.76	.57	.87	356	6/30 (7/34)

RMSD is perhaps more intuitive that X^2, as it correlates with the average difference between the probabilities of the model and patient in each category. Its value can range from zero through one, where zero implies identical distributions. Unfortunately, RMSD cannot be used as the basis of a goodness of fit test, the way X^2 can.

All but one of the nine patients (11%) were matched by the model (two out of thirteen (15%) if we include the four patients who made many unscorable responses). In most cases, the fits were quite good, with the smallest p of any successful fit 0.134. Although the RMSD for the fit to I.O.C. is relatively high, the small number of codable responses means that the differences between the distributions are not significant. Patients P.C. and E.A. pose problems for the model, however. The model's matches for the two patients exhibit many more formal and nonword responses than the patients do, and too few unrelated word responses. One might discount the model's failure to fit E.A., since that patient did not provide a codable response in 21% of the trials, but P.C.'s responses were uncodable only 7% of the time. It seems that these patients' combination of semantic and unrelated errors, appearing with a low nonword and formal error rate, cannot be generated by Dell et al.'s model of global damage in aphasia. These results are consistent with the analysis of Ruml and Caramazza (2000), who showed that five of the twenty-one patients reported by Dell et al. (1997) are inconsistent with the global damage model, and that the model always assumes a high frequency of nonword errors for any patient with low correctness.

Although testing the model's consistency with individual patients is important, it provides only one perspective on the model's performance. We might also want to quantify how well the model is accounting for particular categories of errors when we look across patients. For this task, we can use the 'variance accounted for' metric (VAF). Intuitively, VAF is a measure of the error of the model in comparison to merely predicting the mean of the patient data. Formally,

$$\mathrm{VAF}(M,P) = 1 - \frac{\sum(data - model)^2}{\sum(data - mean)^2}$$

A VAF of 0.5 would mean that the model's error is half of the error we would measure when guessing the mean, while a negative VAF would mean that the model is performing worse than guessing the mean. One would not expect to achieve a VAF of 1.0, due to the sampling error in the patient data, and the metric is perhaps best used to compare models, rather than to test the adequacy of a single proposal.

Table 9 presents the VAF of the global damage model for each category, on both our patients and those presented by Dell et al. (Fits to Dell et al.'s patients were taken

from Ruml and Caramazza (2000).) The four patients who made many 'other' responses were not included in the analysis. In addition, we can calculate summary VAFs, either by taking the mean of the category VAFs or by recalculating VAF using data from all categories (interpreting *mean* in the equation above as the mean of each category as appropriate). The first method weights all response categories equally, while the second implicitly weights each category according to its variance, with high-variance categories such as correct or unrelated receiving more weight than the formal or mixed categories. The table also indicates the sum of the X^2 values for the patients and summarizes the number of patients that the model was unable to match ($p < 0.05$, with figures in parentheses including the four patients who made many 'other' responses). As the rows of the table show, the model behaves very differently on the two groups of patients. Generally speaking, it performs worse on our patients that on those reported by Dell et al.. As might be expected from the use of X^2 as the fitting criterion, the model does best in the correct category, which is where the majority of patient responses fall. But for each patient group, there is at least one category in which the model does worse than guessing the mean. When considering all patients, the model seems to have the most trouble in the semantic and mixed categories.

Possible Error Patterns

One might wonder whether the patients that the model failed to fit, P.C. and E.A., are somehow extraordinary, and whether their performance on the P.N.T. is dramatically inconsistent with the results of patients tested by Dell et al. (1997). To address this question, we can look at the distribution of the two patient groups in the space of possible error patterns. We can also include points corresponding to the distributions that can be generated by the model, to give us a sense of where the patients lie in relation to the model. Of course, we cannot view the six-dimensional space of response distributions directly, so instead we will examine the error patterns two categories at a time. (This scatterplot technique was also used by Ruml and Caramazza (2000).)

Various combinations of categories are shown in Figure 2. The axes of each plot refer to particular response categories, and each mark on a particular plot refers to an error pattern containing responses of those types with the corresponding frequencies. The other categories of each pattern are not reflected in that plot. The small dots represent error patterns generated by the model during the systematic testing of the parameter settings and the subsequent fitting of the patients. Since we varied the parameters throughout their permissible ranges, the distribution of points should indicate the entire variety of patterns the model can generate. Note that, since the relationship between parameter values and the resulting error mix is not straightforward, many different parameter values can yield similar distributions, resulting in clumping in the figure. This is an artifact of the systematic sampling and does not necessarily reflect an important feature of the model. Only the boundary of the region containing dots is relevant for our purposes here. The performances of patients tested and reported by Dell et al. (1997) are plotted using circles. Patients we have reported in this paper are plotted using triangles, except for the four who made many unscorable errors, who are plotted using x's. Since model and patient points that are close together in a particular plot may represent error patterns that differ significantly in response categories that are not shown, these plots present a generous estimate of the coverage of the model. Outliers are labelled with the patients' initials to aid in matching

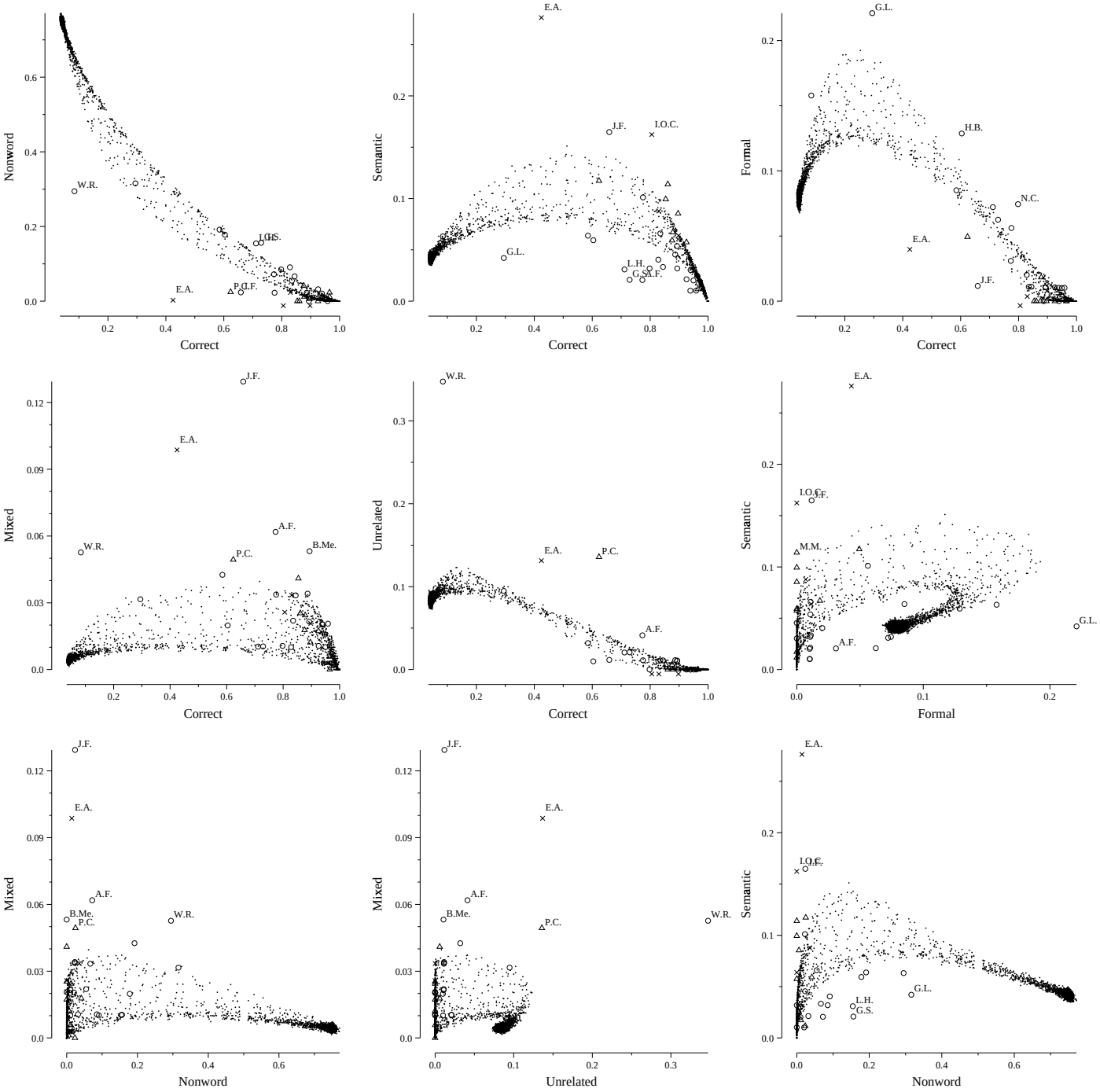

Figure 2. Possible combinations of error frequencies under the global damage model.

across the different projections.

In the upper left panel, for instance, we see that, while most patients have a high level of correctness, a few have levels of correctness of about 60% or less, including some patients reported by Dell et al.. While the model can only generate error mixes in which lowered correctness is accompanied by a proportional increase in nonword errors, several patients, including P.C. and E.A. as well as Dell et al.'s J.F. and W.R., do not follow this trend. Although P.C. and E.A. are outliers in many of the plots, they are joined by other patients reported by Dell et al., such as W.R., J.F., A.F., and G.L. While our patients represent important test cases, they can also be seen as fleshing out trends that would be only weakly suggested if we plotted Dell et al.'s data alone. These scatter plots also ensure that the errors we have detected in the fits of the model to patients do not reflect a systematic minor dislocation of the model's space of possible error patterns. Rather, we can see that the model's space of possibilities needs to be extended much further along several dimensions.

Other Evidence

The failure of the global damage model to fit E.A. or P.C. and the discrepancies between the locations of the patients and the boundaries of the model's capabilities suggest that one or more of the model's assumptions are invalid. But which assumption (or assumptions) should be changed? Ruml and Caramazza (2000) and Rapp and Goldrick (in press) have suggested that an explanation of naming deficits in fluent aphasics solely in terms of global damage is precluded by patients who make semantic errors only in one output modality. Consider patient R.G.B. (Caramazza & Hillis, 1990), for example. He produced 68% correct responses during picture naming, with the remainder semantic errors (either substitutions or descriptions). R.G.B. produced a similar pattern of performance in naming in response to tactilely presented objects and in response to aurally presented definitions. In these and other oral production tasks such as oral reading and spontaneous speech, R.G.B. made many semantic errors but no formal errors. In contrast, R.G.B. performed very well in written naming tasks and never made semantic errors. Furthermore, he performed flawlessly in all word comprehension tasks. The reverse pattern of dissociation to that shown by R.G.B. has also been observed (patient S.J.D., Caramazza & Hillis, 1991; patient R.C.M., Hillis, Rapp, & Caramazza, 1999).

Such patients, exhibiting an absence of semantic errors in one output modality combined with intact comprehension, invite the conclusion of intact semantic processing. Such circumscribed deficits appear incompatible with global damage. We can conclude, therefore, that the globality assumption must be false, and that the assumptions requiring global damage to account for aphasic performance seem good candidates for modification.

A Model of Representation Decay

In light of the evidence for localized damage to the lexical access system, we will modify Dell et al.'s model of aphasic naming to allow different parameter values at different levels of the network. If we retain the assumption that *connection* and *decay* are the theoretically interesting parameters that serve to relate the model of normal processing to the behavior of aphasic patients, then we derive a model in which *connection* and *decay* are both allowed to vary in different parts of the model. Unfortunately, we cannot test

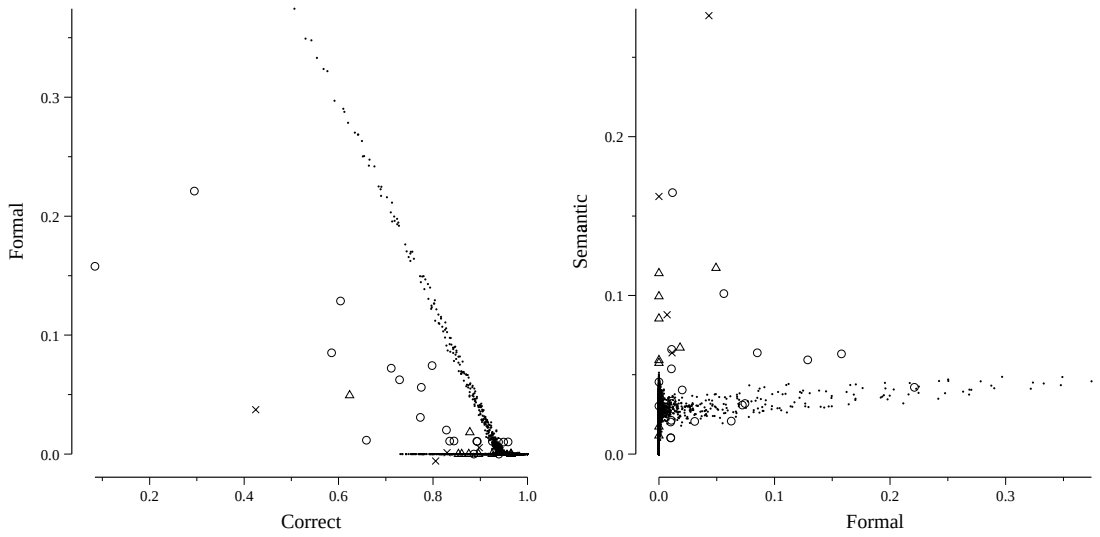

Figure 3. Possible combinations of error frequencies under either of the representation decay models.

this model, since it has five adjustable parameters (a *decay* at each of the three levels, and two *connection*s between them) and we are trying to match six-dimensional distributions (that sum to one). So we will restrict our attention to models that have two adjustable parameters, comparable to Dell et al.'s model. (We will return to the issue of degrees of freedom later, in the final section of the paper.) First, we will consider the model in which we allow the values of *decay* at each level of processing to be specified independently, while keeping *connection* at its default value throughout the model. Then in the next section, we will allow the value of *connection* to vary. By keeping the non-adjustable parameters at their default values, we reduce the number of model components that are varied between models, allowing direct comparisons of closely-related models.

The effect of increasing *decay* should be to quickly reduce the amount of activation at a level of representation. This implies that intrinsic noise and the activation acquired from neighboring representations should play larger roles. (Indeed, Dell et al. (1997) show that increasing *decay* throughout the network gives effects very similar to those obtained by increasing intrinsic noise.)

We simulated two variants of the model of representation decay, each involving only two of the three levels. In the first variant, the values of *decay* of the semantic and lexical levels are allowed to vary, and in the second, the values for the lexical and phonological levels are allowed to change. For both models, we systematically varied the parameters between $\frac{1}{2}$ and 1 in steps of $\frac{1}{64}$. This gave us at least 1,089 combinations for each variant. The results were disappointing. Neither model was capable of producing a wide variety of errors.

Two plots of the possible combinations of error frequencies are shown in Figure 3. Combinations from both variants are shown in both plots. Increasing *decay* for any single

layer alone has very little effect on the performance of the model. Increasing *decay* for both the phonological and lexical layers in tandem causes an increase in nonword errors, up to a maximum of about 7%. Presumably, this reflects the low activation levels at the phonological layer, caused by increased decay at that level and lack of reinforcing activation from the lexical layer. This allows misselection of phonemes, resulting in nonwords. The most interesting response occurs when increasing *decay* at the lexical level when it is already high at the semantic level. This causes a rapid increase in the number of formal errors, accompanied by some semantic, mixed, and unrelated errors, but no nonwords. This reflects the influence of activation maintained at the phonological level feeding back to the lexical layer and influencing lexical selection. This feedback has an exaggerated effect, due to the paucity of activation at the lexical level and lack of reinforcement from activation at the semantic level. The low number of nonwords is a result of the normal functioning of the phonological representations. Enough activation is maintained after lexical selection that misselections at that level are unlikely. Consistent with this explanation, similar behavior was observed when increasing decay at the semantic level when it was already weak at the lexical level.

The restricted range of errors possible under this model of representational decay would make any attempt at fitting patient data futile. (Of course, this does not imply that some form of decay deficit might not be involved in some forms of lexical access deficits. All that the above suggests is that a simple representational decay model cannot account for all the forms of naming deficits.) Instead, we will turn our attention to the other model of localized damage, in which the value of *connection* is varied.

A Model of Transmission Impairment

Under this model, which has been proposed by Foygel and Dell (1999), *decay* remains fixed, and the value of *connection* used between the semantic and lexical layers can be different from the value used between the lexical and phonological layers.[6] Damage is therefore hypothesized to consist of localized impairments in the transmission of information between different representations. This should allow, for instance, errors at lexical selection, while preserving correct selection of the corresponding phonemes. This expectation is based on the intuition that damage localized to the connections between two layers will not prove very damaging to the functioning of the non-involved layer. We now evaluate the ability of this new model to fit the data from our patients.

Fits to Patient Data

In order to start the parameter optimization algorithm in an appropriate part of the parameter space, we systematically sampled the model's behavior, as we did with the model of global damage. We varied the logarithm (base 10) of *connection* on the links between the semantic and lexical layers from -1 to -4, in increments of $\frac{1}{16}$ (corresponding to *connection* values of 0.1 to 0.0001), and similarly for the value of *connection* on the links between the lexical and phonological layers. We kept the value of *decay* at its default value. For each patient, we then fine-tuned the fit.

[6] Foygel and Dell use a *decay* value of 0.6, while we keep *decay* at 0.5 for consistency with Dell et al. (1997). This seems to be the only difference between their proposal and the model we investigate here.

Table 10: Fits of the transmission damage model to the patients.

Patient and	Naming response						Fit		
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	p
A.B.	.97	.02	.00	.01	.00	.01			
top .0806, *bot* .0054	.96	.02	.00	.00	.00	.01	.002	0.2	.999
E.M.	.96	.01	.00	.00	.00	.02			
top .0505, *bot* .0060	.96	.01	.00	.00	.00	.02	.002	0.8	.980
T.H.	.93	.06	.00	.02	.00	.00			
top .0076, *bot* .0379	.93	.05	.01	.01	.01	.00	.008	6.0	.311
R.C.	.88	.06	.00	.02	.01	.04			
top .0075, *bot* .0087	.87	.06	.02	.01	.01	.04	.011	7.4	.190
L.S.	.90	.09	.00	.01	.00	.01			
top .0072, *bot* .0115	.90	.06	.01	.01	.01	.01	.013	6.9	.226
J.R.3	.85	.10	.00	.04	.01	.00			
top .0049, *bot* .0583	.84	.07	.05	.01	.03	.00	.031	24.7	**.000**
L.T.	.88	.07	.02	.02	.00	.01			
top .0067, *bot* .0107	.88	.06	.02	.01	.02	.01	.010	8.6	.128
M.M.	.86	.11	.00	.03	.00	.00			
top .0058, *bot* .0149	.85	.08	.03	.01	.03	.00	.021	15.4	**.009**
P.C.	.62	.12	.05	.05	.14	.02			
top .0034, *bot* .0104	.63	.13	.10	.01	.11	.02	.028	19.8	**.001**
J.R.2	.83	.09	.01	.04	.00	.04			
top .0056, *bot* .0093	.81	.08	.04	.01	.03	.03	.023	18.7	**.002**
J.R.1	.90	.07	.01	.02	.00	.00			
top .0067, *bot* .0264	.90	.06	.02	.01	.02	.00	.010	4.1	.535
E.A.	.42	.28	.04	.10	.14	.01			
top .0018, *bot* .0104	.41	.16	.19	.02	.20	.02	.089	80.0	**.000**
I.O.C.	.81	.17	.00	.03	.00	.00			
top .0052, *bot* .0205	.82	.10	.04	.01	.04	.00	.037	5.7	.333

Table 11: Summary of the performance of the transmission damage model.

Patient	VAF by category						Summary VAF			
Group	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Mean	Wtd.	$\sum X^2$	Failures
Ruml et al.	.99	.75	-2.0	-.42	.85	.92	.17	.87	90	3/9 (5/13)
Dell et al.	.95	.33	.59	-.67	.78	.62	.43	.84	450	10/21
Combined	.95	.52	.57	-.57	.79	.69	.49	.86	540	13/30 (15/34)

The fits found by the regression algorithm are shown in Table 10. The listed parameter settings correspond to values of *connection* for links between the top two layers and the bottom two layers. Three of the nine patients (33%) failed to match at a 0.01 significance level (five of thirteen (38%), if we include the other four patients). In addition to P.C. and E.A., patients J.R.3, M.M., and J.R.2 could not be simulated. The model seemed to consistently make fewer mixed and semantic errors than the patients, and more phonological errors.

Foygel and Dell (1999) present fits of their transmission impairment model to the patients of Dell et al. (unnormalized data). For eight of the twenty-one fits (38%), the corresponding X^2 values are above 12.8, which corresponds to a significance of 0.025. Five of the twenty-one values (24%) indicate a failure to fit at the 0.01 significance level. Using normalized data, our regression algorithm was unable to obtain substantially improved fits, failing on 10 patients at a significance of 0.05. For fourteen of their patients (66%), the model of global damage provides a better fit.

Table 11 gives another view of the model's performance, showing the summary and category VAFs of the model for both our patients and Dell et al.'s. As with the model of global damage, the performance of the model is different across the two groups of patients, although the mixed errors of both groups seem hard to account for. Even when considering all patients, the transmission impairment model has a negative VAF for the mixed error category. This behavior is hidden by the weighted VAF score, which is comparable to that for the global damage model even though this model failed to fit twice as many patients.

Possible Error Patterns

A different view emerges when we view the transmission impairment model's space of possible errors. In Figure 4, we plot the model's coverage of the space of possible response distributions, exactly as we did in Figure 2 for the global damage model. By comparing the two sets of plots, one can see that, in several ways, the transmission impairment model gives broader coverage. In the upper left panel, for instance, we see that the relationship between the frequencies of correct and nonword responses is now much looser, indicating that the model can generate few nonword errors even while generating many errors of other types. The systematic parameter adjustments cause striping in the plot, showing us how adjustment of the *connection* value between the semantic and lexical layers allows the frequency of word errors to be varied without increasing the number of nonword errors. The model's ability to generate mixed errors seems sharply curtailed, however. It cannot generate more than about 2% mixed errors, at any level of correctness, as shown in the

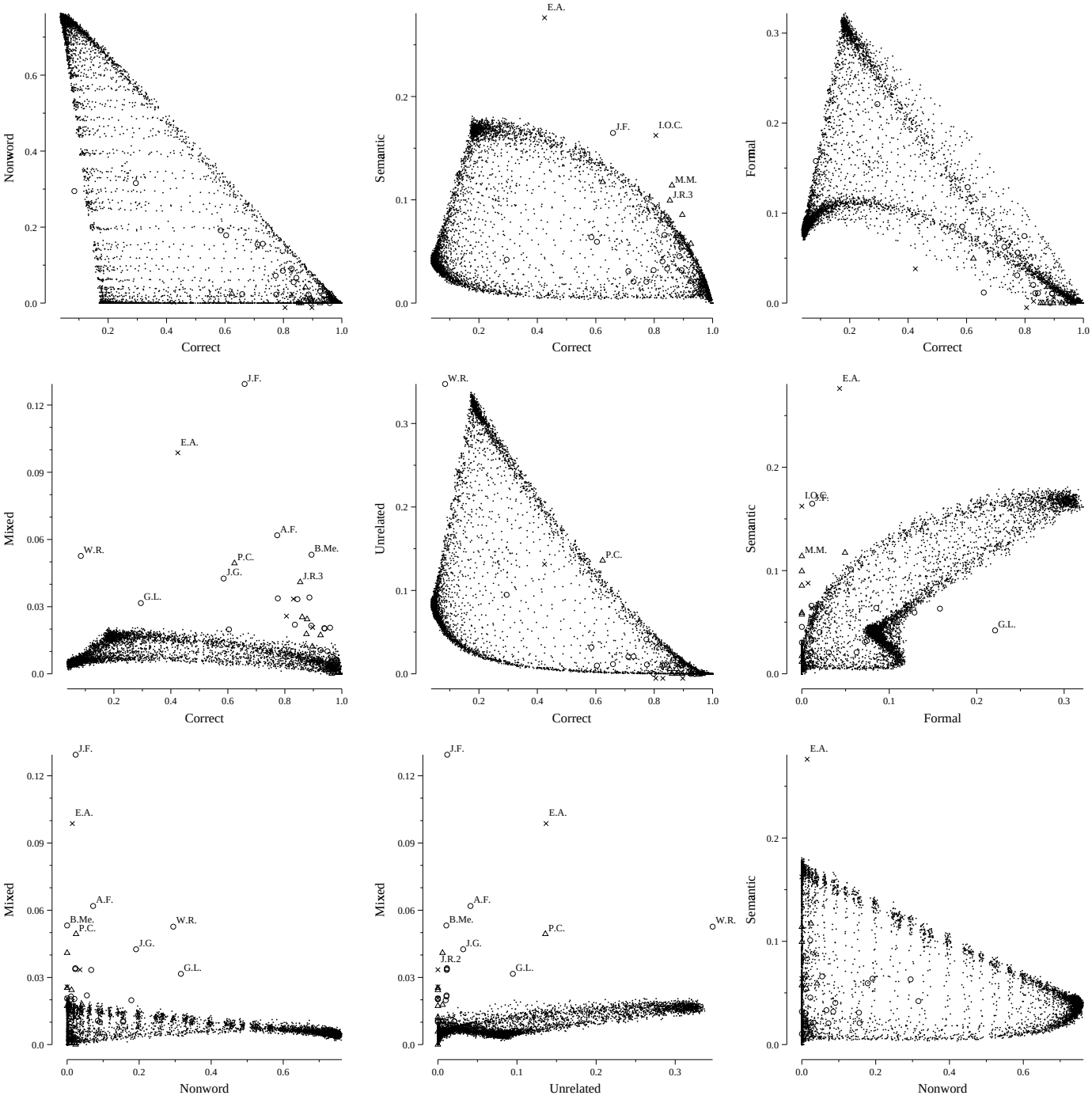

Figure 4. Possible combinations of error frequencies under the transmission damage model.

panels to the lower left. The upper middle panel also shows that the model cannot make a large enough proportion of its incorrect responses semantic errors.

These limitations of the model would not have been clear had we used a single projection of the data. Rather than presenting a collection of multiple projections using various pairs of response categories, we might have chosen a single plane, perhaps using principal components analysis (PCA), that would combine information from all categories. Even though such a slice through the six-dimensional space might be the best possible single projection in terms of capturing the maximum possible variance among the possible error patterns, it would not necessarily capture the relationships between the model and patient data. In Figure 5, we give a simple illustration of how a single projection may obscure additional dimensions along which a model and patient data differ. We plot the model and patient distributions in three-dimensions, rather than two as we did in Figures 2 and 4. In the upper left panel, the major axes are the correct and nonword categories. Subsequent panels show the effect of rotating the plot to reduce the contribution of the nonword dimension and increase the contribution of the mixed dimension. In effect, we are interpolating between the upper left and middle left panels of Figure 4. The axis lines give a rough indication of the contributions of the three dimensions (although, as in other figures, the scales of the axes have been normalized). It is clear that including more than two dimensions in a single projection does not necessarily combine the insights available when looking from multiple views.

Although the transmission impairment model seems to allow a wider range of possible error patterns than the other models we have considered so far, especially in the semantic and nonword dimensions, its inability to match patient data suggests that further modifications are necessary. One possibility is that the model of damage is not the problem, but rather the underlying model is flawed. We turn now to models incorporating different assumptions regarding normal processing.

A Model of Reduced Interactivity

An alternative model of aphasic naming has recently been proposed by Rapp and Goldrick (in press). They consider several models based on the same notion of activation spreading through a network that Dell et al.'s model uses, but varying in the degree to which the levels of representation interact during processing. Using three patients, two of whom made the large majority of their errors as semantic errors, they found that, while the feedback of activation from the phonological representations to the lexical level was necessary to account sufficiently for mixed errors, it was easy to have too much feedback, resulting in a preponderance of formal and nonword errors. They propose a model in which interaction is restricted, and suggest that it can account for their patients' performance.

Rapp and Goldrick's restricted interaction model differs from Dell et al.'s in five ways:

limited feedback: Not only is the value of *connection* on links from lower levels to higher ones smaller than the value in the opposite direction, but the links feeding back from the lexical to the semantic level are entirely absent.

damage via noise: Damage is modeled by varying the amount of noise in the activation functions of representations in a given layer, rather than by changing *decay* or *connection*.

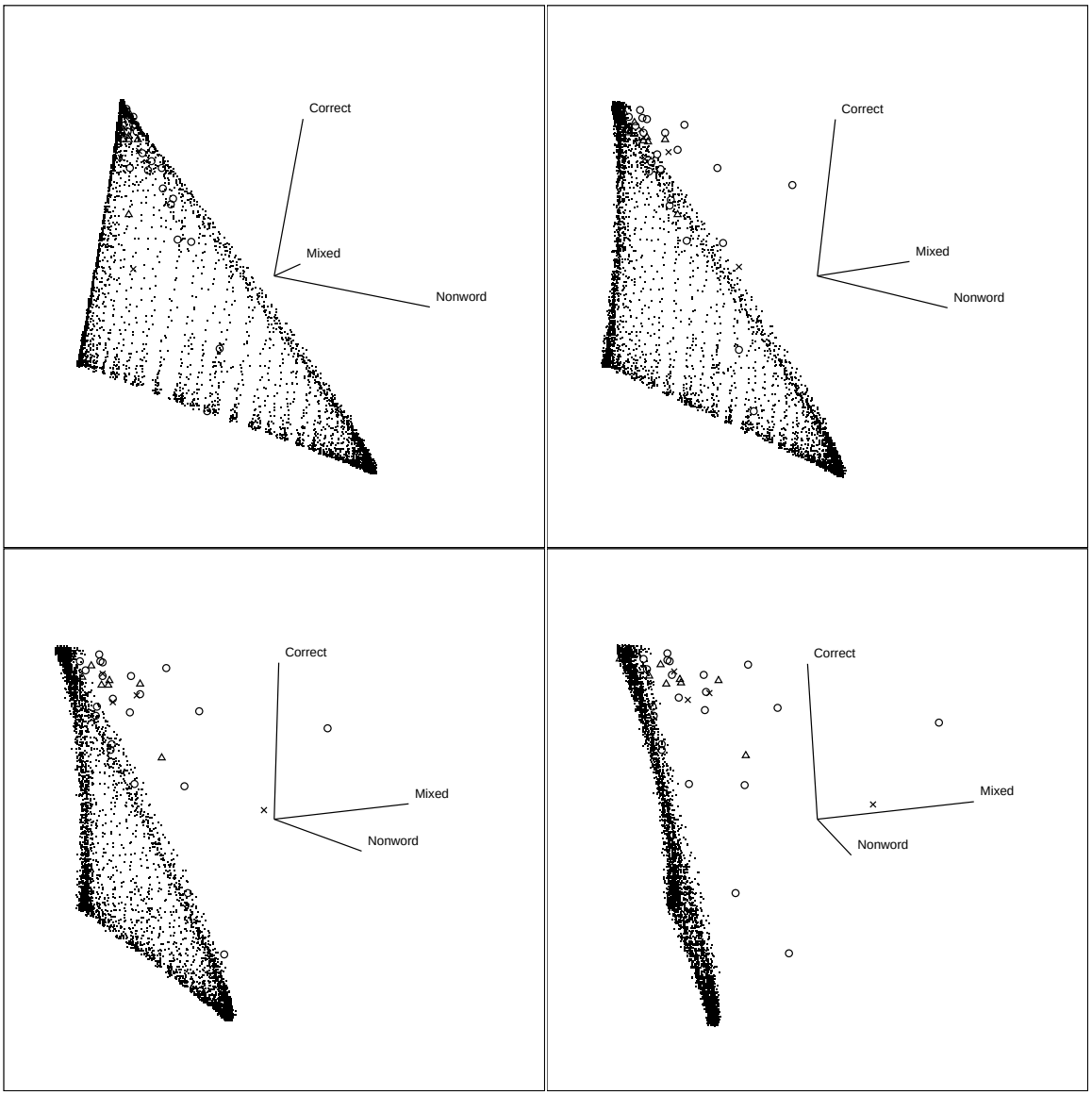

Figure 5. Rotation in three dimensions of data from the transmission damage model, illustrating the importance of multiple views of coverage. Rotation occurs around the *correct* axis, swinging the *nonword* axis out toward the viewer, and bringing the *mixed* axis from behind out to the right.

no ambient noise: The activation function doesn't include any *intrinsic* noise, whose variance is specified on an absolute scale. Rather, noise at each node is always proportional to that node's current activation.

larger lexicon: Only a single network structure is used, and it contains twenty-nine words rather than six. Rapp and Goldrick find that their single large lexicon allows an improved approximation to English error opportunities over Dell et al.'s switching between two small lexicons, particularly with regard to differences in phonological overlap within and between semantic categories.

concepts: The model contains a layer of representation above the semantic layer, corresponding to concepts. Processing therefore includes an additional preliminary selection step, in which, after input to the semantic layer, the most active conceptual representation is chosen. Processing then proceeds to lower levels, as usual, with selection at the lexical and phonological levels.

Although Rapp and Goldrick consider only the first of these assumptions to be theoretically central, we will test all of them, with the exception of the conceptual representation and its additional selection step.[7] Since Rapp and Goldrick's central argument concerns the importance of limited feedback, we will investigate that feature first.

We modified the network to conform to Rapp and Goldrick's restricted interaction assumption, removing the links from the lexical level to the semantic level, and reducing the strength of the connections from the phonological level to the lexical level. To maintain this assumption of reduced feedback, we constrained the value of *connection* on links from the phoneme layer back to the lexical layer to always be a fixed fraction of the value used in the forward direction. We will call this fraction the *feedback attenuation*. Since these reduced feedback assumptions underlie both normal and aphasic processing, we investigated the behavior of the network using both models of localized damage we previously considered (representational decay and transmission impairment) as well as Rapp and Goldrick's idea of noisy representations.

Reduced Feedback and Transmission Impairment

First, we will consider the transmission impairment notion of damage, in which the values of *connection* are decreased. We again systematically sampled the behavior of the model with different combinations of *connection* values for the links from the semantic layer to the lexical layer, and from the lexical layer to the phoneme layer. The feedback attenuation was set to $\frac{1}{10}$.

The resulting error patterns seemed identical to those obtained from the original network under the same damage assumptions (Figure 4). The similarity of the results with the fully interactive model confirms Rapp and Goldrick's observation that feedback links can be included in a model without necessarily playing a large role in determining its behavior. Predictably, similar results were obtained with the feedback attenuation at $\frac{1}{3}$. However,

[7] We exclude investigation of the additional layer because of the multiple additional assumptions it involves. While Rapp and Goldrick introduced their conceptual layer for the purposes of investigating a particular patient, it is more important to us to remain as similar as possible to the other models we are comparing against so we can isolate the factors underlying differential performance.

when we set the feedback attenuation to only $\frac{1}{100}$, (meaning that activation spreads one hundred times more easily from the lexical to phoneme layers as in the opposite direction), the model's performance deteriorated completely, producing only errors at all parameter settings. We hypothesize that this is caused by the diminution of activation in the absence of reinforcing interaction, resulting in the pronounced relative effects of background noise. (We will test this hypothesis later in this paper, by removing background noise.) For our purposes of fitting patient data, it seems this combination of processing and damage assumptions shows little promise beyond models we have already tested.

Reduced Feedback and Representational Decay

Although the assumption of damage as representational decay did not seem promising earlier, combining it with reduced interactivity yielded a model with more interesting behavior, although again the model did not seem suitable for matching patient performance. We allowed the values of *decay* to vary for the representations at the lexical and phonological levels. Feedback attenuation was kept at $\frac{1}{10}$.

Results from systematically varying the two parameters are shown in Figure 6. Either one of phonological or lexical decay alone yields many nonword errors, and both together result in complete breakdown. A very small amount of lexical decay alone results in some semantic and unrelated errors, but no more than 8%, and many formal errors are present as well. The model's performance does not look promising. (Simulations with the feedback attenuation at $\frac{5}{100}$ or $\frac{1}{100}$ were also disappointing, yielding mostly nonword errors at most parameter settings.)

But increasing decay is a rather roundabout way of introducing errors into the model's performance. As we mentioned earlier, this increases the effects of noise and spreading activation, the effects of which can be difficult to predict. Rapp and Goldrick suggest a more direct and intuitive method for introducing errors, which we investigate next.

Reduced Feedback and Noisy Representations

Rapp and Goldrick model aphasic naming by increasing the noise level at the conceptual, lexical, and phonological levels. Because our model does not include their additional layer of atomic conceptual representations, and in order to continue constraining our models to have only two adjustable parameters, we will only consider adding noise to the lexical and phoneme layers. (We carried out limited simulations adding noise only at the semantic level, but only very few semantic errors could be generated without a large number of nonword errors.) We systematically varied the standard deviation of the noise between one and five times its default value. We only changed the noise component which is proportional to the given node's activation; the ambient noise remained at its default level.

First, we experimented with adding noise to a single layer only. Adding noise to the phonological layer predictably results in formal and nonword errors. (This happened at many settings of feedback attenuation). We obtained more formal than nonword errors, which differs from the results reported by Rapp and Goldrick. They obtained almost exclusively nonword errors with noise at the phonological level, probably because their network's lexicon contained many more phonemes than Dell et al.'s, which implies that a larger percentage of the possible combinations of phonemes would result in a nonword error. They may also have used a more aggressive feedback attenuation. Adding noise at the lexical level

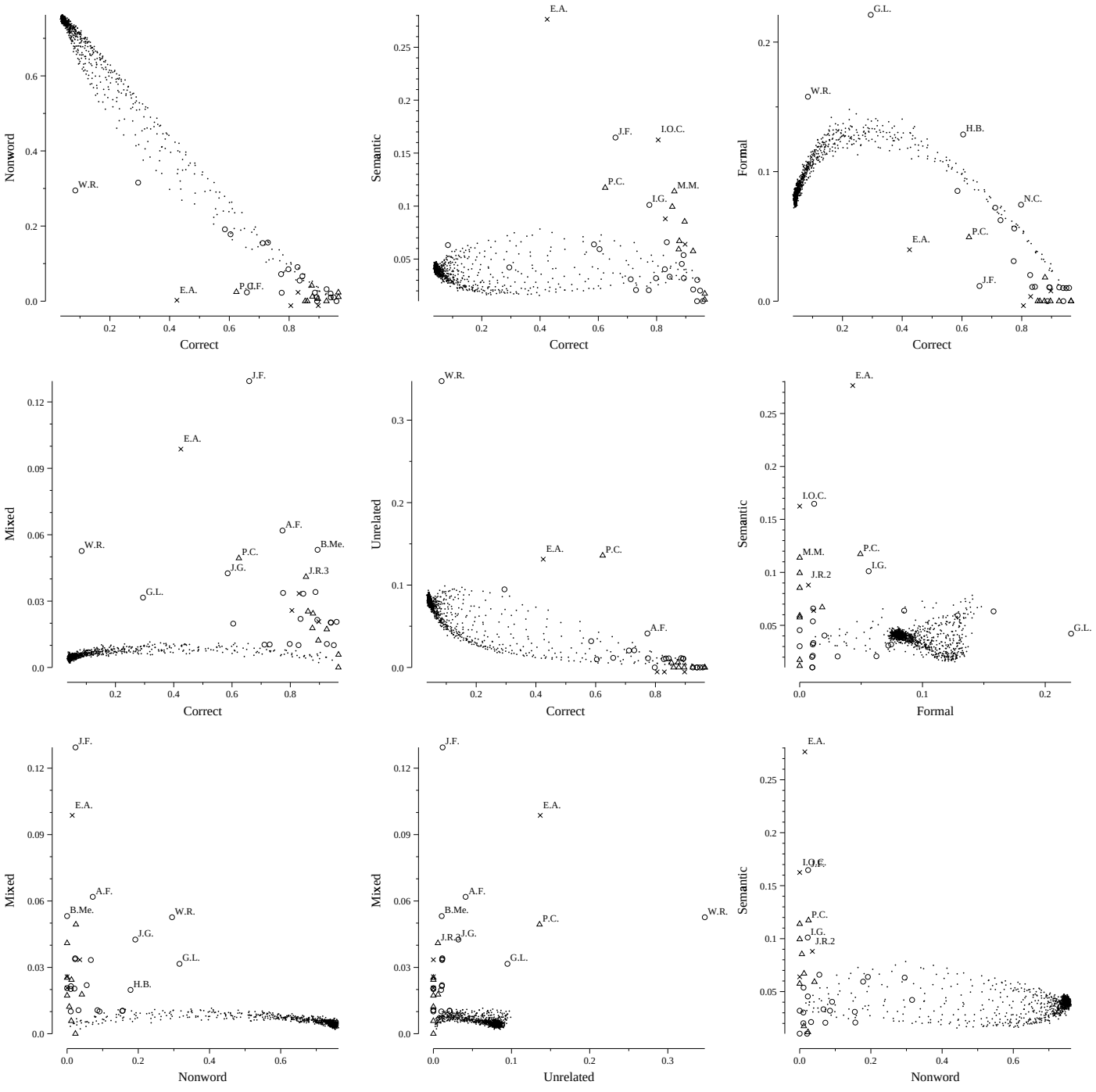

Figure 6. Possible combinations of error frequencies from the restricted interaction network under the assumption of representational decay.

resulted in mostly semantic errors. Increasing the feedback attenuation (from 0.1 to .3) increased the formal errors, while lowering the attenuation (from 0.1 to .001) also reduced correctness (but with a wider variety of errors).

We then investigated the behavior of the model when noise is added at both the lexical and phonological levels simultaneously. This model differs from Rapp and Goldrick's central proposal in only three ways: it retains the lexicon and ambient background noise assumptions of Dell et al., and it does not incorporate a conceptual level of representation. We systematically varied the two noise parameters (lexical and phonological) at several values of feedback attenuation (0.46, 0.32, 0.22, 0.16, 0.1, 0.05, and .001). At the extreme values (0.46, .01, 0.05, and .001), the range of possible patterns did not look promising. At very low levels of feedback, in particular, the highest achievable level of correctness was very low. At the three intermediate levels of feedback, the behavior of the model seemed promising. Figure 7 shows the possible combinations of error frequencies when using a feedback attenuation of 0.22. As Rapp and Goldrick have noted, by using noise at the semantic level alone, the model can produce error patterns with up to 10% semantic errors and no formal errors. The model has a very tightly restricted range of both unrelated and mixed errors for a given level of correctness however.

We fit the model of restricted interaction with noisy representation damage to our patients, using all three intermediate values of feedback attenuation. The most promising model was that at the 0.22 level of feedback. The resulting fits are shown in Table 12. The parameter settings are listed as multiples for the default noise level (ie, 2 represents twice the normal noise). Two of the nine patients (22%) were not matched (four of thirteen (31%) if we include the four patients who made many 'other' responses). This seems comparable to the fit of the other promising model we have seen so far, the original fully-interactive model under the assumption of transmission impairment, which failed on three of our patients (five when including the 'other' four). To gain more confidence in our estimate of the model's ability, we also fit the model to the patients of Dell et al. (1997). The resulting fits are shown in Table 13. The model failed to match six of the twenty-one patients, which seems improved from the transmission impairment model (which failed on ten of the patients).

Table 14 presents a VAF analysis of the noisy representations model. As with the original transmission damage model (Table 11), the model of noisy damage does least well in accounting for mixed errors. It also seems to have general trouble modeling semantic errors, doing only a bit better than guessing the mean. The VAF scores are remarkably little improved over the original transmission damage model, considering the improvement in fitting the patient data.

The other levels of feedback attenuation performed slightly worse. The 0.32 level also failed to fit five of our thirteen patients, and failed on thirteen of Dell et al.'s twenty-one patients, for a total failure rate of $\frac{18}{34}$. The 0.16 level model failed on only three of our patients, but eight of Dell et al.'s, for a total rate of $\frac{11}{34}$. It seems as if the 0.22 level is representative of the model's capabilities.

Reduced Ambient Noise

Given the relative success of a model of damage incorporating noise proportional to each node's activation level, one might wonder whether the intrinsic noise component of the activation is now redundant and dispensable. We modified the restricted interaction model

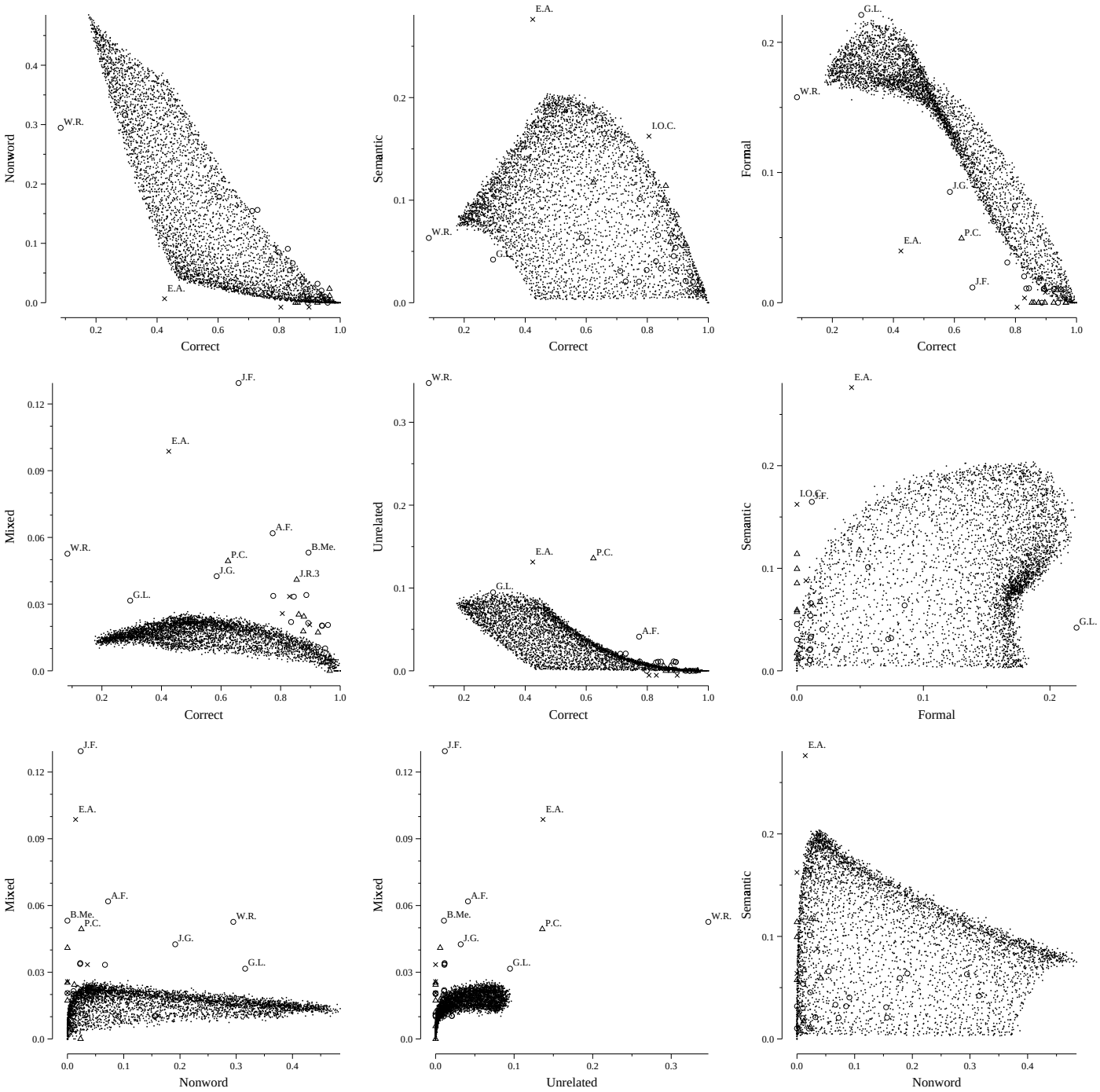

Figure 7. Possible combinations of error frequencies from the restricted interaction network under the assumption of noisy representations. Feedback attenuation was set to 0.22.

Table 12: Fits of the restricted-interaction model to the patients under the assumption of noisy representations. Feedback attenuation was 0.22.

Patient and	Naming response						Fit		
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	p
A.B.	.97	.02	.00	.01	.00	.01			
word 1.12, *phon* 1.87	.97	.01	.01	.00	.00	.00	.007	4.8	.442
E.M.	.96	.01	.00	.00	.00	.02			
word 1.06, *phon* 2.25	.94	.01	.03	.00	.00	.02	.016	6.7	.245
T.H.	.93	.06	.00	.02	.00	.00			
word 1.75, *phon* 1.00	.92	.06	.01	.01	.00	.00	.004	1.8	.870
R.C.	.88	.06	.00	.02	.01	.04			
word 1.75, *phon* 2.31	.85	.06	.05	.01	.00	.03	.022	10.3	.067
L.S.	.90	.09	.00	.01	.00	.01			
word 2.09, *phon* 1.12	.87	.10	.01	.01	.00	.00	.013	3.5	.617
J.R.3	.85	.10	.00	.04	.01	.00			
word 2.09, *phon* 1.00	.87	.09	.02	.01	.00	.00	.016	12.1	**.033**
L.T.	.88	.07	.02	.02	.00	.01			
word 2.00, *phon* 1.81	.86	.08	.03	.01	.00	.01	.010	3.3	.651
M.M.	.86	.11	.00	.03	.00	.00			
word 2.12, *phon* 0.49	.87	.10	.01	.01	.00	.00	.008	3.8	.576
P.C.	.62	.12	.05	.05	.14	.02			
word 4.50, *phon* 1.25	.52	.19	.16	.02	.06	.04	.076	37.3	**.000**
J.R.2	.83	.09	.01	.04	.00	.04			
word 2.19, *phon* 2.09	.81	.10	.05	.02	.01	.02	.023	11.3	**.045**
J.R.1	.90	.07	.01	.02	.00	.00			
word 1.94, *phon* 0.88	.90	.08	.01	.01	.00	.00	.006	1.0	.962
E.A.	.42	.28	.04	.10	.14	.01			
word 5.12, *phon* 0.97	.47	.21	.19	.02	.08	.04	.078	57.0	**.000**
I.O.C.	.81	.17	.00	.03	.00	.00			
word 2.50, *phon* 0.19	.82	.13	.03	.02	.01	.00	.019	1.7	.890

Table 13: Fits of the restricted-interaction model to Dell et al.'s patients under the assumption of noisy representations. Feedback attenuation was 0.22.

Patient and	Naming response						Fit		
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	p
W.B.	.94	.02	.01	.01	.00	.01			
word 1.34, *phon* 2.00	.93	.03	.02	.01	.00	.01	.006	2.1	.828
T.T.	.95	.01	.01	.02	.00	.00			
word 1.62, *phon* 1.62	.93	.04	.01	.01	.00	.00	.017	9.2	.100
J.Fr.	.93	.01	.01	.02	.00	.02			
word 1.37, *phon* 2.25	.90	.03	.04	.01	.00	.02	.018	10.2	.069
V.C.	.92	.02	.01	.01	.00	.03			
word 1.37, *phon* 2.25	.90	.03	.04	.01	.00	.02	.014	4.9	.422
L.B.	.82	.04	.02	.01	.01	.09			
word 1.75, *phon* 2.87	.77	.05	.09	.01	.01	.08	.036	10.7	.057
J.B.	.83	.06	.01	.03	.01	.06			
word 1.94, *phon* 2.37	.81	.07	.06	.02	.01	.03	.023	10.8	.056
J.L.	.85	.03	.01	.03	.01	.06			
word 1.94, *phon* 2.62	.78	.07	.08	.01	.01	.05	.042	17.2	**.004**
G.S.	.73	.02	.06	.01	.02	.15			
word 1.81, *phon* 3.31	.67	.05	.12	.02	.01	.14	.036	10.6	.059
L.H.	.71	.03	.07	.01	.02	.15			
word 1.81, *phon* 3.31	.67	.05	.12	.02	.01	.14	.028	7.7	.174
J.G.	.59	.06	.09	.04	.03	.20			
word 2.22, *phon* 3.56	.57	.07	.14	.02	.02	.19	.025	10.4	.065
E.G.	.94	.03	.00	.02	.00	.01			
word 1.69, *phon* 1.72	.91	.05	.02	.01	.00	.01	.017	9.6	.086
B.Me.	.89	.03	.01	.05	.01	.00			
word 2.47, *phon* 1.25	.81	.13	.03	.02	.01	.01	.056	28.5	**.000**
B.Mi	.88	.05	.01	.02	.01	.01			
word 2.00, *phon* 1.81	.86	.08	.03	.01	.00	.01	.016	7.5	.187
J.A.	.88	.05	.00	.03	.01	.03			
word 1.87, *phon* 2.25	.84	.07	.05	.01	.01	.02	.029	14.9	**.011**
A.F.	.78	.02	.03	.06	.04	.07			
word 2.25, *phon* 2.62	.73	.09	.09	.02	.01	.06	.046	41.1	**.000**
N.C.	.80	.03	.07	.01	.00	.09			
word 1.47, *phon* 2.91	.80	.03	.09	.01	.00	.07	.008	1.2	.942
I.G.	.77	.10	.06	.03	.01	.03			
word 2.25, *phon* 2.12	.78	.11	.06	.02	.01	.03	.008	2.2	.818
H.B.	.61	.06	.13	.02	.01	.18			
word 1.87, *phon* 3.62	.61	.05	.14	.01	.01	.18	.005	0.4	.994
J.F.	.66	.16	.01	.13	.01	.03			
word 3.25, *phon* 1.25	.67	.18	.09	.02	.02	.02	.053	73.3	**.000**
G.L.	.29	.04	.22	.03	.10	.32			
word 4.37, *phon* 3.87	.29	.10	.19	.02	.08	.32	.030	9.1	.106
W.R.	.08	.06	.16	.05	.35	.30			
word 14.40, *phon* 3.87	.20	.13	.22	.02	.15	.27	.104	74.2	**.000**

Table 14: Summary of the performance of the restricted-interaction model with noisy representations.

Patient	VAF by category						Summary VAF			
Group	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Mean	Wtd.	$\sum X^2$	Failures
Ruml et al.	.84	.45	-6.2	.14	.69	.71	-.56	.63	84	2/9 (4/13)
Dell et al.	.96	-.34	.46	-.27	.64	.99	.41	.88	356	6/21
Combined	.95	.036	.34	-.17	.65	.99	.47	.87	440	8/30 (10/34)

Table 15: Summary of the performance of the restricted-interaction model without ambient noise and with noisy representations.

Patient	VAF by category						Summary VAF			
Group	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	Mean	Wtd.	$\sum X^2$	Failures
Ruml et al.	.76	.05	-5.6	.098	.63	.51	-.59	.53	86	2/9 (3/13)
Dell et al.	.94	-1.2	.67	-.33	.39	.92	.23	.83	462	9/21
Combined	.93	-.62	.54	-.22	.43	.93	.33	.82	548	11/30 (12/34)

with noisy representations by removing the ambient noise (represented by the coefficient *intrinsic* in the activation function). Damage was modeled as before, using noise proportional to each node's activation, added at different rates at the lexical and phonological levels. We considered feedback attenuations of 0.32, 0.22, 0.16, and 0.1.

The best of this family of models seemed to be the one with a feedback attenuation of 0.1 (which is lower than the 0.22 we found when including ambient noise.) The possible error patterns are shown in Figure 8. The lack of intrinsic noise seemed to allow a sharper dissociation between error patterns with only semantic errors and only phonological errors. More precisely, the range of possible frequencies of semantic errors that could be generated without producing many formal errors was larger than when using intrinsic noise, and the highest attainable percentage of semantic errors was also higher. As one might expect, since there was less background noise, more noise was also required to achieve a given level of nonword errors.

Unfortunately, these features did not improve the model's ability to match the performance of patients. The model failed to match two of our patients (three out of the full thirteen: J.R.3, P.C., and E.A.), and nine of Dell et al.'s (T.T., J.L., J.G., B.Me., J.A., A.F., J.F., G.L., and W.R.). For our patients, the model's inability to produce many unrelated errors was a severe problem. For Dell et al.'s data, the model's deficiencies were less obvious, although the frequency of semantic errors seems involved. Table 15 gives the VAF scores, which are marginally worse than those of the same model with noise. Again, the mixed error category is troublesome, although semantic even more so.

Models incorporating more feedback (corresponding to feedback attenuations of 0.32, 0.22, and 0.16) matched slightly fewer patients.

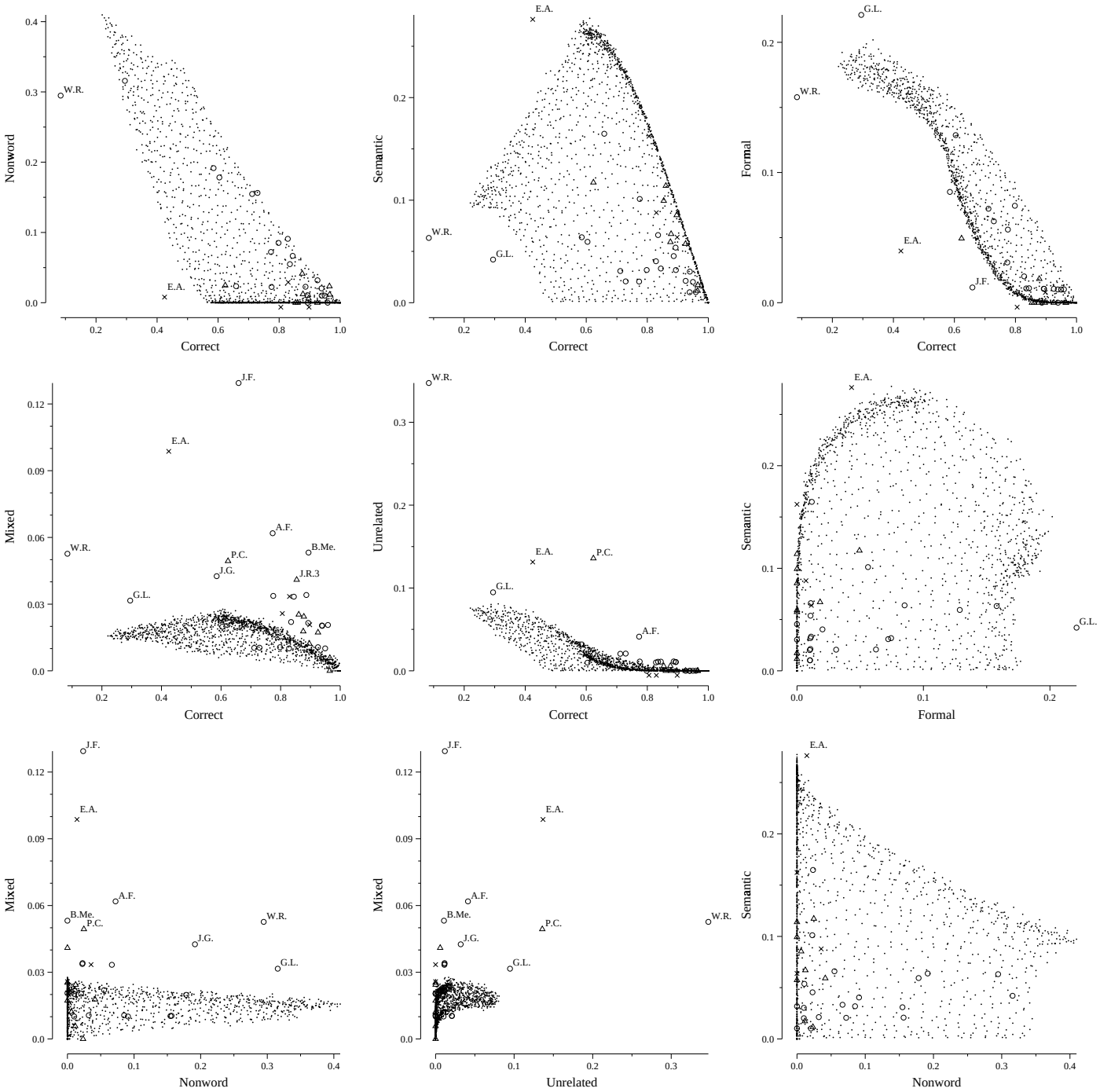

Figure 8. Possible combinations of error frequencies from the restricted interaction network with no intrinsic noise, under the assumption of noisy representations. Feedback attenuation was 0.1.

A Larger Network

The remaining difference between the models of Dell et al. (1997) and Rapp and Goldrick (in press) that we will address is that of the different lexicons. (As we mentioned above, we will not address the addition of a conceptual layer.) Rapp and Goldrick point out that the maximum frequency of mixed errors that can be generated by any model using Dell et al.'s network structures is 10%, since of the two small networks they use, only the one which is used 10% of the time contains a word both semantically and phonologically related to the target. Since both Dell et al. and Rapp and Goldrick construct their lexicons by attempting to match the error opportunities of English (as we discussed earlier), one might think that any differences would be insignificant. In our limited experiments using Rapp and Goldrick's single large network with the assumption of restricted feedback and noisy representations, we were unable to replicate the level of performance we had achieved using Dell et al.'s lexicons.

When using ambient noise, limiting feedback substantially (with a feedback attenuation of 0.1, for instance) resulted in an inability to produce correct responses. When using more feedback (attenuations of 0.32 or 0.22), we were unable to fit at least seventeen of the thirty-four patients, including at least six of our own. At all parameter settings, the model produced many more mixed errors than it had when using the Dell et al. lexicons. Without ambient noise (as in Rapp and Goldrick's proposal), we failed to fit at least sixteen patients (using feedback attenuations of 0.32 and 0.1).

These disappointing results may be merely an artifact of our limited experimentation, but they certainly indicate that the model's behavior is very sensitive to the lexicon used in the simulation. Such details, which are typically subordinated as implementational, have as much impact on model performance as the more theoretically central assumptions, such as the functional impact of brain damage on the lexical system. While the qualitative shape of the model's space of generable patterns remained roughly similar with the larger lexicon, the exact shape and size of the model's coverage changed dramatically. Figure 9 provides a direct comparison, with results using Dell et al.'s networks on the left, and using Rapp and Goldrick's lexicon on the right. The frequency of mixed errors, for instance, tended to be much greater, while maintaining a similar relationship with unrelated errors.

Summary of Empirical Results

We have now traversed the spectrum from the fully-interactive model of Dell et al. (1997), which assumed global damage, to the restricted feedback model of Rapp and Goldrick (in press), which postulated localized noisy representations. Table 16 shows a summary comparison of the models, indicating whether the hypothesized damage affected transmission of activation (*conn*), maintenance of activation (*decay*), or noise in the representations themselves (noise). The table also shows the two types of quantitative evaluation statistics we considered: variance accounted for (VAF, both averaged over categories and computed as a single weighted sum) and the number of patients each model failed to fit (χ^2 test, $p < 0.05$, figures in parentheses include patients who made many 'other' responses). We found that:

1. A model incorporating global damage fails on six of the thirty patients, and is incompatible with documented cases of local damage.

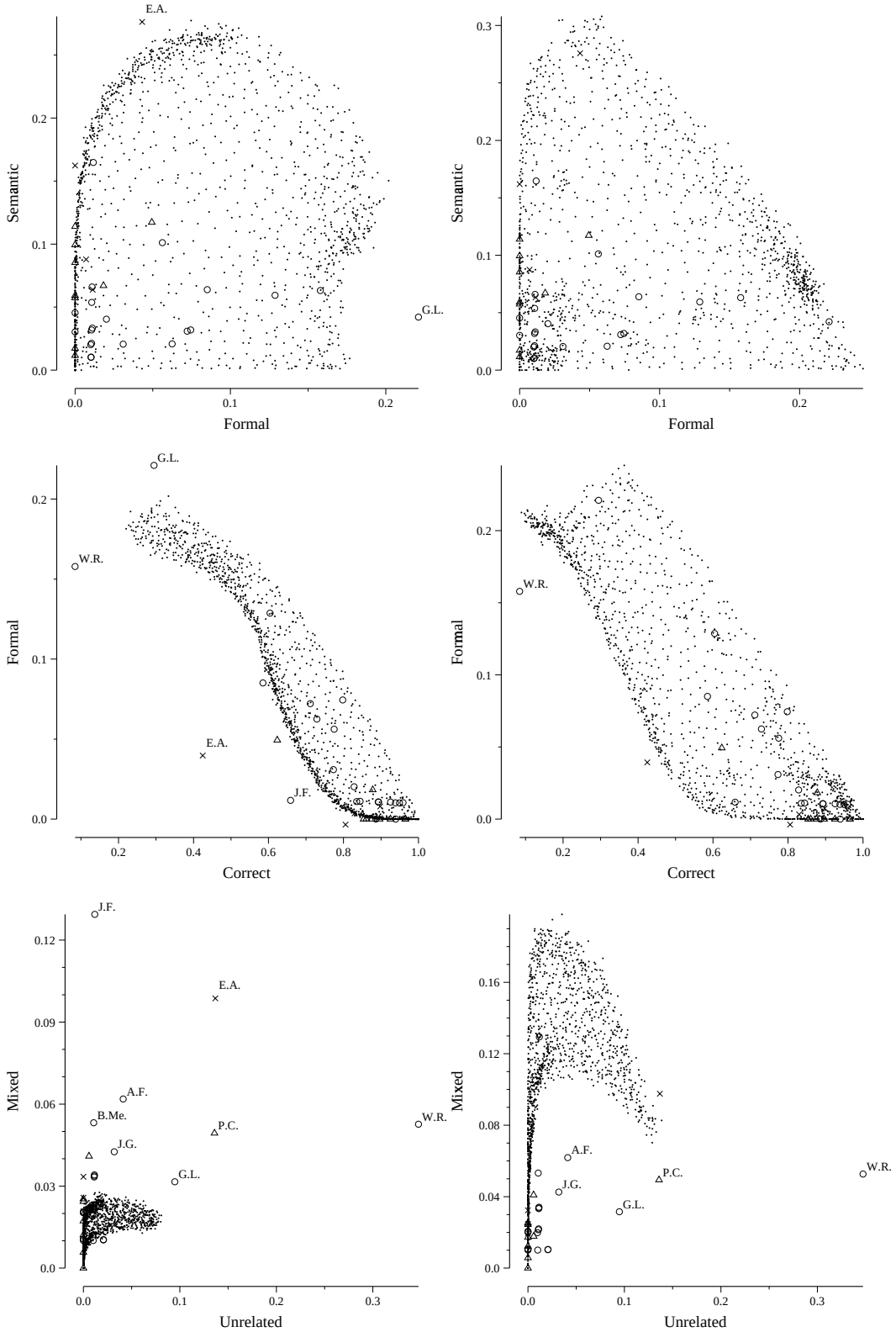

Figure 9. Possible combinations of error frequencies from the restricted interaction model with no intrinsic noise and a feedback attenuation of 0.1, under the assumption of noisy representations. Panels on the left use the Dell et al. lexicons and panels on the right use the Rapp and Goldrick lexicon.

Table 16: Summary of the models considered in this paper.

Network	Damage		Ambient noise	VAF		Failures
	Type	Location		Cat.	Wtd.	
Dell et al. 2 × 6-word	*conn*, *decay*	global	yes	.57	.87	6 (7)
	conn	upper, lower		.49	.86	13 (15)
	decay	sem, lex				poor
		lex, phon				poor
Dell et al. 2 × 6-word, no lex→sem	*conn*	upper, lower				poor†
	decay	lex, phon				poor
	noise	lex, phon	yes	.47	.87	8 (10)
			no	.33	.82	11 (12)
Rapp and Goldrick 29-word	noise	lex, phon	yes			∼ 17
			no			∼ 16

†or similar to fully interactive, depending on feedback attenuation

2. A model incorporating representational decay did not exhibit sufficient variety of possible error patterns to make patient fitting worthwhile.

3. A model of transmission impairment similar to that of Foygel and Dell (1999) fails on thirteen of the thirty patients.

4. A model with reduced feedback seems to perform best when used with a noisy representation brain damage assumption. A model with these features fails on only eight of the patients.

5. Additional models incorporating further assumptions of Rapp and Goldrick, such as a lack of ambient noise or a larger lexicon, matched fewer patients.

Discussion

We systematically explored several of the assumptions included in current computational models of aphasic picture-naming. These included both assumptions regarding interaction during lexical access and assumptions about the effects of brain damage on the performance of a model of normal behavior. We reported the picture-naming performance of thirteen patients, and used those data, in addition to the patterns from twenty-one patients reported by Dell et al., to test the models. (We did not consider data from additional tasks, such as word repetition, or more detailed analyses, such as investigations of a mixed error effect or a grammatical class effect.) No model that we tested could account for all the patients. However, one of the models, a novel intermediate position between proposals of Dell and his collaborators and Rapp and Goldrick, seemed able to match a greater number of patients than the other psychologically plausible models we tested that were closer to those endpoints. Our empirical approach combined multiple views of the possible error patterns a model could generate, along with a formal numerical regression procedure for matching patients' patterns. We showed that both of these methods are crucial for achieving a balanced view of a model's performance, and that evaluations based on single

projections or summary statistics such as variance accounted for (VAF) can fail to detect the shortcomings of a model.

Limitations of Simulation Studies

It is important to note that the conclusions one can draw from the performance of these models are quite limited. In particular, the apparent increase in matching ability of the hybrid model we investigated here must be weighed against the fact that the model still fails to match many patients, and is thus inadequate. One cannot dismiss its failures by appealing to the intuitive overall closeness of the fits. If we were to rely on intuition, when would a mediocre fit become a failure to match? If we are to allow our model to be falsifiable, then we must choose a criterion, in our case, a significance level for a χ^2 test. If certain failures could be directly traced to specific simplifying assumptions, then the direction of further work would become obvious, but this would not excuse the model or provide support for the remaining assumptions.

Pragmatically speaking, the fact that the hybrid model performs better than any other psychologically plausible model we tested suggests that its assumptions may prove useful in constructing a model that can account for all of the patients. This is the true value of simulation work such as ours, or that of Dell et al. or Rapp and Goldrick. As with regression more generally, replicating a trend in a data set is provocative, and in engineering or finance, such partial solutions can be useful end products. But in a scientific setting, an inaccurate model cannot form the basis of further theoretical claims or inferences. It is not clear that such 'partial matches' as we have achieved can provide support for any of a model's assumptions. For instance, if we did not have additional evidence ruling out a theory of aphasia stipulating global damage throughout the lexical system, the simulation results we have seen would lead us to favor the global damage model over the restricted interaction model. It matches more patients than any other model we tested, has a higher mean category VAF, and its weighted overall VAF is likely statistically indistinguishable from the competing model. A suggestive model that is inconsistent with patient data can only help direct future work, and as this example illustrates, partial success may be misleading. It is in this sense that our conclusions, and those of other researchers, must by necessity be limited.

Even if a model were to succeed in accounting for the patient data, however, one would need to be cautious in claiming support for particular assumptions as opposed to others. We have shown, for instance, that although Rapp and Goldrick's proposal for reduced feedback seems promising, the performance of a model incorporating that assumption still depends crucially on the damage assumptions and lexicon that accompany it. In our limited experiments, using a larger lexicon that more closely approximates English resulted in poorer performance. The notion of reduced feedback seems appealing, but even if a model including it were to match the data, one could not claim that those fits supported that assumption. Rather, the entire model, including both the theory of normal processing and the assumptions regarding brain damage, would be validated. Since any of the many assumptions involved could have aided the data-fitting process as easily as hindering it, it would still have remained unclear whether a more accurate model would have continued to enjoy empirical support.

This suggests an important principle for computational modeling: the impact of the

theoretically uninteresting assumptions should be evaluated as carefully as the model components that correspond to the central issues under examination. Such tedious and systematic experiments deserve the same respect as traditional sensitivity analysis in mathematical modeling. Without them, it is premature to draw conclusions in favor of particular assumptions. When a complete examination of the basic assumptions proves impractical to perform, the dependence of the model's performance on all of its components should be explicitly acknowledged.

The Importance of Single-Patient Analysis

Of course, an attempt to account for a patient is predicated on the assumption that the patient's behavior falls within the scope of the model. Even aside from the issue of whether aspects of naming behavior such as grammatical class or descriptive responses are relevant, many studies (including this one) attempt to use data from patients whose impairments are not well understood. If a patient has brain damage to mechanisms other than those involved in lexical access, and that damage affects naming behavior, then modeling that patient's behavior is at best irrelevant, and could possibly be misleading. One might mistakenly modify the structure of the lexicon to correct for a patient's difficulty in visual analysis. Conversely, if the patient's damage is clearly confined to the lexical access system, or if it extends beyond the lexical system only to mechanisms that can safely be assumed to be irrelevant for the tasks at hand, then failure to match must compel rejection of the model as inconsistent with the data. This consideration requires that patients whose performance is used for testing models of naming each be tested extensively on other related tasks to rule out, or at the very least delineate, damage to other relevant systems. The requirement that is used in this paper and by other researchers stipulates only that a patient's output be fluent. This methodological gap opens the possibility that only a subset of the patient groups currently used by researchers for model evaluation is actually relevant.

This situation also provides a good example of the importance of closely examining the match of a model to each patient, as we did with the χ^2 tests. Unlike metrics such as VAF that collapse data across patients, the χ^2 test can point us to particular patients whose performance cannot be explained by the model. In our examination of the global damage model, for instance, we found that the lowest category VAF was for semantic errors. But for P.C., the most troublesome patient for the model, the semantic category was the best fit category (Table 8). As this shows, particular mismatches are obscured by group behavior in a global measure such as VAF. Since the frequency of any particular pattern of patient performance should have no bearing on its ability to disconfirm a theory, metrics with this character are not as useful as those that can detect failures on individual patients. Once we have identified specific problematic patients, we can then reassess the patient's deficits to be sure the model is applicable, and begin the process of uncovering systematic discrepancies across multiple tasks that could lead to theoretical revisions. Collapsing data across patients can obscure those patients that provide the most informative test of a model. (For a more general discussion of the limitations of grouped analysis in neuropsychology, see Caramazza (1986).)

Testing More Complex Models

In this paper, we have limited ourselves to considering models with two adjustable parameters. We imposed this constraint on ourselves to avoid making the data fitting problem trivial. Decisions regarding degrees of freedom are not obvious, however. Certainly, damage to brain areas involved in lexical access can come in many forms, and our models may need to reflect this. Fortunately, the same neuropsychological analysis that we advocated previously for ensuring patient relevancy can be helpful here. By localizing the damaged module or pathway in each patient, we constrain the degrees of freedom we have for model fitting. Once we know that a patient has damage only to certain representations, we must allow changes to our computational model of normal performance only in those representations or pathways connected to them. It is suddenly irrelevant for the purpose of fitting that patient how many parameters the model has for manipulating other features of the lexical system. The model would account for that patient only if it could generate the patient's behavior with any abnormal parameter settings confined to the functional areas that are damaged in that patient. In this way, patients with more clearly defined deficits should play a dominant role in computational modeling of aphasia, since they narrow down the number of parameters that can be adjusted in a modeling attempt. Rapp and Goldrick's work, for example, benefits from such localization constraints. Although they allow noise at each of their four different levels of representation, they only have one or two free parameters at a time, since the functional locus of the damage in the patients they consider has already been localized. When combined with the formal regression approach of Ruml and Caramazza (2000), these *a priori* constraints would allow sound evaluation of very complex models of naming in aphasia.

Inferences in the opposite direction, from models to patients, would require even more stringent tests. If a model that assumed damage to one area of the lexical system were to fit a particular patient better than a model postulating damage in a different location, this cannot tell us anything about the locus of damage in that patient unless we understand the implications of every possible way that the models can be damaged. First, we would have to be confident that the first model can reproduce the new patient's behavior when damaged in the same unknown way that the patient is actually damaged. This assurance could only be gained by extensive prior comparisons with a diverse set of patients whose loci of damage are known. In addition, we would need to be certain that neither model can produce behavior similar to the patient's when damaged in any alternative way (or combination of ways). (This was not the case in the present study, for instance, in which we have seen theoretically divergent assumptions about brain damage result in fits to particular patients that are of similar quality.) Without such detailed knowledge, we cannot be sure that the hypothesized damage that yields the best fit of the model corresponds to the patient's actual damage. Again, evidence from additional related tasks may be of assistance in narrowing down the number of hypotheses that are consistent with the patient's behavior. (Although, of course, such evidence cannot be used during model regression without a commensurately elaborate model that can generate behavior on multiple tasks).

Conclusions

Our systematic exploration of the spectrum of two-step spreading-activation models yielded a new hybrid model which seems to perform slightly better than models closer to proposals of Foygel and Dell (1999) and Rapp and Goldrick (in press). It contains the lexicon and intrinsic noise assumptions of Dell et al. (1997) and Foygel and Dell, and the restricted feedback and damage assumptions of Rapp and Goldrick. We used both a numerical regression algorithm to test the model's fit to patient picture-naming data, and multiple views of its generable response patterns to gain a more intuitive sense of its limitations. It is important to note that no model that we tested could account for all of the patients. Our results are therefore only suggestive of directions for further work, and cannot support strong theoretical inferences. Furthermore, we found that supposedly peripheral modeling assumptions can play at least as large a role in determining model behavior as the assumptions corresponding to the theoretical issues under investigation. We believe that broad neuropsychological assessments of patients can provide constraining evidence that, when used in tandem with close analysis of a model's fit to each patient, will enable sound evaluation of more accurate models in the future.

References

Berndt, R. S., Haendiges, A. N., Mitchum, C. C., & Sandson, J. (1997). Verb retrieval in aphasia: 1. Characterizing single word impairments. *Brain and Language*, *56*, 68–106.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*(1), 41–66.

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex*, *26*, 95–122.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, *349*, 788–790.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Foygel, D., & Dell, G. S. (1999). Models of impaired lexical access in speech production. (Manuscript under revision.)

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders.* Philadelphia: Lea and Febiger.

Hillis, A. E., Rapp, B., & Caramazza, A. (1999). When a rose is a rose in speech but a tulip in writing. *Cortex*, *35*(3), 337-356.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75.

Mitchum, C. C., Ritgert, B. A., Sandson, J., & Berndt, R. S. (1990). The use of response analysis in confrontation naming. *Aphasiology*, *4*, 261–280.

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*(1), 56–115.

Rapp, B., & Goldrick, M. (in press). Discreteness and interactivity in spoken word production. *Psychological Review*.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, *24*, 121–133.

Ruml, W., & Caramazza, A. (2000). An evaluation of a computational model of lexical access: Comments on Dell et al. (1997). *Psychological Review*, *107*(3).

Shallice, T., Glasspool, D. W., & Houghton, G. (1995). Can neuropsychological evidence inform connectionist modelling? Analyses of spelling. *Language and Cognitive Processes*, *10*(3–4), 195–225.

Shelton, J. R., Fouch, E., & Caramazza, A. (1998). Selective sparing of body part knowledge: A case study. *Neurocase: Case Studies in Neuropsychology, Neuropsychiatry, and Behavioural Neurology*, *4*(4–5), 339–351.

Shelton, J. R., & Weinrich, M. (1997). Further evidence of a dissociation between output phonological and orthographic lexicons: A case study. *Cognitive Neuropsychology*, *14*, 105–129.