Are imaging and lesioning convergent methods for assessing functional specialisation?

Investigations using an artificial neural network

Michael S. C. Thomas[1], Harry R. M. Purser[1], Simon Tomlinson[1] & Denis Mareschal[2]

[1]Developmental Neurocognition Lab, Birkbeck, University of London, London, UK

[2]Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck University of London, London, UK

Running head: SBI and lesioning

Word count: 8,969

Address for correspondence:

Dr. Harry Purser
Department of Psychology and Human Development,
Institute of Education,
University of London,
20 Bedford Way, London
WC1H 0AL
Email: h.purser@ioe.ac.uk
Web: http://www.ioe.ac.uk/staff/PHDT/41431.html
Tel.: +44 (0)20 7612 6457
Fax: +44 (0)20 7612 6304

Abstract

This article presents an investigation of the relationship between lesioning and neuroimaging methods of assessing functional specialisation, using synthetic brain imaging (SBI) and lesioning of a connectionist network of past-tense formation. The model comprised two processing 'routes': one a direct route between layers of input and output units, while the other, indirect, route featured an intermediate layer of processing units. Emergent specialisation within the network was assessed (1) by lesioning either the direct or indirect route and measuring past-tense performance for regular and irregular verbs, and (2) by measuring functional activation in each route when processing each verb type (SBI). SBI and lesioning approaches failed to converge when network activation was summed over each route in our SBI approach. Examination of individual network solutions suggested that the verb types might be using the indirect route differently in terms of the *pattern* of activation across the route, rather than in terms of gross activation. A subsequent SBI analysis compared patterns of activation in the indirect route and confirmed that these patterns were more similar between regular-type verbs than between regular and irregular verbs. As the spatial and temporal resolution of neuroimaging techniques improves, the results of this investigation suggest that the key to finding functional specialisation will be to distinguish local coding differences across behaviours that are the results of developmental processes. Other analyses suggest that lesioning data may be limited because, with increasing damage, they reveal the resting activations of a computational system rather than a computational specialisation *per se*.

Keywords: neuropsychology, neuroimaging, synthetic brain imaging, computational modelling, connectionism, neural network, language, past tense, functional specialisation

Are imaging and lesioning studies convergent methods for assessing functional specialisation? Investigations using an artifical neural network

1. Introduction

It is an ongoing concern of neuropsychology to link particular regions of the brain to behavioural variation and, thereby, to particular cognitive functions. For example, the pioneering efforts of 19th Century neurologists, such as Broca, Wernicke, and Lichteim, established a view that language consists of many subcomponents, e.g., naming, repetition, and speaking, each of which was associated with a specific brain area (see Geschwind, 1970). This early work was based largely on the study of the aphasias, using a dissociation-based approach to brain lesions to infer the existence of different functional components of cognition.

A *single* dissociation occurs when a lesion (or an experimental variable) affects performance differently in two different tasks. For example, a lesion to Broca's area is traditionally associated with a pronounced deficit in productive, but not receptive, language (see Grodzinsky & Santi, 2008, for a recent review of the function and definition of Broca's area). A *difference* on some performance measure across tasks is termed by some a *weak* or *impure* dissociation, whereas a *strong* or *pure* dissociation is where the lesion affects performance in one task but not another (see Dunn & Kirsner, 2003, for a thorough treatment of dissociation logic and terminology). A (strong) *double* dissociation (Teuber, 1955) is where a lesion to one brain region affects performance in task A but not task B, whereas a lesion to some other region affects performance in task B but not task A. For example, Wernicke's aphasic patients, who classically present a deficit in receptive but not productive language, might be viewed as forming a double dissociation alongside Broca's aphasic patients. This kind of dissociation-based approach complements notions such as modularity (e.g., Carruthers, 2006;

Fodor, 1983, 2000), functional specialisation (e.g., Caplan, 1981; Shallice, 1988) and localisation of function (see Farah, 1994, for a review).

In contrast to the above *dissociation*-based approach, functional neuroimaging techniques, such as functional magnetic resonance imaging (fMRI), support inferences about the *associations* of particular brain regions with a given cognitive process. fMRI involves (indirectly) measuring changes in blood oxygenation over time, across different brain regions. This 'BOLD' (Blood Oxygen Level Dependent) signal is correlated with synaptic activity (Logothetis, 2002). A common approach adopted for fMRI studies is to obtain a baseline from scanning participants' brains during a control task, and then scan them again during a task thought to involve additionally the cognitive process under investigation. The BOLD signal from the control task is then subtracted from that measured in the main task; brain regions that show a reliable increase in oxygen uptake from the control task to the main task are argued to be associated with the critical cognitive process tapped in that main task.

In broad terms (and setting aside more recent and sophisticated approaches to lesioning and imaging studies, for the sake of clarity), lesion data suggest particular brain regions that are *necessary* for a given cognitive process, whereas neuroimaging data give a broader picture of which regions *contribute* to a cognitive process. A number of caveats to this picture have been proposed, however. For example, if a lesion produces a deficit in some cognitive process, it does not necessarily imply that the lesioned region is computationally involved; instead, it may be on a path of synaptic transmission from one region to another (e.g., the classical notion of conduction aphasia; see Anderson et al., 1999, for a recent discussion), or the lesioned region may have a diaschitic relationship (where one region supports the metabolic balance of another; e.g., Feeney & Baron, 1986) to regions that *are* computationally-involved. If there are redundant systems for a function, a lesion to a

functionally specialised system may produce no behavioural deficit (Price & Friston, 2002). Neuroimaging is an observational and correlative method: imaging cannot determine whether activated regions are necessary for a given function or co-incidentally activated; it cannot reveal areas that may be sufficient to generate a behaviour but that are not activated in a given situation (e.g., again, under the condition of redundant systems). It is worth stressing that devising more sensitive behavioural tests is not a solution to the problems above.

One way of attempting to overcome the limitations of lesioning and neuroimaging data is to look for convergence between the methods: in order to provide support for the mapping of a given cognitive process onto a particular brain region, one would expect to see both supportive double-dissociation data, showing degradation of function in lesioned patients, and an increased BOLD signal in the relevant area of undamaged brains when a task is performed that relies on the cognitive process in question (Chatterjee, 2005). Lesioning would identify necessary components, imaging would identify sufficient components. Encouragingly, the two methods have been broadly convergent in studies of reading (Price at al., 2003) and writing (Menon & Desmond, 2001), but the pattern of results across the two methods has been less clear in an investigation of Broca's area (Davis, Hillis, Bergey, & Ritzl, 2007). Nevertheless, Price and Friston (2002) have argued that the combination of methods might still be limited, because the existence of two or more redundant systems that do not overlap would allow for the possibility where there may be no single necessary system or brain area for a given behaviour.

Crucially, the combination of methods is predicated on the assumption that the findings will converge. What if they do not? There are various circumstances that may give rise to disagreements between imaging and lesioning data. When imaging suggests that a brain

region is central to a cognitive function but lesioning does not, it may be caused by the following:

1) The results reflect a particular strategy adopted by the lesioned patient. For example, there is more than one way to perform the task: imaging reflects the use of Region A, but a Region-A-lesioned patient instead uses Region B.

2) The activity within the brain region in question is due to inhibition rather than excitation. The target region is involved in *not* doing the task (perhaps inhibiting competing processes) rather than participating in doing it.

3) The activation represents the recruitment of a general, rather than task-specific, cognitive resource. In this case, under lesioning, some task performance is achievable without using the general resource.

Conversely, there may be lesioning data suggesting that a brain region is necessary for a cognitive function, but imaging data that do not show increased activation of the region in relevant test conditions. This situation may be caused by the following:

1) The physical characteristics of the region may prevent any increased activation being detected by current imaging methods (i.e., it is too small, or it is diffuse)

2) Performance degradation resulting from lesioning may be due to damage to neuronal pathways passing through a region rather than to damage to the neurons within the region itself

3) Processing is distributed, such that performance on a task can be achieved without requiring above-baseline activation (for example, a different group of neurons in the same region provides the function without causing total activation to exceed that of

the neurons contributing to baseline); nevertheless, the task relies on computational properties that allow it to be selectively impaired following damage to the region.

4) The brain region in question metabolically supports other regions that are computationally involved with task performance, but this metabolic support does not require marked augmentation of oxygen uptake.

In the current article, we use synthetic brain imaging and lesioning of a connectionist network (or artificial neural network, ANN) to explore the relationship between these methods of assessing functional specialisation. The aim is to use the resulting insights to elucidate the relationship between imaging and lesioning in the brain itself, in a simplified model system where all the implementational details are understood. The network to be studied is Thomas and Karmiloff-Smith's (2002) connectionist dual-route model of past-tense processing, in some ways analogous to Pinker's (1984) dual-route theory.

The network was originally developed to investigate the mechanism by which both rule-based (i.e., regular verbs, such as 'walk-walked') and exception handing (i.e., irregular verbs, such as 'go-went') aspects of past-tense acquisition could be learned. In this domain, both lesioning (e.g., Hodges & Patterson, 1995) and functional imaging (e.g., Ullman, Bergida, & O'Craven, 1997) data have been used to support the proposal that the processing of these two verb types relies differentially on separate brain regions. The model comprises two processing 'routes': one is a direct link between layers of input and output units (the 'direct route'), while the other features an intermediate layer of processing units (the 'indirect route'). Using the lesioning approach, Thomas and Karmiloff-Smith demonstrated that the dual-route network had acquired functional specialisations during training of the ANN, such that damage to the direct route was more detrimental to rule-based processing than to exception handling, while damage to the indirect route impacted more on exception handling

than on processing rule-based verbs (see also Westermann, 1998, 2000, for a related model). It is important to note that we do not claim that this model captures the entire richness or complexity of past-tense production in English. The model does not have phonological and semantic components and therefore cannot show dissociations that can be explained by differential reliance on these two systems, unlike some other past-tense models (Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003; Woollams, Joanisse, & Patterson, 2009). Past-tense formation is not the subject matter of this article. The model we use is appropriate for the purposes of this article because it is simple and well understood; past-tense formation is used here only because it is representative of quasi-regular mapping problems with observable dissociations. Thus, Thomas and Karmiloff-Smith's (2002) model may be considered a past-tense-like mapping task rather than a model of past-tense formation proper. Similar simple dual-route models have been used by other researchers as tools for studying dissociations in the domains of reading (Zorzi, Houghton, & Butterworth, 1998) and, indeed, past-tense inflection (Westermann & Ruh, 2009).

As described above, lesioning of ANNs can readily reveal functional specialisations (see also Weems & Reggia 2006, for a simulated lesion model of aphasias). However, an alternative method has been used to uncover such specialisation within connectionist models that relies on activation levels in the intact network. This is analogous to functional imaging methods and indeed has been labelled *synthetic brain imaging* (henceforth SBI; Arbib, Bischoff, Fagg, & Grafton, 1995). In an ANN, the 'neural' activity occurs across a layer of simple processing units and can be represented by a vector of values, with each vector element corresponding to a given processing unit's activation level. The connectivity between layers is represented by a matrix of excitatory and inhibitory connection strengths between individual units. SBI is a method of graphically representing processing inside the network, in which the strength of

the internal connectivity is combined with the size of the activation values exploiting this connectivity for any given processing pattern. SBI therefore distinguishes the ways in which different processes (e.g., producing regular vs. irregular verbs) exploit the same fixed network connectivity structure to drive behaviour (see also Sanger, 1989, for a related approach called *contribution analysis*, and Shultz, 2003, for an application of this method to understand the emergence of neural network representations across development).

Recently, Cangelosi and Parisi (2004) used SBI to investigate the functional specialisations that emerge within an ANN trained to process nouns and verbs. They found that processing these different types of words produced different foci of activation within the network: network elements that had different functional properties supported the emergence of different language behaviours. Additionally, Westermann and Ruh (2009) have recently demonstrated that employing the past-tense domain and dual-route model similar to the one used here, SBI could reveal differential use of the two routes by regular and exception verbs. However, neither set of researchers conducted parallel lesioning of the network elements in order to determine whether the dissociation method would reveal the same pattern of functional specialisation as that revealed by SBI.

Thus, although previous studies have explored functional specialisation of connectionist networks with either lesioning or SBI, none has directly compared the two methods to assess the degree of convergence. The aim of this article is to do exactly that, with a view to clarify *why* real-world lesioning and imaging might sometimes fail to tell the same story. There are several advantages of studying connectionist networks to pursue our research question: unlike real brains, we can control and measure every aspect of a connectionist network. It is easy to image *and then* lesion the very same networks. We are emphatically *not* claiming that the past-tense model studied in this article represents a fully adequate model of real neural

structures. Nor are we claiming any correspondence between either route and any particular neural structure: the model is at an abstract level and involves two routes, informed by recent work that identifies at least two neural routes for reading (e.g., Richardson, Seghier, Leff, Thomas, & Price, submitted). We acknowledge that the backpropagation algorhithm is generally viewed as biologically implausible, although it may also be viewed as an approximation of a Hebbian-based algorithm that uses bidirectional connections to diffuse error signals through a network (Xie & Seung, 2003). Furthermore, we are not concerned here about elucidating mechanisms of language; instead, language is used as an example domain that allows us to explore the relationship between lesioning and imaging data.

Later, we will be careful to distinguish the key properties of the model from the simplifications. Here, it suffices to highlight one key property of the model: the network's representational states are the outcome of a developmental process. Horwitz and colleagues (e.g., Horwitz, Tagamet, & McIntosh, 1999) conducted SBI of more biologically realistic neural network models (of the visual system), including explicit modelling of the BOLD function in order to map the modelled imaging onto observed fMRI data. That model emphasised biological realism at the expense of simplifying the origin of the representational states: the architecture was pre-wired. Since our goal was to compare the functional specialisation of neurocomputational structures as revealed by different measurement techniques, it was crucial that these specialised processes be the result of a plausible developmental process (specifically, the result of structure-function correspondences during learning) rather than a pre-wired solution encoded by the modeller.

2. Method

2.1. Network

The model comprised a feedforward, pattern-associator network. This included two routes connecting the input and output layers. A direct route linked 90 input units with 100 output units and an indirect route connected these same units via an intermediate layer of 20 hidden units. The model was trained to produce the correct past tense of English verbs when the relevant verb stem was presented at its input.

Computational theory indicates that the direct route should have sufficient computational resources for learning the input-output mappings for regular verbs, whereas the indirect route is more suited to learning the input-output correspondences for irregular verbs (Rumelhart & McClelland, 1986). Broadly, the two-layer network (or direct route) can learn a regular function, but the additional computational resources of the hidden layer in the indirect route are required to handle further mappings that are inconsistent with the regular function. Nevertheless, the two-layer network can successfully tolerate a low proportion of exception patterns in a training set, particularly if the network is given a disproportionately large amount of training on those patterns. Hence, the degree to which one might expect exceptions to be handled by the indirect route is a function of how inconsistent or 'exceptional' those input-output mappings happen to be. Emergent functional specialisation might be expected within the dual-route model of past-tense learning, because it employs an error-correcting backpropagation algorithm; there is therefore competition between the routes to contribute to minimising the disparity between output and target activations. Should one route eliminate this disparity, there will be no error left over to drive plastic weight changes (i.e., learning) in the other route. If one route learns a pattern, there is no reason for the other to do so. Further details of the mechanism by which the model achieves emergent specialisation of function can be found in Thomas and Karmiloff-Smith (2002).

The training set consisted of the stems of 508 monosyllabic English-like verbs paired with their past-tense forms, as specified in Plunkett and Marchman (1993) and used by Thomas and Karmiloff-Smith (2002). Each verb consisted of three phonemes represented over six binary articulatory features, corresponding to a vector representation with 18 elements. Given the intention to evaluation functional specialisation by lesioning the network, a larger representation was desirable because the results obtained from network damage can be artefactual if the network is trivially small (Bullinara & Chater, 1995). Therefore, each element from the original code was replicated five times to create a total vector of 90 elements to be presented to the 90 units of the network's input layer, with the code for each replication involving a 20% chance of inversion of binary features (implementing noisy redundancy). The output layer employed a similar code but with the addition of 10 elements representing two articulatory features to capture the affix "–ed" to form a regular past tense.

Four types of verbs were represented within the training set: (1) Regular verbs whose past tense took the +ed form (e.g., walk-walked; later referred to as *Regular*); (2) Exception verbs whose past tense was the same as the present tense form (e.g., cut-cut; referred to as Exception Pattern 1 or *EP1*); (3) Irregular verbs whose past tense was formed by changing the internal vowel of the stem tense (e.g., dig-dug; *EP2*); and (4) Irregular verbs whose past tense was entirely dissimilar to the stem tense (e.g., go-went). The verbs in category 4 were presented to the network three times as often as the others, because learning would not be possible without repeated exposure to these forms (Plunkett & Marchman, 1993). These verbs will be referred to as *EP3f*, to mark the effect of both the highest level of 'exceptionality' but also the effect of high token frequency. The training set consisted of 410 regular verbs, 20 no-change verbs, 68 irregular vowel-change verbs, and 10 irregular arbitrary past-tense verbs.

To establish the performance of the network in forming generalisations based on the training set, a set of novel verbs was presented to the network after training. The generalisation data set consisted of a group of 572 novel verbs of which the majority rhymed with a member of the four categories of the training set (i.e., two of their three phonemes were identical to their training set counterpart). The composition of the generalisation set was as follows: 410 novel verbs rhyming with regular verbs ('Regular-Rhyme'), 20 novel verbs rhyming with no-change irregulars, 76 novel verbs rhyming with vowel-change irregulars, 10 novel verbs rhyming with arbitrary irregulars, and 56 novel forms which did not rhyme with any verbs (i.e., shared no more than a single phoneme with any of the verbs in the training set). The members of the generalisation set that shared characteristics with elements of the training set tested the network's ability to extend the past tense 'rule' to novel stems or exhibit some other pattern of generalisation (such as irregularisation). Our results will primarily focus on the novel verbs that rhymed with regulars and which one would therefore expect to be handled by the same structure that processes regular verbs in the training set. These will be referred to as *Rule*. We focus on *Rule* verbs of the generalisation set because these give an indication of whether the network (or part of the network) has *abstracted* the past tense 'rule', rather than merely learned associations specific to the training set (it should be noted that all generalisation of connectionist models is similarity- rather than rule-based, and also that the network was able to generalise to novel exemplars from all verb classes of the training set; these are omitted here for the sake of brevity).

Each network began with connection weights initially randomised between ±0.5.The network was then trained for 500 presentations of the complete training set (epochs) and then tested on both the training set and the generalisation set. At each epoch, the training set was presented in a different random order. The learning rate and momentum were set to 0.1 and

0.0, respectively. Twelve replications were run with different initial random seeds. Following Thomas and Karmiloff-Smith (2002), a nearest neighbour method was implemented to assess the network's accuracy in its responses: the phoneme with the least Euclidean distance between itself and corresponding element of the output vector was taken as being the network's intended output. If all phonemes matched the target output, the verb was given a score of 1 otherwise it was given a score of 0. The behavioural metric was expressed in terms of the percentage of each verb type that the network outputted correctly.

2.2. Lesioning

Emergent specialisation within the network was assessed by lesioning either the direct or indirect route, i.e., removing a percentage of the weights between the units comprising the route (by setting the weights to zero), then determining the percentage of correct response returned by the network for each category of verb within the training or generalisation set (where, for the purposes of this model, the 'correct' novel response was taken to be application of the past-tense rule). Weights were lesioned rather than hidden units, because this approach allows for more distributed damage, thereby avoiding artefactual small-scale model behaviour (see Bullinaria & Chater, 1995). Performance was assessed at the intact level and then by removing 50% of the weights from one of the two routes (both before and after the hidden layer in the indirect route). Several lesioning levels were piloted. The 50% level was chosen in order to avoid ceiling and floor effects: at this level of damage, all verb types fell below 100% accuracy of the intact network but none was at 0%. There was no crossover of the sensitivity functions for the routes, with varying lesion level. For each of the 12 replications of the model, 10 different lesions were applied to each route. By averaging across 10 lesions for each replication, any evidence of specialisation would owe to properties of each route as a whole, rather than to the particular weights that were removed in any single

lesion. In terms of neurological patients, the simulations would therefore correspond to 240 cases of brain damage.

For each verb category, the degree of specialisation towards either the direct or indirect route was defined as the difference in the degradation of performance that resulted from lesioning each route. For example, a comparison was made between the percentage of regular verbs that were correctly associated with their past tense form after a direct lesion level of 50% and the percentage accuracy on the same set of verbs when the indirect route experienced a 50% lesion. If performance declined more when the direct route was lesioned, this was taken as evidence that the direct route had greater specialisation for this verb class. Formally, we defined the specialisation level for each verb type as the difference between the two levels of degradation following a single route lesion. Numerically, this was equivalent to subtracting the level of performance after direct route damage from the level of performance after indirect route damage. A positive value indicated a specialisation towards the direct route and a negative value indicated a specialisation towards the indirect route.

2. 3. Synthetic Brain Imaging

To assess emergent specialisation by Synthetic Brain Imaging, consideration was given to the level of activation of each unit within the network and the connection weights that had developed between the units by the end of training. At this point, we distinguish between the *unit activations* (computed by summing the net input to each unit and passing the value through the unit's non-linear threshold function) and *functional activation states* (henceforth FAS). The FAS correspond to how hard each connection is being driven and are calculated by the products of unit activations and connection strengths. FAS vectors can be computed for processing routes either with or without an intermediate layer of processing units.

In terms of the architecture of the current dual-route network, the FAS can be depicted by six vectors: two vectors for the direct route and four vectors for the indirect route. The vectors were of two types: (1) Sending vectors, representing the total activation being driven from a layer of units and (2) Receiving vectors, representing the total activation being received at a layer of units. The direct route consisted of one sending vector of activations being driven from the input layer to the output and one receiving vector of activations being received at the output layer from input. The indirect route consisted of two sending vectors and two receiving vectors: one sending vector and one receiving vector each for input-to-hidden and hidden-to-output. Each vector had one element for the FAS value of each unit in the relevant network layer.

The value of each element in a *sending vector* was calculated by multiplying the activation level of a sending unit with absolute size of each weight emanating from this unit and summing the product. The vector depicts the functional activation being driven along a pathway by a given layer. This calculation is expressed in the following equation:

$$FAS_j = \sum_{1}^{i} a_j \times (| w_i |) \qquad \text{(Equation 1)}$$

where *FAS* is the functional activation state of sending unit *j*, a_j is the activation of unit *j*, and w_i are the *i* weights emanating from unit *j*.

Similarly, the value of each element in a *receiving vector* was calculated by multiplying the absolute size of each of the weights arriving at the receiving unit by the unit's activation level, and then summing these products. The vector depicts the functional activation by which a given layer is being driven by a pathway. The calculation is expressed in the following equation:

$$FAS_k = \sum_{1}^{i} a_k \times (| w_i |) \qquad \text{(Equation 2)}$$

where *FAS* is the functional activation state of receiving unit *k*, a_k is the activation of unit *k*, and w_i are the *i* weights arriving at unit *k*.

It is important to note that the *absolute* values of the connection weights were used in these calculations, so that both high levels of excitation or inhibition along a connection would yield the same value for a vector element. The rationale for the use of absolute values is that within functional brain imaging, it is not possible to directly observe neural output, but only to observe indirectly the level of resource used (i.e., the BOLD signal). Given that resource use could represent the activity of either excitatory or inhibitory neurons during the performance of a task, it is therefore appropriate to simulate this with absolute connectivity values within the model. Because in the model we know which connections are excitatory and inhibitory, we later compare the picture given by separating the influence of these two types of functional activation state.

First, we briefly compare our method for computing SBI values with prior studies. Some previous SBI algorithms have summed activity over a particular time window, corresponding to some scan duration (e.g., Arbib et al., 2000; Cangelosi & Parisi, 2004). For example, Arbib et al. (2000) summed receiving activation (firing rate multiplied by connection strength) for a region of interest, combining the activation of sending processing units with the strength of their efferent connections. Cangelosi and Parisi (2004) used a similar measure to that used by Arbib and colleagues, but also recorded activity *within* a layer. Horwitz et al. (1999) also integrated the absolute value of activity within connections over the time course of the study and within the different areas of the model, but did not give details of the algorithm adopted. The current study's algorithm is similar to Arbib et al.'s, insofar as unit activations and

connection strengths are combined, but our algorithm yields a snapshot of the network's state at the end of training, where the processing of a given input occurs in a single pass, rather than integrating across a time window. Additionally, we consider sending activation in addition to receiving activation.

Given the FAS elicited when the dual-route network processed a given pattern (or set of patterns), it was then necessary to derive a measure of specialisation to each route. For this, a subtractive method was employed. A baseline activation was calculated separately for direct and indirect route by averaging across all the FAS elements of that route, for the five verb types of principal interest in the current study: *Regular*, *Rule*, *EP1*, *EP2* and *EP3f*. This baseline was used rather than some simulation of a 'resting state' baseline in order to for the baseline to be clear to the reader and also to avoid artefactual results that might arise from an arbitrary method of simulating a resting baseline[1]. The difference between the activation value for a particular verb class and this baseline value showed the relative change in contribution made by one or other route when the pattern was presented. Comparing the relative changes in contribution along both routes allowed the degree of specialisation to be calculated. The inference made was that a higher relative level of activity along one of the routes indicated the emergence of specialisation for processing that set of patterns by that route. The specialisation between the routes, then, was calculated by subtracting the relative change in activation for the indirect route from the relative change in direct route, for each class of verbs. A positive value of this figure indicated a specialisation towards the direct route and a negative value indicated a specialisation towards the indirect route.

[1] The simulations in this article were repeated with a baseline intended to be like a 'resting state' baseline. This baseline was created by inputting noise into the model. The pattern of results found was very similar to that reported; the arguments and analysis of the article were unchanged.

Finally, to aid interpretation, visualisations of the functional activation states within the network were produced. The 6 vectors coding the two routes were spatially extended using interpolation into 24 steps in both direct and indirect route, and a colour coding used to represent FAS levels. This spatial extension and smoothing of the vectors made use of local averaging of FAS values and served to allow readier visual interpretation of these FAS maps.[2] To reflect common practice in fMRI, diagrams show the baseline subtraction used to derive the regions differentially activated by a given verb class.

3. Results and Discussion

Our primary research question was to address whether lesioning and synthetic imaging methods produced a convergent picture of functional specialisation in the model; and if they diverged, to identify the neurocomputational basis for this divergence.

3.1. Main Analysis

3.1.1. Lesioning

Figure 1 shows the network's performance in producing correct past-tense forms for *Regular* and *EP3f* exception verbs, following 50% connectivity lesions to the direct or indirect route. These classes correspond to the most regular and irregular verb types. For *Regular* verbs, intact performance was 100%. Averaged over the 12 networks and 10 replications per lesion site (as noted above, by averaging across 10 lesions for each replication, any evidence of specialisation would relate to overall route properties, rather than to the particular weights removed in any particular lesion), the network's performance fell to 55.8% after a lesion to the direct route and to 76.5% after a lesion to the indirect route. Thus, the network's specialisation for *Regular* verbs was +20.7% towards the direct route. For *EP3f* verbs, once

[2] The algorithm used to generate the FAS diagrams is available as an Excel file at: http://www.psyc.bbk.ac.uk/research/DNL/techreport/SBI_spreadsheet.xls

again intact performance was 100%. The network's performance fell to 47.3% after a direct route lesion but to 29.0% after a lesion to the indirect route. The network's specialisation for *EP3f* verbs was -18.3%, that is, 18.3% to the indirect route.

3.1.2. Synthetic Brain Imaging

In order to determine the level of route specialisation for the different verb types in the intact networks, the percentage divergence from mean FAS was computed for each type. First, a baseline was calculated for each route, by averaging the FAS across the five verb types *Regular*, *Rule*, *EP1*, *EP2*, and *EP3f*. Second, the percentage difference from this mean was calculated for each type of verb, for each route. Percentages were used rather than absolute values to adjust for differences in average FAS levels in the two routes (FAS levels in the direct route tended to be higher because the magnitude of the weights was larger). Third, the percentage difference for the indirect route was subtracted from that for the direct route to give the index of specialisation. Finally, the values for each of the 12 networks were averaged (see Table 1). For example, for one network, the baseline activations for direct and indirect routes were 49.6 and 20.4, respectively. For *Regulars*, the percentage disparity from these baselines was -2.1% and -5.9%, respectively. The difference between these values was +3.8%, indicating specialisation to the direct route.

The 'Difference' column of Table 1 reveals that the FAS modulations by verb type were relatively small. Inasmuch as the verb types represent different behaviours, these different behaviours were achieved by relatively modest modulations of the functional activation states in the artificial neural network. The strongest specialisation was for *EP3f* exception verbs, in favour of the direct route. This contradicts the result found in the lesioning analysis, where *EP3f* verbs showed a strong specialisation towards the indirect route. Figure 2a shows a set of SBI images depicting the absolute FAS levels induced in each route by *Regular* and *EP3f*

verbs for one run of the model. Baseline FAS levels were calculated by averaging across all verb types, and the subtraction between these FAS maps for the two verb types to are shown to indicate their respective route specialisations. The figure confirms how both verb types use both routes; for *EP3f*, both routes are driven harder, but the direct route comparatively more so.

The Regular-Baseline subtraction clearly shows the three phonemes and inflection driving the output of the network in the direct route. The EP3f-Baseline subtraction shows an uneven use of the input, with particular reliance on the central phoneme in the direct route. It is notable that the structure of the SBI image for Regulars is not apparent in the FAS map for these verbs (Figure 2a, top left panel), but emerges only following the subtraction of baseline FAS (Figure 2a top right panel). This offers some support for the use of the subtraction method of imaging distributed representations, in addition to localist representations which would more obviously benefit from this approach.

3.2. Comparing the functional specialisation measured by lesioning and SBI

The percentage specialisation indicated by lesioning was plotted against the specialisation from imaging, for each of the 12 networks, for each verb type (Figure 3). Lesion specialisation is on the horizontal axis; imaging specialisation is on the vertical axis. For both axes, positive values indicate specialisation towards the direct route; negative values indicate indirect route specialisation. If the two techniques give convergent results, the scatterplot points should be within the upper-right and lower-left quadrants of each diagram; and a regression line through the points should have a positive gradient. The results indicate the following: First, the individual networks varied in their specialisation on both lesioning and imaging measures. Second, the pattern was modulated by verb type. Third, some indices of specialisation were consistent for a given verb type: lesioning showed preferential

specialisation of both Regulars and Rule to the direct route, and specialisation of EP2 and EP3f to indirect. However, when excitation and inhibition were conflated, no verb type exhibited consistent route specialisation of FAS, even where the overall group mean indicated some degree of specialisation. Finally, although three of the verb types showed some consistency in the direction of lesion and SBI measures (i.e., the regression lines had positive gradients), these effects were relatively small compared to the variance in the data.

Why should the synthetic imaging method produce such inconsistent results? Based on the FAS maps in Figure 2a, the SBI method appears to be a useful method for distinguishing the differential processing requirements of the two verb types we have focused on. Therefore, the disparity between SBI and lesioning methods does not arise from shortcomings in SBI itself, but perhaps from averaging across FAS vector scores, or perhaps from averaging across individual networks which have adopted different solutions. A further possibility is that the disparity arises because excitatory and inhibitory activations were deliberately conflated in the method. This is because we followed the constraints of real-world fMRI, where it is currently not possible to separate out these types of activity. However, in the model, it is relatively straightforward to separate excitatory and inhibitory contributions to specialisation.

3.3. Splitting Functional Activation States into excitation and inhibition

Table 2 shows the percentage divergence from baseline FAS levels for each verb type, for excitation and inhibition individually. It can be seen that analysing excitation and inhibition alone indicated a similar lack of evidence for route specialisation to that shown by the above analysis, which conflated excitation and inhibition (Table 1). However, as in Table 1, these data are averaged across all 12 networks – if the networks differ in their solutions to producing correct past-tense forms across verb types, it may be more informative to look at SBI images for an individual network.

Figures 2b and 2c show FAS maps of excitation and inhibition, respectively, for one run of the model. The most striking difference is that the excitation subtractions appear to be more structured than those for inhibition, with excitation showing a similar pattern for Regulars to that seen in the images that conflated excitation and inhibition (Figure 2a). The excitation subtraction for EP3f verbs shows that the direct route is driving the first three phonemes of the output rather than the inflection, which is consistent with the fact that this verb type requires no inflection production (e.g., *go-went*). Notably, the inhibitory FAS reveal that the two verb types are using the indirect route in a different way: the hidden receiving areas show red peaks of activation in different regions, while the output receiving areas show a different pattern of peak FAS for Regulars and EP3f verbs, somewhat like barcodes, that largely do not align when superimposed. The possibility that SBI was generating different information for the two routes prompted further investigation.

3.4. Further investigation of SBI data

Here, rather than using the preceding measure of route specialisation, we simply examined how hard each route was driven in processing each verb type. For each network, summed FAS was calculated for the verb types for the direct and indirect routes separately and this was correlated against specialisation as measured by lesioning. In keeping with the above analysis, this was first conducted for conflated FAS (see Figure 4), then for excitation and inhibition separately. The results were as follows. Conflated: a *negative* correlation for the direct route, R^2 = .29, no clear correlation for the indirect route, R^2 = .05. Excitation: again, a *negative* correlation for the direct route, R^2 = .37, but no clear correlation for the indirect route, R^2 = .03. Inhibition: no clear correlation for either direct, R^2 = .07, or indirect, R^2 = .06, routes. Figure 4 shows that the reason for this negative correlation is that the exception verbs – *EP1*, *EP2*, and *EP3f* – were driving the direct route harder than the *Regular* and *Rule* verbs.

However, on this measure of summed activation, the indirect route was not more strongly activated by either exception or non-exception verbs.

Figure 2b suggested that Regulars and EP3f verbs might have used rather different patterns of activity in the indirect route, because the SBI images showed what appear to be largely non-overlapping patterns of peak FAS for the two verb types. If so, then simply looking at strength of activation would fail to reveal route specialisation. As an alternative, then, we computed the angle (broadly equivalent to the correlation) between the FAS vectors for each verb type in the indirect route, where 90 degrees would represent completely orthogonal FAS patterns and 0 degrees would be perfect superposition. This was performed separately for Input-sending, Hidden-receiving, Hidden-sending, and Output-receiving vectors, for each combination of verb types (e.g., *Regular* vs *EP1*, *Rule* vs *EP2*, etc.), for each of the 12 networks. These angles were then averaged across the 12 networks. For the sake of brevity, we present only the angles computed from the Hidden-sending vector, which are most relevant because they reflect any re-representation of the inputs operating in the hidden layer. Table 3 shows descriptive statistics for angles between Hidden-sending vectors, for conflated FAS (very similar results were obtained for both excitation and inhibition). Paired-samples t-tests revealed that the angle between *Regular* and *Rule* verbs was reliably smaller than that between *Regulars* and *EP1* verbs, t (11) = 5.33, $p < .001$, $d = 2.19$, *Regulars* and *EP2* verbs, t (11) = 7.63, $p < .001$, $d = 2.88$, and also *Regular* and *EP3f* verbs, t (11) = 3.54, $p < .005$, $d = 1.25$.

Thus, there is clear evidence that the indirect route is re-representing *Regular* and *Rule* verbs more similarly than *Regulars* and *exception* verbs. Thus while the direct route separates verb types by the strength of activations, the indirect route allows the development of relatively non-overlapping representational codes at the hidden layer in order to appropriately activate

the output layer for each verb type. An appropriate analogy might be that the direct route is more 'analogue', where brute strength of connections – and therefore FAS – drives output units above their thresholds to switch on, whereas the indirect route is more 'digital', with degree of overlap of re-represented codes associated with different resulting output layer behaviour. Critically, measuring either relative (i.e., Figure 3) or absolute (Figure 4b) values of FAS could not detect such 'digital' specialisation of coding, because functionality did not rely on differences in net FAS summed over many units, but on the *pattern* of FAS across units. Lesioning, however, was blind to whether specialism was in terms of net FAS differences or re-representational codes.

In sum, our consideration of the disparity between synthetic imaging and lesioning methods for assessing functional specialisation suggests the following. Lesioning can reveal the location of the essential functional elements. By contrast, imaging, perhaps unsurprisingly, reveals where the most functional activity is. While lesioning indicated that the indirect processing route was most important for exceptions, synthetic imaging pointed to the opposite. This is because imaging was unable to pick up specialisation achieved via distributed representational encoding in the indirect route. It only sufficed to measure brute activity levels in the direct route.

3.5. The emergence of activation biases following lesioning

Thus far we have focused on possible conceptual difficulties with the imaging approach through conflating excitation and inhibition, and through collapsing across distributed representational codes. Are there, however, potential limitations with the lesioning approach? In particular, we wished to examine the transparency assumption (Caramazza, 1986; Shallice,

1988); that is, the idea that damaged structure is directly responsible for lost function while intact structure is directly responsible for remaining function. In our model system, lesioning connection weights from the input layer serves to progressively reduce the influence of input information (verb stems) in driving the output of the network. However, in the absence of input, processing units do not return to zero activation. In an artificial neural network, processing units have activation biases (sometimes referred to as the *threshold* or *bias* of a unit). The normal response of units is a combination of these activation biases with input information. As input information is reduced, the activation biases progressively contribute to driving behaviour. In the limit, with no input information arriving, the response of the system is fixed and driven only by the activation biases. Where a network produces different types of behaviour (here, the verb classes), the possibility exists that the activation biases may incline the network to one type of behaviour. In turn, this raises the possibility that a lesion does not straightforwardly remove the behaviour normally produced by the lost structure, but might also reveal underlying functional biases of the resting state of the system to produce certain behaviours.

To examine this possibility, we explored the boundary condition of 100% lesion to remove the influence of the input on network behaviour. To reveal the contribution of activation biases in the output layer, we lesioned all weights arriving at the output layer. To reveal the combined contribution of hidden unit and output unit activation biases, we lesioned connections from the input to output and input to hidden, but retained the connections from hidden to output. We then evaluated whether the (fixed) output activations were closer to the responses required for any of the verb types.

Figure 5 shows the RMS error in past-tense verb production for *Regular* and *Ep3f* verbs of a 100% lesioned network ('Output only'). It can be seen that there is a small but clear bias

towards correctly producing *Ep3f* verbs over *Regular* ones (RMS errors of .498 and .484, respectively, a difference of .14, t (11) = 5.61, $p < .001$, $d = 2.87$). Figure 5also shows the accuracy of responses driven by the combination of hidden unit and output unit activation biases ('Output + hidden'). The bias towards correctly producing Ep3f past-tense forms over Regular ones (.492 vs .451, a difference of .41, t (11) = 5.89, $p < .001$, $d = 2.71$) is now much larger (i.e., there was a reliable interaction between lesion extent and verb type, $F(2, 11) = 10.624$, $p < .01$, $\eta_p^2 = .491$). The activation biases are determined by the frequency of input-output mappings in the training set: units tend to reside in the most probable response state. It is therefore unsurprising that the *EP3f* verb class, marked out by its higher token frequency, influences the rest state activations more. However, the greater bias from the activation biases of the hidden units implies that as this contribution is reduced (by lesioning the connections from the hidden to output layer), *EP3f* verbs will be differentially impaired. Lesioning, therefore, does not simply remove the computational structures responsible for certain functions; it produces a computational state with an atypical balance between the influence of activation biases and the influence of input patterns. In this sense, the transparency assumption is violated and lesioning cannot directly index the functional specialisation of structures under normal conditions.

4. General Discussion

In this article, we have directly contrasted lesioning and imaging approaches of assessing functional specialisation, using a connectionist model of past-tense formation. Our analyses suggest that conventional neuroimaging data are unlikely to give a full picture, because they are generated by differences in net activity summed over a great many neurons, but are insensitive to the *pattern* of activity across those neurons. Lesioning data may also be limited

because, with increasing damage, they reveal the activation biases of a computational system rather than a computational specialisation *per se*.

The lesioning analysis indicated strong direct route specialisation for *Regular* verbs and also clear indirect route specialisation for *EP3f* exception verbs. This was expected prior to running the model, based on computational principles of structure-function correspondences. However, the imaging analysis was inconsistent, indicating a direct route specialisation for exception verbs. The SBI metric of specialisation was based on a difference between how hard each route was being driven (the product of processing unit activations multiplied by the absolute strength of the associated connection weights). We suggested that the disparity between lesioning and synthetic imaging could have arisen because our imaging analysis deliberately conflated excitatory and inhibitory activation. However, the anomalous result held even when the analysis was repeated for excitation and inhibition separately. The implication is that conflation of excitation and inhibition per se was not the principal limitation of the SBI method.

When we considered the two routes separately, it appeared that exception verbs were distinguished from regular verbs in the direct route: exception verbs drove the direct route harder. This appeared to resemble a form of analogue coding for regularity. However, the verb types were indistinguishable in the extent to which they activated the indirect route. Examination of individual network solutions suggested that the verb types might be using the indirect route differently. This led us to compute the angle between the FAS vectors for each verb type in the indirect route. This analysis revealed greater angles between *Regular* and exception verbs than between *Regular* and *Rule* verbs. The analysis confirmed that the patterns of activation through the indirect route were more similar for *Regular* and *Rule* verbs than for *Regular* and exception verbs. The presence of an intermediate layer of processing

units in the indirect route enabled the network to re-represent input patterns to begin to distinguish regular from exception verbs. To do so, it used a type of discrete, or 'digital', code.

One reason for the disparity between lesion and synthetic imaging methods then becomes clear. We initially relied on difference measures between the routes to assess specialisation: different level of impairment after damaging each route vs. different levels of FAS in each route during normal processing. For the latter, however, the levels of FAS only differed between the verb types in the 'analogue' direct route, making it seem like this route was specialised to exception verb processing. Only a consideration of FAS *patterns* yielded the role of the indirect route in exception processing that was readily revealed by lesioning. Thus, our analyses suggested that conventional neuroimaging approaches may be limited by their reliance on comparing gross activation of regions.

However, our results suggest that lesioning is also limited as a method of investigating functional specialisation, because damaging regions that are computationally involved in producing behaviours serves to reveal the influence of activation biases of the 'downstream' areas served by those regions. In the current study, we found a bias towards producing high-frequency exception verbs in a network with no input. Such a bias cannot indicate *functional*, *computational* or *processing* specialisation because an inputless network cannot perform any computation or processing. As sketched in the introduction, classic neuropsychology seeks to draw conclusions about normal and intact cognitive processes from the patterns of performance seen in people with brain damage. Such conclusions rely on the *transparency assumption*, which entails that the cognitive system in a damaged brain is the same as that in a normal brain except for the local modification caused by the damage (e.g., Caramazza, 1986; Shallice, 1988). The part of our network that remains after lesioning all input cannot be

described as ‘normal’, because of the output bias. This suggests a novel reason for questioning the transparency assumption (see Farah, 1994, for other reasons to question the assumption).

It is reasonable to question the suitability of the dual-route model for the purposes of our investigations. Is it too simple to be informative? Although simple, it is important that functional specialisation was a product of learning rather than a built-in assumption. The model’s solution of separate of ‘analogue’ and ‘digital’ coding into the direct and indirect routes stems from an experience-dependent processes acting over a network of simple integrate-and-fire units. While recognising that artificial neural networks are fairly abstract analogues of real neurocomputational systems, we believe the complexity of the representational states emerging from simple processing assumptions is a strength of the model. Lesioning suggested different areas were key for two different cognitive processes, while imaging suggested that just one of these areas was particularly important for both processes (the direct route). Our simulations demonstrated that this situation can arise because imaging is insensitive to patterns of activity within a region.

In the Introduction, we suggested that lesion data indicate particular brain regions that are *necessary* for a given cognitive process, whereas imaging data give a broader picture of which regions *contribute* to a cognitive process. Our exploration suggests that imaging can only reveal contributing regions that are more active overall during the particular cognitive process under scrutiny. Conventional neuroimaging is blind to regions that shift their patterns of activation depending on condition but not their overall activation (assuming that the method does not have the resolution to distinguish the local coding differences). This is certainly a candidate explanation for any lack of agreement between lesioning and imaging techniques. As the spatial and temporal resolution of neuroimaging techniques improves, the

current findings suggest the key to finding functional specialisation will be to distinguish local coding differences across behaviours that are the results of developmental processes. There has been a recent growth in pattern-information (multivoxel) analysis of fMRI data (see, e.g., Mur, Bandettini, & Kriegeskorte, 2007, for an introduction). For example, in a study involving both native English and Japanese speakers, Raizada et al. (2010) investigated the hypothesis that if fMRI patterns elicited by /ra/ and /la/ stimuli were separable from each other, then the listener would be perceptually able to distinguish /ra/ and /la/. They found that the separability of neural patterns did correlate with behavioural performance, not only across groups but also across individuals within each group. Critically, the average amount of fMRI activation did not differ for /ra/ and /la/; only the pattern of activation. As noted by Raizada and Kriegskorte (2010), pattern differences that are found to correlate with behavioural differences support the notion that pattern differences reflect brain processes that are *functionally* important.

Further new directions are found in some recent studies that involve multivoxel fMRI of lesioned brains (e.g., Teipel et al., 2007, Alheimer's; Saur et al., 2010, stroke). Such studies have identified not only pattern differences that correlate with behavioural differences, but also some that correlate with future clinical outcomes (e.g., changes in symptom severity). Saur et al. (2010) found that pattern differences predicted subsequent language recovery in stroke patients better than did behavioural language measures. The computational approach used in the current article could informative for this research area: modelling could elucidate how and why patterns of fMRI activation might change after lesions. Finally, the work reported here underscores the role that implemented computational models can have in elucidating theoretical and methodological issues in the empirical literature (see Mareschal et

al., 2007, for further discussion). Although simplifications of real brain systems, such models enable users to explore the effects of various interventions on a neural system.

Acknowledgements

This work was supported by European Commission grant NEST-029088(ANALOGY), ESRC grant RES-062-23-2721, and a Leverhulme Study Abroad Fellowship to MT.

References

Anderson, J. M., Gilmore, R., Roper, S., Crosson, B., Bauer, R. M., Nadeau, S., Beversdorf, D. Q., Cibula, J., Rogish, M. III, Kortencamp, S., Hughes, J. D., Gonzalez Rothi, L. J., & Heilman, K. M. (1999). Conduction aphasia and the arcuate fasciculus: a reexamination of the Wernicke-Geschwind model. *Brain and Language, 70,* 1–12.

Arbib, M. A., Bischoff, A., Fagg, A. H., Grafton, S. T. (1995). Synthetic PET: Analyzing Large-Scale Properties of Neural Networks. *Human Brain Mapping, 2*, 225-233.

Bullinaria, J. A., & Chater, N. (1995). Connectionist modelling: Implications for neuropsychology. *Language and Cognitive Processes, 10*, 227-264.

Cangelosi A., Parisi D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, *89*, 401-408.

Caplan, D. (1981). On the cerebral localization of linguistic functions: logical and empirical issues surrounding deficit analysis and functional localization. *Brain and Language, 14,* 120-137.

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*, 41–66.

Carruthers, P. (2006). *The architecture of the mind: massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

Chatterjee, A. (2005). A madness to the methods in cognitive neuroscience? *Journal of Cognitive Neuroscience, 17*, 847–849.

Plunkett, K. & Marchman, V. (1993) From rote learning to system building: acquiring verb morphology in children and connectionist nets. Cognition, 48, 21-69.

Davis, C., Hillis, A., Bergey, G. and Ritzl, E. (2007) Who needs Broca's area? Comparisons from lesion and fMRI methods. *Brain and Language, 103*, 14-15.

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

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the "locality" assumption. *Behavioral and Brain Sciences, 17*, 43-104.

Feeney, D. M. & Baron, J. (1986). Diaschisis. *Stroke, 17*, 817-830.

Fodor, J. A. (1983). *The modularity of the mind: an essay on faculty psychology.* Cambridge Massachusetts: MIT Press.

Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge Massachusetts: MIT Press.

Geschwind, N. (1970). The organization of language and the brain. *Science, 170*, 940–944.

Grodzinsky, Y & Santi, A. (2008). The battle for Broca's region. *Trends in Cognitive Sciences, 12*, 474-480.

Hodges, J. R. & Patterson, K. (1995). Is semantic memory consistently impaired early in the course of Alzheimer's disease? Neuroanatomical and diagnostic implications. *Neuropsychologia, 33*, 441-459.

Horwitz, B., Tagamet, M. A., & McIntosh, A. R. (1999). Neural modeling, functional brain imaging, and cognition. *Trends in Cognitive Science, 3*, 91–98.

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

Logothetis, N. K. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society B*, 357, 1003-137.

Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism, Vol. I: How the brain constructs cognition*. Oxford, UK: Oxford University Press.

Menon, V. & Desmond, J. E. (2001). Left superior parietal cortex involvement in writing: integrating fMRI and lesion evidence. *Cognitive Brain Research, 12*, 337- 340.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI - an introductory guide. *Social Cognitive and Affective Neuroscience, 4*, 101-9.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA, USA: Harvard University Press.

Plunkett, K. & Marchman, V. (1993) From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21-69.

Price, C. J., & Friston, K. J. (2002). Degeneracy and cognitive anatomy. *Trends in Cognitive Sciences*, *6*, 416-421.

Price, C. J., Gorno-Tempini, M. L., Graham, K. S., Biggio, N., Mechelli, A., Patterson, K., & Noppeney, U. (2003). Normal and pathological reading: converging data from lesion and imaging studies. *NeuroImage, 20*, S30-S41.

Raizada, R. D. S., & Kriegeskorte, N. (2010). Pattern information fMRI: new questions which open it up, and challenges which face it. *International Journal of Imaging Systems and Technology, 20,* 31-41.

Raizada, R. D. S., Tsao, F.M., Liu, H.M., & Kuhl, P.K. (2010). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: Prediction of individual differences. *Cerebral Cortex, 20*, 1–12.

Richardson, F., Seghier, M., Leff, A., Thomas, M. S. C., & Price, C. J. (submitted). How does reading access the amodal language system? Testing predictions from neurology and functional imaging using dynamic causal modeling. Unpublished manuscript.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models*. Cambridge, Mass.: MIT Press.

Sanger, D. (1989). Contribution analysis: A technique for assigning responsibilities to hidden units in connectionist networks. *Connection Science, 1*, 115-138.

Saur, D., Ronneberger, O., Kümmerer, D., Mader, I., Weiller, C., & Klöppel, S. (2010). Early functional magnetic resonance imaging activations predict language outcome after stroke. *Brain, 133*, 1252-1264.

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

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Teipel, S. J., Born, C., Ewers, M., Bokde, A. L., Reiser, M. F., Möller, H. J., & Hampel, H. (2007). Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment. *Neuroimage, 38*, 13-24.

Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology, 6*, 267-296.

Thomas, M. S. C. & Karmiloff-Smith, A. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioral and Brain Sciences, 25,* 727-788.

Thomas, M. S. C. & Karmiloff-Smith, A. (2003). Modelling language acquisition in atypical phenotypes. *Psychological Review*, 110(4,) 647-682.

Ullman, M., Bergida, R., & O'Craven, K. (1997). Distinct fMRI activation patterns for regular and irregular past tense. *NeuroImage, 5*, S549.

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

Weems, S. A., & Reggia, J. A. (2006). Simulating single word processing in the classic aphasia syndromes based on the Wernicke–Lichtheim–Geschwind theory. *Brain and Language, 98*, 291–309.

Westermann, G. (1998) Emergent Modularity and U-Shaped Learning in a Constructivist Neural Network Learning the English Past Tense. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pp. 1130-1135. Hillsdale, NJ: Erlbaum.

Westermann, G. (2000) A Constructivist Dual-Representation Model of Verb Inflection. *Proceedings of the 22th Annual Conference of the Cognitive Science Society*, pp. 977-982. Hillsdale, NJ: Erlbaum.

Westermann, G., & Ruh, N. (2009). Synthetic brain imaging of English past tense inflection. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1364–1369). Austin, TX: Cognitive Science Society.

Woollams, A. M., Joanisse, M., & Patterson, K. (2009). Past-tense generation from form versus meaning: Behavioural data and simulation evidence. *Journal of Memory and Language*, 61, 55-76.

Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation, 15*, 441–454.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). The development of spelling-sound relationships in a model of phonological reading. *Language and Cognitive Processes, 13*, 337-371.

Table 1. Percentage divergence from mean activation for each verb type, based on overall activation. Positive values in the 'difference' column indicate specialisation to the direct route; negative values indicate specialisation to the indirect route.

Verb type	Direct	Indirect	Difference
Regular	-2.5%	-2.8%	+0.3%
Rule	-1.2%	-1.5%	+0.3%
EP1	-1.0%	-0.1%	-1.0%
EP2	1.0%	2.0%	-1.0%
EP3f	3.7%	2.3%	+1.4%

Table 2. Route specialisation for each verb type, separately for excitation and inhibition. Positive values in the 'difference' column indicate specialisation to the direct route; negative values indicate specialisation to the indirect route.

	Excitation			Inhibition		
Verb type	Direct	Indirect	Difference	Direct	Indirect	Difference
Regular	-3.7%	-4.5%	+0.8%	-1.1%	-1.4%	+0.3%
Rule	-1.8%	-0.7%	-1.1%	-0.5%	-2.2%	+1.6%
EP1	-0.5%	2.1%	-2.6%	-1.7%	-1.8%	+0.1%
EP2	2.0%	0.8%	+1.2%	0.0%	+3.1%	-3.1%
EP3f	4.0%	2.3%	+1.7%	3.3%	+2.3%	+1.0%

Table 3. Angles between Hidden-Sending vectors, for each pairwise verb comparison. 0 degrees would indicate completely overlapping vectors, 90 degrees would indicate entirely orthogonal vectors.

Comparison	Mean Angle	*SD*
Regular vs Rule	19.0	2.8
Regular vs EP1	26.5	4.1
Regular vs EP2	33.2	6.6
Regular vs EP3f	23.1	4.0
Rule vs EP1	22.0	4.6
Rule vs EP2	37.0	8.0
Rule vs EP3f	24.6	5.0
EP1 vs EP2	30.1	6.2
EP1 vs EP3f	21.1	2.7
EP2 vs EP3f	26.4	6.8

Figure 1. Network performance in producing correct past-tense forms for *Regular* and exception (*EP3f*) verbs, following a 50% lesion of connectivity to either the direct or indirect route. Error bars depict standard errors of the means across 12 networks.

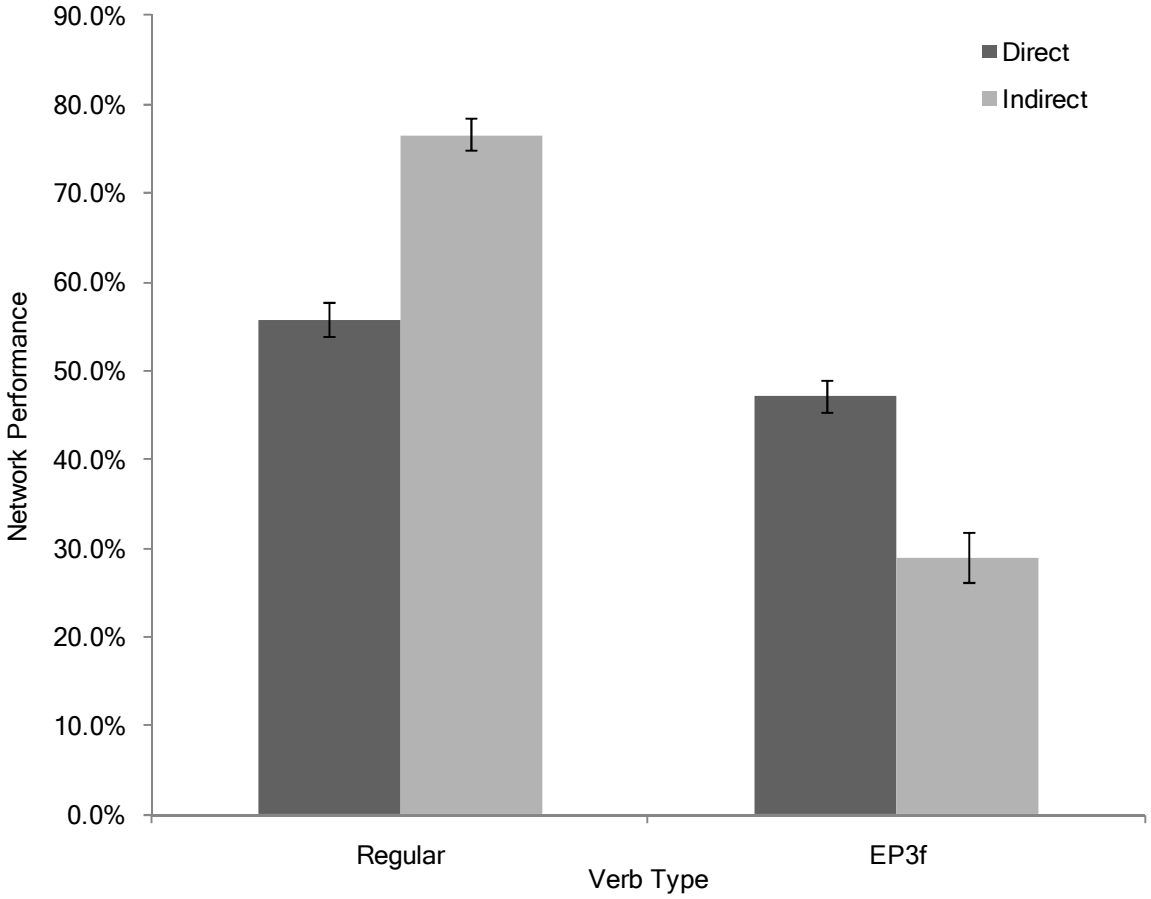


Figure 2. SBI images for regular and exception (*EP3f*) verbs. The direct route is the left half of each image, the indirect route is the right half. The bottom is the input layer, the top is the output layer. The FAS diagrams were created by interpolating values between the 6 vectors depicting functional activation states. Interpolated values were smoothed by local averaging where the vectors were of different dimensionality.

Figure 2a. FAS activation.

Figure 2b. FAS activation, showing only excitation.

Figure 2c. FAS activation, showing only inhibition.

Figure 2a. FAS activation.

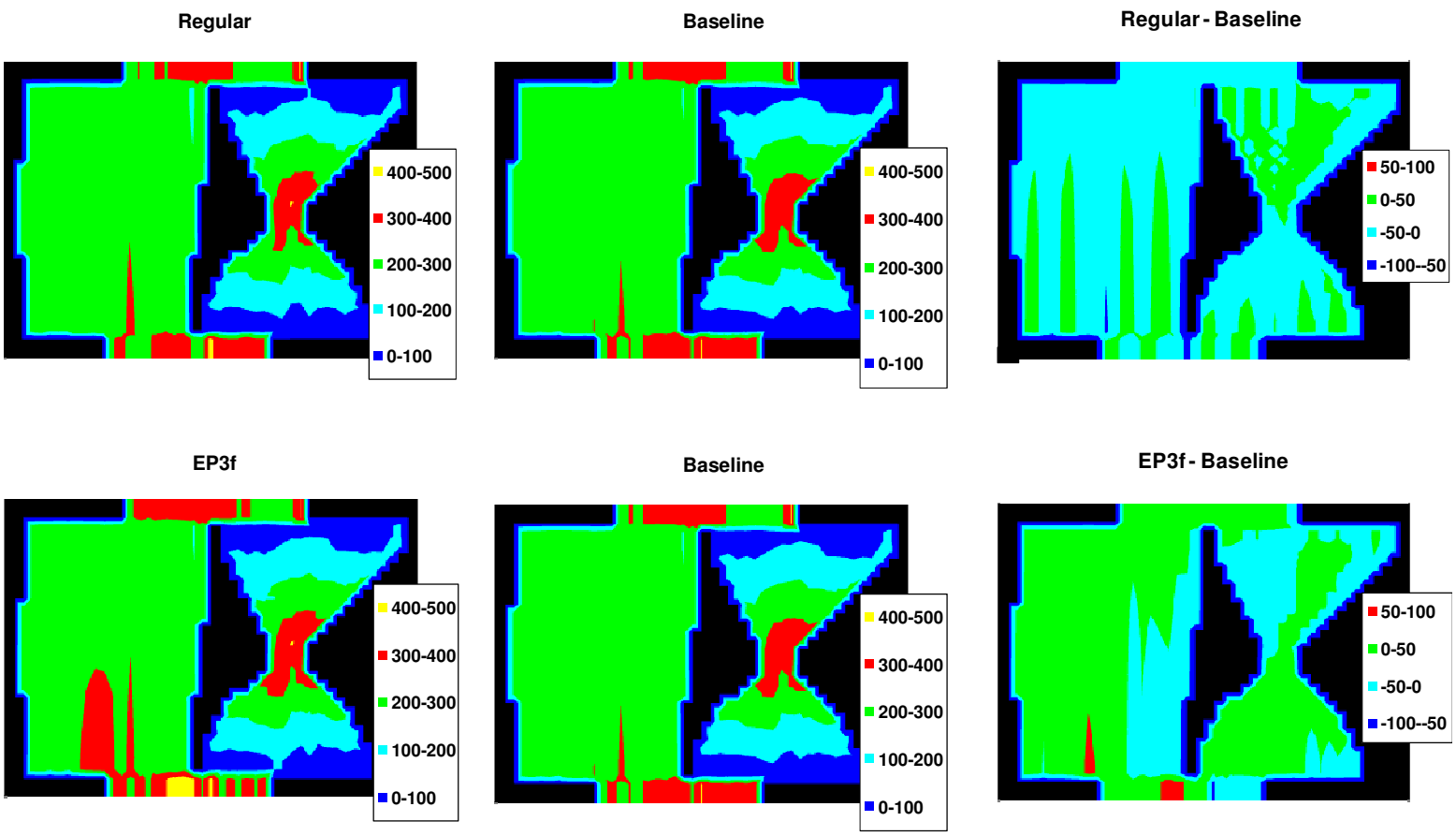


Figure 2b. FAS activation, showing only excitation.

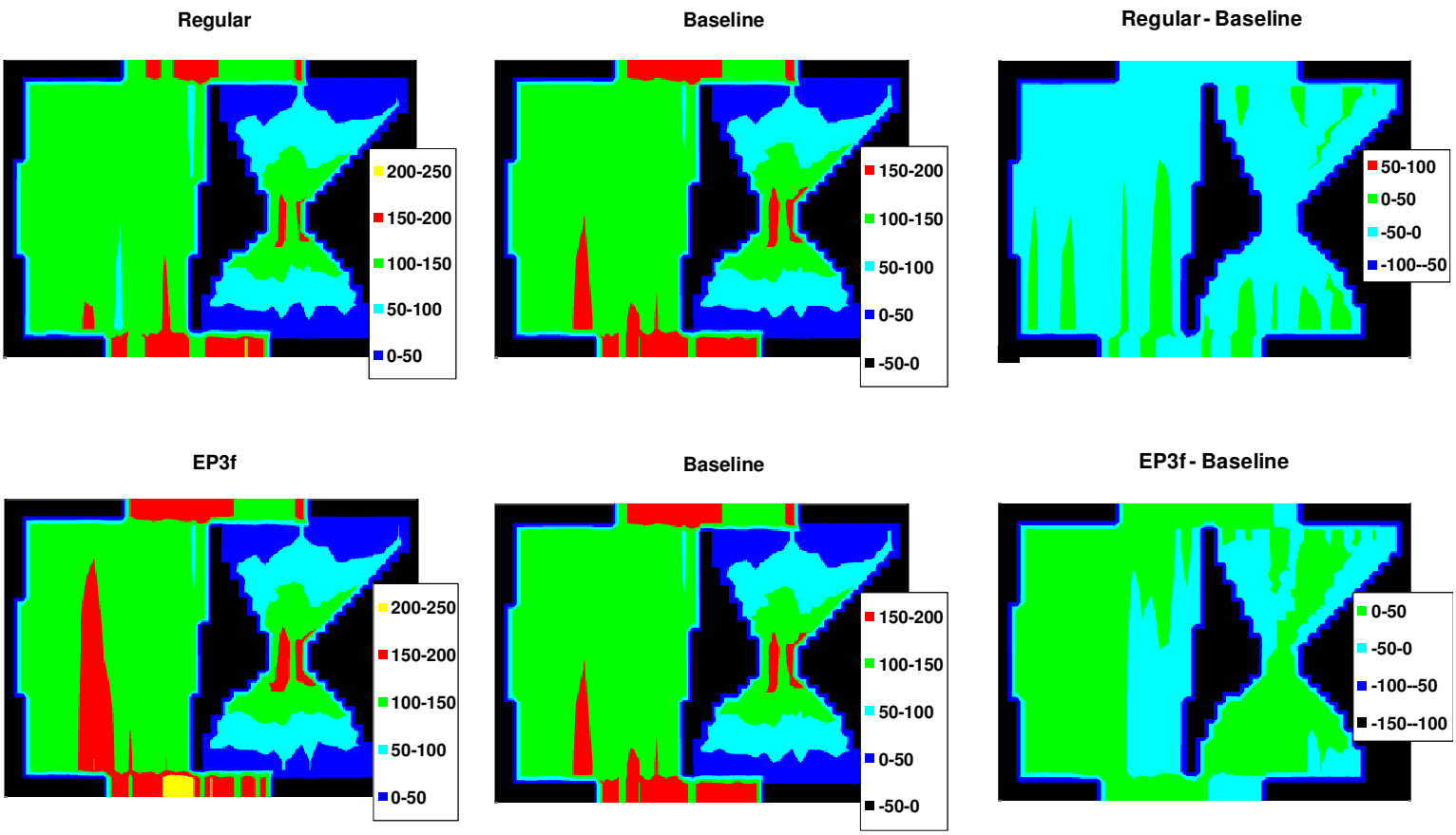


Figure 2c. FAS activation, showing only inhibition.

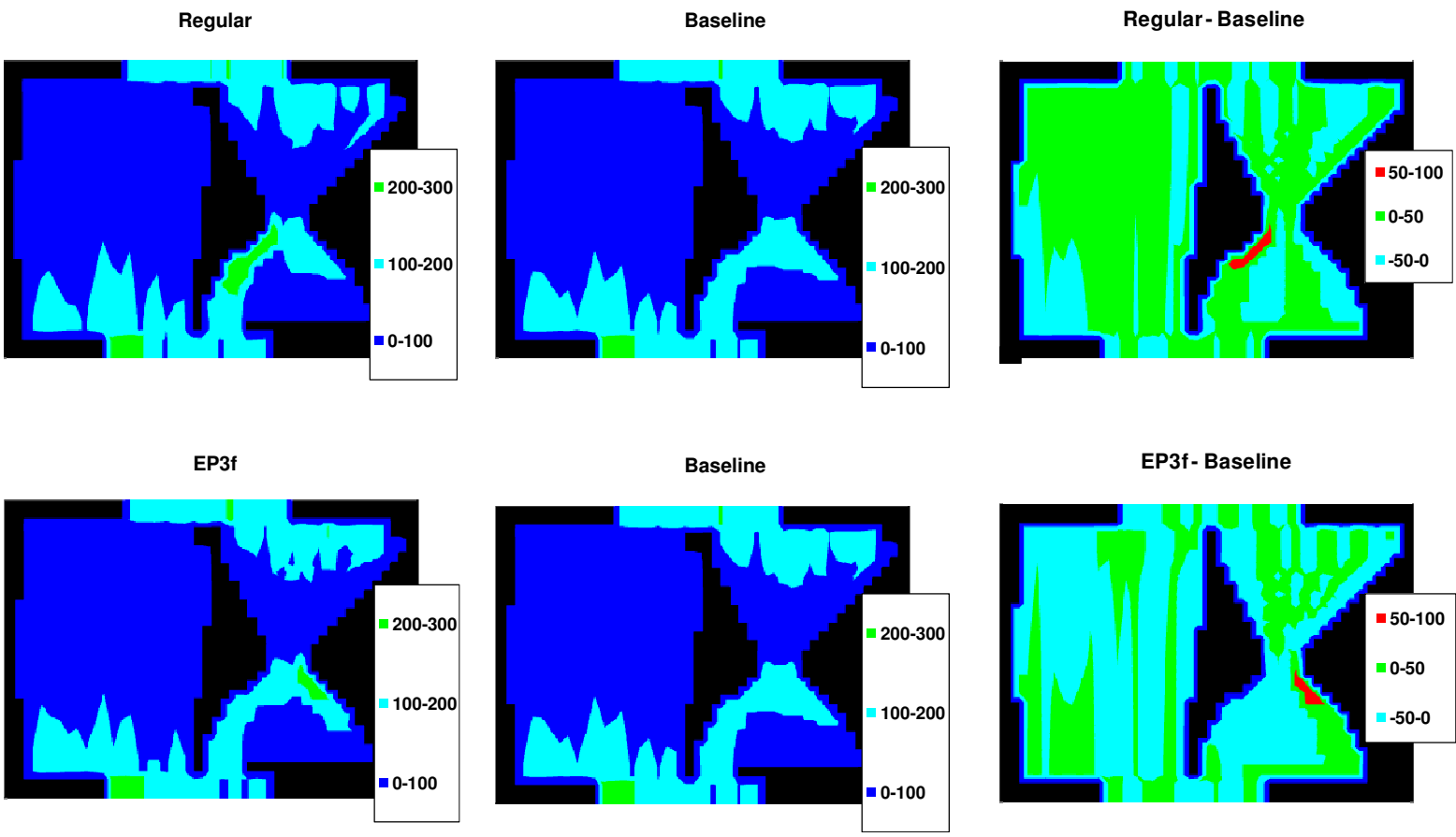


Figure 3. Scatterplot showing percentage specialisation indicated by lesioning and imaging, for each of the 12 networks, for each verb type. Lesion specialisation is on the horizontal axis; imaging specialisation is on the vertical axis. For both axes, positive values indicate specialisation towards the direct route; negative values indicate indirect route specialisation.

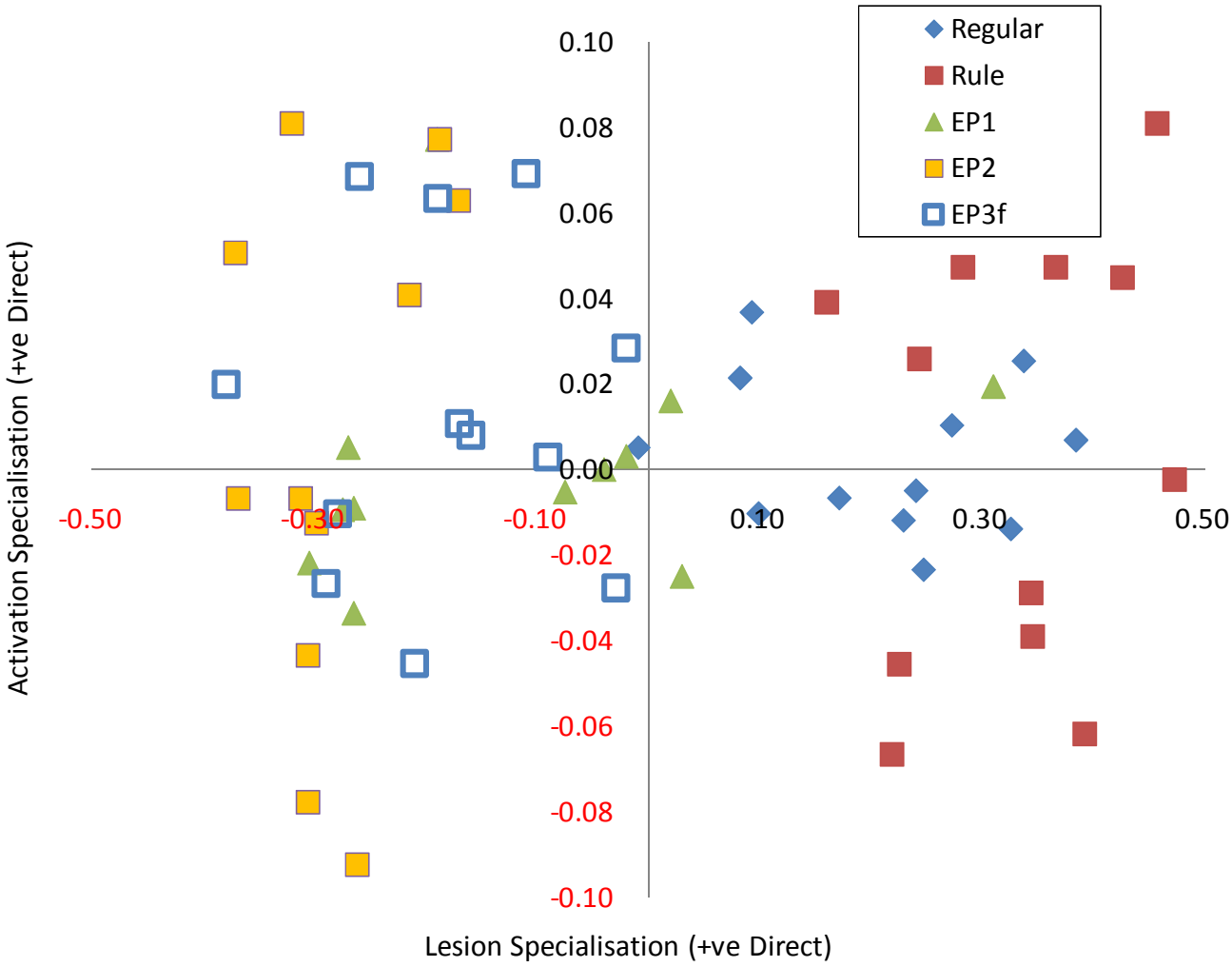


Figure 4. Scatterplot showing percentage specialisation indicated by lesioning and summed FAS for imaging, for each of the 12 networks, for each verb type, and separately for direct and indirect routes. Lesion specialisation is on the horizontal axis; imaging specialisation is on the vertical axis. For both axes, positive values indicate specialisation towards the direct route; negative values indicate indirect route specialisation.

Figure 4a. Direct route

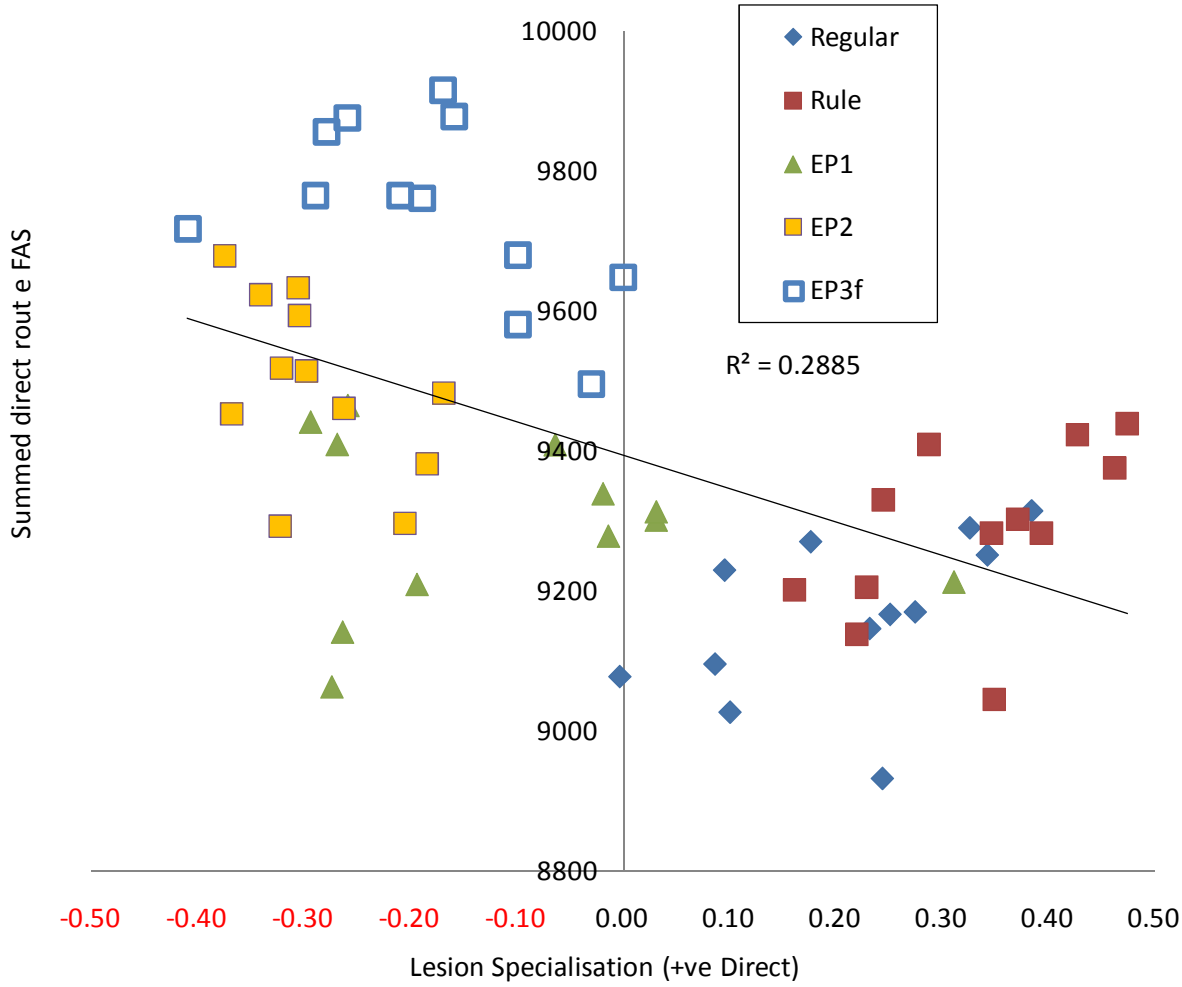


Figure 4b. Indirect route

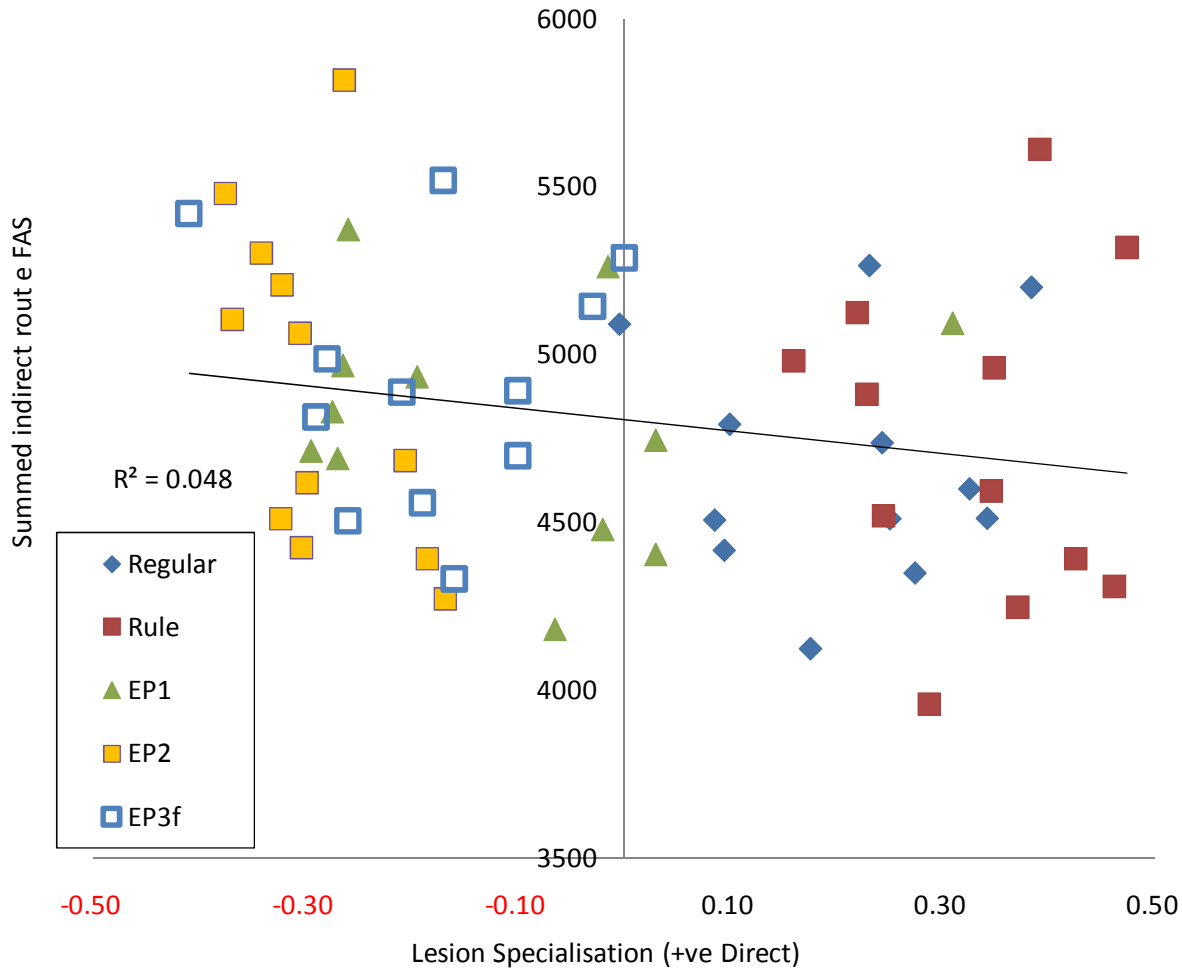
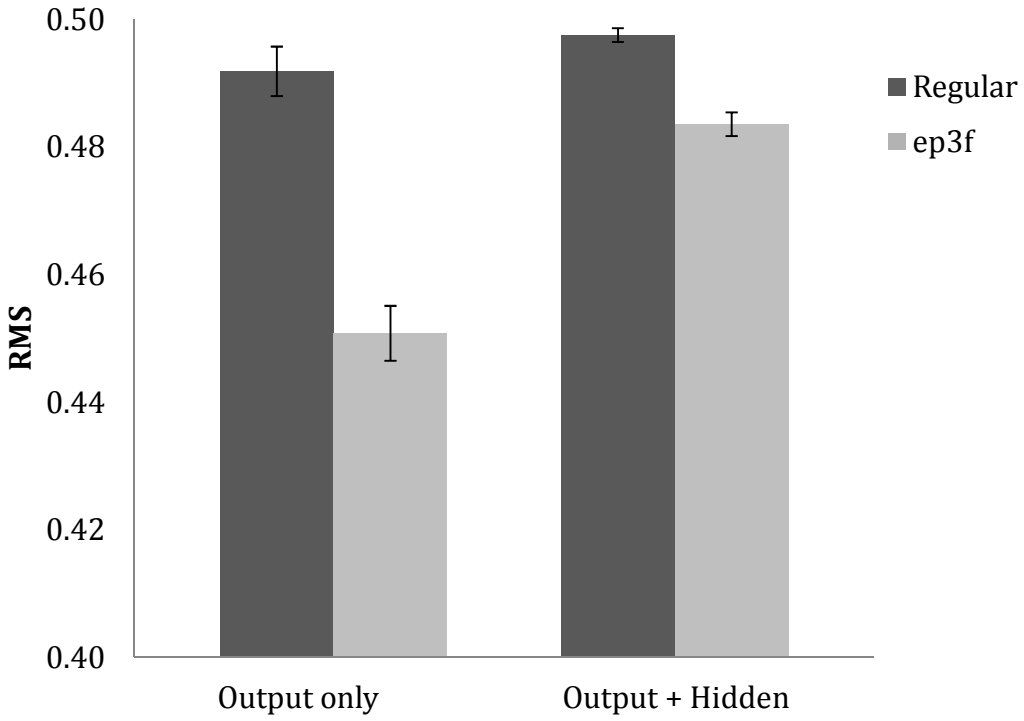


Figure 5. RMS error in producing correct past-tense forms, for Regular and Ep3f exception verbs. ‘Output only’ denotes 100% lesion of *all* connection weights, revealing a bias in the resting state of the output units. ‘Output + Hidden’ denotes 100% lesion of all weights apart from those connecting the hidden and output layers. Higher scores indicate stronger bias and error bars depict standard errors of means.

Figure 5.


0.50
0.48
0.46
0.44
0.42
0.40
RMS
Output only
Output + Hidden
Regular
ep3f