Double Dissociations Emerge in a Flat Attractor Network

Ihintza Malharin (i.malharin@bcbl.eu)
BCBL. Basque Center on Cognition, Brain & Language, 69 Mikeletegi 20009 Donostia-San Sebastián. Gipuzkoa, Spain

Simona Mancini (s.mancini@bcbl.eu)
BCBL. Basque Center on Cognition, Brain & Language, 69 Mikeletegi 20009 Donostia-San Sebastián. Gipuzkoa, Spain
Ikerbasque. Basque Foundation for Science, Bilbao, Spain

James S. Magnuson (j.magnuson@bcbl.eu, james.magnuson@uconn.edu)
BCBL. Basque Center on Cognition, Brain & Language, 69 Mikeletegi 20009 Donostia-San Sebastián. Gipuzkoa, Spain
Ikerbasque. Basque Foundation for Science, Bilbao, Spain
Department of Psychological Sciences, University of Connecticut, Storrs, Connecticut 06269 USA

Abstract

Double dissociations were long considered a gold standard for establishing functional modularity. However, Plaut (1995) demonstrated that double dissociations could result without underlying modularity. He damaged attractor networks with separate orthographic and semantic layers (as well as a hidden layer with feedback connections from semantics) that were trained to map orthography to semantics. Damaging connections coming from either the orthographic layer or recurrent semantic connections (to and from cleanup units) could both yield double dissociations, with some models exhibiting greater relative deficits for abstract words, and others for concrete words. We investigated whether double dissociations would emerge in a simpler attractor network with 2 sets of units (orthographic and semantic) and 2 layers of connections (orthographic-to-semantic and recurrent semantic connections). Random damage to orthographic-semantic connections yielded double dissocations (some damaged models showed stronger relative deficits for abstract words, while others showed stronger relative deficits for concrete words). Semantic-semantic damage led only to concrete deficits. The presence of double dissociations given different degrees of damage in each model reconfirm Plaut's (1995) findings in simpler, "flat" attractor network (O'Connor, Cree, & McRae, 2009), with less potential for modularity. The tendency for concrete impairments given damage to the semantic attractor level is at once surprising and revealing; it demonstrates a division of labor (and partial modularity) that emerges in this network. We will discuss theoretical implications, as well as next steps in this research program.

Keywords: double dissociations; concrete and abstract words; attractor network

Introduction

Behavioral double dissociations provide strongly suggestive evidence for modularity of functions, especially when supported by consistent neuroanatomical evidence (e.g., distinctly different lesion locations for the two selective impairments). For example, one patient with a selective impairment in processing abstract words and another with a concrete deficit would suggest separable representations and/or functions.

The logic behind complementary selective impairments is that a 'single dissociation' of function A (impairment in just that function) does not demonstrate modularity because it may be, for example, that functions A and B rely on the same system, but A breaks down before B does. For example, a person with a damaged knee might exhibit normal walking (function B) but impaired running (function A). This does not demonstrate that the knee is not important for walking; this pattern would emerge simply because function B puts greater demands on the system than function A.

Similarly, in a cognitive domain, we cannot be satisfied that a single impairment identifies a separable function. For example, if we observed a patient with a long-term memory deficit but no deficit in short-term memory, we could not conclude that long-term memory is a separate system from short-term memory – it could just be that long-term memory puts more demand on a single system, or that long-term memory depends upon short-term memory. If we observe a patient with a complementary deficit (impaired short-term memory but unimpaired long-term memory), we would have a double dissociation, and initial evidence for separable functions. Ultimately, we would want to support this with anatomical evidence (e.g., evidence that the two patients have damage in different locations).

It is also important to note that double dissociations are not limited to cases of pure impairment (function A impaired, function B 100% intact). Researchers also view 'disproportionate' impairments (functions A and B are both impaired, but A is much more impaired) as relevant cases (Caramazza & Mahon, 2003). There are risks to over-interpreting such cases, but nonetheless, many researchers view double dissociations in disproportionate impairment (patient 1: strongly impaired in function A, weakly in function B; patient 2: weakly impaired in function A, strongly impaired in function B) as relevant and potentially strong evidence for modularity. While there remains vigorous debate about what constitutes a double dissociation, and various ambiguities in attempting to interpret them (Davies, 2010), the modal view is that double dissociations provide a gold standard for initial evidence of modularity.

In a review of different theories on the organisation of conceptual knowledge aiming to explain category-specific deficits, Caramazza and Mahon (2003) provided an account of the main hypotheses and linked them to functional neuroimaging evidence available at that time. While the three

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

theories they discuss – the Sensory/Functional hypothesis developed by Warrington and McCarthy (1987), the Domain General hypothesis by Caramazza and Shelton (1998), and the Conceptual Structure account by Tyler, Moss, Durrant-Peatfield, and Levy (2000) – made predictions that went against some patient-based evidence, some of the assumptions of each model seemed very promising. Both the Sensory/functional theory (Warrington & McCarthy, 1987) and the Domain Specific hypothesis (Caramazza & Shelton, 1998) postulate that the semantic system is organised into modality-specific semantic subsystems. However, most of these model's predictions have been disproven by patient data, as patients showed deficit patterns that could not be explained by these models or fully opposed their predictions (see Caramazza & Mahon, 2003, for a comprehensive review of these models).

On the other hand, the Conceptual Structure account by Tyler et al. (2000) assumes random damage to an amodal conceptual system to explain category-specific semantic deficits. The assumptions originally presented by these authors relate to the categories of living and non-living things, but as some of our assumptions are similar, we will expand them to concrete and abstract words (the focus of our modeling below). Specifically, Tyler et al. (2000) assume that living things have more shared features than non-living things. Additionally, they suppose that the features of living things are more highly correlated with *shared* perceptual properties whereas for artifacts, functional information is more highly correlated with *distinctive* perceptual features. Finally, they assume that highly correlated features are more resistant to damage than less correlated features.

This motivates interesting computational questions. If double dissociations predict functional separation, and functional separation predicts anatomical separation, modeling double dissociations would require an underlying model with separate components (modules) implementing the separable functions. Conversely, the double dissociation logic predicts that we should not be able to *simulate* double dissociations in a nonmodular system, nor by damaging a single subcomponent in a modular system.

In fact, there have been multiple computational demonstrations that double dissociations can emerge without modularity, challenging the standard logic. Most relevant for us is a demonstration that abstract vs. concrete double dissociations can emerge when an attractor network with separate orthographic and semantic layers is damaged. In this study, Plaut (1995) attempted to model deep dyslexia by damaging a connectionist model of reading. Plaut's network used 7 sets of units and 13 layers of connections (although only 4 unit layers and 5 layers of connections were relevant for his simulations), including separate orthographic, semantic and orthographic layers, as well as clean-up layers and intermediate (hidden) layers. Plaut hypothesized that damaging connections between the orthographic layer and the intermediate layer would lead to better performance on concrete words compared to abstract words, while damaging the connections between the semantic layer and the clean-up would lead to better performance on abstract words than concrete words.

Damage to the orthographic-to-hidden pathway led mainly to abstract deficits, while damage to a semantic-to-cleanup pathway led mainly to concrete deficits. However, random damage to either pathway could result in *either* kind of deficit (note that here we refer to 'disproportionate' impairments, where one category is substantially more impaired than the other, even if both are degraded). Because only one patient with a complementary selective impairment is sufficient to support a modularity hypothesis on classic double dissociation logic, the finding that random damage to the same pathway (in different networks) could lead to different deficits supports the conclusion that double dissociations could result without modularity (although different layers in Plaut's network might be considered modules, damaging connections from a single layer – orthographic *or* semantic – yielded double dissociations).

However, as described earlier, the model elaborated by Plaut (1995) was complex, and interesting divisions of labor emerged: damage to connections from the orthographic layer yielded mainly abstract deficits, and damage to semantic-to-cleanup connections yielded mainly concrete deficits, even though damaging either location could result in either selective abstract or selective concrete deficits. Would such a pattern still emerge in a simpler architecture that would have less potential to develop division of labor (by having fewer layers of nodes and connections)?

We investigated whether double dissociations would emerge in a simpler network (O'Connor et al., 2009), using the same words and semantic features (slightly expanded) as Plaut. O'Connor et al. (2009) aimed at simulating the structure, computation and temporal dynamics of basic-level and superordinate concepts using a non-hierarchical model that would treat these concepts identically. The absence of a hidden layer alleviated the risk that their model could encode some level of hierarchy in hidden nodes. In our case, avoiding the use of a hidden layer would ensure minimal potential for architectural modularity.

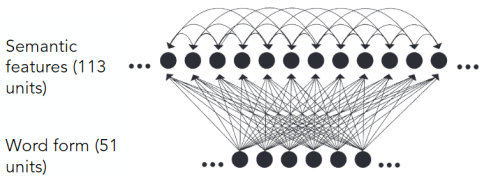

Figure 1: Attractor network adapted from O'Connor et al. (2009). Word form patterns are orthographic.

Our network was composed of a orthographic layer and a semantic layer and 2 sets of connections (Fig. 1). Connections between the orthographic and the semantic layer would be our first damage location, and recurrent semantic connections would be the second. Our main hypothe-

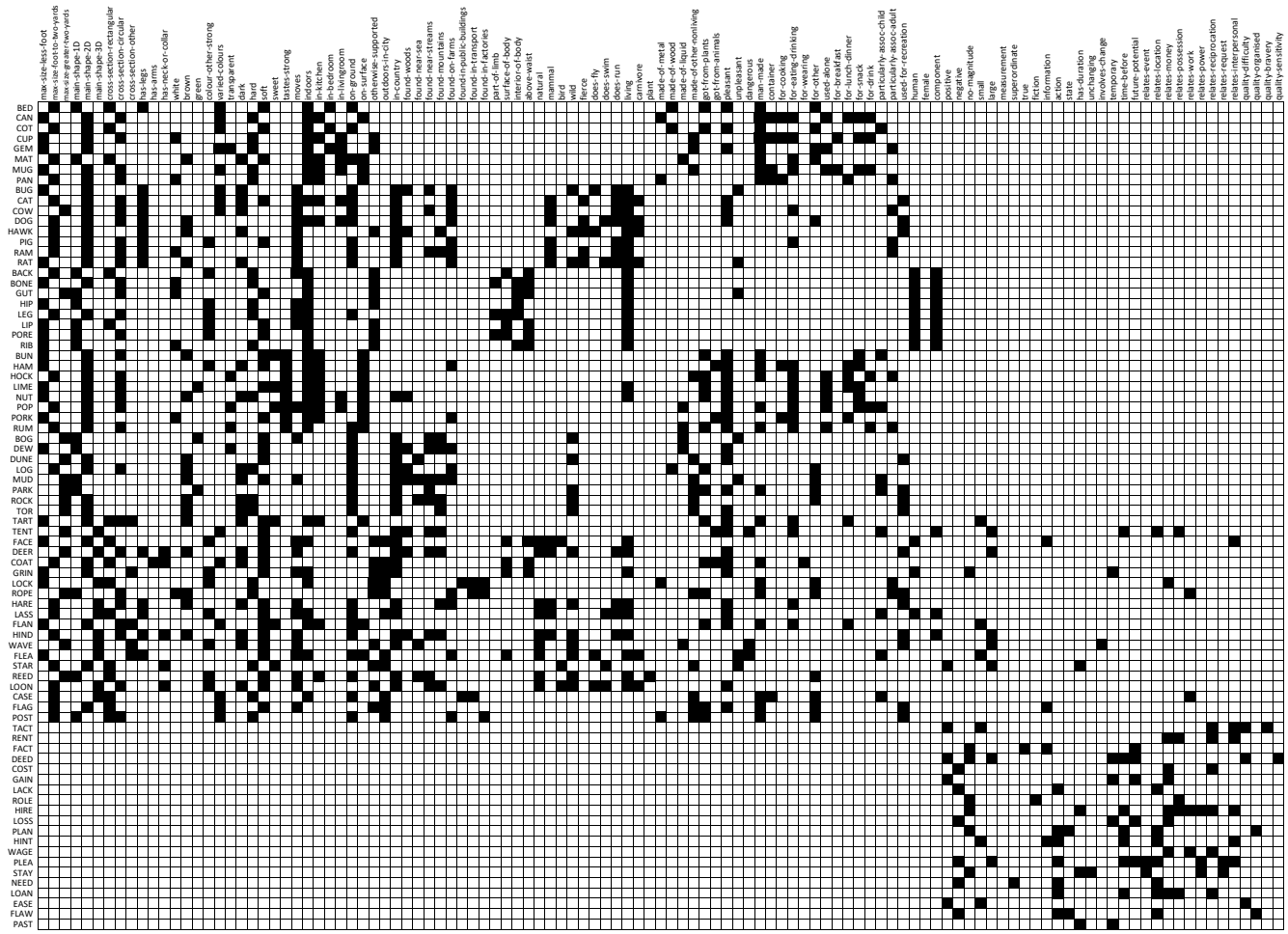

Figure 2: The 80 words used and their corresponding pattern of 113 semantic features (based on Plaut & Shallice, 1993, with some additions).

sis was that if Plaut's (1995) findings reveal a general principle about connectionist systems mapping form to meaning, we should repclicate his double dissociations following each kind of damage. Additionally, following Plaut's results (1995) we expected that on average, a larger deficit of abstract words rather than concrete words would follow damage to the orthographic-semantic connections, and a larger deficit of concrete words compared to abstract words would follow damage to the recurrent semantic connections.

Method

Architecture of the model

Our network (Fig. 1) was based on O'Connor et al.'s attractor network (2009) and was composed of 2 sets of units and 2 layers of connections. Our network takes orthographic features representing multiple phonemes simultaneously as input, and maps them directly to a semantic layer that has recurrent connections. Every semantic node has a connection to every other semantic node, and after input is applied, activation cycles 10 times.

The orthographic layer is composed of 51 nodes for the orthographic features (letters) that compose the 80 words used. There are 4 subsets of orthographic features representing the possible letters in each position for the words in the lexicon (15, 8, 17, and 11 in position 1-4, respectively).

The semantic layer contained semantic patterns which were constructed from 113 semantic features developed mainly by Plaut and Shallice (1993). 60 words were classified as concrete and 20 as abstract words, following Plaut and Shallice (1993). The words and their semantic patterns are presented in Fig. 2. Note that there was an imbalance in the distribution of semantic features in words, as concrete words tend to have more semantic features than abstract words, and some features are more likely in concrete words, while others are more likely in abstract words. These imbalances are intentional, and are motivated by theoretical assumptions about the representations of abstract vs. concrete words (with the latter assumed to have more features, and more correlations in shared features with other concrete concepts), similar to the logic for living vs. non-living things discussed above

(cf. Tyler et al., 2000). The simulations were conducted with PDPTool (McClelland, 2015; McClelland & Rumelhart, 1988) in Matlab (version 2013b).

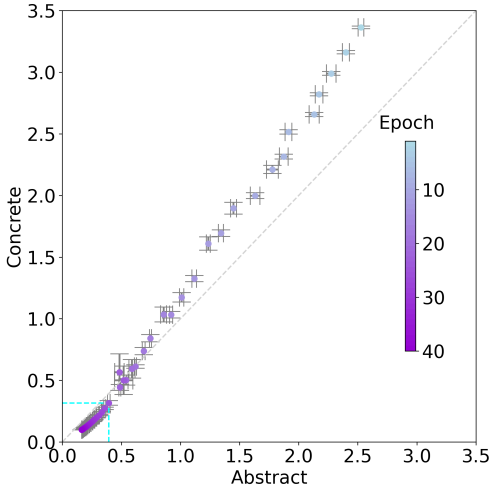

Figure 3: Training progress indexed by mean Euclidean distance of semantic outputs from Abstract vs. Concrete targets. Mean for the 10 models is plotted for each epoch (epoch 1 towards the upper right, epoch 40 near the origin). Horizontal and vertical error bars indicate standard error for the 2 sets of concepts. Cyan lines near the origin intersect at epoch 25, which we chose as the 'stable' point for lesion simulations. Epoch 25 is both where the initial advantage for Concrete items disappears and where standard error plateaus for both dimensions.

Training networks

We created 10 versions of the network with different initial weight randomizations. We then trained the resulting 10 networks using backpropagation for 40 epochs (where 1 epoch is 1 pass through all items in the lexicon, randomly ordered in each epoch). The learning rate was set to 0.001 and momentum to 0.9. Training progress is plotted in terms of mean Euclidean distance from abstract words to concrete words in Fig. 3. Euclidean distance was calculated by comparing the defined output pattern for a word with the network's actual output. As can be seen in Fig. 3, concrete words were learned more quickly than abstract words, but by approximately epoch 25, performance was similarly high for both categories.

Damaging networks

By epoch 25, performance became equally good for abstract and concrete words, and highly stable across networks. We thus selected epoch 25 as an appropriate point to compare networks with increasing levels of damage. First, we created 10 copies of each of the 10 trained networks at epoch 25. We then damaged each of the resulting 100 networks to different levels. This approach offers the possibility of examining how important the starting state is (i.e., the 10 individuals represented by 10 different sets of saved weights from epoch 25) vs. amount of damage. It is convenient to think of these 100 networks as 100 simulated patients.

The damage was done by removing or "masking" 10-80% of randomly-selected connections in each layer, in incremental steps of 10% (i.e, 10%, 20%, 30%...). We expected that the distance between the target word vector and the activated word vector would increase with increasing damage. Again, there were 2 layers of connections that could be damaged. We conducted lesion simulations with each of these separately: orthographic-semantic connections and recurrent semantic connections.

Analysis

Word recognition was evaluated by computing the Euclidean distance between the target word's vector and the activated word's vector for both concrete and abstract words, in each of the 10 instances of each of the 10 trained models.

A double dissociation would be observed if following one type of damage and at any masking level, some networks would show worse word recognition (i.e, higher Euclidean distance) for one type of words, while other networks would show the opposite pattern of results.

Results

Damage to the orthographic-semantic connections

Fig. 4 shows the results of damaging the connections linking the orthographic layer to the semantic layer, in increments of 10% from 10% to 80%. As Fig. 4 clearly shows, the distance between the target word vector and the activated word vector increased with increased level of damage. Note that there is 1 point for each of the 100 networks. However, we only indicate which original network each simulated 'patient' comes from. So for example, there are 10 bright yellow points in each panel – one for each of the 10 versions of Model 10.

Notably, the networks appear to be randomly distributed. While all networks perform slightly differently, none deviates greatly from the others. We also see that the starting point (which of the original 10 models a simulated patient is based on) is not crucial; the outcomes depicted by the points for patients based on Model 10 appear to be randomly intermixed with the simulated patients based on other models. Of course, how much the starting point matters will depend on factors like how much training is done before damage. As we noted above, we selected epoch 25 because it was a point where we observed similar (good) performance for abstract and concrete words, and low variability between models.

Fig. 4 shows that on average, the Euclidean distance of abstract words is usually *greater* than for concrete words (i.e., more points are below the identity line, indicating larger distance [poorer performance] for abstract words). However, while more models show a disproportionate impairment in abstract words, there are instances of the opposite pattern at each level of damage, where some simulated patients show a

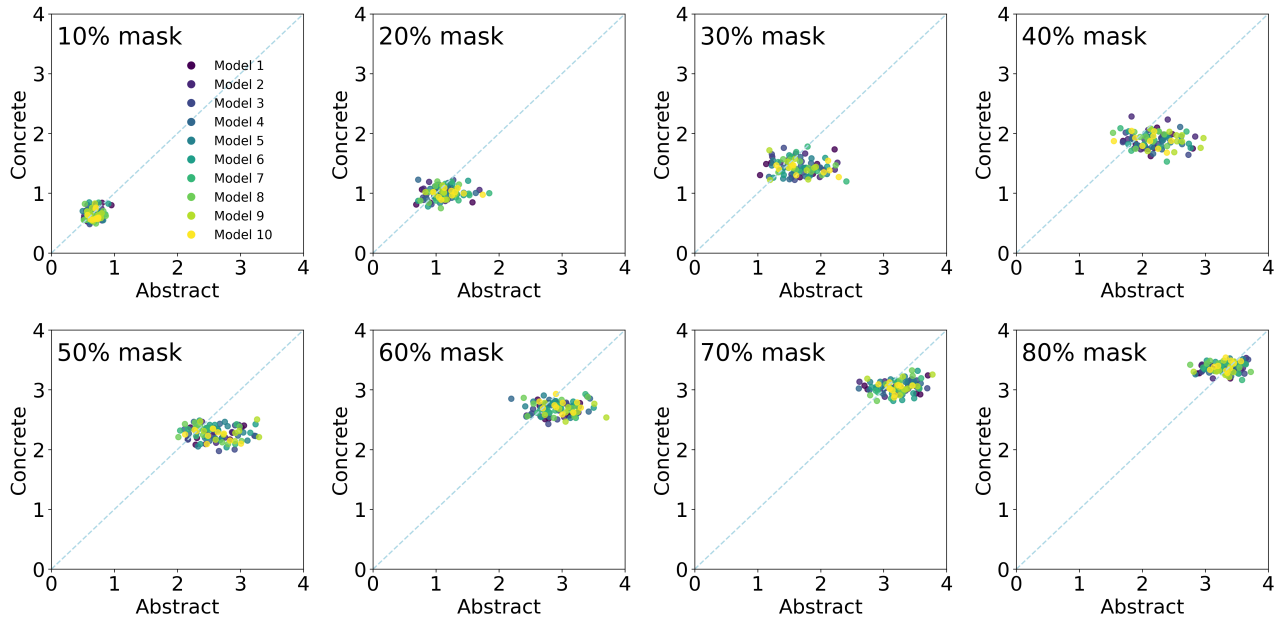

Figure 4: Euclidean distance of concrete and abstract words across masking levels following damage to the orthographic-semantic connections.

disproportionate impairment in concrete words (points above the identity line). As we discussed earlier, to establish a double dissociation, we only need 1 instance of each selective impairment, and disproportionate impairments like the ones we observe here in our simulated patients have routinely been interpreted as evidence for selective impairments. Therefore, as random damage to the orthographic-to-semantic connections can lead to complementary selective impairments, we replicate the core result of Plaut (1995): we observe double dissociations following random damage to a single 'module' (layer) of our network (i.e., ***double dissociations without modularity***).

Damage to the recurrent semantic connections

Now we consider the outcomes when we damage the recurrent connections of the semantic layer. Fig. 5 is organized exactly like Fig. 4, showing the performance of the 100 simulated patients on abstract and concrete words given masking of randomly-selected weights, in increments of 10%, from 10% to 80%.

Fig. 5 shows that the distance between the target word vector and the activated word vector increased with increasing level of damage. Again, the 100 patients appear to be randomly distributed. Interestingly, there is less variability in the models' performance when they are damaged in the recurrent semantic connections rather than in the orthographic-to-semantic connections.

Most importantly, however, Fig. 5 also shows that only one kind of selective impairment emerged from damaging recurrent semantic connections: aside from a few borderline cases at low levels of damage, models always exhibited greater impairment in *concrete* words than abstract words. This suggests that the recurrent connections become specialized to 'clean up' or reinforce semantic distinctiveness for concrete concepts, revealing an interesting division-of-labor in the flat attractor network.

Discussion

Our goal was to investigate whether Plaut (1995)'s finding of double dissociations without modularity would replicate in a flat attractor network (O'Connor et al., 2009) with minimal modularity and therefore less potential to develop hierarchical representations or divisions of labor. While our network was composed of only 2 layers, we used a slightly expanded version Plaut's materials for training our model, and we retained the same rationale for damaging the network.

Damaging the orthographic-to-semantic connections provided a robust replication, with each level of damage providing instances of models with disproportionate impairments in either concrete or abstract words, at all levels of damage (*i.e., double dissocations without modularity*). However, the most surprising result was the absence of double dissociations following damage to the recurrent semantic connections, regardless of the level of damage. Such damage only led to selective impairments for *concrete* concepts. This pattern suggests that the semantic recurrent connections provide a distinctive function, such as cleaning up or reinforcing concrete semantic patterns. This result is partially consistent with Plaut's finding that damaging semantic-cleanup connections in his network led to primarily concrete deficits (though also occasional selective impairments for abstract words). Again, the lexicon as defined may impose this pressure on the network,

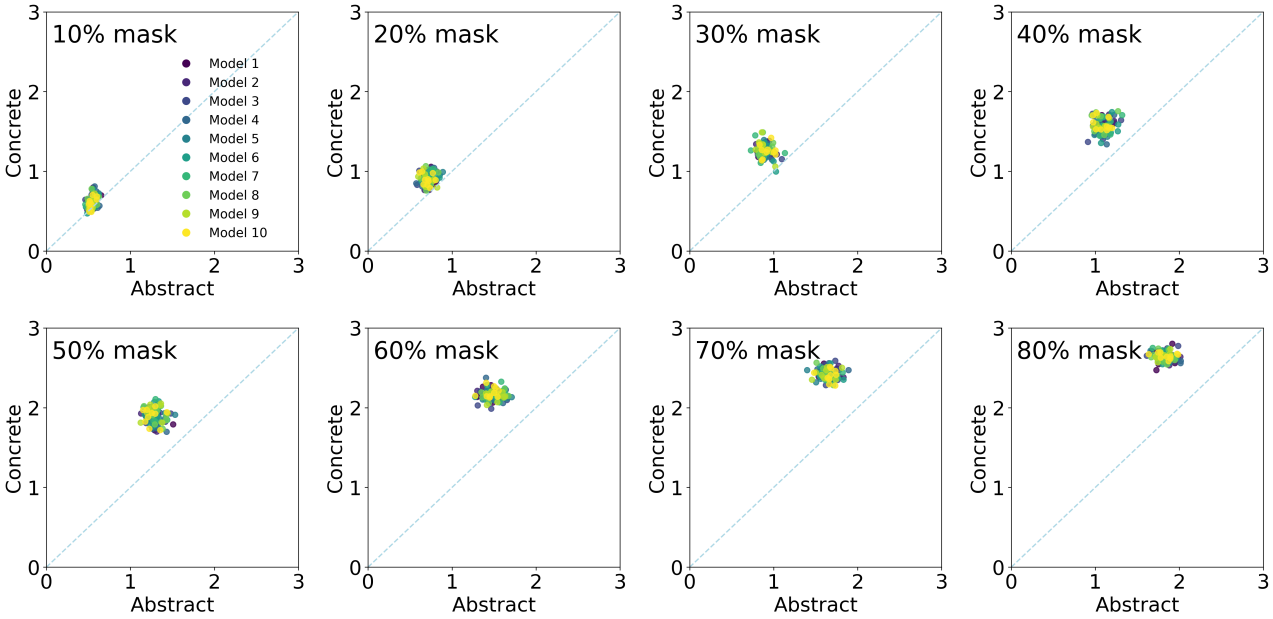

Figure 5: Euclidean distance of concrete and abstract words across all masking levels following damage to the recurrent semantic connections.

since concrete concepts on the one hand have more features, but on the other hand, potentially greater overlap with other (concrete) concepts' features. As we pointed out earlier, there is a theoretically-motivated imbalance in the number of concrete and abstract words as well as in the number of concrete and semantic features. There were more concrete words than abstract words (i.e, 60 compared to 20) and concrete semantic features than abstract semantic features (i.e, 82 concrete semantic features and 31 abstract semantic features).

Thus, concrete words have more features on average than abstract words, which could make them more robust to damage. In addition, since concrete concepts have more features, concrete semantic patterns may be less distinctive than abstract semantic feature patterns. This imbalance leads to concrete words having denser semantic features than abstract words, also allowing concrete words to be more likely to have overlapping features and clusters of features with other (concrete) words than abstract words. In contrast, the abstract concepts simultaneously use fewer features on average, and there is a smaller set of features unique to abstract concepts.

Our results raise challenging questions (or in some cases echo questions raised by Plaut's original findings). Why should damaging form-meaning connections predict more difficulties with abstract concepts, while recurrent semantic weight damage systematically yields uniquely concrete impairments? We tentatively speculate that concrete concepts are less distinctive in this lexicon than abstract concepts, leading the recurrent semantic connections to separate concrete feature patterns. In other words, highly overlapping / correlated semantic features for concrete words would be more sensitive to damage than less correlated semantic features found in sparse semantic patterns such as abstract words' patterns. This interpretation of our results is opposite to that of the Conceptual Structure account's assumptions mentioned earlier, which postulates that semantic features that are highly correlated with other features will be more resistant to damage (Caramazza & Mahon, 2003). While this theory was developed with focus on category-specific semantic deficits relating to living vs. non-living things, the logic regarding correlations among features also applies to the concrete vs. abstract distinction.

Conclusion

Our aim was to test whether we could observe double dissociations without modularity in a simpler network (with fewer potentially separable functions) than the network used by Plaut (1995). Our simple network, with phonology-to-semantic and recurrent semantic networks, exhibited robust double dissociations with minimal modularity in the form to semantic mapping.

The presence of double dissociations given different degrees of damage in each model reconfirm Plaut's (1995) findings in a much more "flat" architecture, with less potential for modularity. The tendency for concrete impairments given damage to the semantic attractor level is at once surprising and revealing; it demonstrates the division of labor (and partial modularity) that emerges in this network. Ongoing and future directions include assessing to what degree these results are robust for different semantic and orthographic representations.

Acknowledgements

This project was supported in part by National Science Foundation grant PAC 2043903 (PI JSM), by the Basque Government through the BERC 2022-2025 program, and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project PID2020-119131GB-I00 (PI JSM).

References

Caramazza, A., & Mahon, B. Z. (2003, August). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *TRENDS in Cognitive Sciences*, *7*(8).

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of cognitive neuroscience*, *10*(1), 1–34.

Davies, M. (2010). Double dissociation: Understanding its role in cognitive neuropsychology. *Mind Language*, *25*(5), 500–540.

McClelland, J. L. (2015). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises (second edition).* Stanford, CA.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Boston, MA: The MIT Press.

O'Connor, C. M., Cree, G. S., & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. *Cogn. Sci.*, *33*(4), 665–708.

Plaut, D. C. (1995, April). Double dissociation without modularity: evidence from connectionist neuropsychology. *J. Clin. Exp. Neuropsychol.*, *17*(2), 291–321.

Plaut, D. C., & Shallice, T. (1993, November). Deep dyslexia: A case study of connectionist neuropsychology. *Cogn. Neuropsychol.*, *10*(5), 377–500.

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*.

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