Connectionist Modeling of Neuropsychological Deficits in Semantics, Language and Reading

Christine E. Watson
Department of Psychology, University of Pennsylvania

Blair C. Armstrong
Department of Psychology, Carnegie Mellon University

David C. Plaut
Department of Psychology, Carnegie Mellon University

To appear in M. Faust (Ed.), *Advances in the neural substrates of language: Toward a synthesis of basic science and clinical research.* New York: Wiley-Blackwell.

Contact information:
David C. Plaut
Department of Psychology
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213-3890
USA
Email: plaut@cmu.edu

Introduction

Representations of linguistic information and the neural substrates that underlie them are incredibly complex. This chapter illustrates how connectionist modeling has furthered our understanding of normal and impaired processing in three related domains – semantic memory, knowledge of grammatical class, and word reading – and how the development of these models engenders a reciprocal relationship between theoretical and empirical research. In particular, we highlight the value of employing domain-general learning and information-processing principles to derive explicit accounts of the ways in which factors which make contact with, or are central to, linguistic abilities can interact to give rise to a range of behaviors. We also detail how the connectionist approach has provided for theoretical advancement that would not have occurred using the double-dissociation rubric of traditional cognitive neuropsychology and how it has allowed the exploration and development of ideas that would have been difficult, if not impossible, to formulate verbally.

Semantic memory impairments

Developing a theory of how semantic memory is represented and processed is central to understanding key aspects of human cognition, as this knowledge is required for a variety of language-related tasks and beyond. A key source of empirical data for theory-development is derived from the study of patients with various neurological impairments (e.g., cerebral infarction, viral infection, dementia) who exhibit consistent and specific patterns of semantic memory impairment. The present section focuses on two

contrasting types of “pure” semantic memory deficits – that is, impairments to semantic memory in which other cognitive functions such as lower-level perception and short-term memory have (in some cases, at least) been documented as being relatively intact – in which there is either selective loss of knowledge for particular semantic categories or a more uniform loss of knowledge which spans all semantic categories. These data have been particularly challenging for and central to the development of recent theories. (Note that we use the term “category” to refer to both broad superordinate-level categories, such as *living things*, as well as narrower basic-level semantic categories, such as *birds*.)

The first type of impairment consists of so-called category-specific semantic deficits (CSDs), in which knowledge for one category is substantially impaired while other categories are relatively preserved, albeit with some important exceptions. These impairments tend to manifest themselves as increased commission or category co-ordinate errors, wherein the incorrect responses patients’ produce to items from the impaired category are also members of the impaired category (e.g., in a picture naming task, participants with impaired knowledge of living things might respond “dog” to an image of a sheep; Lambon Ralph, Lowe, & Rogers, 2007). Two key patterns of selective impairment have emerged in the literature (see Capitani, Laiacona, Mahon, & Caramazza, 2003, for a review of over 100 case studies of these impairments). The most frequently reported pattern consists of a selective loss of living thing knowledge while knowledge of non-living things is preserved (e.g., Warrington and Shallice, 1984). Associated with this general loss of living thing knowledge are several exception categories. Knowledge of musical instruments, for example, tends to be lost along with

living things, whereas knowledge of exception categories such as body parts and manufactured foods is preserved. The opposite pattern of impairment – selective loss of nonliving thing knowledge, which is accompanied by loss of knowledge for exception categories such as foods and body parts – has also been reported (Warrington & McCarthy, 1983, 1987; Caramazza and Shelton, 1998), though far less frequently. Note that of these cases, though there have been a number of "pure" selective impairments without any apparent additional impairments to particular sensorimotor modalities (e.g., Caramazza & Shelton, 1998), a large number of cases have also been reported in which knowledge for particular sensorimotor modalities and semantic categories have both been impaired (e.g., McCarthy & Warrington, 1988; Maginé, Ferreira, Giusiano, & Poncet, 1999).

CSDs have been linked to a variety of etiologies, with approximately half of the known cases being associated with damage from herpes simplex virus encephalitis (HSVE) and the remaining cases being mainly associated with various forms of dementia and cerebrovascular accident. Cases of selective loss of living thing knowledge are primarily associated with HSVE and damage to the left temporal lobe; there is less specificity and consistency in impaired brain regions associated with the less frequently observed loss of nonliving thing knowledge (Capitani et al., 2003).

In contrast to the CSD cases, there have also been many documented cases of general semantic impairments in which all categories of knowledge have been documented as being equally affected, with the impairment usually taking the form of errors of omission (that is, patients are unable to make any response to a probe stimulus; Lambon Ralph et al., 2007). These general semantic impairments are

primarily associated with *semantic dementia*, a disease which selectively affects the anterior and inferolateral temporal cortex. This region largely overlaps with the regions associated with HSVE and living thing deficits, though semantic dementia may uniquely involve more lateral regions and the HSVE more medial regions of the temporal lobe (Noppeney et al., 2007).

Traditional accounts of semantic impairments

Employing double-dissociation logic, Caramazza and Shelton (1998; see also Sartori & Job, 1988; Santos & Caramazza, 2002) argued that the separate cases of category-specific semantic impairments for living things and nonliving things suggest that these categories are subserved by anatomically distinct neural substrates; they further provide a post-hoc evolutionary basis for this view. Their perspective is known as the *domain-specific* hypothesis. Accounting for category-specific deficits is trivial under this framework – selective lesions to brain regions subserving each category would leave the other categories intact. Though Caramazza and Shelton did not attempt to explain uniform deficits to semantic knowledge, their likely account for this phenomenon is straightforward, as well: uniform semantic deficits would result from equal damage to each semantic module.

The domain-specific hypothesis has obvious intuitive and theoretical appeal. This theory is also the only one to date that is able to account for the extreme selective impairment of particular categories while leaving the others completely intact (patient E.W. being a highly-controlled example of these effects, Caramazza & Shelton, 1998). Nevertheless, despite these high level successes, we find the domain-specific

hypothesis to be lacking in several key respects. First, the theory is in essence nothing more than a recapitulation of the data. As a result, though it may be able to *explain* some phenomena, it is not a very useful tool for making new *predictions* with which to expand our understanding of semantic memory. Second, though the domain-specific hypothesis succeeds in accounting for some extreme cases not currently captured by reductionist accounts and their supporting connectionist models, the theory also shows no promise of parsimonious incorporation of the broader semantic deficit literature and cognitive neuroscience research which suggests that knowledge is partially organized by modality (Martin & Chao, 2001). In particular, cases in which the impaired category is also associated with a differential impairment of knowledge from a particular sensorimotor modality (e.g., McCarthy & Warrington, 1988; Lambon Ralph, Howard, Nightingale, & Ellis, 1998), or in which knowledge of particular *grammatical* categories is impaired (discussed in the next section) appear difficult to reconcile within this framework. To account for these data, the domain-specific hypothesis would likely need to be expanded so as to have independent processing modules associated with the semantic knowledge of each category for each modality, a postulation that substantially complicates the account without providing independent evidence warranting this complexity. Finally, the domain-specific hypothesis offers no insight into the qualitative differences in terms of the types of incorrect responses participants make (e.g., errors of commission vs. error of omission) when suffering from different types of semantic memory impairment.

Reductionist accounts of semantic impairments

Several attempts to account for the various semantic impairments have also been made using a reductionist approach. These theories differ in terms of whether there is an explicit semantic store (e.g., Farah & McClelland, 1991) or merely a routing hub for completing tasks that require mapping information from one sensorimotor or linguistic modality to another (e.g., Rogers, Lambon Ralph, Garrard, Bozeat, McClelland, Hodges, & Patterson, 2004) and whether there are innate modality-specific subdivisions of semantic knowledge (e.g., Warrington & Shallice, 1984; Farah & McClelland, 1991), a single amodal semantic store (e.g., Tyler, Moss, Durrant-Peatfield, & Levy, 2000), or some combination thereof (Simmons & Barsalou, 2003). However, all of this work shares the principle that different semantic impairments emerge as the result of an interaction between the rich statistical structure of the semantic knowledge associated with different categories and sensorimotor modalities (Cree & McRae, 2003), and basic architectural constraints in semantics – *not* from explicit category-specific semantic stores. To evaluate the plausibility of these accounts, several researchers have implemented their proposals in connectionist models which can be used to study and explore the semantic memory store, both before and after it has been subjected to simulated impairment.

The present discussion centers upon a recent theory and model of normal and impaired semantic memory outlined by Rogers et al. (2004). Though this account is not a comprehensive model of semantic impairments, it nevertheless captures many important aspects of the data and has generated interesting predictions which have

advanced the empirical characterization of semantic deficits. In realizing this, the Rogers et al. account makes several alterations to classical theories in which semantics is an explicit knowledge storehouse, and instead characterizes semantics as an inter-modal routing hub while also synthesizing important principles described in previous work – notably, the sensory-motor theory outlined by Warrington and Shallice (1984) and related computational implementation by Farah and McClelland (1991), and the notion of similarity-based organization in an amodal semantic store (Tyler et al., 2000).

The implemented model of Rogers et al.'s (2004) theory is shown in Figure 1 and is organized into three main groups of features – a pool of visual feature units, a pool of semantic units, and a pool of verbal units. The verbal units are further divided into several sub-groups representing different types of features: the name and the perceptual, functional, and encyclopedic features of a given concept. Both the verbal and visual feature units can serve as either inputs or outputs to the model and are therefore considered to be "visible" units. In contrast, the states of the semantic units are "hidden" from the external environment and are determined by the activation they receive from the units with which they have connections. Rogers et al. thus instantiate the claim that rather than semantics serving as an abstract storehouse of the semantic properties we encounter in the world, it is a learned communications hub which allows for the mapping of information from one pool of units to another as required during the completion of various day-to-day tasks. It is also worth noting that in making this assumption, Rogers et al. are able to circumvent the challenging issues surrounding how and what information should be represented in the semantic storehouse (Cree & McRae, 2003). Instead, they are able to focus on the external properties of objects

such as their visual features or the verbal descriptions individuals provide thereof which are more amenable to empirical research.

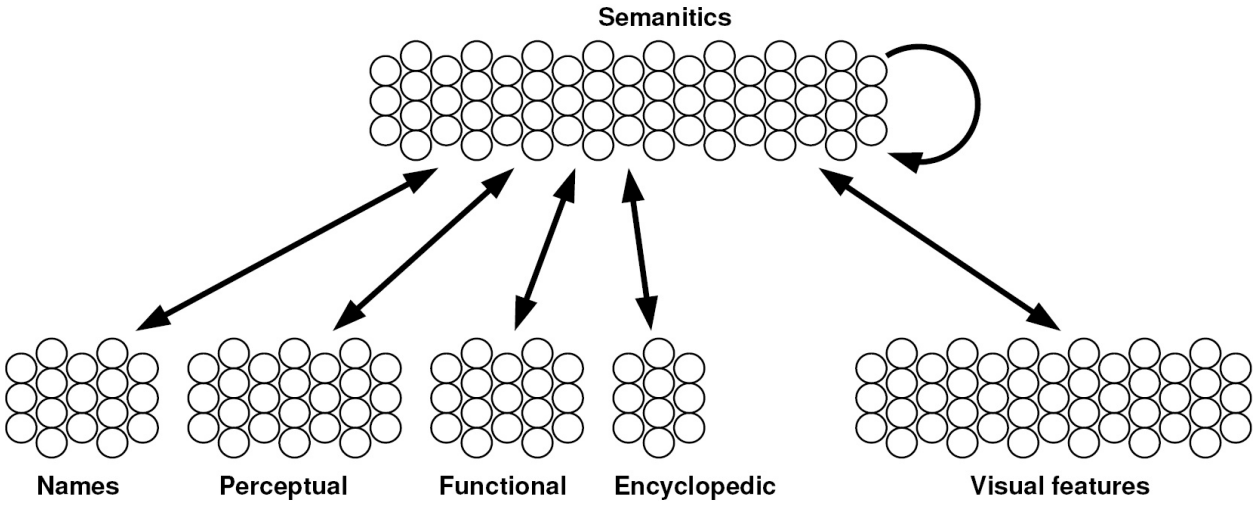


Figure 1. Network architecture of the Rogers et al. model. The network was divided into three pools of units representing the visual, semantic, and verbal features of concepts. The verbal units were further partitioned to represent name, perceptual, functional, and encyclopedic features. Note that whereas the visual and verbal unit activations for a given concept were pre-specified based on normative data, the hidden unit activations were an emergent characteristic of the network as it learned to complete different tasks requiring inter-modal mappings. (Adapted from Rogers et al., 2004)

Recognizing the value of bringing their model in line with existing empirical data, Rogers et al. (2004) carefully crafted the input and output representations used to train their model in two main ways. First, they attempted to capture the detailed statistical structure of the different categories via an analysis of the convergent results of visual and verbal norming studies, as this structure has previously been associated with categories' respective robustness to impairment. For instance, their representations captured the finding that living thing representations tend to be more similar overall than

nonliving thing representations. Second, the outputs that the model needed to produce for a given set of inputs were modeled after a series of semantic tasks analogous to those employed in testing patients: picture naming, word-to-picture matching, sorting, drawing and copying, and category-matching. Each of these tasks consisted of employing one pool of input features to activate another pool of output features via the hidden semantic layer (e.g., in picture naming, the model was trained to activate the correct verbal “name” units for a given concept when its visual feature units were presented as input).

The Rogers et al. (2004) model was successfully employed by Lambon Ralph et al. (2007) to model a subset of the category-specific deficits literature – specifically, how HSVE could be responsible for a living things deficit. Motivated by the behavioral finding that HSVE patients tend to make errors of commission, they speculated that HSVE causes an increase in signal noise in the semantic memory system. Patients are thus able to activate approximately the correct target, but the noise may incorrectly lead them to an incorrect response within the correct category. By virtue of the high similarity between living thing exemplars, this noise would lead to a differential increase in errors for living things. Computational simulations of HSVE in which random noise was added to the values of the weights replicated this pattern of effects – the simulated patients made an increased number of commission errors, and differentially more errors were observed for the living things category.

Rogers et al. (2004) also investigated the application of another type of damage to the same model with the aim of accounting for the uniform semantic impairments

associated with semantic dementia. Starting from the observation that semantic dementia patients tend to make errors of omission, Rogers et al. hypothesized that semantic dementia causes a “dimming”, or reduction in size, of the weights in the semantic memory system which would decrease the overall activity therein. This would lead to insufficient semantic activity to elicit a response on a given task. Computational simulations of this hypothesis first confirmed that weight dimming caused an equal impairment to knowledge in all domains, as in semantic dementia. Furthermore, Rogers et al. noted that though the overall error rates for the different categories were identical, the types of errors that the model made were different across the living and non-living thing categories. In particular, though errors of omission were high overall, there was a relatively greater likelihood of omitting knowledge for non-living thing knowledge, whereas there was a relatively greater likelihood of commission errors for living thing knowledge. They attributed the former to the low similarity between non-living thing exemplars causing each exemplar to be represented fairly separately from all others using a relatively sparse set of large-magnitude weights; a reduction in even a small portion of the weights necessary to activate a particular representation would therefore still direct the network’s state towards the sparsely populated region of semantic space where that exemplar was represented, but the reduced weights were insufficiently strong to drive semantic activation above the response threshold. They attributed the latter case to the fact that living thing exemplars are highly similar to one another, and resultantly a large numbers of moderately-sized weights therefore operate in unison to drive semantic activation above a response threshold. Dimming therefore did not lead to as substantial increase in errors of omission; however, as in the model of category-

specific deficits, the reduction in magnitude of the weights in semantics led to a loss in the fine-grained distinctions between the living thing exemplars and caused an increase in errors of commission. Motivated by these unexpected findings, Rogers et al. conducted follow-up behavioral studies on patients which corroborated the model's predictions.

Taken together, the reductionist account and accompanying connectionist model outlined by Rogers et al. (2004) represents an important advance in our understanding of semantic memory deficits. This is true in two main respects: first, this work provides a highly detailed account of the patient data at a level not offered by other theories to date. Second, in addition to accounting for existing data, this model also has furthered the understanding of semantic memory impairments by producing novel predictions which were later validated via neuropsychological evaluation. Additionally, all of this work was realized within a single model which readily lends itself to parsimonious extension to other semantic memory impairments (e.g., accounting for cases wherein knowledge of a particular category and a particular modality are simultaneously impaired).

Nevertheless, several issues remain. For instance, the current modeling framework has not been applied to understanding how non-living things could be differentially more impaired relative to living things. However, Rogers et al. (2004) speculate that adding representations for additional modalities which may be particularly salient for these domains and not others (e.g., motor representations capturing how objects are handled), combined with modality-specific biases on where simulated connection impairments occur, might allow the model to simulate this type of deficit.

Past modeling work by Farah and McClelland (1991), who selectively damaged particular regions of an explicit semantic memory store, which contained architecturally-separated subdivisions of sensory and functional knowledge and observed either living thing or nonliving thing deficits, lends empirical support to this intuition. The current framework also does not offer a detailed treatment of the patterning of exception categories observed in the patient literature – a key characteristic of category-specific semantic impairments, though the analysis of semantic feature norms by Cree and McRae (2003) indicates that the exception categories tend to have similar distributional characteristics as the main category of impairment; damage which affects the main category should also exert a similar pressure on the exception categories (e.g., musical instruments may group with living things in part because their auditory features are particularly useful for distinguishing amongst them).

Summary

Though neither the traditional domain-specific nor reductionist connectionist accounts of semantic memory impairments are all-encompassing at present, there appears to be additional theoretical value associated with the connectionist approach. The traditional account is able to address certain extreme patterns of data (e.g., selective impairment of non-living things), but appears to lack a parsimonious way of extending itself to the broader semantic-memory literature, and beyond (e.g., the grammatical category deficits discussed in the following section). By virtue of being a recapitulation of the data, the traditional account also does not offer any new predictions

against which we could further evaluate it and extend our empirical understanding of these phenomena.

In contrast, the reductionist connectionist account outlined by Rogers et al. (2004) is grounded in domain-general and independently verifiable principles and has directly contributed to expanding our understanding of semantic impairments. We therefore remain optimistic that future research along these lines, in conjunction with continued interleaving of neuropsychological assessment and the study of individuals who have not suffered from neural damage, will lead to the development of theories and models capable of accounting for the full gamut of deficits.

Grammatical category deficits

Interestingly, categorical impairments are not unique to the semantic domain; these kinds of deficits have also been reported for *grammatical* categories of words (e.g., nouns worse than verbs, De Renzi & diPellegrino, 1995; verbs worse than nouns, Berndt, Mitchum, Haendiges, & Sandson, 1997). Paralleling the semantic category literature, these impairments have been explained either with the traditional neuropsychological logic of the double dissociation or with attempts to reduce the impairments to damage of other, underlying knowledge. But despite the successes of computational models in accounting for semantic category deficits, they have yet to impact our understanding of noun- and verb-specific impairments in the same way. The present section evaluates recent advances in the computational literature and suggests how further research in this direction will yield insights into the nature of grammatical category deficits.

Traditional accounts of grammatical category impairments

While verbs may be inherently more susceptible to loss after damage for a variety of reasons (e.g., later acquisition during development in many languages; Gentner, 1982), the existence of patients with noun-specific deficits suggests a double dissociation between these two types of knowledge (but see Mätzig, Druks, Masterson, & Vigliocco, 2008). As such, one explanation of noun/verb impairments is that grammatical category is the relevant principle of organization, either among lexical representations (Caramazza & Hillis, 1991; Hillis & Caramazza, 1995) or morphological ones (Shapiro, Shelton, & Caramazza, 2000). To account for patients whose noun or verb deficits are modality-specific (i.e., restricted to particular *linguistic* input and output modalities; e.g., a deficit for verbs only in written naming), each input or output lexicon is also assumed to have representations organized by grammatical category (Hillis & Caramazza, 1995; Caramazza & Hillis, 1991). Because some patients have no obvious semantic impairment, these grammatical distinctions are presumed to be represented independently of semantic knowledge (Caramazza, 1997). For this reason, there is no predicted effect of *semantic* similarity among the impaired words; only a word's grammatical category is relevant for predicting its loss after damage.

This explanation of noun- or verb-specific deficits can account for most of the reported patterns of data, largely because these patients motivated the theory in the first place. Nevertheless, the prediction that all words in a grammatical category should be impaired irrespective of meaning has not been confirmed. Berndt, Haendiges, Burton, and Mitchum (2002) tested patients on abstract and concrete noun and verb reading,

but the results are not unequivocal: although two patients with verb deficits were indeed poorer at reading concrete *and* abstract verbs relative to nouns, no patients with noun deficits were tested, and so a difficulty effect cannot be ruled out.

Another problem with the grammatical category hypothesis is that it is inherently post-hoc, motivated by the observation that some patients were worse with nouns or verbs. When modality-specific grammatical category deficits were also observed, the theory was extended to include input and output lexical representations organized by grammatical category. Together with the standard neuropsychological account of semantic category deficits, the picture that emerges is one of areas specialized for each linguistic modality and semantic and grammatical category. We find such an organization needlessly complex, especially in the absence of any complementary reasons to stipulate these subdivisions. Furthermore, the observance of new patterns of categorical deficits would not easily be accommodated by the theory – except to propose the existence of even more specialized representations.

Reductionist accounts of grammatical category impairments

The existence of a high correlation between grammatical category and semantic category suggests that there may be another explanation for noun/verb impairments. To wit, many verbs are actions, and many nouns are objects – especially the nouns and verbs used to test patients – so some researchers have proposed that noun- or verb-selective deficits are the result of semantic damage to knowledge about objects or actions (Gainotti, Silveri, Daniele, & Guistolisi, 1995; Vinson & Vigliocco, 2002; Bird, Howard, & Franklin, 2000).

One such account (Bird et al., 2000) was motivated by Warrington and Shallice's (1984) modality-specific account of semantic knowledge as well as the overlap noticed by Gainotti et al. (1995) between lesions resulting in noun and living things deficits (left temporal lobe) and verb and non-living things deficits (left frontal and parietal lobes), respectively, (though see Capitani et al., 2003 for evidence suggesting that lesions resulting in non-livings things deficits are more widespread). Bird et al. (2000) propose that the meanings of verbs, like those of non-living things, are weighted more heavily towards functional semantic features relative to perceptual ones; nouns, in contrast, pattern like animate things and show the reverse weighting. Damage to sensory or functional information, then, is predicted to produce apparent category-specific deficits, and data from three patients with noun deficits who were also worse at naming living relative to non-living things supported their hypothesis. The verb deficit patients they tested, however, showed no grammatical category effect after the lower average imageability of verbs was taken into account, suggesting that many apparent verb deficits are the result of poorly designed testing materials.

On another reductionist account of grammatical category deficits, Vinson and Vigliocco (2002) used a computational model of semantic space to show that the meanings of words cluster by meaning and not grammatical category. Speaker-produced feature norms for nouns and verbs were used as the input to Kohonen maps, a class of connectionist models that learn without supervision to re-represent input patterns over an output "map" of lesser dimensionality (Kohonen, 1997). A winner-take-all learning procedure adjusts the weights between an input pattern and the output unit that responds maximally to it, but, critically, the weights of the winning unit's neighbors

are also adjusted. After learning, similar input patterns will cluster together topographically in the output map. In Vinson and Vigliocco's (2002) model, nouns and verbs clustered according to similarity of meaning rather than grammatical category; that is, action nouns ("the bombardment") were closer in distance to action verbs ("to bombard") than to object nouns. Lesions to specific areas on the map or to types of semantic features (e.g. visual features) produced disproportionate object or action word deficits in the context of a mild, medium, or severe overall impairment.

Critically, many behavioral, neurophysiological, and other neuropsychological studies independently support the idea of a general object/action distinction in the brain (see Milner & Goodale, 1995 for a review). While some neuroimaging studies contrasting nouns and verbs have yielded inconsistent results (finding effects of grammatical category, Shapiro et al., 2005; Perani et al., 1999; or not, Tyler, Russell, Fadili, & Moss, 2001), these discrepancies may also be a result of the meaning/grammatical category correlation. When nouns and verbs were matched for semantic properties (e.g., all words referred to actions), no difference between grammatical categories was found (Siri et al., 2008). Similarly, when manipulable objects and actions involving manipulation were compared, there was an effect of manipulability but not grammatical category (Saccuman et al., 2006).

However, neither traditional neuropsychological nor reductionist accounts of grammatical category deficits have adequately addressed the way in which the adult structure is learned, a research question well-suited to a computational modeling approach, and to connectionist models, in particular. Because models with learned internal representations do not require prior commitment to the representational

structure necessary to complete the task at hand, they provide a way to explicitly investigate the acquisition of knowledge and the learnability of a hypothesized organization. Of course, for either account, the question of learnability could be circumvented with the claim that such an organization is innate – but this claim raises further questions about the selective pressures that could produce such an innately specified organization. Instead, computational models can be used to test the hypothesis that grammatical category deficits may emerge after damage to semantic knowledge shaped only by domain-general learning mechanisms, the characteristics of nouns and verbs in the environment, and the way in which people respond to them.

A connectionist model of grammatical category impairments

Watson (2009) implemented a distributed connectionist model with learned semantic representations and no explicit instruction on the grammatical category of words; the architecture of the model is shown in Figure 2. As in Rogers et al.'s (2004) account of semantic category deficits, “semantics” in the model (the four oval-shaped groups) was not conceived of as an amodal store of features but instead as a learned set of representations that develop under pressure to perform various linguistic and conceptual tasks successfully (see also Plaut, 2002). In this model, phonological output was required in response to the visual or auditory input associated with it (i.e., naming from vision or audition). Additionally, if the word had an action associated with it (i.e., “to kick”), the model was required to produce that action in addition to the action’s name. The role of the hidden semantic units, then, was to use a domain-general

learning algorithm to develop representations that enabled the model to successfully complete the tasks at hand.

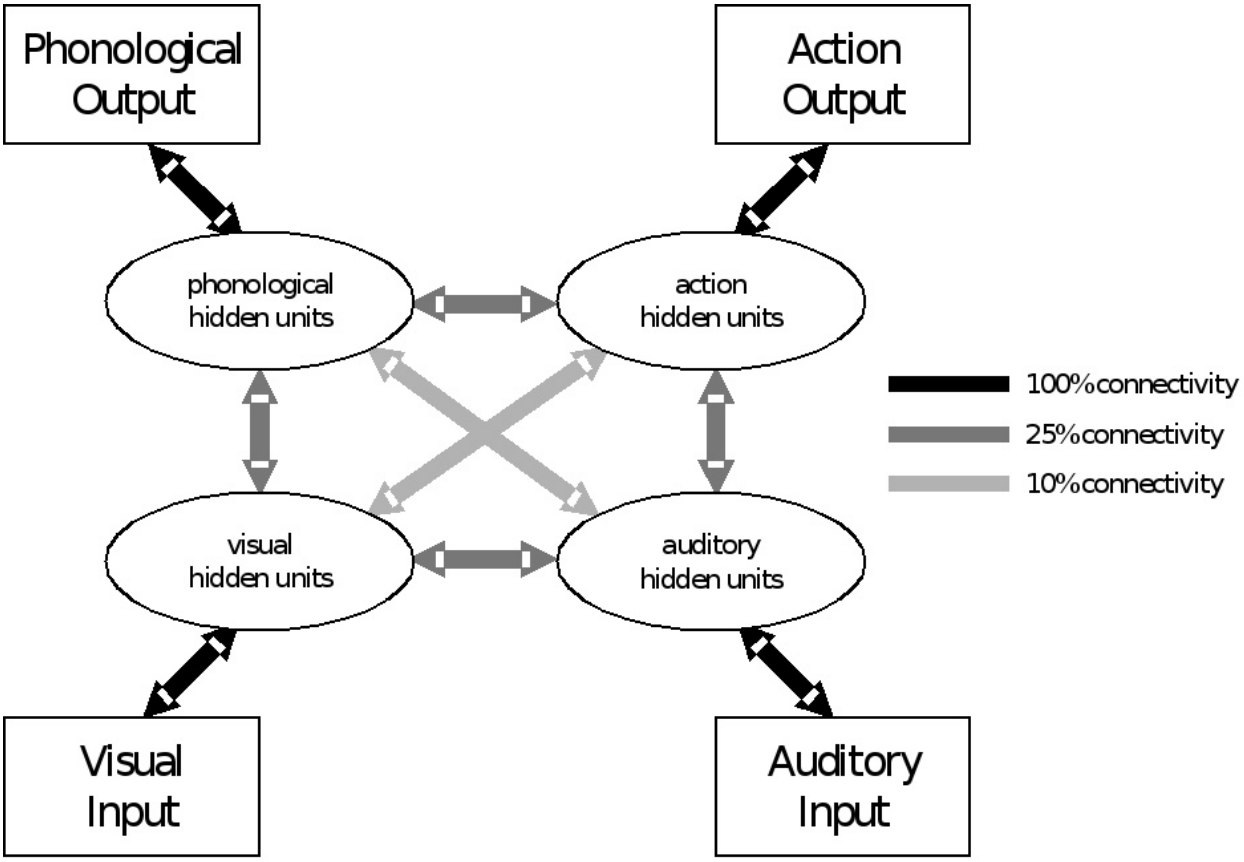


Figure 2. Network architecture of the Watson (2009) model. Input and output layers (26 units each) are represented by rectangles; hidden semantic layers (60 units each) are represented by ovals. The density of the connectivity between two layers is represented by the increasingly dark shades of grey: darker is equivalent to denser connectivity.

These semantic representations were further shaped by a topographic bias (Jacobs & Jordan, 1992; Plaut, 2002) on the connectivity between units: units in layers close to one another were fully interconnected, while units in layers far from one another had sparser connectivity. In the model, groups of hidden units that communicated with

particular input or output modalities were fully connected to them. However, the connectivity *between* groups of hidden units was constrained by distance; for instance, only 10% of the "action hidden units" were connected to the hidden group farthest from it, the "visual hidden units" (see Figure 2). This constraint instantiated computationally properties of neurons in the brain: because of size limitations, there is pressure for connections between neurons to be local rather than long-distance (Nelson & Bower, 1990). As a result, neighboring neurons will come to be strongly interconnected and to respond similarly.

Part of the motivation to include this representational pressure came from the success of a computational simulation of a different aphasic syndrome, *optic aphasia* – a selective impairment in naming visually presented objects (Plaut, 2002). In this model, imposing a topographic bias on the semantic hidden units caused units closest to particular input or output modalities to become "functionally specialized" for representing modality-specific aspects of knowledge. As a result, damage to connections from vision to the semantic units that were partially specialized for representing phonological information produced a selective impairment for naming objects presented visually (i.e., optic aphasia).

In the context of grammatical category deficits, if nouns and verbs learn to rely on different sets of functionally specialized semantic units, damage to these units could produce disproportionate deficits for one category or the other. But how do modality-specific areas become differentially important for naming nouns or verbs? Essentially, particular input or output modalities offer sources of information that may be more or less reliable during learning, and these sources of information are hypothesized to vary

between grammatical categories, on average. Previous behavioral (e.g., Myung, Blumstein, & Sedivy, 2006) and neuroimaging (e.g., Hauk, Johnsrude, & Pulvermüller, 2004; Martin, Haxby, Lalonde, Wiggs, & Ungerleider, 1995) data have suggested that visual and action knowledge participate in the meanings of nouns and verbs to differing degrees, and the representations and tasks given to the model reflected these differences. Because verbs often refer to actions, the model was required to produce both the name and the associated action during naming of a verb. Additionally, the visual representations of nouns were richer than those of verbs, reflecting the assumption that there is more detailed (e.g., color, form) and more consistent (in time and across instances) visual knowledge associated with objects than with actions.

The result of using realistic tasks and representations in conjunction with the topographic bias on connectivity between semantic hidden units was that particular groups of units became more important for one grammatical category or the other – and this difference yielded grammatical category deficits after damage (Figure 3). In particular, the group of hidden units strongly connected to action output became specialized for representing action knowledge, and this same action knowledge was recruited during naming. As a result, damage here produced relative verb deficits (top left panel). On the other hand, the hidden units strongly connected to visual input became specialized for representing visual knowledge. Given the richness of visual knowledge for nouns and the lack of other associated information (i.e., action), nouns were particularly affected by damage to the visual hidden units (bottom right panel). Although a difference between grammatical categories was not predicted after damage to the hidden units closest to phonological output or auditory input, a slight advantage

for verbs emerged; this pattern was attributed to the higher average difficulty of nouns and was predicted to disappear if the visual representations of verbs more closely mirrored the variable and inconsistent visual input associated with actions in the world.

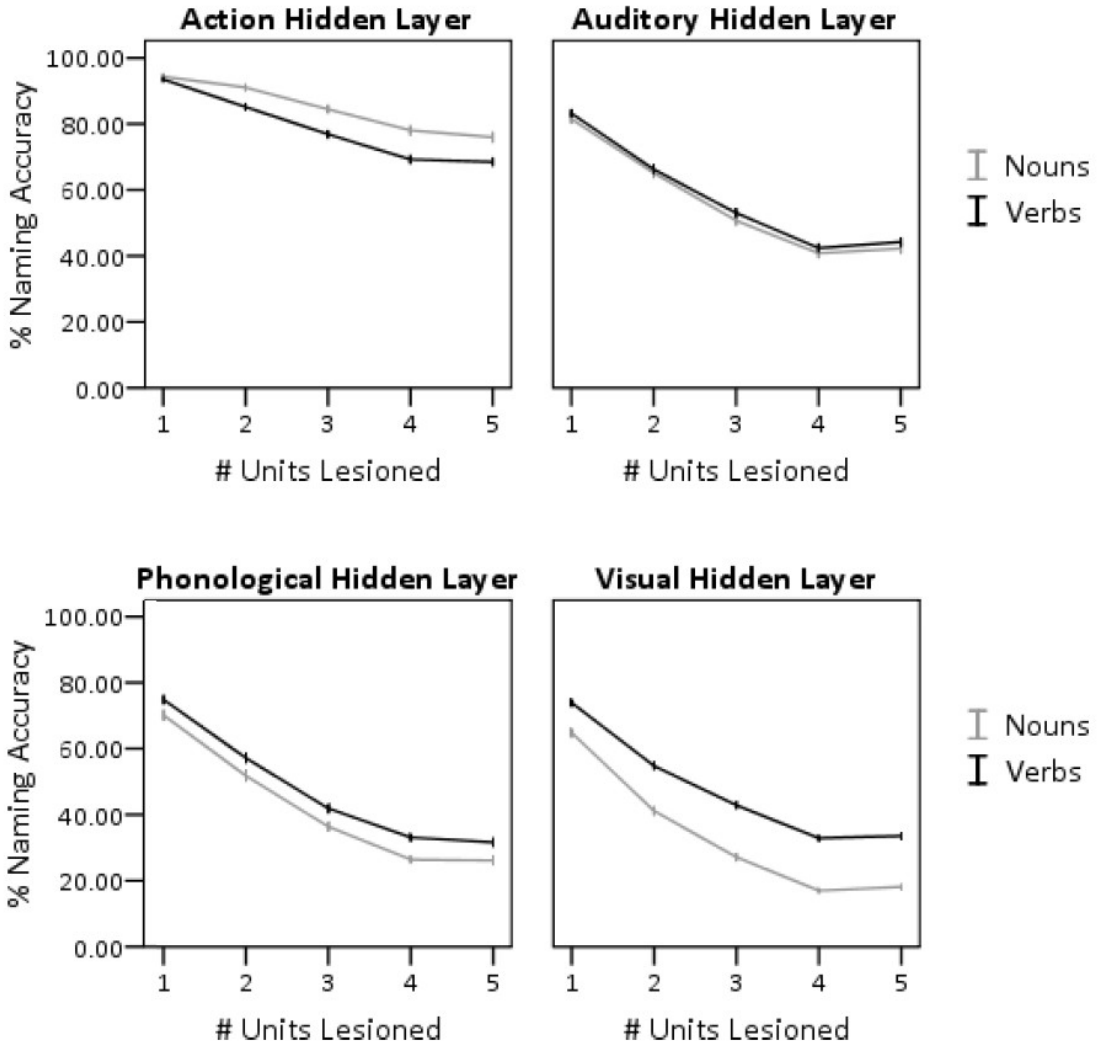


Figure 3. Noun and verb naming accuracy after removing individual hidden units in each semantic layer of the Watson (2009) model. Results are from naming from visual input, and only the phonological output was considered for correctness. Error bars represent +/- 1 standard error of the mean.

These modeling results support a more parsimonious and principled alternative to the view that representations must be organized by grammatical category to produce grammatical category deficits and addresses the major failing of grammatically-based accounts – the problem of the way in which the adult structure comes to be. Although this domain, in particular, will benefit from continued computational investigations, the results of this model show that armed only with domain-general principles of learning and processing, grammatical category effects can emerge from a model required to learn stimuli and tasks similar to those encountered by people.

Selective impairments in word reading

Semantic representations play a key role in connectionist accounts of both category-specific and grammatical-class impairments, as covered in the first two sections of this chapter. The same turns out to be true in the third domain we consider – impairments in single word reading resulting from brain damage, known as the *acquired dyslexias*.

The traditional account of oral reading is the *dual-route* model (Coltheart, 1978), which posits that there are two separate mechanisms involved in translating print to sound. The first, termed the *nonlexical* pathway, captures the systematic relationships between spelling and sound, typically in the form of grapheme-phoneme correspondence (GPC) rules (e.g., G → /g/; M → /m/; A_E → /ei/; V → /v/). Such rules generate correct pronunciations for *regular* words like GAVE, as well as for pronounceable nonwords like MAVE. However, about 20% of English words are *irregular* or *exceptions* (e.g., HAVE), in that the GPC rules yield a mispronunciation, termed a *regularization* error (“haive”). Since skilled readers can pronounce HAVE

and other irregular words correctly, dual-route theories posit a second, *lexical* pathway that translates written words onto spoken words directly. However, because it relies on word-specific knowledge, the lexical pathway cannot pronounce nonwords. Thus, on a dual-route account, skilled readers need a nonlexical route to pronounce nonwords and a lexical route to pronounce irregular words. Interestingly, on most dual-route accounts, semantic representations are not directly involved in pronouncing words. The most influential computational implementation of a dual-route account is the Dual-Route Cascaded (DRC) model of Coltheart, Rastle, Perry, Langdon, and Ziegler (2001).

Traditional accounts of the acquired dyslexias

At first glance, the dual-route model would seem to receive strong support from the patterns of impairments in word reading that occur following brain damage. Patients with *phonological dyslexia* read both regular and exception words well, but make many errors on nonwords, often producing an incorrect word in response, termed a *lexicalization* error. By contrast, patients with *surface dyslexia* read regular words and pronounceable nonwords well but make regularization errors on exception words, particularly those of low frequency (e.g., PINT → "pihnt"; FLOOD → "flude"). The straightforward dual-route account is that phonological dyslexics have damage to the nonlexical pathway (impairing nonwords), whereas surface dyslexics have damage to the lexical pathway (impairing exception words).

As it turns out, the dual-route account of each of these patterns of performance runs into some difficulties. One of the challenges facing these efforts is that, as Coltheart et al. (2001) point out, "to simulate extreme versions of these two acquired

dyslexias is trivial with the DRC model" (p. 242) – disabling the lexical pathway produces regularizations of *all* exception words while leaving regular words and nonwords unaffected, whereas disabling the GPC rules *eliminates* nonword reading while leaving word reading unaffected. The problem is that, although these patterns are natural to produce on a dual-route account, neither has ever been observed: impairment of exception words is always modulated by word frequency, and severe nonword reading impairment is always accompanied by some impairment to word reading.

We will focus on surface dyslexia in the remainder of this chapter as it bears more directly on the role of semantic representations in word reading; see Nickels, Biedermann, Coltheart, Saunders, and Tree, (2007) for attempts to simulate phonological dyslexia using the DRC model, and Harm and Seidenberg (1999) and Welbourne and Lambon Ralph (2007) for relevant connectionist simulations.

A connectionist account of surface dyslexia

In contrast to dual-route accounts of word reading, the "triangle" model (Harm & Seidenberg, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989) posits that word reading is supported by cooperative and competitive interactions among orthographic, phonological, and semantic representations (often depicted as the three points of a triangle; see Figure 4). Each type of information is encoded as patterns of activity over a group of neuron-like processing units, and the knowledge that governs their interactions is instantiated by weights on connections between them (via additional groups of "hidden" units). Within the triangle framework, there are no word-specific representations; rather, the system learns to make the

orthographic, phonological, and semantic patterns for each familiar word a stable configuration over the entire network. Although interactions among orthography and phonology capture systematic spelling-sound knowledge, and interactions of each with semantics capture word-specific knowledge, the entire network participates in processing all types of stimuli: regular words, irregular words, and nonwords.

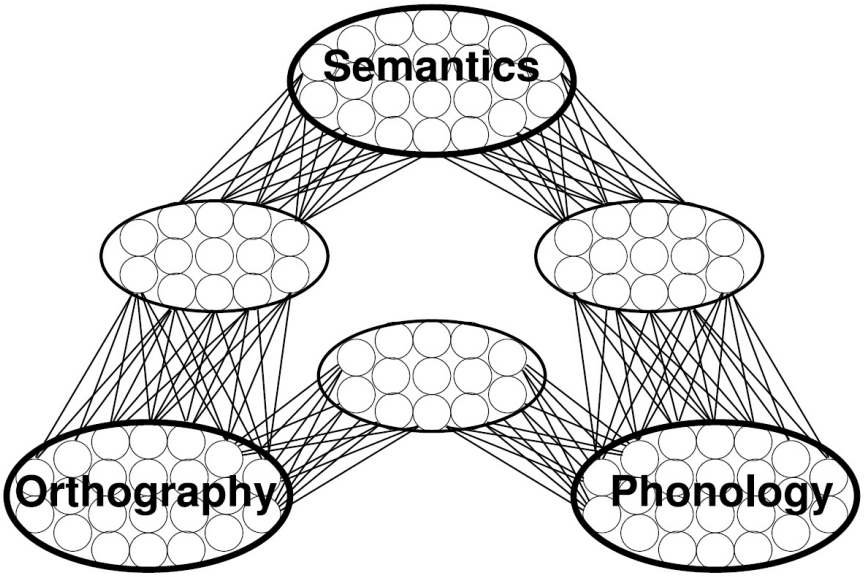
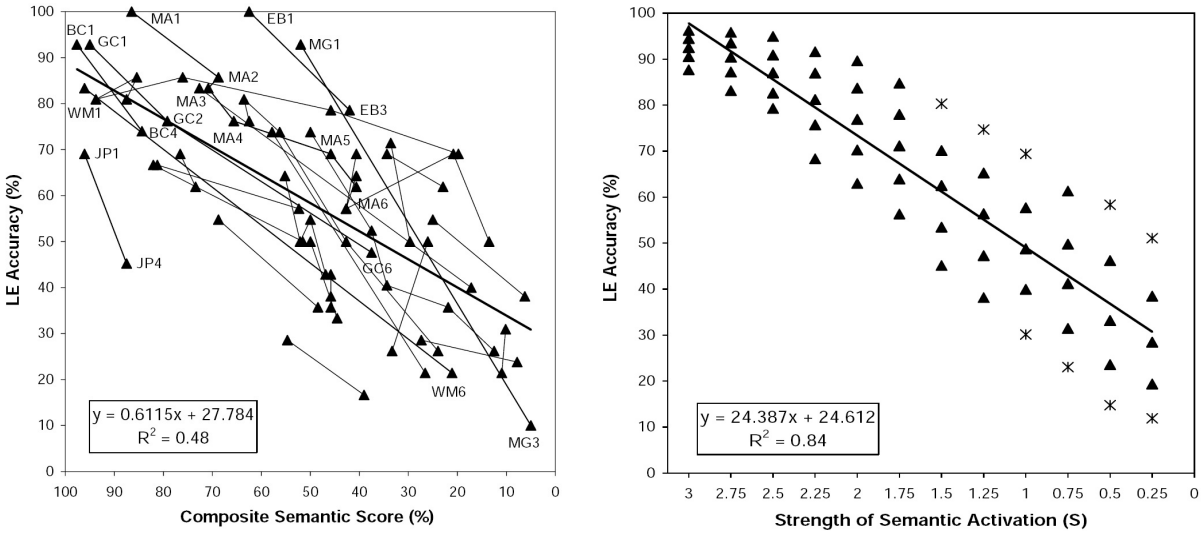


Figure 4. The triangle framework of word reading, in which patterns of activity in orthography, phonology and semantics interact and mutually constrain each other (via intermediate groups of "hidden" units) in processing both words and nonwords. (Adapted from Plaut et al., 1996)

Plaut et al. (1996) presented a number of connectionist simulations in which networks based on the triangle framework learned to map from orthography to phonology for both regular and irregular words, and yet also generalized well to nonwords. These results belied dual-route claims that good performance on both irregular words and nonwords requires separate mechanisms. However, none of the

networks, when damaged, provided a good match to surface dyslexia. A much better account was provided by more closely approximating the full triangle framework: training an orthography-to-phonology network in the context of support from semantics, and then damaging the model by progressively weakening this semantic support.

Thus, one of the main points of contention regarding surface dyslexia is whether it is due to semantic damage or, as the dual-route model claims, to lexical damage separate from semantics. Woollams, Lambon Ralph, Plaut and Patterson (2007) presented 100 observations from 51 patients with a form of progressive semantic deterioration known as semantic dementia, and showed a systematic relationship between the severity of semantic impairment and irregular word reading (see Figure 5a). Critically, patients who initially showed a semantic impairment but normal exception-word reading (e.g., MA, EB) – which might seem inconsistent with the triangle model account (Blazely, Coltheart & Casey, 2005) – went on to exhibit surface dyslexia along with the rest of the patients on further testing. Woollams et al. also showed that versions of the Plaut et al. (1996) simulation of surface dyslexia, varying in magnitude of semantic support during training, exhibited a similar distribution of performance with the progressive removal of semantics (see Figure 5b).



(a) (b)

Figure 5. (a) Performance in reading low-frequency exception (LE) words of 100 observations of 51 patients, as a function of a measure of semantic integrity. Longitudinal observations from the same patient are connected by lines. b) Performance of parametric variations of a connectionist simulation of word reading (varying in strength of semantic support during training) as a function of semantic damage (i.e., weakening of semantic support). (Adapted from Woollams et al., 2007)

In response to Woollams et al.'s (2007) empirical findings and simulations, Coltheart, Tree and Saunders (2010) presented DRC simulations in which they generated 40,200 versions of the model by administering all possible combinations of severity of damage to the lexical pathway (removing the X lowest-frequency orthographic word units) and the nonlexical pathway (removing the Y least-frequently used GPC rules). They then identified the 100 versions that most closely matched Woollams et al.'s 100 observations. Unfortunately, this data fitting was done without regard to possible measurement error and without regard to the consistency of the ascribed damage to longitudinal observations of a given patient. The result is that, not only does the work provide no explanation of why exception-word reading is related to

semantic impairment, but it also leads to highly implausible claims regarding the progression of damage in some patients. For example, according to the DRC fits for one patient, lexical damage decreased from 45% to 34.5% before increasing again to 46%, while nonlexical damage steadily decreased from 47.5% to 20% to 9.5%. Thus, Coltheart et al.'s (2010) ``account'' of this patient is that, despite having what is undeniably a degenerate disease, the patient's nonlexical route recovered nearly completely over the course of testing (see Woollams, Lambon Ralph, Plaut, & Patterson, 2010, for further discussion).

In summary, the connectionist “triangle” framework for word reading provides insight into why the pronunciation of low-frequency exception words in patients with surface dyslexia should be related to the severity of their semantic impairment. By contrast, even when an implementation of a traditional dual-route model can be made to fit the same data, it fails to provide the same insight into the phenomenon.

Conclusion

Connectionist modeling is based on the belief that certain computational principles of neural systems are fundamental to understanding both normal and disordered cognition. The value of modeling is not so much to fit particular patterns of observed data, as to provide a vehicle for exploring the implications of a set of theoretical claims concerning the representations and processes underlying cognition, and how they are impacted by brain damage. We have illustrated the value of this approach in three domains: semantic impairments, grammatical-class impairments, and acquired dyslexia. In each case, the traditional account reifies the relevant behavioral distinction in the

structure of the system itself, with the result that the account provides little insight into why the phenomena pattern as they do.

By contrast, the alternative, connectionist accounts attempt to explain the observed patterns of data as resulting from more basic representational or processing commitments. Although the existing modeling work certainly has limitations, in each case the approach holds the promise of providing a deeper understanding of how brain processes support cognitive processes, both in neurologically intact individuals and in those who have suffered brain damage.

CHRISTINE E. WATSON is a Postdoctoral Research Fellow in the Department of Psychology and the Center for Cognitive Neuroscience at the University of Pennsylvania in Philadelphia, PA, USA. She uses behavioral studies of normal and brain-damaged individual, in conjunction with computational modeling, to elucidate the relationship between action and event comprehension and verb knowledge, as well as the neural bases of solving verbal and non-verbal analogies. Email: watsonc@mail.med.upenn.edu

BLAIR C. ARMSTRONG is a graduate student in the Ph.D. program of the Department of Psychology at Carnegie Mellon University in Pittsburgh, PA, USA. His research involves using behavioral studies and computational modeling to examine the temporal dynamics of word comprehension and decision processes, and how these adapt in response to feedback. Email: blair.c.armstrong@gmail.com

DAVID C. PLAUT is a Professor in the Department of Psychology and the Center for the Neural Basis of Cognition of Carnegie Mellon University in Pittsburgh, PA, USA. His research involves using computational modeling, complemented by empirical studies, to investigate the nature of normal and disordered cognitive processing in the domains of reading, language, and semantics. Email: plaut@cmu.edu

References

Berndt, R.S., Haendiges, A.N., Burton, M.W., & Mitchum, C.C. (2002). Grammatical class and imageability in aphasic word production: their effects are independent. *Journal of Neurolinguistics, 15*, 353-371.

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

Bird, H., Howard, D., & Franklin, S. (2000). Why is a verb like an inanimate object? Grammatical category and semantic category deficits. *Brain and Language, 72*, 246-309.

Blazely, A. M., Coltheart, M. & Casey, B. J. (2005). Semantic impairment with and without surface dyslexia: Implications for models of reading. *Cognitive Neuropsychology, 22,* 695-717.

Capitani, E., Laicona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology, 20*, 213-261.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*, 177-208.

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

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). New York: Academic Press

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

Coltheart, M., Tree, J., & Saunders, S. J. (2010). Computational modeling of reading in semantic dementia: Comment on Woollams, Plaut, Lambon Ralph and Patterson (2007). *Psychological Review, 117,* 256-272.

De Renzi, E. & di Pellegrino, G. (1995). Sparing of verbs and preserved, but ineffectual reading in a patient with impaired word production. *Cortex, 31*, 619-36.

Gainotti, G., Silveri, M. C., Daniele, A., & Giustolisi, L. (1995). Neuroanatomical correlates of category-specific semantic disorders: a critical survey. *Memory, 3*, 247–264.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj (Ed.), *Language development: Vol. 2. Language, thought, and culture*, 301-334. Hillsdale, NJ: Erlbaum.

Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*, 491-528.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*, 301-307.

Hillis, A.E. & Caramazza, A. (1995). Representations of grammatical categories of words in the brain. *Journal of Cognitive Neuroscience, 7*, 396-407.

Jacobs, R.A., & Jordan, M.I. (1992). Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience, 4*, 323-336.

Kohonen, T. (1995). *Self-organization maps*. New York: Springer-Verlag.

Lambon Ralph, M.A., Howard, D., Nightingale, G., & Ellis, A.W. (1998). Are living and non-living category-specific deficits causally linked to impaired perceptual or associative knowledge? Evidence from a category-specific double dissociation. *Neurocase, 4*, 311-338.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain, 130*, 1127-1137.

Magnié, M., Ferreira, C. T., Giusiano, B., & Poncet, B. (1999). Category specificity in object agnosia: preservation of sensorimotor experiences related to objects. *Neuropsychologia, 37*, 67-74.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology, 11*, 194, 201.

Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L., & Ungerleider, L. G. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science, 270,* 102-105.

Mätzig, S., Druks, J., Masterson, J., & Vigliocco G. (2009). Noun and verb differences in picture naming: Past studies and new evidence. *Cortex*, *45*, 738-758.

Milner, A. D., & Goodale, M. A. (1995). *The visual brain in action*. Oxford: Oxford University Press.

McCarthy, R. A., & Warringon, E., K. (1988). Evidence for modality-specific meaning systems in the brain. *Nature*, 334, 428-430.

Myung, J., Blumstein, S.E., & Sedivy, J.C. (2006). Playing on the typewriter, typing on the piano: manipulation knowledge of objects. *Cognition, 98*, 223–43.

Nickels, L., Biedermann, B., Coltheart, M., Saunders, S. & Tree, J. J. (2007). Computational modelling of phonological dyslexia: How does the DRC model fare? *Cognitive Neuropsychology, 25,* 165-193.

Noppeny, U., Patterson, K., Tyler, L. K., Moss, H., Stamatakis, E. A., Bright, P., Mummery, C., & Price, C. J. (2007). Temporal lobe lesions and semantic impairment: a comparison of herpes simplex virus encephalitis and semantic dementia. *Brain, 130*, 1138-1147.

Perani, D., Cappa, S. F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M. M., & Fazio, F. (1999). The neural correlates of verb and noun processing: A PET study. *Brain, 122,* 2337–2344.

Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19*, 603-639.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22*, 253 336.

Pulvermüller, F., Härle, M., & Hummel, F. 2001: Walking or talking?: Behavioral and electrophysiological correlates of action verb processing. *Brain and Language, 78,* 143-168.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P. A., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111*, 205-235.

Saccuman, M. C., Cappa, S. F., Bates, E. A., Arevalo, A., Della Rosa, P., Danna, M., & Perani, D. (2006). The impact of semantic reference on word class: An fMRI study of action and object naming. *NeuroImage, 32*, 1865–1878.

Santos, L. R., & Caramzza, A. (2002). The domain-specific hypothesis : A developmental and comparative perspective on category-specific deficits. In E. M. E. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind*, 1-23. New York: Psychology Press.

Sartori, G., & Job, R. (1988). The oyster with four legs: A neuropsychological study on the interaction of visual and semantic information. *Cognitive Neuropsychology, 5,* 105-132.

Shapiro, K., Shelton, J., & Caramazza, A. (2003). Grammatical class in lexical production and morphological processing: Evidence from a case of fluent aphasia. *Cognitive Neuropsychology, 17*, 665-682.

Shapiro, K. A., Mottaghy, F. M., Schiller, N. O., Poeppel, T. D., Fluss, M. O., Müller, H. W., Caramazza, A., & Krause, B. J. (2005). Dissociating neural correlates for nouns and verbs. *NeuroImage, 24*, 1058–1067.

Simmons, K., & Barsalou, L. W. (2003). The similarity in topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, 20, 451-486.

Siri, S., Tettamanti, M., Cappa, S. F., Della Rosa, P., Saccuman, C., Scifo, P., & Vigliocco, G. (2008). The neural substrate of naming events: effects of processing demands but not of grammatical class. *Cerebral Cortex, 18,* 171-177.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75*, 195-231.

Tyler, L. K., Russell, R., Fadili, J., & Moss, H. E. (2001). The neural representation of nouns and verbs: PET studies. *Brain, 124,* 1619 –1634.

Vinson, D.P, & Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: Semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics, 15*, 317-351.

Vigliocco, G., Vinson, D.P., Lewis, W., & Garrett, M.F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48*, 422-488.

Warrington, E. K., & McCarthy, R. A. (1983). Category specific access dysphasia. *Brain, 106*, 859-878.

Warrington, E. K., & McCarthy, R. A. (1987). Categories of knowledge: Further fractionations and an attempted integration, *Brain, 110*, 1273-1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*, 829-854.

Watson, C. E. (2009). *Computational and behavioral studies of normal and impaired noun/verb processing.* Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University.

Welbourne, S. R. & Lambon Ralph, M. A. (2007). Using PDP models to simulate phonological dyslexia: The key role of plasticity-related recovery. *Journal of Cognitive Neuroscience*, 19, 1125-1139

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review, 114,* 316-339.

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Patterson, K. (2010). SD-squared revisited: Reply to Coltheart, Tree, & Saunders (2010). *Psychological Review.*