# Re-evaluating Computational Models of Semantic Cognition

Olivia Guest

Department of Psychological Sciences
Birkbeck, University of London

Thesis submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy (Ph.D.) at Birkbeck, University of London

September 2014

**Abstract**

Several computational models of the semantic cognitive system have been developed. This thesis considers four such models: the hub-and-spoke, the conceptual structure, the modality-specific, and the conceptual topography model. These models account for both generalised and category-specific semantic impairments. The models encapsulate different, and partially mutually exclusive, theoretical positions, but still account for similar, if not overlapping, semantic impairments. However, no single theory explains the full spectrum of both healthy and impaired semantic cognition. In order to better understand the space of theories and their inter-relations, this thesis reports results from (re)implementing the four theories and attempting to simulate both types of semantic impairment within each implementation. These four implementations shed light on the various computational and modelling assumptions and implications of each theoretical position. Compatibilities (and incompatibilities) between each theory (and model) are also discussed. It is additionally argued that some assumptions within each account, even though superficially different, are shared, and, conversely, some seemingly minor or background assumptions are centrally important. Examples of the latter include: limiting attention within any model to just two input modalities; evaluating model and patient behaviour with non-naturalistic semantic tasks; ignoring input from executive and affective systems; or (within the hub-and-spoke model) appealing to emergent properties of the implementation.

**Declaration**

I declare that the work presented in this thesis is my own. Where it builds on other people's work or ideas this is clearly marked.

**Acknowledgements**

I would like to thank my supervisor, Professor Richard P. Cooper, for being supportive, helpful, generous with his time, and believing in me. Also thanks to my second supervisor, Dr. Eddy Davelaar, who has also been supportive throughout.

I would like to express my gratitude to the members of staff I have worked with over the last four years, especially Dr. Marie Smith, Dr. Alex Shepherd, Dr. Linda Thomson, and Prof. Ulrike Hahn. I am grateful to my department for providing me with a graduate teaching assistantship without which I would have had a poorer Ph.D. experience in more ways than one.

Thanks to my fellow research students who not only taught me a lot about their own work, but also have been good friends. This is especially true for Eleonora Vagnoni, Renata Sadibolova, Berna Sarı, Nick Berggren, Constantinos Mitsopoulos, Katerina Pishiris, Ramiro Joly-Mascheroni, and Philip Pell; all of whom I will always consider friends, despite geographical, ideological, and/or academic separation. I would like to express extra appreciation for Nick Sexton, Christopher Brand, and Nicole Cruz de Echeverria Loebell — their support has been instrumental in the last year of my Ph.D. This is also true for the Cognition and Computation M.Sc. students: Anna Peng, Suzanne Pahlman, and Sandy Vrontissis.

Special thanks to my two best friends and confidantes in the cognitive modelling world: Dr. Christina Bergmann and Burcu Arslan. Neither of whom have been in the same country as me for more than a few days, but they have nonetheless inspired and supported me in ways they probably cannot imagine.

Whilst teaching at Birkbeck, I have also met some excellent undergraduate students both in terms of their academic abilities and their character, especially Samir De Marchi, Maiya Zargaryan, Panchita Braune, Samira De Robbio, Idill Abdourahman, Sofia Ciccarone, Katerina Pappa, Paula Logan, Johanna Thea Farquha, Anna Szajda, Abdoulie Jarra, and many more. All of whom have motivated me by showing me that my efforts at teaching and marking are appreciated.

I am grateful to: Yianna Lazarou, Avery Dorko, Laurelai Bailey, Alexadra Brody Salazar, Niovi Phinopoulou, Nadia Kornioti, Magda Čeng, Margarita Iosif, Chris Price, Cary Bass-Deschênes, Hayes Hartman, Christopher Steinmuller, Lucas Price, László Szerémi, Allen Slea, Haley Goshert, Iris van Wanrooij, Henry Dickerson, etc.

Finally, I would like to extend gratitude to Alex, Erica, and David Guest who have been very supportive.

**Dedication**

I dedicate this thesis to my grandparents, Anthoulla Koumoullis, who wanted to go to university but was not allowed by her father, and Andreas Koumoullis, who did go despite the racism and other struggles immigrants must cope with. Even though they do not know me, I owe them everything.

# Contents

# Chapter 1

# Semantic cognition

## 1.1 Overview

The term *semantic memory* refers to a part of human long term memory consisting of a collection of abstract facts about the world. It is subsumed as part of the *semantic cognitive system*, which also comprises semantic executive functions — control mechanisms applied over semantic memory in order to access specific memories. Semantic facts underpin linguistic meaning, providing a substrate for reasoning and inference, for categorisation, and for the creation of prototypes or exemplars. It intuitively appears that semantic memory is an abstraction or generalisation over a set of experiences collected gradually over time, as first proposed by Collins and Quillian (1969, 1972).

In this chapter, the neuropsychological dimensions of semantic cognition will be described. Attention will be drawn specifically to disorders that affect this system; from historical investigations carried out near the turn of the previous century, to more current clinical and neuropathological details. In addition, the various theoretical positions pertaining to semantic cognition, its proposed function in healthy people and the patterns of breakdown in patients, will be discussed. Finally, computational models, borne out of some of the theories described, will be reviewed. These three sources of knowledge — healthy and patient participant data, verbal theories, and computational modelling — provide the introduction for the rest of the thesis.

## 1.2 Neuropsychological investigation

### 1.2.1 Frontotemporal lobar degeneration

#### 1.2.1.1 Historical details

Near the turn of the previous century, a number of neurologists and psychiatrists described dementing diseases that comprised symptoms such as behavioural changes, aphasia, language disturbances, and apraxia, and after post-mortem examinations, were seen to be correlated with neurodegeneration of the temporal and frontal lobes (Dejerine & Sérieux, 1897; Mingazzini, 1913–1914; Pick, 1892, 1901, 1904, 1906; Rosenfeld, 1909; Sérieux, 1893)[1]. Following on from that, during the 1930s and 40s, many papers were published focussing on these types of patients, often referred to as having Pick's disease (see Hodges, 1994, for a synopsis of these investigations). However, English-speaking interest waned after the Second World War, shifting focus away from neuropsychology to the neuropathology and neurochemistry (although continental Europe continued researching both the clinical and pathological aspects) of the disease. Some researchers even proposed that these patients were clinically identical to cases of Alzheimer's dementia (Hodges, 1994). In fact, this was a separate form of dementia, now known to be caused by *frontotemporal lobar degeneration* (FTLD): a spectrum of degenerative aetiologies affecting the temporal and frontal lobes and correlating with various clinical syndromes (Goedert, Ghetti, & Spillantini, 2012; Snowden, Neary, & Mann, 2002).

A large part of the credit for the reignition of interest in the neuropsychological examination and clinical classification of FTLD in the English literature goes to Elizabeth Warrington. In the mid nineteen seventies she described a semantic-specific impairment of cognition, further enforcing the dissociation between semantic and episodic memory proposed by Tulving (1972, 1987). She studied patients who showed a distinct loss of object knowledge, word-finding difficulties, and severe anomia. These patients, however, seemed to have otherwise healthy cognition, and their speech was grammatical (Warrington, 1975; Warrington & Shallice, 1979; Warrington, 1981; Warrington & McCarthy, 1983; Warrington & Shallice, 1984; Warrington & McCarthy, 1987; Warrington & Cipolotti, 1996). Schwartz, Marin, and Saffran (1979) also reported a patient with a progressive dementia who made severe semantic errors (e.g., identifying dogs as cats), as well as errors which would come to be seen as hallmarks of surface dyslexia (Graham, Patterson, & Hodges, 2000; Schwartz, Saffran, & Marin, 1980; Patterson et al., 2006; S. M. Wilson et al., 2009; Woollams, Ralph, Plaut, & Patterson, 2007). These symptoms are now known as part of the spectrum of dementing disorders caused by FTLD, specifically of

[1] Translations for Pick (1892, 1901) can be found in Girling and Berrios (1994, 1997).

a subtype called semantic dementia (SD, c.f., Snowden et al., 2002, 2001). At around the same time, Mesulam (1982) studied patients with language, but no significant behavioural, disturbances: their language-specific loss culminated in reading, writing, grammatical, and comprehension difficulties. These patients became known as suffering from Mesulam's syndrome or, more commonly, primary progressive aphasia. After the details of these behaviour-, language-, and semantic-specific syndromes were published, an explosion of research into these kinds of dementias took place (e.g., Cummings & Duchen, 1981; Cummings & Benson, 1983; Cummings, 1991; De Renzi, Liotti, & Nichelli, 1987; Graff-Radford, Damasio, & Hyman, 1990; Gustafson, 1987; Hagberg, 1987; Holland, McBurney, Moossy, & Reinmuth, 1985; Katzman, 1986; Kirshner, Tanridag, Thurman, & Whetsell, 1987; Knopman, Christensen, & Schut, 1989; B. Miller, Cummings, & Villanueva-Meyer, 1991; Munoz-Garcia & Ludwin, 1984; Orrell & Sahakian, 1991; Poeck & Luzzatti, 1988; Wechsler, 1977; Weintraub, Rubin, & Mesulam, 1990; Wisniewski, Coblentz, & Terry, 1972).

Perhaps predictably, the neuropathological investigation to further classify and explain these disorders went hand-in-hand with the reignited interest in neuropsychological and clinical research. Gustafson (1987) and Neary, Snowden, Northen, and Goulding (1988) initiated the development of clinical and pathological diagnostic criteria for the three different kinds of dementias that correlate with FTLD. Their work culminated in the current clinical/neuropsychological diagnostic criteria described in Neary et al. (1998) and Pijnenburg (2011) and the neuropathological criteria for diagnosing FTLD found in Cairns et al. (2007). These three disorders: behavioural-variant frontotemporal dementia, primary progressive aphasia, and semantic dementia; are collectively referred to as frontotemporal dementias (FTD) when describing their clinical/neuropsychological aspects, or as FTLD to denote their underlying neuropathology, or sometimes even by their histology (e.g., tau positive inclusions, often referred to as Pick-bodies); however, the genetic and histological details underpinning FTLD are outside the scope of this thesis (for genetic underpinnings see: Laforce, 2013). The concern here is with behavioural deficits.

It is important to note (as have others, e.g., Hodges, 1994; Westbury & Bub, 1997), that during the earlier stages of the scientific investigation of the spectrum of the dementias caused by FTLD, the terms used were quite confusing. This is because consensus on both the definitions of technical terms, e.g., for the names of the diseases, and the diagnostic criteria had not yet been reached. Some areas of the literature are still in need of standardisation, see section 1.2.1.2 for more details. Confusion is perhaps further compounded by the fact that the clinical/neuropsychological diagnostic criteria changed slightly from Anonymous (1994) to

their revised current state in Neary et al. (1998) and Pijnenburg (2011, criteria updated only with regards to the behavioural variant of FTD) along with proposed amendments (only with regards to the primary progressive aphasia subtype) in Gorno-Tempini et al. (2011).

#### 1.2.1.2 Subtypes of frontotemporal lobar degeneration

As mentioned previously, there are three syndromes that share the same or similar locus of degeneration[2] and histopathological characteristics (although FTLD aetiologies are pathologically and clinically heterogenious, Cairns et al., 2007; Gorno-Tempini et al., 2011; Rabinovici & Miller, 2010), but that have specific, often completely dissociable, cognitive repercussions. In more detail, these syndromes are:

**Behavioural-variant frontotemporal dementia** (bvFTD), also referred to as frontotemporal dementia[3] without a specifier and as frontal-variant FTD, is the most common dementia after Alzheimer's, comprising 70% of the patients with non-Alzheimer frontotemporal lobar degeneration (Hodges & Miller, 2001; Hodges, Miller, et al., 2001; Neary et al., 1998; Rabinovici & Miller, 2010; Snowden et al., 2002).

BvFTD is also sometimes called Pick's disease (see Hodges, 1994, for an overview of the history and changing nomenclature) — although this is often a misnomer, as bvFTD does not directly imply the existence of Pick-type histological changes, and conversely Pick's does not directly imply bvFTD, as other types of clinical FTLD syndromes can also correlate with the same Pick-type histology (Cairns et al., 2007; Hodges et al., 2010; Rabinovici & Miller, 2010; Schroeter, Raczka, Neumann, & von Cramon, 2007). Pick-bodies (tau positive inclusions), in fact, are relatively rare in patients, being found in only a quarter of cases (Snowden et al., 2002).

BvFTD is primarily a behavioural disorder, hence the name, in which patients suffer severe changes in their character and social awareness; thus dissociating it from the other two types of FTLD mentioned below. These behavioural changes correlate with lesion damage that is centred more on the frontal than the temporal lobes (Schroeter et al., 2007).

**Primary progressive aphasia** (PPA) is a language-specific degenerative disorder, clearly dissociating a linguistic faculty from the rest of cognition (M. Bonner, Ash, & Gross-

[2] In the latter stages of degeneration, the lesions in patients are usually both temporal and frontal, although at initial stages, so if caught early enough, lesions will be localised to one or the other.

[3] Care must be taken to not confuse "frontotemporal dementia" meaning this specific frontal lobar behavioural subtype of dementia caused by FTLD with "frontotemporal dementia" used as a blanket term for all three of the dementias caused by FTLD.

man, 2010). PPA was first described by Mesulam (1982), who divided patients into fluent and non-fluent subtypes. However, PPA patients have come to be separated into: semantic-, logopenic-, and nonfluent/agrammatic-variant primary progressive aphasia — a classification introduced by Gorno-Tempini et al. (2004) due to patients not conforming to the simpler bifurcation (M. Bonner et al., 2010; Grossman et al., 1996; Gorno-Tempini et al., 2011).

As with the dissociation of the different kinds of FTLD, the classification of the different subtypes of primary progressive aphasia causes some confusion when reading articles written before their introduction. To further compound these classification issues, the semantic variant of PPA is often called semantic dementia. Some authors believe them to be two distinct syndromes, claiming that semantic dementia patients display a visual deficit for objects and faces, while those with the semantic variant of PPA do not (e.g., Mesulam et al., 2009). However, most authors do not differentiate between the two (Adlam et al., 2006; M. Bonner et al., 2010; Gorno-Tempini et al., 2011, 2004; Grossman, 2010; Knibb, Xuereb, Patterson, & Hodges, 2006; Westbury & Bub, 1997). This is mainly because both groups of patients, given time, progress to having the same clinicopathological and neuropsychological characteristics symptomatic of semantic dementia (even under the definition of Mesulam et al., 2009); and due to the nature of the literature, "because very few progressive aphasics have had the semantic testing necessary to differentiate them from patients with semantic dementia" (Westbury & Bub, 1997, p. 382).

It must be noted that a further source of confusion also exists when examining the literature on PPA. When it is referred to as progressive nonfluent aphasia — as named originally by Grossman et al. (1996) and in line with the nomenclature of the official diagnostic criteria (Neary et al., 1998) — instead of as PPA, the implication is that it is the logopenic and the nonfluent/agrammatic and not the semantic variant. So when the term progressive nonfluent aphasia is used the form of FTLD involves both frontal and temporal cortices and can be differentiated (both neuropsychologically and using neuroimaging) from that of bvFTD and SD (Schroeter et al., 2007). In contrast, when the broader term PPA is used it cannot be differentiated as easily from SD, for the reasons mentioned previously (M. Bonner et al., 2010).

**Semantic dementia** (SD), a temporal degenerative disease, is characterised by a progressive loss of conceptual knowledge (Davies et al., 2005; Rabinovici & Miller, 2010; Schroeter et al., 2007). SD is also sometimes known as progressive fluent aphasia (Hodges, Patterson,

Oxbury, & Funnell, 1992; Grossman, 2010; Lambon Ralph, Howard, Nightingale, & Ellis, 1998), semantic-variant FTD (Pijnenburg, 2011; Seeley et al., 2005; Snowden et al., 2002, 2001), and semantic-variant PPA (M. Bonner et al., 2010; Gorno-Tempini et al., 2011, 2004; Mesulam et al., 2009). SD patients show a loss of the meaning of words, objects, faces, non-verbal sounds, smells, tastes, and somatosensory stimuli, but a preservation (at the early stages) of the rest of their cognitive faculties. It appears to be a degenerative disorder mainly affecting semantic cognition, hence the name — although once the disease progresses behavioural disturbances become more common (Rosen et al., 2006).

The psychological and the pathological lines of investigation paved the way for all the FTLD variants to be studied more closely. The FTLD sub-syndrome of interest in this thesis is the semantic-specific neurodegenerative disease called semantic dementia — with importance placed on the patterns of preservation and loss of semantic cognition shown by SD patients during specifically created semantic tasks, which will be addressed in the next section.

### 1.2.2 Tasks used to assess semantic cognition

As mentioned before, patients with specific deficits on tests of semantic cognition were first described by Warrington (1975). Her patients, who were in their early sixties, were tested on many aspects of their cognitive functioning in order to identify their deficit as one of pure semantics and not one of an intellectual, perceptual, or linguistic nature. Specifically, three patients were tested on their verbal and performance IQs, their comprehension of syntactically complex phrases, their ability to discriminate shapes and fragmented letters, their competence in matching different views of faces and objects, and their memory capacity for visual imagery. It was apparent that her patients had normal to above average cognitive functioning, were well-oriented in space and time, had normal vision and visual acuity, were sufficiently articulate and fluent in their responses, and had normal digit spans[4]. However, their ability to discriminate familiar items amongst distractors, and to recall specific properties of objects was significantly impaired.

The severe agnosias (or conceptual difficulties, Simmons & Barsalou, 2003) seen in the patients described by Warrington (1975) were uncovered using: *a*) a forced choice object-recognition test, during which the participants had to indicate which, amongst a triplet of coloured drawings, was the picture that corresponded to a certain category (e.g., "Which one

[4]The three patients described by Warrington (1975) seem to be affected by surface dyslexia. This has been inferred from the fact that patient A.B. had trouble reading non-standard, yet high-frequency, words such as "nephew", but had no trouble with "classification". See Warrington (1975, p. 638-9) for details on this task and Woollams et al. (2007) for more recent investigations.

is an animal?"), to an attribute (e.g., "Which one is used the kitchen?"), or to an associative property (e.g., "Which one is the heaviest?"); *b*) a task consisting of 40 photographs that probed the patients' knowledge for the same property information as in the first test, but as there were no distractors, the questions used previously were reformulated to require a Boolean-valued response (e.g., "Is it an animal?" or "Is it a swan or a duck?"); *c*) an auditory equivalent to the previous task; *d*) a word definition task that focused on the distinction between concrete and abstract words; *e*) a task in which patients had to discriminate nonsense words from real ones, and absurd sentences from sensible ones; *f*) a qualitative test of their comprehension of frequently used proverbial phrases (e.g., "Strike while the iron is hot"); and *g*) an identification task for sounds (e.g., the sound a dog makes). All of the visual and auditory stimuli used in these tasks were selected so as to be balanced in regards to their categorical and other properties, so there were equal numbers of both living and non-living things, and of concrete and abstract words, and so on. This allowed for clear differences between the patients' performance on separate groups of items to be identified.

The battery of tasks developed by Warrington (1975) revealed a pattern of deficiencies in the three patients that hint at dissociations within the semantic system. In tasks *a*), *b*), and *c*) patients were relatively unimpaired at classifying objects into the available superordinate categories (e.g., animals); in contrast to this, they were very poor at answering the questions about objects' subordinate[5] attributes and associations (e.g., deciding if an item is made of metal, or if it is heavier than a telephone directory), and at chance level at matching an item's name to its picture. The results from task *d*) show a "particularly striking" set of responses from patient A.B. He was able to provide sufficiently apt definitions for abstract words (e.g., supplication), but completely unable to do so with concrete words such as cabbage, poster, and needle; his respective responses were: "Eat it", "No idea", and "Forgotten". During task *e*) the patients

showed they had the ability to correctly read the words and sentences; however, they were unable to discern the meaningless from the meaningful, scoring slightly better at doing so in the case of words than whole sentences. In *f*), the patients could not indicate any knowledge of the overarching meaning of proverbs, despite, in A.B.'s case knowing what the constituent concrete

[5]Super- and sub-categories are not defined or probed in the same way by all authors. In the case of Warrington (1975), a superordinate category is a group such as animal, plant or inanimate object or, a more specific subset of the previous, such as mammal, tree, or liquid. Whereas a subordinate category involves sorting the same objects into four legs, green, or juicy. In other words, superordinate categories pertain to the various levels of semantic categories as they are used in most of the current literature, however subordinate categories as in Warrington are a different and slightly orthogonal classification involving somato-sensory features of the items themselves (e.g., colour, consistency, or number of legs) rather than some general label given to them (e.g., fruit, bird, or tool). In contrast, Rogers et al. (2004) uses "subordinate category" to mean subdivisions within domains, see footnote 11.

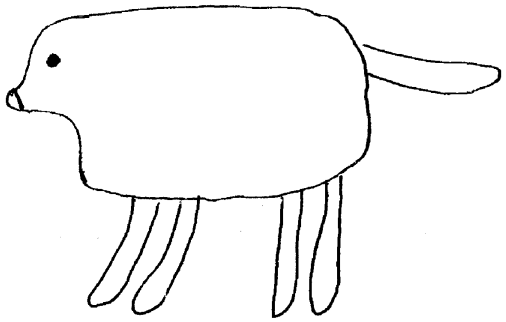

Figure 1.1: A duck drawn by patient JH (Bozeat et al., 2003, fig. 1). More examples of drawings like hers can be seen in Lambon Ralph et al. (1998).

words meant. Finally, *g*) indicated that two out of the three patients were severely agnosic to sounds; the exception was patient E.M. who scored at ceiling, despite being as incapable at other semantic tasks as the other two.

A common battery of neuropsychological tests have since been developed that assess semantic cognition aptitude. These *semantic tasks* usually comprise: *confrontation naming*, where an appropriate verbal name must be provided for a picture; *word-to-picture matching*, where a linguistic label must be paired with its corresponding picture from a selection that includes distractors; *sorting*, where a selection of words or pictures must be classified under hierarchical categories; and *drawing, copying and delayed copying*, where three sketches must be created, the first recreated purely from memory in response to a word, the second by direct copy from a line-drawing, and the third from memory a short time after the direct copying subtask (Farah & McClelland, 1991; Gotts & Plaut, 2002; Rogers et al., 2004; Warrington, 1975).

### 1.2.3 Semantic dementia

The term *semantic dementia* was coined by Julie Snowden and colleagues in 1989, to mean impairments of the semantic cognitive system that are caused by temporal cortical atrophy. The syndrome's characteristics comprise impairments to language, memory, and recognition of objects while leaving other cognitive faculties largely intact (Snowden, Goulding, & Neary, 1989). She examined three patients with these symptoms and inferred that their neurodegenerative illness has cognitive repercussions that can be dissociated from those of Alzheimer's dementia[6], and thus they can be classed into a separate form of neurological impairment. More patients have since been diagnosed as having semantic dementia. In addition, various cases reported at the turn of the previous century, mentioned in section 1.2.1.1, have been identified

[6]This is because in Alzheimer's dementia (AD) visuospatial and episodic memory deteriorates at the same pace as semantic memory. Also, in AD neurodegeneration is broad, not just focussed on the temporal lobes as in SD. Although, semantic deficits occur in both dementias (c.f., Garrard, Patterson, Watson, & Hodges, 1998).

retrospectively (Compston, 2011; Fletcher & Warren, 2011).

Patients diagnosed with SD perform better when tested on familiar or typical items as opposed to novel or exceptional ones. For example, if asked to classify a bat (an atypical mammal), patients are expected to have difficulty in giving the correct response; similarly, a zebra may be misidentified as a horse, despite the presence of stripes (Tyler, Moss, Durrant-Peatfield, & Levy, 2000). Additionally, patients often extend typical labels to semantically related objects, like labelling all prototypical mammals as dogs or, inversely, constricting the definition of mammals to just four-legged furry creatures. In much the same way, they effectively normalise objects by applying frequently occurring sets of features to them; for example, an SD patient created a drawing of a swan with a human-like face and four legs (McClelland & Rogers, 2003, fig. 2d). Similarly, another SD patient's drawing of a duck with four legs can be seen in Figure 1.1.

### 1.2.4 Category-specific semantic deficits

While SD normally involves a general semantic deficit, some patients perform significantly better when tested on certain categories of objects over others. Such intra-semantic selective deficits are important as they shed light on the internal structure of semantic memory. This pattern of dissociations has come to be known as *category-specific deficits*.

So in addition to the largely global semantic degeneration observed in SD, other (sub)populations of patients show deficiencies that appear to be limited to a category or a modality, or are at an intersection of the two (e.g., Bunn, Tyler, & Moss, 1998; Capitani, Laiacona, Mahon, & Caramazza, 2003; Carbonnel, Charnallet, David, & Pellat, 1997; Farah, Hammond, Mehta, & Ratcliff, 1989; Lambon Ralph et al., 1998; Rosazza et al., 2003). For example, patients who are diagnosed with Herpes Simplex Virus Encephalitis (HSVE), an infection that causes a decrease in grey matter in the anterior temporal lobe, sometimes demonstrate a selective loss of knowledge for animals, with relatively unimpaired knowledge of artifacts (Bunn et al., 1998; De Renzi & Lucchelli, 1994; Laiacona, Capitani, & Barbarotto, 1997; Moss, Tyler, & Jennings, 1997; Noppeney et al., 2007; Sheridan & Humphreys, 1993; Warrington & Shallice, 1984). Alzheimer's dementia patients also show the same dissociation (Garrard et al., 1998; Silveri, Daniele, Giustolisi, & Gainotti, 1991).

In other words, these groups of patients perform significantly worse under semantic testing on living things in comparison to inanimate objects. On the other hand, global dysphasic, global cerebral degenerative disorder, temporal lobe stroke, and head injury patients have been occasionally reported to show the reverse impairment: performing significantly better at se-

mantic tasks involving living than non-living items (Capitani et al., 2003; Hillis & Caramazza, 1991; Moss & Tyler, 2000; Sacchett & Humphreys, 1992; Shallice & Cooper, 2011; Warrington & McCarthy, 1983, 1987). This demonstrates a doubly dissociable category-specific property of the semantic system.

Other similar double dissociations have also been reported, such as the "particularly striking" example of AB studied by Elizabeth Warrington, and other patients, who can define abstract but not concrete words (Breedin, Saffran, & Coslett, 1994; Warrington, 1975, 1981; Yi, Moore, & Grossman, 2007) and of patients with the complementary deficit (Tyler, Moss, & Jennings, 1995). Patients have also been documented with category-specific deficits that appear to transcend the living/non-living distinction; for example, selective losses of knowledge for parts of the body (Dennis, 1976; Suzuki, Yamadori, & Fuji, 1997) or for fruit and vegetables (Rogers et al., 2004; Sheridan & Humphreys, 1993). This raises the question as to whether body parts, and fruit and vegetables are mentally classified under living, non-living, or under separate domains of knowledge; or perhaps mentally represented in a different way altogether. In fact, category-specific deficits, while intriguing, are a source of debate in the literature. Consensus has not been reached yet as to which, if any, lesion locations correspond to which type of conceptual loss.

### 1.2.5 Access impairments

To add to the properties of semantic memory, a partially orthogonal dissociation to the various category-specific deficits was identified by Shallice (1987): degraded representations versus impaired access. On the one hand, there are patients that seem to have permanent loss of semantic knowledge. SD patients, patients with dementia of the Alzheimer's type, and HSVE patients are usually found to be consistent in their semantic deficits (e.g., their inability to name objects), dubbed *degraded store patients*. This indicates that the relevant knowledge is completely irretrievable, seemingly because its neural representation has been irrevocably damaged (Breedin et al., 1994; Chertkow, Bub, & Seidenberg, 1989; Warrington, 1975).

On the other hand, some patients show signs of intermittent access to their store of conceptual knowledge by producing inconsistent testing results (Cipolotti & Warrington, 1995; Gotts & Plaut, 2002; Warrington & Shallice, 1979; Warrington & McCarthy, 1983). *Access patients* are documented as entering a refractory-like state after being tested on a particular item; such as for example, patient VER, studied by Warrington and McCarthy (1983), who was a global dysphasic with category-specific deficits as well as an access/refractory pattern of semantic degeneration (hence the partial orthogonality). When these patients are tested on an item (e.g.,

a cow), all its semantic neighbours (in this case, other mammals) are temporarily detrimentally affected, while leaving unaffected the knowledge of unrelated concepts (e.g., tools). It therefore appears as if these patients' semantic abilities vary as a function of the tests themselves. Specifically, the temporal interval between probing a particular conceptual category consecutively needs to be long enough to allow some kind of recovery, which would result in regaining access to the relevant memories. Otherwise, if the time delay between tests is sufficiently small, access patients are unable to respond accurately, presumably due to semantic interference.

### 1.2.6 Discussion

A variety of semantic patients with a range of aetiologies has been documented in the literature, providing the impression of a complex and sophisticated system, that even after neurodegeneration retains the ability to perform domain-level classification. Specifically, the nature of SD causes patients' semantic skills to disappear in a process akin to the reverse of learning. This unlearning appears to reverse the semantic developmental stages an individual goes through: concepts learnt later in life are normally lost first (Mandler, 2000; McClelland & Rogers, 2003; Warrington, 1975). This, along with the other characteristics of each semantic disorder — category-specific and access refractory deficits — hint at some form of hierarchy within the system, in which structural damage is intrinsically linked with, and gives rise to, functional deficiencies.

The spectrum of semantic disorders provides evidence for a relatively functionally distinct semantic memory system with certain categorical and access/refractory properties. These findings are invaluable to neuroscience and neuropsychology in general, but also specifically to the task of understanding the operation of the semantic system, supporting the development of theories to account for its form and function, and modelling it using artificial neural networks. The next few sections will explore these two related efforts.

## 1.3 Theories of semantic cognition

### 1.3.1 Overview

As has been touched upon, the semantic cognitive subsystem creates and stores concepts and derives associative, taxonomic, and causal (c.f., Fenker, Waldmann, & Holyoak, 2005) relations over them. This section aims to give an account of theories concerning the form and function of the semantic system. The following theories draw their inspirations both from general se-

mantic deficits and from the deficits of category-specific patients, who show partial or complete preservation of some categories over others, and thus suggest dissociations within the system. Not all theories will be presented here, as looking at every account is beyond the scope of this thesis. Specifically, theories (such as those proposed by: Arbib, 2008; Gallese & Lakoff, 2005; Glenberg & Robertson, 2000; Zwaan, 2004) on the stronger end of the embodied cognition scale are left out, as are those on the other extreme end of the spectrum (e.g., Collins & Loftus, 1975; Landauer & Dumais, 1997; Levelt, 1989).[7]

In this section, the theories of semantic cognition will be presented in chronological order of publication — with the oldest theories being the modality- and domain-specific, then the conceptual structure theory, and finally, the newest being the hub-and-spoke and the conceptual topography theories.

### 1.3.2 (Pre-)semantic features

Before looking at the details of each theory, *(pre-)semantic features* need to be considered. For the purpose of clarity, henceforth, angle and corner brackets will be used to denote features and concepts respectively, i.e., ⟨*feature*⟩ and ⌜*concept*⌝ (similar to the notation used by Sartori & Lombardi, 2004; Sartori, Gnoato, Mariani, Prioni, & Lombardi, 2007). Features, also known as properties or attributes, — although some theorists (e.g., Rogers et al., 2004; Simmons & Barsalou, 2003) differentiate between these words in order to use them to refer to different levels of processing or modalities within the semantic system — are grounded in modalities and appear to be shared amongst semantic concepts. Features can be: *modal* when they describe some perceptually derived property, such as ⟨*is red*⟩, ⟨*has wings*⟩; or *functional* when they describe the use or functionality of an object, like ⟨*makes music*⟩ or ⟨*is edible*⟩. Sensory features are further subdivided into visual, auditory, olfactory, gustatory, motor, somatosensory, and so on. Functional features were originally defined as the function or purpose of an object which "may be a product of a more abstract or more schematized representations based (at least during acquisition) on sensory/motor information" (Warrington & McCarthy, 1987, p. 1292); although sometimes functional is interpreted as a synonym for motor feature (e.g., Simmons & Barsalou, 2003). Modal features are derived from sensory information, while functional properties are likely to be passed on linguistically and through extended experience with using the object. Linguistic or lexical input to the semantic system is also taken into account by some theories,

[7] The theories and models that are discussed can still be considered as secondary embodied or weak embodied on the scale proposed by Meteyard, Cuadrado, Bahrami, and Vigliocco (2012) because they involve some sensory-motor information being manipulated in order to create higher-level amodal concepts. However this classification is a little unwarranted given that the proponents of such theories, both embodied and otherwise, would probably not agree with such a classification.

however it is less clear how these features relate to the sensory/functional dichotomy; Rogers et al. (2004), for example, see linguistic/verbal features and phonological input as two other modalities. Despite minor disagreements, most of the following theories of semantic cognition adhere to features being either sensory or functional. However the interpretation of the latter is not consistent and thus functional features are sometimes seen as another perceptual modality and other times as a relation over and above sensory features.

On the surface, features seem to be used to form concepts in the same way letters of the alphabet are used to write words. The applicability of this metaphor is a matter of perspective as, according to Solomon and Barsalou (2001), features have globally accessible instances, but locally they are correlated to specific concepts. What this means is that features' local forms differ, so the shared global form is selected on the basis of which local form is the most dominant (generalisable, archetypal). So taking an example from Solomon and Barsalou (2001), the following have $\langle mane \rangle$ as a feature: $\ulcorner horse \urcorner$, $\ulcorner pony \urcorner$, $\ulcorner lion \urcorner$, but experimentation shows that $\ulcorner horse \urcorner$'s $\langle mane \rangle$ is the dominant feature with $\ulcorner pony \urcorner$'s $\langle mane \rangle$ closely related to it, however $\ulcorner lion \urcorner$'s $\langle mane \rangle$ is a weaker instance of $\langle mane \rangle$ (presumably due to the differences in perceptual form between the animals). Such intra-feature dissociations give rise to, for example, the $\langle mane \rangle$ of $\ulcorner horse \urcorner$ impeding the activation or verification (the process of determining if a property belongs to an object) of $\ulcorner lion \urcorner$'s identically-named feature. Thus features are affected by the same priming, facilitation, interference, amplification, and context effects as are the fully fledged concepts (Solomon & Barsalou, 2001). Perhaps, obviously so, since concepts are defined as a set of features.

These findings blur the distinction between features and concepts. Nevertheless, the line between the two can be determined by the coarseness of the current view of the semantic system. On the one end of the spectrum: "a fully integrated concept is one that can be accessed [independently] of any particular context" (Antonucci & Alt, 2011, p. 551); and on the other: a purely sensory (e.g., visual) set of features are accessible only when perceiving or imagining (e.g., viewing or visualising) the item that possesses them. (Pre-)semantic features are discussed further below, as each of the following theories has their own take on their role, origin, and if they are outside or within the semantic store.

### 1.3.3 Modality-specific theory

Realising the important role features play for semantics, Elizabeth Warrington proposed that either modal (e.g., visual, olfactory, auditory, etc.) or functional features are responsible for forming concepts and categories (Warrington & McCarthy, 1983; Warrington & Shallice, 1984;

Warrington & McCarthy, 1987, 1994). Through examining features' contribution to the description of objects, she put forward that man-made objects (e.g., tools) have a dependence on functional properties, while living things (e.g., food and animals) are defined by their physical/sensory properties. This dissociation Warrington and McCarthy (1987) named the physical/functional dichotomy to underpin and update the previously apparent animate/inanimate distinction, which had been documented in Warrington and Shallice (1984).

The *modality-specific theory* follows on from the neurological tradition of associating each area of the brain with a specific sensory modality, this theory proposes that knowledge of the categories is stored in or near to the modal areas that contribute most to each domain or category of knowledge. So according to this theory, the area responsible for animals will be in or around the visual cortex, as visual features are what ultimately give rise to animate concepts. The implication is that the semantic system is physically dissociable into perceptual/functional areas due to the inherent storage structure of these modalities in the brain.

The modality-specific theory predicts that focal damage to a modality-oriented area will damage the features stored therein and thus detrimentally affect processing of the concepts that depend on them. So in the case of damage to visual features, the remaining semantic categories will be damaged only to the extent that they depend on this modality. In other words, assuming the store for functional features is left intact, the damage to tools (which are proposed to depend least on perceptual features) will be relatively small. On the other hand, damage to sensory areas will produce a deficit for the animate domain, and leave the inanimate domain largely intact.

The original proposal that some categories are rich in certain kinds of modal/functional features, although refined since the 1980s, has found support in feature studies (e.g., McRae & Cree, 2002), has been implemented in Farah and McClelland (1991)'s computational model, and has been largely accepted as common ground for other theories of semantic cognition (e.g., Humphreys, Forde, et al., 2001; Simmons & Barsalou, 2003). However, this functional/perceptual dissociation does not provide an account for all exceptional cases: such as for a patient with Alzheimer's dementia who had a loss of conceptual knowledge for living things but did not show a dissociation between knowledge of visual and functional features, and another patient with SD who had impaired knowledge of visual attributes of stimuli but did not show a category-specific deficit for animals (c.f., Lambon Ralph et al., 1998). So while providing insight into the functioning of semantic memory and providing a basis for other theories (e.g., the hub theory, see subsection 1.3.6) this theory only provides a partial account of the spectrum of disorders seen in patients.

### 1.3.4 Domain-specific theory

In contrast with the modality-specific outlook, Caramazza and Shelton (1998) propose what they call the *domain-specific*[8] *theory*, which states that the brain has areas that have been evolutionarily shaped to handle specific categories (also see: Capitani et al., 2003; Caramazza, 1998; Caramazza & Mahon, 2003). This theory proposes that there is innate connectivity to handle specific domains and categories. In particular, the domain-specific theory states that due to the evolutionary advantages of tools they are stored differently presumably due to pre-set wiring of certain brain areas with others. In other words, certain areas of the brain are pre-wired in order to deal with certain categories. Caramazza and Shelton (1998) use their theory to explain why the ventral stream is connected innately to motor areas: in order to process tools effectively and efficiently.

An updated version of the original domain-specific theory is outlined by Mahon and Caramazza (2011), wherein a slightly different theory is described (to perhaps address the under-specification of their originally proposed version of domain-specific theory): that brain regions involved in semantics are proximal to the area known for extracting relevant features (which appears to have moved closer to the modality-specific theoretical position). For example, Mahon and Caramazza (2011) posit that the medial fusiform gyrus (claimed to be disproportionally activated by tools) derives its location from its proximity to the parietal cortex (involved in manipulating objects). This is a more plausible account of connectivity but it does not offer an explanation for the patterns of activations seen in imaging during semantic tasks, nor an explanation for the lesion sites that give rise to category-specific affects (over and above those provided by the modality-specific theory).

However, Caramazza and colleagues maintain support for innate connectivity, e.g., that pathways between the ventral stream and motor areas are evolutionarily predetermined to facilitate tool use, although perhaps developmental connectivity is more plausible, as it allows for more environmental contribution to the formation of the structures that support the semantic system. This allows us to make a distinction between what is dubbed here as strong domain-specific theory (as found in: Caramazza & Shelton, 1998; Caramazza, 1998; Caramazza & Mahon, 2003) and weak domain-specific theory (described in Mahon & Caramazza, 2008, 2011; Martin, 2007); the former asserts evolutionary pressure creates loci that are domain-oriented, whereas the latter relaxes the domain-specificity to allow for the existence of areas that are to some extent sensory- as well as domain-specific, as well as slightly weakening the assumptions

[8] The proponents of the theory give it the name "domain-specific", but as they define tools as a separable class they seem to be operating lower down in the conceptual hierarchy, on the level of categories (see: Caramazza & Shelton, 1998).

regarding specifically-evolved domain-specific cortical regions.

As seen, the original domain-specific theory has some issues relating to its assumptions; but it does predict some of the patients' patterns of dissociations, given their form of damage. For example, focal lesions to semantically involved areas are expected to give rise to single categories being mal-processed, which indeed holds for some cases but not others. Either way this evidence does not necessarily support the part of the theory that claims that some areas have evolved specifically to process certain categories[9]. Moreover, it seems unlikely that evidence pertaining to an innate organisation of the semantic system, and neocortex in general can be found in the literature. In other words, "the organization of neocortex, which is commonly assumed to be a prime anatomical substrate for unique cognitive modules in the human brain, exhibits no robust signs of localized anatomical specialization above and beyond specific sensory and motor connections, and their polymodal interactions. [And in fact there is] little substantive research into how the de novo, evolved functions of the mind can be distinguished from the phenotypic consequences of individual, social, and cultural learning experiences." (Panksepp et al., 2002, p. 106)

In conclusion, a strong version of the domain-specific theory is hard to defend given the problems with asserting that innate domain-specific cortical regions exist. Either way, the predictions this theory makes are, in some sense, the same as those of the modality-specific theory's: focal damage causes (sub-conceptual) modal features to be lost, which in turn serve to support a single category or domain, and thus will manifest as the loss of semantic knowledge in a specific category/domain. Neither of the two theories can account for all the patient data as they currently stand.

### 1.3.5 Conceptual structure theory

Before proposing the domain-specific theory, Alfonso Caramazza and colleagues proposed a theory called the organised unitary content hypothesis (OUCH; Caramazza, Hillis, Rapp, & Romani, 1990); which, as shall be seen in following sections, has been further refined and modelled by Tyler et al. (2000) and Tyler and Moss (2001). The OUCH family of theories hold that semantic concepts are created through the passing of features into a single feature-space; the statistical properties of the input cause the semantic space to be segmented into categories and domains. This means that physical properties of objects, and thus by extension semantic features of concepts, drive the hierarchical branching of semantic knowledge: two

[9]Evolutionary psychological accounts are potentially misleading, even dangerous, without bearing critiques in mind (such as: Panksepp & Panksepp, 2000, 2001; Panksepp, Moskal, Panksepp, & Kroes, 2002).

concepts that share many features are more likely to be in the same category, while concepts that only have little in common are likely to be very distinct. As such, when concepts share many features with each other (e.g., ⌜*dog*⌝ and ⌜*cat*⌝) they are more likely to be preserved in semantic patients compared to concepts, and thus categories, without these internal correlations. Whereas distinctive features (e.g., ⌜*zebra*⌝'s ⟨*stripes*⟩) that can define a concept are easily lost (due to not being uniquely correlated with any other property or cooccuring over many similar concepts) causing the concept to merge into its neighbours (e.g., ⌜*zebra*⌝ might decay to ⌜*horse*⌝, as only ⟨*stripes*⟩ renders them separable). Another factor that defines the robustness to damage of categories and concepts is co-occurrences over features (e.g., ⟨*feathers*⟩ and ⟨*beak*⟩ are mutually inclusive, except in very rare cases like ⌜*turtle*⌝); so correlations over different features with a concept, and over the same features between concepts create robust representations.

The *conceptual structure theory* proposes that lesions damage the weakly correlated features of categories, this giving rise to dissociations — because, by definition, weakly correlated items are more vulnerable in such a system. So post-damage the patients' abilities deteriorate in such a way that different domains are affected to different extents. Because correlation implies preservation, the many shared features that living things have are better preserved, but cause generalisation with damage (all four-legged animals end up being alike). While on the other hand, as Tyler et al. (2000) claims (in line with the modality-specific theory), functional properties, which define artifacts, are robust to damage due to being highly correlated with perceptual properties. In other words co-occurring (sets of) features are likely to reinforce each other and thus protect each other from damage, which is consistent with patients' performance on semantic tasks, which shows a frequency effect on the category-, feature-, and concept-levels.

All semantic system theories agree with the conceptual structure theory that features form a backbone for concept creation, and are a mechanism for organising concepts into categories. To certain extents most other theories also support that the hierarchical structure that semantic knowledge can be presented in is dictated by the correlations of features both within and between concepts; meaning that shared features create categories and distinctive features define individual items. However, the conceptual structure theory, in its strongest version, states that the pattern of co-occurrences of features or the lack thereof is solely what determines their preservation, and ultimately the preservation of the concepts and categories downstream. Evidence of the intra- and inter-categorical statistical structure of features possessing the ability to drive the semantic system's organisation and to dictate the pattern of relative loss and preservation, can be seen in most neural network models (as co-occurrences inherently have this effect), where the input to semantics is a set of high-level features that describe each object

(e.g., Rogers et al., 2004; Tyler et al., 2000).

Empirical support for the conceptual structure theory can be found in almost all patient data; as has been seen subsection 1.2.3, it is most definitely the case that SD patients lose the commonest of features and concepts last, and the rarest and most distinctive of features first. This lends a lot of credibility, to the claims of this theory; specifically, that feature-structure helps shape the semantic space and split knowledge into domains and categories. Additionally, the distribution of features can explain how archetypes and exemplars are represented, e.g., an average ⌜*bird*⌝ or a stereotypical ⌜*vehicle*⌝ — by means of a generalised concept that has all features activated according to the baseline frequency of said category. So for example, due to repeated exposure to ⌜*dog*⌝ (assuming dogs are the most common mammal humans meaningfully interact with) ⌜*animal*⌝ is defined as the set of features ⌜*dog*⌝ most often activates.

To summarise, on the one hand the conceptual structure theory is appealing as it is compatible with most other theoretical positions; certainly so with the two previous theories. While on the other hand, (in isolation) the theory provides no robust lower-level prediction of the nature of the localisation of damage. This means that the conceptual structure theory must be paired with another theory for patient lesion data to be predictive. This in and of itself need not be considered a direct criticism. Especially since the semantic cognitive system and the cortical structures underpinning it should, in theory at least, be dissociable, given they exist on different levels of analysis.

### 1.3.6 Hub-and-spoke theory

The *hub theory* claims that an amodal semantic hub area exists in the temporal lobes. The hub is bidirectionally connected to modality-specific spokes, which are what grant it its ability to form amodal concepts, as can been seen in Figure 1.2. Thus using perceptual input it can form abstract attractors (like in a recurrent neural network). The structure and processing of this theory are based on the successes of the implementation, the Rogers et al. (2004) model, and are also particularly inspired by the modality-specific approach and the conceptual structure theory. The conceptual structure theory and the modality-specific theory are compatible with the hub theory. In the case of the latter theory, modality-specific areas exist as they form the pre-semantic/feature input to the hub, also known as the spokes, and the former accounts for the decay of features and concepts both in the hub and its spokes.

The neural network hub model, whose structure and emergent properties directly inspire the hub theory's account, shows that reciprocal connections between the central hub area and the perceptual processing areas are sufficient to give rise to attractors which appear to decay in

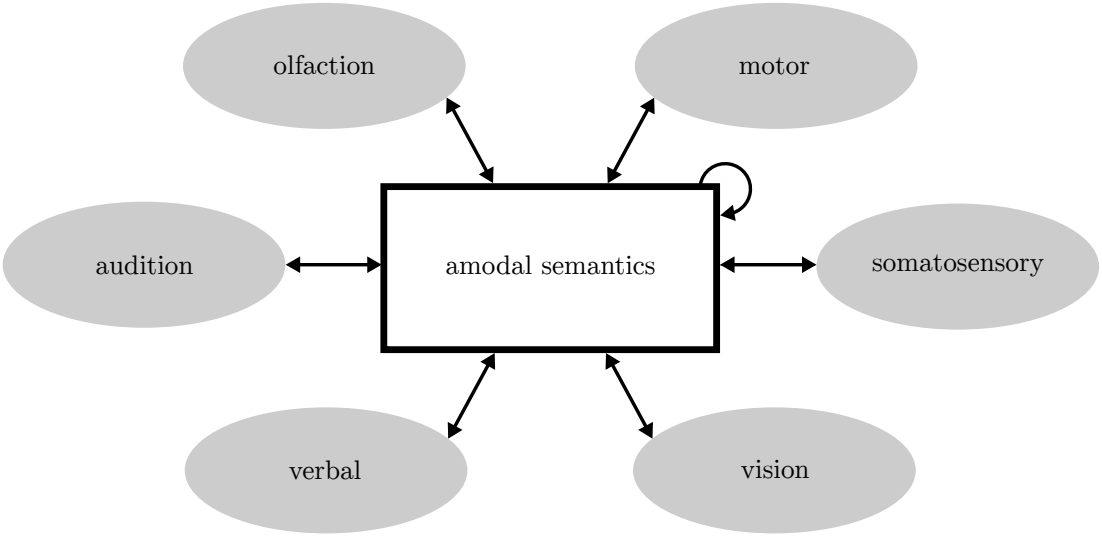


Figure 1.2: An overview of the hub and its modal spokes, based on fig. 2, Lambon Ralph et al. (2007).

such a way as to imitate patients with general semantic deficits and category-specific semantic deficits (Lambon Ralph et al., 2007; Rogers et al., 2004). The outputs of all the high-level sensory processing areas (visual, auditory, somatosensory, etc.), according to the hub theory, are the features that are derived from perceptual processing pathways. These are fed into the hub which uses them to form high-level amodal concepts. The attractors that underpin each concept, when damaged, decay in such a way that: the domain of animals collapses into a super-attractor that can recognise animals but not access their unique individual features; inanimate objects, conversely, cannot form a superordinate attractor easily due to being too dissimilar, instead they maintain some access to individual features and do not easily become confused.

This theory is dependant on certain assumptions: that all modal areas provide high-level pre-semantic features at their interface with the semantic hub; that the neuronal connectivity required for a hub in the temporal lobes indeed exists; that this connectivity is bidirectional; and that the contents of the hub are completely amodal. These are not wholly untestable, although it is hard to determine how much of the evidence in favour of a hub does in fact support the semantic hub theory and not any of the other contenders since many of their predictions overlap. The hub theory does offer a prediction that the others do not: damage to the hub, proposed to be in the anterior temporal lobes in an area known as the anterior temporal pole (ATP), should invariably give rise to semantic deficits that do not show a dissociation based on modality. So if a patient with focal damage to the ATP is found their pattern of deficits should be purely amodal. This does require some very careful clarification on behalf of the

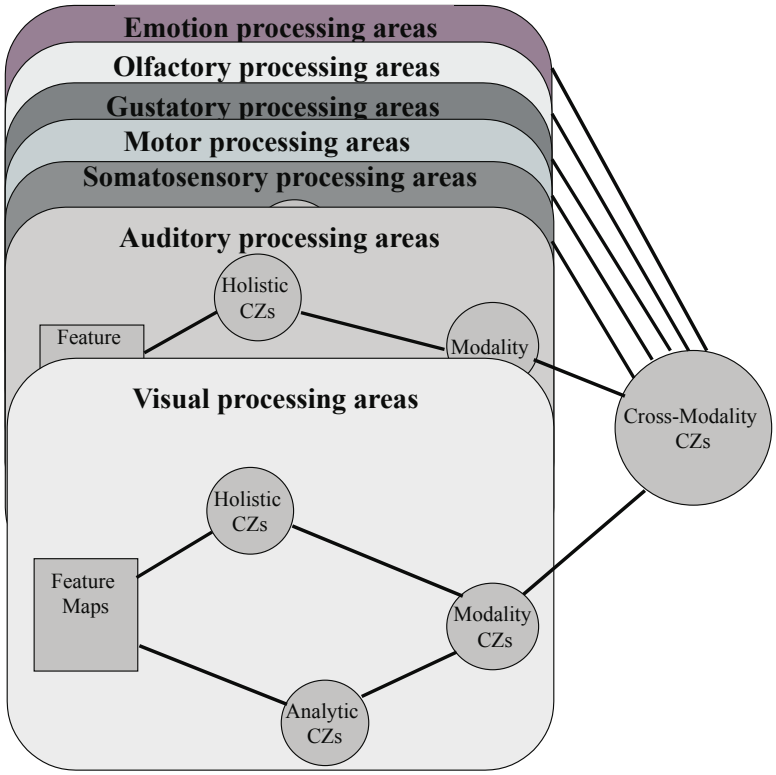


Figure 1.3: An overview of how the conceptual topography theory proposes CZs are organised hierarchically (Simmons & Barsalou, 2003, fig. 3).

proponents of this theory, especially since the hub model itself as presented in Rogers et al. (2004) does show category- and modality-specific dissociations, because the features it learns affect the organisation of the hub. Meaning that even though it is amodal in some senses, damage to it does not manifest as such. This issue is discussed further below.

### 1.3.7 Conceptual topography theory

Simmons and Barsalou (2003) propose the *conceptual topography theory*, which combines, and addresses some of the pitfalls of the above theories, and is inspired by the convergence zone (CZ) theory, developed by Damasio (1989) and Damasio and Damasio (1994). This semantic cognition theory aims to bridge the gaps in the theoretical account of the semantic system and claims the aforementioned views are largely reconcilable.

The conceptual topography theory proposes that the semantic system is composed of layers of neurons that each are in charge of extracting increasingly complex features from perceptual input, giving rise at the top of the hierarchy to highly complex concepts. The cognitive system in accordance with the CZ theory, is split into layers each dealing with extracting patterns of activation in neurons on the level below, whose neurons are what constitute a convergence

zone. At the bottom of the hierarchy of CZs within the semantic system there are perceptual feature maps that detect low-level perceptual information (e.g., line orientation, colour, etc.), then on a level above another convergence zone captures the patterns of activation over the features maps of a single modality, and further downstream a super-modal convergence zone encodes patterns over all the modalities. Once so-called conjunctive neurons — neurons that link active neurons in feature maps together, thus acting like the hidden units in an artificial neural network — have captured a specific feature map pattern they are able to recreate it without the bottom-up original input. This outlines the most basic version of the CZ theory, the additions that Simmons and Barsalou (2003) propose in order to adapt the theory to explain the semantic system are: the similarity-in-topography (SIT) principle which states that the "spatial proximity of two neurons in a CZ reflects the similarity of the features they conjoin[, as the] two sets of conjoined features become more similar, the conjunctive neurons that link them lie closer together in the CZ's spatial topography" (Simmons & Barsalou, 2003, p. 457); and the variable dispersion principle which claims that in a CZ "the proximity of the noncontiguous clusters [of conjunctive neurons] for a category reflects the similarity of its instances[, so as] the instances of a category decrease in similarity, its noncontiguous clusters of conjunctive neurons become increasingly dispersed in the CZ's spatial topography" (Simmons & Barsalou, 2003, p. 459).

While this theory is perhaps the most detailed, especially with respect to neuronal connectivity, and the most powerful, it still remains to be seen if the required evidence in support of the connectivity proposed to create conjunctive neurons is found (Simmons & Barsalou, 2003). Although, the required auto-association is possible in neural networks, so this is, in theory, a plausible mechanism, but more research would be required to discover how biological neuronal networks carry this out. Either way it can be modelled easily using an auto-associator network.

It is important to note, that this view of the semantic system does not provide any explanations or predictions that can be tested using patient semantic task data. What it does do is propose a complex organisational structure and a mechanism of how concepts might be instantiated on the neural level.

### 1.3.8 Discussion

With regards to the general assumptions behind each theory, Caramazza (1998) makes a clear distinction between, what he calls, the *categorical* and the *reductionist* explanations of semantic organisation. The former proposes that cognitive faculties are modules dedicated to a distinct task, while the latter is the view that the mind is something akin to a universal Turing machine.

Caramazza (1998) also asserts that the categorical theory "is open to the possibility that there are distinct neural mechanisms for the perception and storage of different semantic categories; [as opposed to the reductionist position that] is committed to interpreting category-specific effects as emergent properties of the (sensory/motor) structure of the members of semantic categories" (p. 267).

There appears to be a false dichotomy between categorical and reductionist views as described by Caramazza (1998). In other words, these two views are not mutually exclusive: a classical computer is made up of parts that are highly purpose-specific — modular at every level of analysis — and yet still possesses Turing completeness. The levels of the purpose-built hardware, purpose-written firmware, and user-facing software are discretely separable, both between and within their individual levels, because the system they exist in has been engineered. In contrast, the cognitive system is not designed in such a way, i.e., with clearly delimited levels of analysis. In fact it is not designed purposefully at all, rather it is shaped by evolutionary and environmental forces that do not care or benefit from clearly defined layers of analysis. However that does not preclude their existence, especially since there is evidence for both specialised parts (e.g., speech production in Broca's area) and generalised parts (e.g., when whole sub-networks of neurons are involved in giving a lecture).

Perhaps the *lumpy-feature space theory*, which proposes that the cognitive system is not differentiated into discrete modal/categorical loci, but as a global set of activations per domain or category (touched on in: Caramazza, 1999; Martin & Chao, 2001; Simmons & Barsalou, 2003) can be seen as on a spectrum between the general reductionist and specialised categorical views. It can also be seen as reconciling the best parts of these two views, although perhaps at the detriment of its own clarity. Unfortunately, lumpy-feature space perspectives on semantic cognition are presently not very well specified. The literature itself is very sparse on the intrinsic details that would form the basis for such a theory and how that would differentiate it from the other accounts presented here[10].

However, the degree to which such mechanisms are pre-set or predetermined by evolution is a contentious issue, as the evidence proposed by the progenitors of the domain-specific theory seems to be open to interpretation. Additionally, it also must be stressed that studies like those by Ishai, Ungerleider, Martin, Schouten, and Haxby (1999) and Koenig et al. (2005), which use an arbitrary or novel category (that hence cannot be argued to have any evolutionary purpose), do indeed document that large-scale cognitive networks, as opposed to domain-specific

[10]In fact in Martin and Chao (2001) the three references provided for further reading yield little more information.

loci, are what give rise to categorisation and other semantic processes. Bearing in mind that the domain-specific theory states that "evolutionary pressures have resulted in specialized (and functionally dissociable) neuronal circuits dedicated to processing, perceptually and conceptually, different categories of objects" (Caramazza & Mahon, 2003, p. 356), we can deduce that the domain-specific view has another pitfall in that there are too many categories for there to be a specific brain area dedicated to each one, so dedicated category-specific regions are impossible (Martin & Chao, 2001). And as previously mentioned, the problem that such evolutionary-based theories face is a lack of evidence for innate cortical regions (Panksepp & Panksepp, 2000, 2001; Panksepp et al., 2002). Further criticism and discussion of the domain-specific view can be found in letters between the proponents of two competing theories (Mahon & Caramazza, 2003; Moss & Tyler, 2003).

That being said, localisation by category, to a certain extent, cannot be ruled out. Given what is known about how neuronal networks function, the brain should be using the features that have been extracted from each percept in order to semantically process a concept. These features should contain enough information to perform classification, association, and identification of the items they describe, while at the same item give rise to compressed stored representations, for example by generalising features over more than one concept, like $\langle eyes \rangle$ for $\ulcorner dog \urcorner$, $\ulcorner human \urcorner$, $\ulcorner fish \urcorner$, and even for $\ulcorner needle \urcorner$. Based on the fact that these features need to be stored somewhere, whether distributed or not, some form of category- or domain-based localisation will occur as features are by definition the only candidate mechanism that can give rise to semantic categories. Even if we assume evolutionary pressure created loci per domain, the selection would still have to be based on sorting based on input. However, these areas are still highly contentious in both the neuroimaging literature and across neuropsychological patient studies, as consensus has yet to be reached as to which areas are more likely responsible for which categories.

On the other hand, the conceptual structure theory is compatible with all other theories: the modality-specific theory, if a particular category were to have more visual than auditory perceptual features this would ultimately affect which areas of the brain are recruited to create that category; the domain-specific theory, if we use a weak version and allow for conceptual structure to drive localisation; the hub theory, as it does encompass feature structure much like that of the conceptual structure theory; the conceptual topography theory, which allows for the distribution of features to drive organisation. This is because the theory deals with the concepts that are represented by the semantic system, and does not make any specific claims about the system itself.

For the conceptual topography theory, the impression is that it is in need of further formalisation and modelling work. The SIT principle was created in-part in order to reconcile the various theories of conceptual deficits (Simmons & Barsalou, 2003). Unfortunately, for conjunctive neurons there is not enough evidence yet, but for CZs there is extensive compelling evidence in many perceptual cortices (see Simmons & Barsalou, 2003, for more details).

Thinking back to the nature of features themselves, it is assumed that the creation of pre-semantic features is carried out by other parts of the temporal lobe, and therefore semantic system, by all of the aforementioned theories. This indicates that perhaps it is unlikely that any single semantic theory as presented can explain everything about semantic memory, unless the creation of features is abstracted away from semantic memory and thus, the operations required to turn sensory information into high-level features is outside the remit of semantic memory by definition. This does not seem like a justified treatment of features, as after all features themselves arguably carry semantic meaning. Moreover, the grounding of features in sensory input is not enough to explain their origin, since some features are clearly linguistic, or least created by some form of top-down process.

## 1.4 Computational models of semantic cognition

### 1.4.1 Overview

In the following section, influential computational approaches to semantic memory shall be described. For clarity, the models will be referred to by one of their defining features: *1*) the model reported in Rogers et al. (2004) will be called the *hub model*, because it theorises a central semantic hub that is recurrently linked to modal brain areas; *2*) Gotts and Plaut (2002) will be the *neuromodulation model*, due to the fact it incorporates a model of neuromodulatory processes that limit refractory-like effects, called synaptic depression; *3*) Tyler et al. (2000) the *conceptual structure model*, as the authors claim that features are correlated with each other in a way that allows for the emergence of semantic categories and domains[11]; and *4*) Farah and McClelland (1991) is henceforth referred to as the *modality-specific model*, as it is one of the first models to put forward an account that describes how categorical structure can be derived

[11] *Domain* is used to refer to the topmost-level distinction seen in patients with category-specific impairments (i.e., significant differences in their accuracy on living and non-living stimuli). Therefore, there appear to be two basic domains: animals and artefacts. While, *category*, on the other hand, normally refers to the sub-groupings within animals and artefacts, such as mammals or vehicles. Although, there are exceptions of categories, such as fruit, gemstones, and musical instruments, that appear to contradict the established mutually exclusive distinction of items into either living or not (Shallice & Cooper, 2011). For the purposes of consistency and clarity, in this document, domain is used to denote the top-level classification of concepts into animals and artefacts, and category is used for all the other (sub)divisions found in semantic space.

from input consisting of "only" perceptual and functional properties.

In general, semantic models suggest at least a tentative similarity between the emergent patterns of activation in their computational units and those found in the neuronal networks of the brain. All appear to reproduce effects qualitatively similar to those seen in patients under semantic testing, albeit in some cases completely disjoint populations of patients, as explained in section 1.2. Models are able to derive, or at least learn, representations of semantic facts, modelled as real-valued vectors in the range of $[0, 1]$ (or $[-1, 1]$ in the case of the *modality-specific model*). Post-lesioning, these models appear to account for the dissociations seen in the patient groups they aim to emulate. In all models, it should ideally be possible to foresee how relevant patients might score on a semantic task they have yet to be tested on. Additionally, if the form and extent of damage to a patient's semantic system is known, provided some form of equivalence mapping between the model's pools of connectionist units and cortical regions exists, it should be possible to predict a set of semantic test scores.

The paradigm used by the consensus of such models to recreate patients' semantic cognition deficiencies is as follows: initially recreate the semantic system as in a healthy individual, and then modify some aspect of the model to reflect the effects of brain damage found in patients. In other words, first the neural network is taught a set of training patterns paralleling the accumulation of knowledge from childhood to maturity[12]. Next, the network is tested to determine whether it has reached the level of semantic skill seen in healthy adults. Subsequently, the model is purposefully damaged using an appropriate analogue to the form and extent of lesioning found in patients. Finally, it can be semantically tested once again, in order to determine the link between changing the properties of low-level network components and the high-level distortion and eventual loss of semantic concepts.

As before in section 1.3, the semantic cognition models are presented in chronological order.

### 1.4.2 Modality-specific model

Consider first the *modality-specific model*, which is a neural network implementation for category-specific deficits. Farah and McClelland (1991) postulate that the categorical structure apparent in these semantic patients is due to a modal, as opposed to a categorical, organisation of concepts, an idea initially proposed by Warrington and McCarthy (1983) and Warrington

[12]Some models, such as the one described in Farah and McClelland (1991), clarify that the process of training is not comparable in any way to the process of semantic maturation seen throughout an individual's life. However, it must be stressed that a model for semantic memory would be obviously far greater in its explanatory powers if it also included developmentally-inspired stages. Taking this into account, some models do indeed show signs of following the learning phases humans do. In other words, some concepts are learned before others, and the relative stability in a model's learnt semantic space is punctuated by short bursts of transitional learning stages, see McClelland and Rogers (2003, p. 314) for an overview of such models.

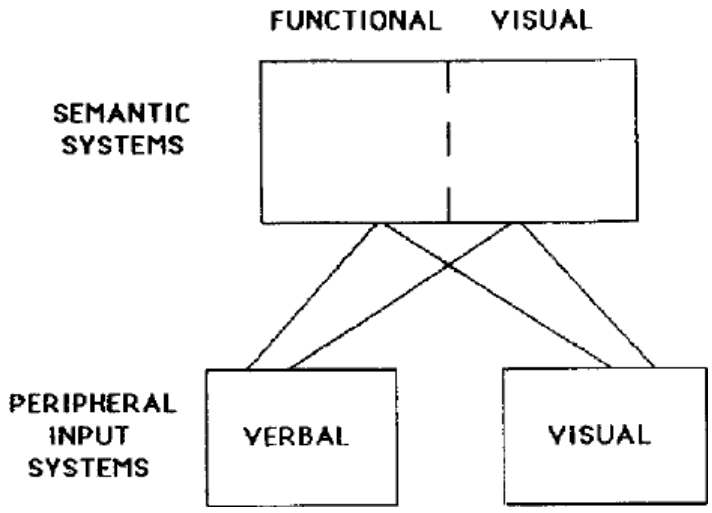


Figure 1.4: Architectural structure of *modality-specific model* (Farah & McClelland, 1991, fig. 1).

and Shallice (1984). Specifically, functional properties dominate the conceptual descriptions of artefacts, while visual perceptual properties define living things. Therefore, this reduces category-specificity to underlying damage to sensory or functional semantic processing. The model consists of a simple recurrently connected neural network, see Figure 1.4, with three pools containing: *1*) 60 semantic visual units and 20 semantic functional; *2*) 24 perceptual visual units; and *3*) 24 perceptual verbal units. Once their network was trained it was tested on two semantic task analogues: confrontation naming, and word-to-picture matching (see subsection 1.2.2); combined with two forms of damage: disconnecting units by zeroing the weights on their connections, and removing the units themselves by setting them to a fixed activation value of 0. The latter form of lesioning is applied only to units within the semantic pool and is seen to give rise to category-specific deficits of living things. The former damage is restricted to just affecting the connections between the perceptual verbal units and the visual semantic units and creates the pattern of category- and modality-specificity seen in a single patient, described by McCarthy and Warrington (1988).

The architecture used in the *modality-specific model* reflects the theoretical position that visual and functional properties are part of, or indeed may constitute in and of themselves, the core of semantic knowledge. This creates a clear distinction between visual semantics and (high-level) visual perception. This is to say, the authors see the semantic system, on some level, as composed of regions dedicated to a sensory-motor modality. In conclusion, this model shows that the semantic system does not need categorical structure in low-level components (such as in features, or by creating category-specific brain regions), to give rise to functionally dissociable category-specific effects on higher levels (such as in semantic testing, or

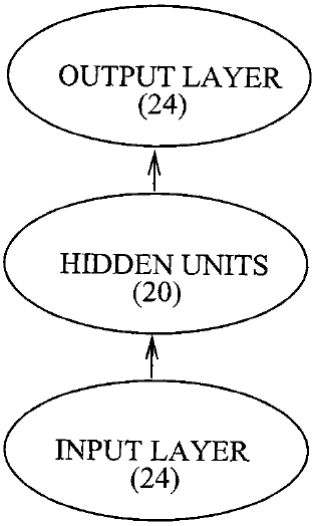


Figure 1.5: Architectural structure of *conceptual structure model* (Tyler et al., 2000, fig. 2).

introspection). The *modality-specific model*'s success, however, is conditional on the existence of correlations between: the somato-sensory pathways that encode a percept; and the category to which the concept corresponding to that percept belongs.

### 1.4.3 Conceptual structure model

Next is the *conceptual structure model*. This model also aims to provide a theoretical account of category-specific deficits. Tyler et al. (2000) postulate that the distribution of distinctive and shared features varies across semantic domains and thus, supports the creation of a categorical structure. On one hand, animals are relatively homogeneous, but are nonetheless distinguishable by a healthy semantic system because it can access the properties unique to each individual (e.g., telling apart a zebra from a horse by accessing the property "has stripes"). On the other hand, inanimate objects are differentiated by their functional properties, which are not shared over many artefacts (e.g., a hammer is completely functionally distinct from a spanner, despite the fact they are both tools). The *conceptual structure model* learned a training set that reflects the distinctive features within patterns, and the correlations occurring both within and between patterns. To implement their model Tyler et al. (2000) used a simple feedforward network, see Figure 1.5, with three layers composed of : *1*) 24 input and *2*) 24 output units in charge of in/output of verbal and functional features; and *3*) 20 hidden units. This modelling architecture demonstrates qualitatively the same category-specific losses as in patients, after global applied damage in the form of removal of a random subset of connections between units.

The *conceptual structure model*, as the *modality-specific model* before it, explicitly uses correlations and other relationships over (sets of) features to model (pre-)semantic space, and

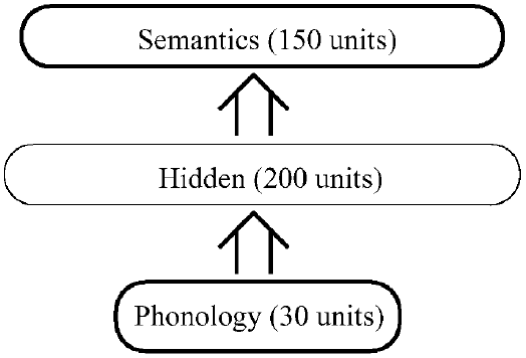


Figure 1.6: Architectural structure of *neuromodulation model* (Gotts & Plaut, 2002, fig. 1).

provides evidence for the possibility that semantic memory is not in itself category-specific. The key to this model's success relies on using the ability of neural networks to implicitly analyse input data by storing correlated features more robustly than exceptions. In other words co-occurring features (such as those found in living things) are likely to reinforce each other and thus "protect" each other from lesioning damage: correlation implies preservation; with the opposite being true for distinctive and rare features. So post-damage such a system decays in a way that means different domains are affected to different extents. For example, the many shared features within living things are preserved and appear to generalise with damage. Thus distinctive features like stripes are lost and all four-legged animals are treated equally. In contrast, the representations of non-living things share fewer features and so decay more easily after a certain level of lesioning; before that level, due to artefacts' richness in correlations between functional and perceptual properties, they are more robust as single concepts. Thus, showing both sides of the double dissociation involved in category-specificity. This model's aptitude at modelling patient performance is a result of the interaction between the input data and the neural network.

### 1.4.4 Neuromodulation model

The third model to be discussed is the *neuromodulation model*, which supports the Gotts and Plaut (2002) view that the various access, and damaged store disorders lie on a spectrum: ranging from displaying exclusively refractory-like deficits, to combinations of the two, to solely losses within the semantic store. Additionally, Gotts and Plaut propose that damage to the neuromodulatory system, whose purpose is to suppress the automatic neural refractoriness caused by synaptic depression and to enhance neuronal activation signals, is sufficient to cause access effects during semantic testing. For this reason, the authors combine neurobiologically-derived equations for synaptic depression and for neuromodulation (adapted from Varela, Song, Turri-

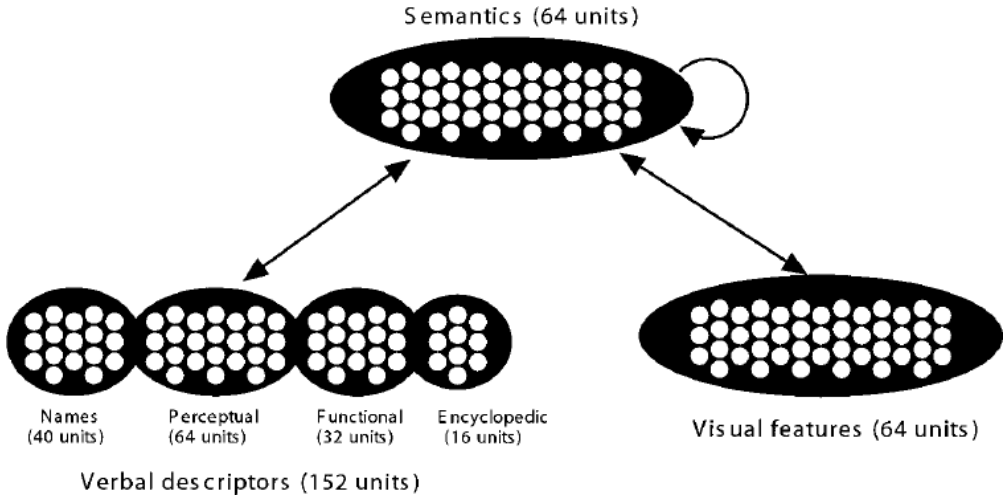


Figure 1.7: Architectural structure of *hub model* (Rogers et al., 2004, fig. 1).

giano, & Nelson, 1999) with a purely feedforward architecture. The network's three layers, see Figure 1.6 are composed of: *1*) 30 phonological input units; *2*) 200 hidden units, corresponding to undefined intermediate brain regions; and *3*) 150 output units for semantics. The two forms of possible lesioning, neuromodulatory damage and disconnecting the hidden and semantic units, give rise to access/refractory and degraded store patterns of deficits respectively, as seen when the model performs the word-to-picture matching task.

When neuromodulation breaks down whole sub-networks of units enter a state of mutual inhibition if probed repeatedly on semantically similar concepts. Thus, the model displays the characteristics of access/refractory patients, since the groups of units that have been affected require a recovery period and the ones that have not are still able to function adequately. Degraded store results are dependant on the network being trained in a way that takes into account the frequency of items encountered in the environment, so when the model is lesioned concepts are disproportionally affected. This means that familiar concepts are preserved but categorical outliers are lost. Combining the two forms of damage intermediate patterns of impairments can be modelled, thus capturing some exceptional patient cases. Thus, the *neuromodulation model*'s success is dependant on the variable frequency of training patterns and how this fact is encoded in the network (i.e., by assigning intersecting subsets of units for neighbouring concepts), and on the neurotransmitter-detector add-on each unit is equipped with.

### 1.4.5 Hub-and-spoke model

The fourth model to be examined is the *hub model* of Rogers et al. (2004) (see also Lambon Ralph et al., 2007). This model's creators postulate that the interactions of *attractors*,

which are used to represent amodal concepts within semantic space, can account for both healthy and deficient semantic cognition. Attractors are specific network configurations (vectors of the states of hidden units) that emerge when recurrent connectivity exists, and when sufficient time to attain equilibrium is provided. These stable network states exercise attractive power over a set of neighbouring network states, collectively known as their basin, which means that if the network is in any of these "nearby" states it will ultimately arrive at the attractor itself. These properties, according to Rogers et al. (2004), are also found in semantic memory. The model consists of a neural network with three bidirectionally connected pools, see Figure 1.7, containing: *1*) 40 phonological name in/output units; *2*) 62 visual in/output units; *3*) 61 perceptual, 32 functional, and 18 encyclopedic units, which together constitute verbal in/output; and *4*) 64 semantic hidden units; and they train it using a variant of backpropagation though time[13]. The *hub model*, due to its large number of in/output modality analogues, is able to emulate all the semantic tasks mentioned in section 1.2.2. Lesioning is simulated by globally removing random connections, which induces SD-like deficits.

As has been touched upon, Rogers et al. (2004) parallel the emergence of attractors with the learning of concepts, and they put forward that such knowledge is amodal: the somatosensory input from the various modality-specific pathways is processed by the hidden units, which thus form semantic representations that are amodal. This is how they believe semantic memory encodes experiences over time. For the case of the category-specific deficits seen in some of their SD patients, they appeal to the attractor basins' properties post-lesioning. They show that animals are a tight cluster of similar concepts, thus consisting of many neighbouring attractors; while inanimate objects are distal (to the average central point of their domain) and, thus, they from distinct conceptual points in semantic space, and therefore their attractors are further apart. When connections are removed the attractor basins for living creatures decay to form a larger super-attractor, which has a combined attractive power; meaning categorisation of input as an animal is possible, but access to individual features might be lost. However, non-living things' basins do not merge; instead they maintain their individual attractors, albeit with slightly distorted basins, allowing them to perform slightly better in this domain. The *hub model* functions in such a way that percepts form the substrate for which amodal concepts are built upon; nonetheless, its achievements are conditional on the training set and algorithm allowing the attractors to form in the necessary way.

[13]Fig. 1, p. 207 in Rogers et al. (2004) reports contradictory number of units per network layer to fig. 3, p. 212, it is assumed the latter is correct.

### 1.4.6 Discussion

The above four models share some basic assumptions about the semantic system. Arguably, the most important aspect of their modelling strategies and a point they all agree upon is the general distribution of pre-semantic perceptual (and functional) features: animals and plants are closely perceptually related to each other, due to the fact they have evolved from a common ancestor and thus are composed of slightly different body parts; tools, vehicles, and other inanimate objects are not similar to each other, they have been created by humans to solve different problems, so by definition inanimate objects are distinct from both living things and from each other. Without the training sets that encode patterns in this specific way, all of the models would be incapable of producing a good fit to patient data. Thus, it appears as if the features, whose extraction from the environment itself is not modelled, play a pivotal role in giving rise to the semantic system's structure. This is to say that, perceptual input to the semantic system drives its organisation and dictates the way memories will decay.

Despite a superficial consensus, each of the four models appears to address the characteristics of semantic memory using different explanatory mechanisms. This implies that the models may not necessarily be mutually compatible. For example, the qualitative trend in the *conceptual structure model* showing a transition from performing better with artefacts to performing better with animals after a certain amount of damage does not seem to apply to the *hub model*, which is arguably a more sophisticated model; this means that the former model can account for the double dissociation seen in category-specificity, while the latter model can only account for one side of the coin. Such problems might be resolved, allowing for a collective explanatory power, by analysing the properties of each of the models, thus giving rise to a unified account. From this fact, a key research question is derived: To what extent are these models theoretically compatible and if they are, can they somehow be merged within a single model?

As the *modality-specific model* and the *conceptual structure model* appear to address the same problem from a slightly different angle – category-specific deficits – they may be seen as compatible. They may even be seen as being the same basic idea that has been extended in the latter model to capture both category-specific patterns of deficits. Due to this, a closer look at the latter model would be possibly more useful in regards to taking away any general principles, as they are both relatively simple models and the *conceptual structure model* was trained 300 times. So the findings in Tyler et al. (2000) appear to be particularly robust, in principle. The *modality-specific model* makes the assumption that features (or possibly relations over features, that constitute some kind of meta-features) are part of semantics, which the *conceptual structure*

*model* is less clear on.

The *neuromodulation model* is the only model which addresses the patterns of deficiencies seen in access/refractory and degraded store patients. As such, this model is unique and thus, does not lend itself well to direct comparison with the other three. However, its method for simulating familiarity with certain commonly occurring concepts is compatible with any neural network training algorithm, and its mechanism for capturing the access dissociation is also theoretically generalisable to any kind of neural network unit. Hence its mechanism for neuromdoulation can be seen as a possible add-on to any of the other architectures.

The most complex model, the *hub model*, can emulate a variety of semantic tasks, apparently giving a close fit to patient data, and a qualitative account of how the attractor-space is shaped by the pre-semantic feature sets. It has been used to account specifically for both the general deficits of SD (Rogers et al., 2004) and the category-specific deficits of HSVE (Lambon Ralph et al., 2007) patients, though it uses an analogue for HSVE lesioning – adding Guassian noise to connection strengths – that appears to be retrofitted onto the model. This puts the *hub model* in a position of direct contradiction with the results obtained from the *conceptual structure model* (or the *modality-specific model*) as the former needs Gaussian noise on connections to model category-specificity, while the latter does not. Additionally, it is theoretically distinct from the other models as it makes a strong case for semantic memory being completely amodal; implying all features[14] are external to semantics. This means that modality- and category-specific impairments would be closer to a disconnection syndrome, which Warrington (1975) is strongly against. It, thus, raises questions such as: Is there really an amodal semantic concept for a song? Arguably, if a painting is a concept (as used by Warrington, 1975, p. 650), then so is a song. Are both their semantic instantiations amodal? Does it makes sense for such an amodal area to exist if most brain regions appear to be modality-oriented? It would perhaps be sensible to allow for both amodal and modal semantics at such early stages of exploration.

A general concern for all neural network models is that even though the results obtained are highly dependant on the training set used, it is still a matter of debate what the constituent functional properties are and where they come from. Moreover, no model explicitly takes into account any form of top-down signal. For example, the Sapir-Whorf hypothesis, or linguistic relativity, states that language up to a certain degree is able to shape the creation of distinct concepts. Such linguistic labels applied to concepts can, arguably, be construed as what the *hub model* does with its so-called name units, but Rogers et al. (2004) do not consider their effect

[14]The word is ambiguous, as features could be *semantic features*, which are amodal and consist of (cor)relations over and above perceptual features, or they could be modality- derived features, which are high-level perceptual properties.

on the model. In conclusion, while all models for semantic cognition described here attempt to solve very similar, if not identical, modelling problems, they constitute somewhat incompatible accounts. It appears as if a unified framework, explaining semantics on a more general but also more consistent level, is needed.

## 1.5 Summary

In this chapter, the neuropsychological investigation of semantic memory, and semantic cognition more generally, has been described. Patients with frontotemporal lobar degeneration, and specifically the semantic dementia subtype, are a good source of information on the relationship between functional and structural damage to semantic cognition. The performance on semantic tasks from patient groups with damage to their temporal lobes, such as stroke, Alzheimer's, and herpes simplex virus encephalitis patients, also provide valuable details for the understanding of semantic cognition. This is because these patients sometimes display deficits that only affect a very specific subset of semantic concepts, e.g., a selective loss for animals over artefacts, or for verbal over visual features.

A subset of the theories that attempt to account for these phenomena have been presented. These theories provide a broader explanation for the function and structure of the semantic system that incorporates patient data and dissociations within concepts. Unfortunately, these theories for semantic cognition fail to capture the spectrum of disorders. This is partially because even within a specific subtype of patient their performance on semantic tasks are not homogeneous – due to the nature of neurodegeneration.

Finally, some of the computational modelling work that addresses the aforementioned theories has been discussed. The models presented here are implemented using neural networks and use the input vectors of the networks to encapsulate semantic features and some, but not all, models use the set of all features to represent concepts. (The hub model does not do this. It represents concepts amodally as the abstract vector the set of hidden units compute.) The models are able to emulate patient testing. Specifically, they reproduce the pattern of decay of conceptual knowledge when tested using semantic tasks. As such, they allow the theories they represent to be executed, which means that the implications of the theories themselves can be compared and contrasted. These differences are important both theoretically and with respect to the implementations of these models as shall be seen in the next few chapters.

# Chapter 2

# Four models of semantic cognition

## 2.1 Overview

In the following chapter, two previously described models — the Rogers et al. (2004) hub model, and the Tyler et al. (2000) conceptual structure model — are reimplemented both faithfully and using alternative implementation and modelling decisions. Also presented are: a model that combines the (pre-)semantic features of the hub model with the conceptual topography theory by Simmons and Barsalou (2003); and a model that combines the modelling paradigm used in DISLEX (Miikkulainen, 1997) and the modality-specific theory with the (pre-)semantic feature structure of the hub model. The fundamental elements of these four different families of implementations will be described in this chapter as they will be used to model semantic deficits, as well as healthy semantic cognition, in the chapters to come.

## 2.2 Implementations, models, and theories

The four types of implementations of models of the semantic system will be discussed, while bearing in mind that theories, models, and implementations are distinct entities. The first two should be clearly different: a *theory*, as defined by the Collins concise dictionary (second edition, 1990) is "a system of rules, procedures, and assumptions used to produce a result"; whereas a *model* is "a [computational] representation, usually on a smaller scale, of a [cognitive] structure". On the other hand, an *implementation* is a specific instantiation of a model (usually involving programming code, or at least an algorithmic or mathematical commitment). However an implementation also contains details (e.g., specific time/space complexity of the code itself) that are neither part of the model proper nor part of the overarching theory (Cooper, Fox,

Farringdon, & Shallice, 1996).

At each level of analysing an implementation it should be possible to differentiate details that pertain only to the implementation, details that are model-dependant, and details that constitute the core of the theory components. If an implementation detail (e.g., a very specific learning rate range) is found to be central to the model working, it should be elevated to the model-level. Likewise a model-level property (e.g., attractors within a neural network framework, see section 2.4), if found to be integral, should be promoted to the theory level. If explaining the theory requires explaining specific model- or implementation-level details, then these properties are part and parcel of the theoretical account, as well as the models and implementations that may be derived from it. This implies that the level of analysis that a detail belongs to is highly context dependent, and not something intrinsic to the detail itself. In other words, it is very difficult to untangle the features of an implementation, model, or theory in the abstract; but given a specific implementation it should be possible.

## 2.3 Conceptual structure model

### 2.3.1 Theory details

The theory described in Tyler et al. (2000) is a version of the conceptual structure theory, which supports the notion of a unitary, i.e., non-modality-specific, semantic system. The conceptual structure theory (previously discussed in subsection 1.3.5), proposes that the contents of concepts define the structure of the categories. In other words, features constitute the building blocks of semantic memory and these correlate with each other to form the hierarchy of concepts, categories and domains.

Nevertheless, features are split into perceptual and functional, in line with a common approach to semantic memory modelling, inherited from Warrington and McCarthy (1987) — even though, as mentioned previously, Tyler et al. (2000) do not agree with the theoretical position of Warrington and McCarthy (1994). Functional features are described as giving an "object a meaning in a dynamic environment characterized by cause and effect [and they are] more resilient to brain damage than other kinds of information" (Tyler et al., 2000, p. 201). Each artefact has a single functional property which correlated with very specific perceptual features (e.g., the concept $\ulcorner fork \urcorner$ contains the feature $\langle has\ prongs \rangle$, of which possession is visually discerned but also dictates function and is distinctive). Animals also have functional properties, dubbed their biological function, which are determined by their "mode of interaction with the environment" (Tyler et al., 2000, p. 202) (e.g., eyes are for seeing). Tyler et al.

| Artefacts | | | | | |
|---|---|---|---|---|---|
| Distinctive perceptual | | Shared perceptual | | Functional | |
| 1 0 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 0 0 | 0 0 0 0 |
| 0 1 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 0 1 0 0 | 0 0 0 0 |
| 0 0 1 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 0 0 1 0 | 0 0 0 0 |
| 0 0 0 1 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 0 0 0 1 | 0 0 0 0 |
| 0 0 0 1 | 0 0 0 0 | 0 1 0 1 | 0 0 0 0 | 0 0 0 1 | 0 0 0 0 |
| 0 0 1 0 | 0 0 0 0 | 0 1 0 1 | 0 0 0 0 | 0 0 1 0 | 0 0 0 0 |
| 0 1 0 0 | 0 0 0 0 | 0 1 0 1 | 0 0 0 0 | 0 1 0 0 | 0 0 0 0 |
| 1 0 0 0 | 0 0 0 0 | 0 1 0 1 | 0 0 0 0 | 1 0 0 0 | 0 0 0 0 |
| **Animals** | | | | | |
| Distinctive perceptual | | Shared perceptual | | Functional | |
| 0 0 0 0 | 1 0 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 1 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 0 1 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 0 0 1 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 0 0 1 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 0 1 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 0 1 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |
| 0 0 0 0 | 1 0 0 0 | 0 0 0 0 | 1 0 1 0 | 0 0 0 0 | 1 0 1 0 |

Table 2.1: Table showing the 16 patterns used to train the models (Tyler et al., 2000, fig. 1). Each line corresponds to a concept with 24 binary features.

(2000) claim that cooccuring functional features tend to become highly correlated with, and thus interdependent on, perceptual features and vice versa. An example of a simple set of concepts based on these ideas can be seen in Table 2.1.

### 2.3.2 Model details

Based on their theoretical position, Tyler et al. (2000) set out to show that a unitary system, without localised feature stores and without an imbalanced distribution of different types of of features per domain, can nonetheless give rise to category-specific patterns of deficits. To accomplish that, they create a feedforward network with a high-level architecture (see Figure 1.5). The model has three layers composed of input and output units that represent verbal and functional features. Patterns are applied to the input layer, which propagates activation to a layer of hidden units, which in turn activates the output layer in order to function as an autoassociator. However as the design of the network is such that it autoassociates whole patterns only, the only semantic task that it can approximate is a simplified form of the word-to-picture matching task, although the parallels with the word-to-picture task are purely qualitative, as there is no word or picture representation, just a set of features.

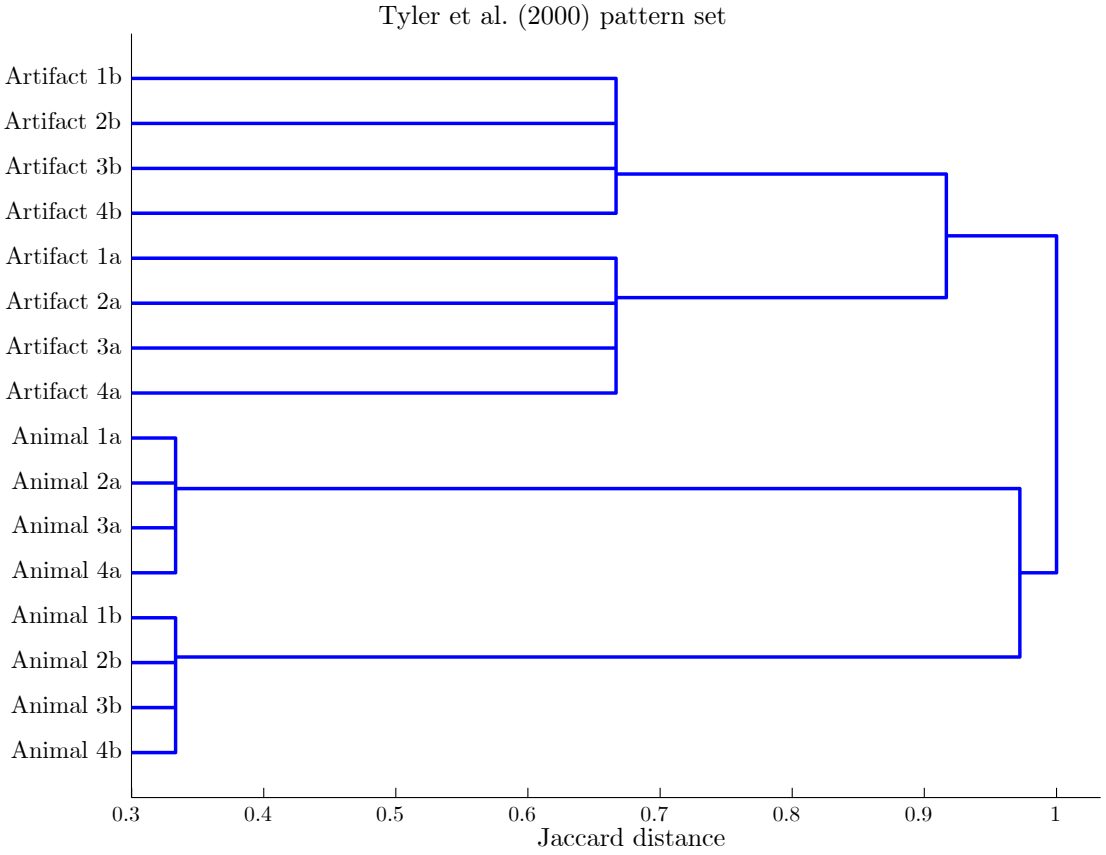


Figure 2.9: A hierarchical cluster dendrogram depicting the original Tyler et al. (2000) patterns' relationship to each other using Jaccard distance.

Domain and category structure in the model is proposed to emerge through the correlations, or lack thereof, over the features. Specifically, the design of the pattern set (see Table 2.1) is such that living things have more shared and more highly correlated features while artefacts have fewer correlated and more distinctive (unique) features. These statistical properties are uncovered by asking participants to list properties for concepts, and then using the statistical structure of these properties to determine the within-category, across-category/within-domain, and across-domain distances for the living and non-living patterns (see: McRae, Seidenberg, & de Sa, 1997).

### 2.3.3 (Re)implementation details

Two (re)implementations of the conceptual structure model were developed. The first was a direct reimplementation of that described by Tyler et al. (2000) (i.e., a three-layer feed-forward network), while the second augmented that implementation with recurrent connections within the hidden layer (i.e., a simple recurrent network). Following Tyler et al., both networks

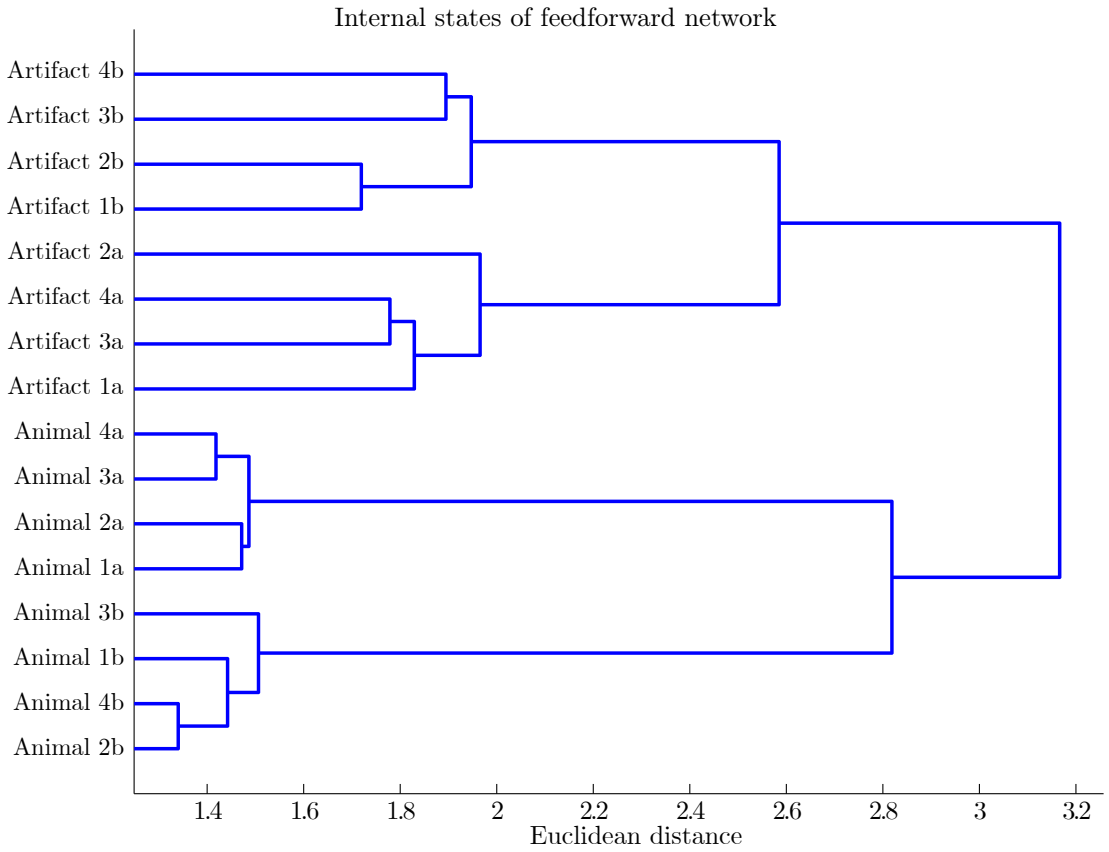


Figure 2.10: A hierarchical cluster dendrogram depicting the hidden unit states of the feedforward implementation.

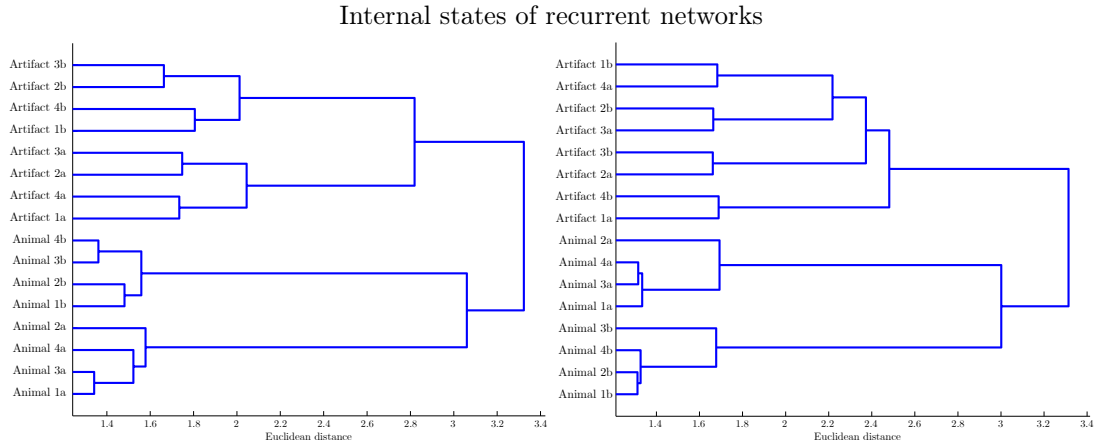


Figure 2.11: Two hierarchical cluster dendrograms showing the individual differences between recurrent implementations, differing only in the values of their initial connection weights.

consisted of 24 input and output units and 20 hidden units, and both networks were trained as auto-associators with the patterns in Table 2.1. In both cases, weights were initialised to uniformly distributed values within the range $[-0.001, 0.001]$. The feed-forward network was trained using standard back-propagation of error (see, e.g., Rumelhart, Hinton, & Williams, 1986), both with a learning rate of 0.25 and a momentum of 0.9 to conform to Tyler et al., and with a learning rate of 0.05 and no momentum[1]. In order to calculate error, in both networks, cross entropy was used (see, e.g., Plunkett & Elman, 1997). For both networks, weights were updated after every epoch.

Training was stopped when the sum squared error per pattern was no larger than 0.01, which as in the description of Tyler et al. occurred within $1,000$ training epochs. The simple recurrent network was trained using back-propagation through time (as described in appendix A), with the network settling in 28 cycles, no momentum and a learning rate of 0.05. The same criterion was used to terminate training (i.e., all units were within 0.01 sum squared error of their targets), though due to the recurrence this took substantially longer (approximately $15,000$ epochs).

The feed-forward network was an exact replication of the network described by Tyler et al. (2000), while the simple recurrent network was developed to explore the potential role of attractors within the conceptual structure model, and in particular whether the behaviour of the implementation was affected by the existence of attractors. Moreover, the recurrent network implementation may be directly compared with the reimplementation of the hub-and-spoke model to be discssed in section 2.4. The recurrent conceptual structure implementation can be interpreted as a simple version of the hub model, because it conforms to the appropriate topology, but instead it is trained on the much simpler patterns developed by Tyler et al. (2000).

As in Tyler et al. (2000) the networks learned to associate the input pattern with the same output pattern. The hidden units (see Figure 1.5) do not represent concepts themselves, and so are outside the scope of the conceptual structure theory — they are merely part of the implementation: a feedforward or recurrent network. However, the states that were derived by the network in its hidden layers can be examined. This allows for an overview of the structure of the categories and domains within the network, which is part of the conceptual structure theory in a qualitative way, since one of the central claims is that the internal feature structure of the patterns drive the categorical structure of the semantic system. In Figure 2.9, the Jaccard

[1]This is in order to evaluate whether the performance of the model is a function of the initialisation and training parameters or of the error function, see appendix C.

distances[2] between patterns and in Figure 2.10, the Euclidean distances[3] between the hidden unit states, can be seen. These two figures give an indication of how the feedforward model represents the patterns in its hidden layer, namely as a function of the structure of the pattern. Conversely, Figure 2.11, shows that this internal representation of patterns in the recurrent model is a little more complex. Comparing the left with the right subfigures in Figure 2.11 it is clear that the structure of the patters does not determine the internal structure of all recurrent networks. The left network is able to create attractors that have a clear parallel to those in Figure 2.9, while the network on the right has clustered category $a$ and $b$ of each domain in an unorthodox way. Such fine details of the internal representation of the networks cannot be determined by looking merely at the output error, as all networks presented here scored at ceiling.

The two implementations are able to also perform a sort of naïve category or concept fluency task, wherein the specific required subpattern is given as input, and the network provides whatever concepts it has learned that fit that description. For example 1010 0000 can be applied to the shared perceptual input (at input positions 9-16, see Table 2.1) and noise applied to the remaining units. Once the network is settled the output will be of the patterns that contains the shared perceptual features 1010 0000.

This model, implemented in two different ways, will be further used to model semantic deficits in the following chapters. However this theory cannot explain or predict general semantic deficits, such as those seen in semantic dementia. This is a very important criticism of Tyler et al. (2000), and the conceptual structure account of semantic memory in general. In the next section, three implementations of the hub-and-spoke model, that purportedly accounts for both global semantic deficits and category-specific patterns of dissociation, will be described.

## 2.4 Hub-and-spoke model

### 2.4.1 Theory details

According to the hub theory, modality-specific perceptual inputs (e.g., visual, aural, motor, somatosensory, etc.) are reciprocally connected to a central amodal hub, as shown in Figure 1.2. The information passed between the hub and its spokes allows for retrieval of semantic associations (e.g., thinking of $\ulcorner dog \urcorner$ based on hearing a bark), identification (e.g., calling a picture of

[2] Jaccard distance is a metric that respects the fact that absent features should not contribute to the similarity of concepts, e.g., neither $\ulcorner dog \urcorner$ nor $\ulcorner screwdriver \urcorner$ come with the feature $\langle has\ wheels \rangle$ set to 1, but that does not mean they are close to each other in semantic space.

[3] The Euclidean metric is more appropriate here as the hidden units take on real values; they do represent features.

a ⌜*dog*⌝ "dog"), categorisation (e.g., classifying a poodle as ⌜*dog*⌝, ⌜*mammal*⌝, and ⌜*animal*⌝), and generation (e.g., describing, drawing, or imitating a ⌜*dog*⌝). Damage to the connectivity within the amodal semantic hub, and between the hub and the modal spokes, is proposed to give rise to analogues of the deficits seen in patients.

The mechanism that underpins concepts — both in the hub theory and in the connectionist implementation — is the emergence of attractor states. Such states arise in dynamical systems that have recurrently connected components. Given partial input the system state gravitates towards the centre of a basin of attraction, thus recreating the full multi-modal experience of the concept. The hub theory proposes that, as a result of lesioning connections, neighbouring attractor basins coalesce, creating larger more generalised concepts. Attractors that are proximal in semantic space merge to represent a more general concept, e.g., ⌜*horse*⌝, ⌜*pony*⌝, ⌜*donkey*⌝, ⌜*zebra*⌝, etc., merge into ⌜*horse*⌝, or even more generally into ⌜*animal*⌝.

In addition to theoretical commitments to attractors, the proponents of the hub-and-spoke theory also make some predictions regarding the cortical localisation of the semantic system. Rogers et al. (2004) propose that the hub can be located in the anterior temporal pole, and that connections to it from surrounding cortical regions, along with its contents, give it its status. The claim that the hub is found in a specific part of the temporal lobe is a useful one because it allows the gathering of evidence for or against its existence. There has been some recent work using lesion and neuroimaging studies aimed at determining if the required connectivity and activations are present (e.g., Binder & Desai, 2011; Campo et al., 2013; Hoffman, Jones, & Ralph, 2012; Hoffman & Lambon Ralph, 2011; Jefferies, 2013; Lambon Ralph, 2014; Lambon Ralph, Sage, Jones, & Mayberry, 2010; Lambon Ralph, Pobric, & Jefferies, 2009; Pascual et al., 2013; Patterson, 2007; Pobric, Jefferies, & Lambon Ralph, 2010; Skipper, Ross, & Olson, 2011; Tranel, 2009; Tsapkini, Frangakis, & Hillis, 2011). If found, this would add further support to the hub theory.

### 2.4.2 Model details

The hub model consists of units with real-valued time-varying activations, which are divided into 215 visible and 64 hidden, shown in Figure 2.12. The visible units are divided into three in/output pools consisting of: 40 name units, 64 visual feature units, and 111 verbal (61 perceptual, 32 functional, and 18 encyclopaedic) descriptor units. The hidden layer of the network comprises fully connected hidden units that receive activations from each other and the input modalities; these units learn to encode semantic representations by abstracting over their perceptual input.

Name units represent natural language labels (e.g., "car"), visual units code for visual perceptual features (e.g., ⟨*is blue*⟩), and verbal units assume the role of general verbal properties that are perceptual (e.g., ⟨*makes noise*⟩), functional (e.g., ⟨*can cut*⟩), and encyclopaedic (e.g., ⟨*is living*⟩), see Figure 2.13. In order to avoid information (categorical or otherwise) being encoded within a name sub-pattern, names are defined orthogonally. Thus name sub-patterns are a set of binary units, of which only one may be active per pattern. Rogers et al. (2004) argue that this labelling strategy parallels natural language in as much as, e.g., the word "robin" does not in itself carry any information about the bird it refers to; the mapping from "robin" to, for example, an image of the animal in question is purely arbitrary. In contrast, the visual and verbal sub-patterns represent perceptual and linguistic information, and therefore must conform to predefined prototypes. The structure of these latter two sub-patterns results from the high-level processing performed at the final stages of their respective sensory pathways (Rogers et al., 2004). Visual properties and verbal descriptors represent statements like ⟨*has a red breast*⟩, ⟨*can fly*⟩, and, according to the pattern details in Rogers et al., statements such as ⟨*is a bird*⟩ and ⟨*is living*⟩ are also included. This explicit classification of an item, i.e.,⟨*is a vehicle*⟩ is superfluous, as the classification of objects into groups, along with the creation of the categories themselves, can be accomplished based on the directly observable features of the objects (Small, Hart, Nguyen, & Gordon, 1995), although such redundant encodings may nonetheless exist in the semantic system.

In addition to the modal-based division, patterns are differentiated into two domains: those representing living creatures and inanimate objects. These are further separated into: mammals and birds; and vehicles, tools, household objects, and fruit. In spite of this distinction, the category of fruit does not conform to the living/non-living classification as clearly as the rest of the groups do; this may be due to the fact that fruit is a natural kind, i.e., not man-made, and it is considered an inanimate object. Each category is related to a probabilistic prototype used for constructing visual and verbal sub-patterns that belong therein. Names are assigned to patterns regardless of their categorical classification, since they function merely as labels. Rogers et al. (2004) claim that the various types and levels of severity of semantic deficits arise due to these intrinsic distinctions in the binary encodings for each object, in conjunction with the nature of the damage imposed on connections (Lambon Ralph et al., 2007).

To produce an output response the network must first receive and semantically process a representation of an object in the training set, thus effectively performing pattern completion. This latter computation involves cycling through semantic states until a stable, settled, state is reached; meaning the hidden units do not change their individual activations given more time.

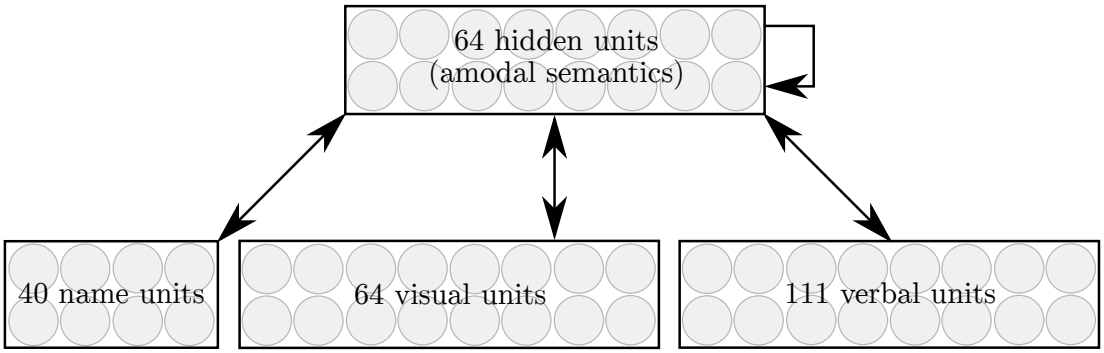


Figure 2.12: The hub model's recurrent network topology, based on figures 1 and 3 in Rogers et al. (2004).

In other words, settling is the process whereby the network is run until it attains some form of equilibrium; thus it can be considered as having decided on the meaning of its input (i.e., if it relates to output).[4] Once the trained network has settled, the network's state conforms to the real-valued pattern of an implicitly learned attractor basin – an internal configuration that is reachable due to the recurrent connectivity of the hidden units. This in turn activates the output units, thereby completing the input pattern.

### 2.4.3 (Re)implementation details

Our (re)implementations, following Rogers et al. (2004), were real-valued recurrent networks consisting of three pools of input/output units: 40 name units, 64 visual units, and 111 verbal units (further subdivided into 61 perceptual, 32 functional, and 18 encyclopaedic units). These units were bidirectionally connected to 64 fully recurrent hidden units, as shown in Figure 2.12. Activation spread from one or more spokes to the hub and from the hub back to every input/output pool.

Rogers et al. (2004) provide a probabilistic template for generating appropriate training sets (reproduced in Figure A.86). We used this template to create a pattern set equivalent in structure to the original of Rogers et al. (see section A.2), as evidenced by comparing hierarchical cluster dendrograms, for training and testing our reimplementations. According to the template, mutually exclusive subsets of visual and verbal features underpin the main distinction between man-made and inanimate objects, as shown in Figure 2.13. Other structural properties are: that the two domains are subdivided into 6 categories (mammals, birds, fruit, vehicles, household objects and tools); that verbal sub-patterns include a single feature present to denote category and domain membership; and that names consist of a single uniquely activated unit, thus

[4] Alternatively, it may oscillate ad infinitum and hence not settle or produce stable output.

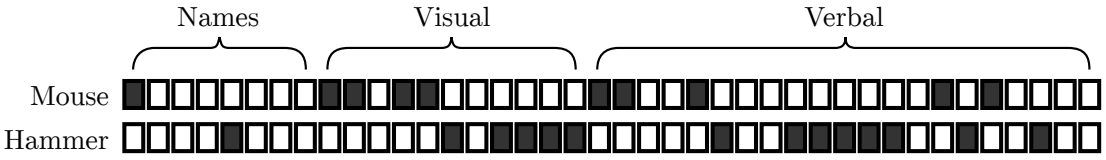


Figure 2.13: Two examples of simplified patterns. Solid rectangles represent activated features in the visual and verbal sub-patterns (e.g., ⟨*has fur*⟩), while empty ones represent features that are not present.

creating 40 orthogonal name bit vectors. Some names are shared between certain visual/verbal sub-patterns in order to create category-level names, thus giving rise to archetypal patterns (e.g., calling a ⌜*bird*⌝ "bird" as opposed to "sparrow" or "robin").

The elements of the training set were binary vectors each with 215 bits. Each vector had the following bits set: *a*) the individual visual or verbal features it possesses (e.g., ⟨*is red*⟩, ⟨*has legs*⟩); *b*) the localist orthogonal bit vector that constitutes the name sub-pattern (e.g.,a representation of "robin"); and *c*) the localist category and domain membership units within the verbal sub-pattern (e.g., ⟨*is mammal*⟩, ⟨*is tool*⟩. In Figure 2.13 these are represented by the last 7 units. In Figure 2.14 the Jaccard distance between the patters we created can be seen, and in Figure 2.15 are the distances between the patterns used in the original Rogers et al. (2004) model sent to us by T. Rogers (personal communication, August 21, 2012). The reimplementation was We trained on both of these pattern sets.

The first network was trained using epochwise back propagation through time (BPTT: R. Williams & Zipser, 1989, 1995), following the procedure of Rogers et al. (2004) where specified. BPTT is a variant of back propagation that involves *unrolling* a multi-layered feed-forward version of the recurrent network and training the weights using back propagation and time-averaging[5] within this new unrolled network. When the learning phase is completed the network is reverted back to its normal recurrent state. Following Rogers et al., the network was settled for 28 steps during training. As in Rogers et al., the input units were clamped (i.e., forced to take on their target values) for twelve of these steps. We refer to this method of training as $\text{BPTT}_1$.

An alternative method, which we refer to as $\text{BPTT}_2$, is to clamp the targets to the outputs for the full 28 settling steps, and not use time-averaging, every other aspect of this training procedure is identical to $\text{BPTT}_1$ (see appendix A). This reduces the noise in the error signal during training resulting in an order of magnitude fewer epochs to learn the training set, which is also what time-averaging aims to do.

[5]Without time-averaging the network did not converge, see appendix A.

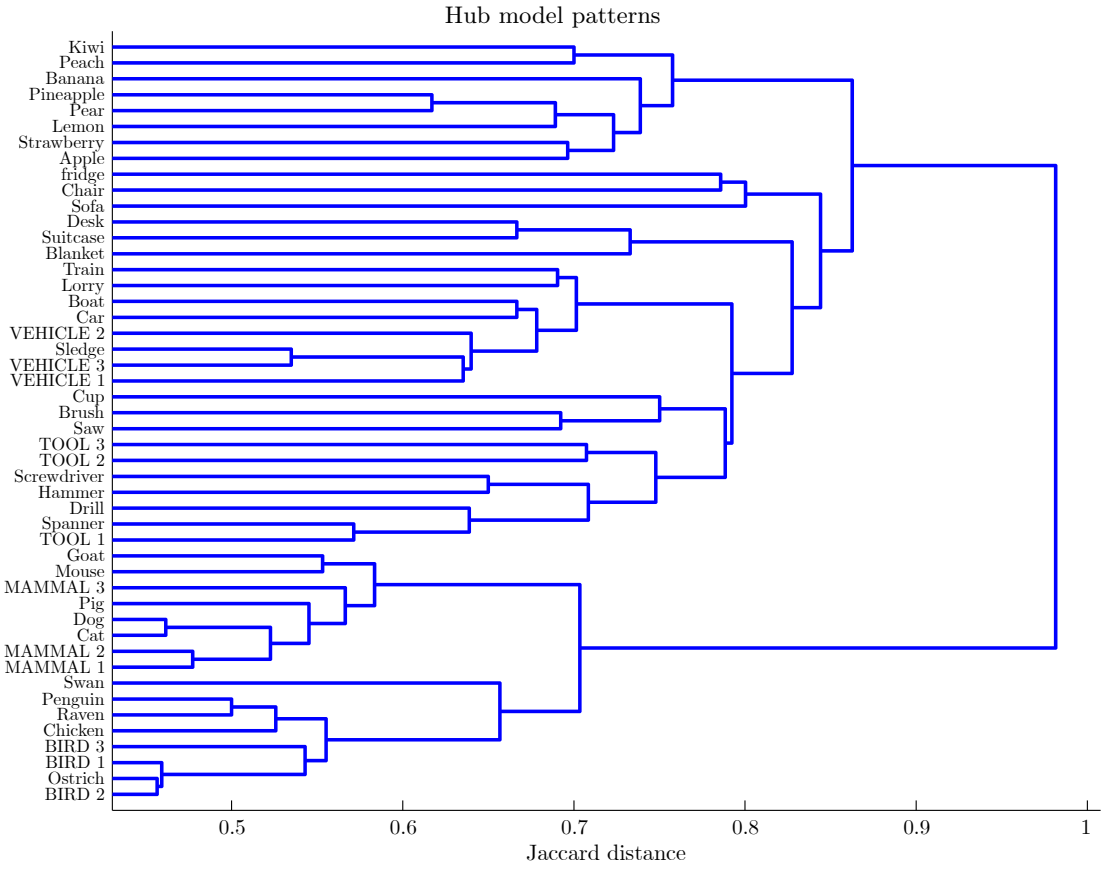


Figure 2.14: A hierarchical cluster dendrogram depicting the Rogers et al. (2004) patterns' relationship to each other using Jaccard distance, created based on the prototype given by the original authors.

A radically different way of implementing the model is to use a Boltzmann machine (BM). BMs are a type of binary-valued recurrent stochastic network. This kind of network is able to conform to the topology required by the hub theory and permits the emergence of attractors (Hinton & Sejnowski, 1986). Training involves minimising the difference in unit activations between the network settled with all inputs clamped, known as the plus state, and the network settled on each sub-pattern (e.g., just the verbal features clamped), called the minus states (Ackley, Hinton, & Sejnowski, 1985). Using a BM instead of a recurrent network trained with back-propagation through time allows for testing whether the behaviour of the model (including when lesioned) is dependent on the specific network architecture (Plaut & Shallice, 1993) or if it is a more universal property of the hub model architecture.

After training, our replications of the Rogers et al. (2004) network robustly map names to visual and verbal sub-patterns and vice versa. Thus, given the name subpattern of ⌜*chicken*⌝, the visual and verbal units of the network take on patterns (once the network has settled) that correspond to the visual and verbal features associated with ⌜*chicken*⌝. Similarly, when given

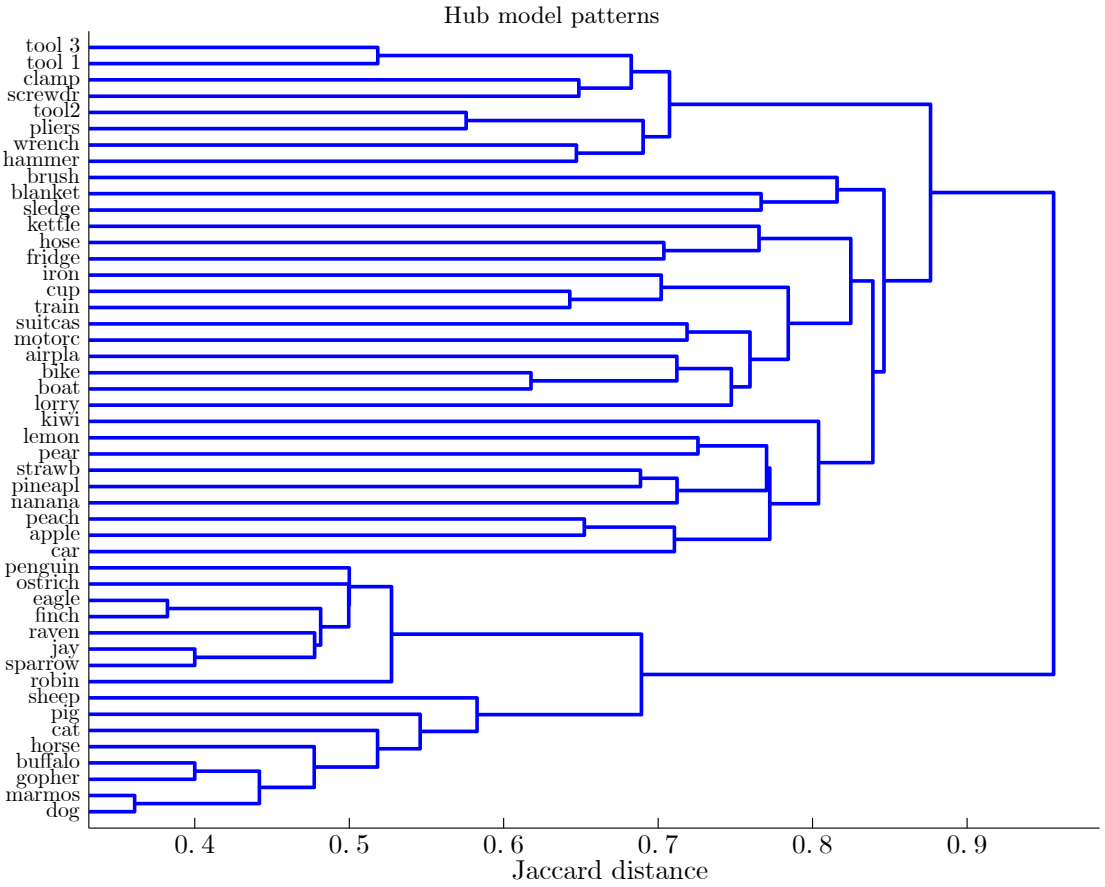


Figure 2.15: A hierarchical cluster dendrogram depicting the original Rogers et al. (2004) patterns' relationship to each other using Jaccard distance (T. Rogers, personal communication, August 21, 2012).

the visual or verbal features of that pattern, the other visible units take on values associated with the name and verbal or visual features of the pattern. More critically, when given a superordinate name (e.g., "bird") the sets of units corresponding to visual and verbal sub-patterns take on states that amount to the weighted average of the three nondescript patterns that share that same name if the network is $\text{BPTT}_1$ or $\text{BPTT}_2$. Conversely, when provided with the visual or verbal descriptors of any of these patterns the network activates the general-level name. For the BM, if a superordinate name is applied to the input, the verbal and visual output is one of the three base-level patterns at a probability reflecting the frequency with which the network was exposed to them during training. This means that each time the network is settled the probably of a specific verbal visual configuration is one third.

Once trained, all network types reach ceiling scores in drawing and copying and word-to-picture matching. However, healthy naming and sorting scores are not possible in the BM, but are in the BPTT networks. This is due to the inherent stochastic nature of BMs, the need for

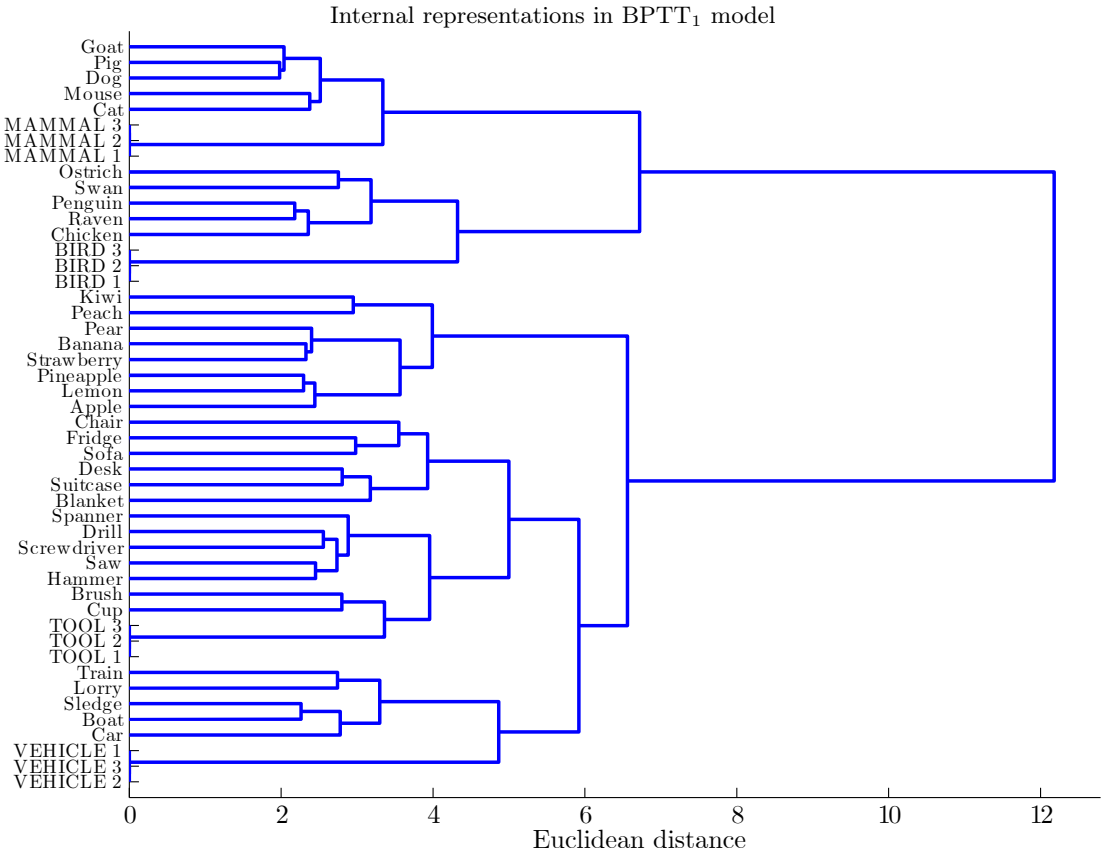


Figure 2.16: Dendrogram for BPTT$_1$ internal semantic states sampled 100 times per name sub-pattern input. Category names, shared over three patterns, are capitalised.

extra training to better learn the mapping between visual input and name output, and the design of the two tasks. The naming and sorting tasks make use of localist interpretations for names and category membership, which are not part of the BM's learning strategy (it needs extra minus states presumably). This issue might be addressed given longer training, although with the training given in this study the BM does score at ceiling on the word-to-picture matching and the drawing, copying, and delayed copying tasks, which use distributed representations.

As required by the hub theory, all the networks have internal states that allow for the mapping of the perceptual inputs to the output modalities, thus completing each of the four semantic tasks. Fundamentally, the internal semantic space must mirror the categorical and domain structure of the training set. This attractor-space can be represented using a dendrogram as in Figure 2.16, which shows the Euclidean distance between both individual concepts and between categories and domains. This allows a comparison between the intended categories and those that arise from the structure of the learned attractor states (see figure 5 in Rogers et al., 2004).

Rogers et al. (2004) provide a list of qualitative properties that their model's internal rep-

resentations possess. As shown in Figure 2.16, our implementations also conform to this list: Firstly, the two domains, animals and artefacts, are clearly separated from each other, as are to a lesser extent the six categories. Secondly, the model's representation of category-level names (e.g., "BIRD 1") are classed within their category cluster. And finally, fruit are classed under the domain of inanimate objects, but are in a distinct cluster to the the rest of the artefacts.

The hub-and-spoke (re)implementations contain a number of assumptions that have been carried over and adapted from previous theories. In fact, the modality-specific theory and the sensory/functional dichotomy, have already been modelled (i.e., without a hub, e.g., Farah & McClelland, 1991). A model with fewer assumptions, and fewer complex components, that can still account for the same groups of patients is a more parsimonious account. In other words, if all else is equal, a modality-specific account is preferable to a hub-and-spoke account.

## 2.5 Modality-specific model

### 2.5.1 Theory details

The modality-specific theory claims the semantic system consists of dissociable stores that each contain modality-based features. Together these different perceptual and functional feature stores represent concepts. Warrington and McCarthy (1994) claimed that two animate/inanimate domains of knowledge have a dissociable dependence on sensory or functional features respectively (Warrington & McCarthy, 1983; Warrington & Shallice, 1984; Warrington & McCarthy, 1987). These two related assumptions have been useful in shaping the accounts of semantic memory that came after and importantly form the theoretical backbone of other more recent semantic theories.

Although serious criticisms can be brought against this theory (see section 2.7), modality-specific accounts live on in both the conceptual topology and in the hub-and-spoke theories, so understanding how this, perhaps naïve, theory works will hopefully shed light on the assumptions carried over from this theory into subsequent related theoretical accounts.

### 2.5.2 Model details

In this section, a model inspired by the work of Miikkulainen (1993) will be used to show that a version of the modality-specific theory can be used to model both general semantic deficits and category-specific effects. The modality-specific theory was originally proposed by Warrington and McCarthy (1983, 1994) who argued that dissociable semantic stores exist for the various

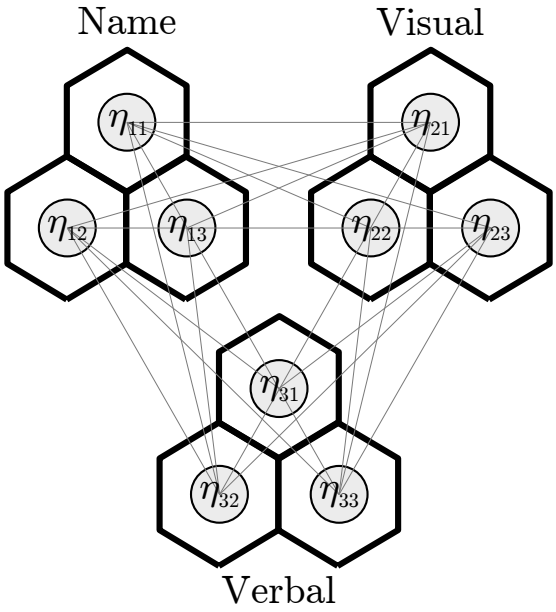


Figure 2.17: This is a schematic of the modality-specific model, simplified to only contain a few units. Each of the modal SOMs is bidirectionally connected to the other two allowing for input and output from each SOM to the rest.

modality-based features. A model of this theory was created by Farah and McClelland (1991), which gives a computational account of how modality- and category-specific deficits can arise.

The model presented in this section is inspired by DISLEX (DIStributed feature map LEXicon, Miikkulainen, 1990, 1993, 1997). DISLEX is a model of the lexical system, focused on the processing of ambiguous words (homographs, homonyms, and synonyms), and giving a computational account of lexical deficits in dyslexia and aphasia (for patient details see: Caramazza, 1991; Hagoort, 1993; Milberg & Blumstein, 1981; Milberg, Blumstein, & Dworetzky, 1987). DISLEX provides separate stores for each input and output modality: orthography (which can be seen as the visual modality's contribution), phonology (the auditory modality's contribution), and semantics. The training data is similar to that used in the two previously discussed models, except that DISLEX uses real-valued patterns instead of binary ones. To clarify, DISLEX is not a model within any one theory described in this thesis, because it is not a model of the semantic system per se. It does however conform in its topology to what is required for a modality-specific semantic memory model since it has three pools of units connected to each other, see Figure 2.17, where these three can be renamed and trained as the three subpatterns of name, visual, and verbal features shown in Figure 2.13.

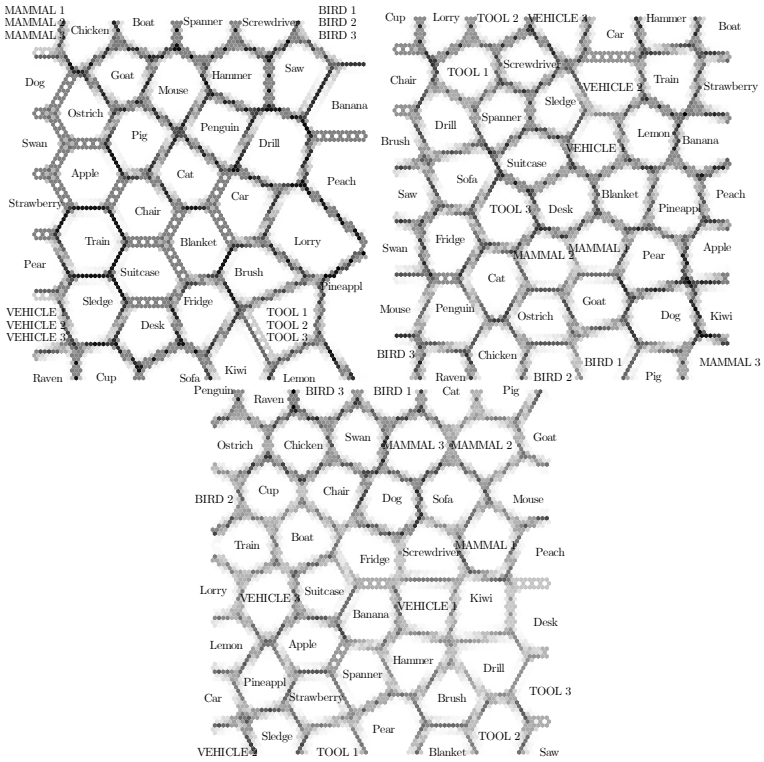


Figure 2.18: The U-matrices of each of the SOMs with their BMUs labelled. Very simply, a U-matrix represents the distances between units of on the map: darker hexagons represent longer distances (for more details see appendix B).

### 2.5.3 Implementation details

Three distinct self-organising maps (SOMs; see appendix B) are used to represent the three modalities. As seen in Figure 2.18, the three SOMs represent different cognitive (sub)faculties, and are connected to each other using linear connections (see Figure 2.17). The connections between modal maps represent the equivalent connections that occur within the semantic system that allow for (what appears to be) amodal or, in this case, distributed panmodal representations of concepts. Input to the model is provided via the three SOMs' surfaces directly, depending on what subpattern is needed as input. So for example to present a word to the model the name-SOM is given the input pattern that represents, e.g., the word "dog", and then activation spreads via the connections between the name-SOM and the visual- and verbal-SOMs, thus activating units on their surfaces that represent the visual and verbal parts of $\langle dog \rangle$.

To input a pattern the best-matching unit (BMU: the unit on the surface of the map that comes to represent a pattern) and the BMU's neighbourhood (with a pre-specified radius) is activated directly on the surface of the appropriate SOM. This requires the hybrid SOM units to form a halo of activations around the BMU. In other words, if, for example, unit 12 in Figure 2.17 is the required BMU then $\eta_{12}$ is calculated using Equation B.10 (and so on for the


Figure 2.19: The activations on the surface of the name and verbal SOMs with the BMUs labelled. When the name SOM is given the input it causes the verbal SOM to produce the appropriate output. Grey hexagons represent units that are off. Coloured units are activated.

neighbourhood, $N_{12}$ with radius $r = 1$ during testing), then it is used as name input which will in turn provide visual and/or verbal output. These outputs, see Figure 2.19, are interpreted as described in appendix B.6.3.3.

The hybrid SOM network was trained using epoch-wise weight updates both for the SOMs and the supervised connections. Each SOM is a toroidal $48 \times 40$ map (see appendix B.7) of hexagonal cells with a Gausssian neighbourhood function (see Equation B.4). For the supervised weights the Widrow-Hoff learning rule (Equation B.15) was used with a standard logistic function for computing the post-synaptic unit states. The targets for each layer varied during training as the radius decreases to one (see Equation B.10). The targets started off very large, since the SOM training radius was large, and took on real values (since the Gaussian neighbourhood kernel was used), which decayed slightly on each epoch with a lower bound of one (see Figure B.98). This results in the SOM honing its BMUs through competitive learning, as usual, whilst at the same time the supervised connections between units also learn to associate appropriate BMUs on all three maps.

In this section we presented an implementation espousing the sensory/functional dichotomy within the classical modality-specific theory. The defining feature of this account is the existence of separate semantic stores, and the imbalanced contribution of modalities to domains of knowledge. Both of these assumptions have been used as a basis for other theories, especially the conceptual topography theory, which can be seen as the natural next step in the evolution of this theory. A model within this account is presented in the next section.

## 2.6 Conceptual topography model

### 2.6.1 Theory details

In contrast to the account given by the hub-and-spoke and the conceptual structure models, which involve (or at least tolerate) the existence of amodal representations, are accounts such as those proposed by embodiment theorists. Their theories usually claim that the semantic system (and the cognitive system in general) is devoid of purely amodal representations, and that processing semantic concepts requires only relevant sensorimotor cortical regions (Arévalo, Baldo, & Dronkers, 2012; Aziz-Zadeh & Damasio, 2008; Hauk & Tschentscher, 2013; Pulvermüller, 2013; Zwaan, 2004). Proponents of distributed, embodied, extended, grounded, situated cognition (Robbins et al., 2008) — collectively referred to as embodied cognition (Barsalou, 2008) — usually claim that what appears to be non-modality-specific is instead (in)directly grounded in the senses. Bearing in mind (one way of representing) the spectrum of embodiment positions within semantic cognition proposed by Meteyard et al. (2012), there are few strong embodied computational accounts.

In other words, while there is consensus that sensory and motor information contributes to the formation of meaning, embodied semantic cognition accounts rule out any purely amodal concepts being computed. Unfortunately, there are very few computational models to support or enhance this account:

> Another limitation of current [grounded cognition] work is the relative lack of formal and computational accounts. It is fair to say that current empirical research on grounded cognition heavily reflects demonstration experiments. As philosophers of science note, when a new area emerges, demonstration experiments dominate to justify the area's importance. Eventually, mechanistic theories develop that stimulate new generations of research, distinguish between mechanistic accounts, and elaborate mechanistic accounts further. Mechanistic accounts of grounded cognition have existed for some time and continue to emerge increasingly (e.g., Cangelosi & Riga, 2006; Farah & McClelland, 1991; Feldman, 2006; Pezzulo & Calvi, in press; Plaut, 2002; Wennekers, Garagnani, & Pulvermüller, 2006). Some preexisting systems have much potential for development as grounded theories (e.g., O'Reilly & Norman, 2002; Ullman, Vidal-Naquet, & Sali, 2002). In addition, various preformal architectures have potential for development as computational systems (e.g., Damasio, 1989; Simmons & Barsalou, 2003). In general, though, it is clear that

> much further theoretical development remains, and that such developments will move the area forward significantly.
>
> (Barsalou, 2010, p. 719)

As mentioned previously in subsection 1.3.7, and above by Barsalou (2010), the convergence zone theory is an account of embodied semantics that lends itself to computational modelling. Convergence zones are hierarchical (sensorimotor) feature maps that capture associations between features using so-called conjunctive neurons. Once convergence zones capture a set of features that make up a concept they can recreate it in the absence of sensorimotor input. Within the conceptual topography theory, convergence zones organise themselves based on the similarity-in-topography principle, which claims that neurons within a convergence zone become proximal as a function of the similarity of the features they represent, and on the variable dispersion principle, which states that conjunctive neurons in charge of a category are not strictly contiguously located, but the more similar the instances that make up the category are the lower the dispersion of the conjunctive neurons that represent it (Simmons & Barsalou, 2003).

In general terms, the conceptual topography theory can be seen as subsuming the basics of the modality-specific theory (initially proposed by Warrington & McCarthy, 1983, see subsection 1.3.3) of semantic memory, which supports separate stores of modality-derived features, but adds on the possibility of cross-modality connections once these features have been refined through the levels of modal convergence zones. In other words the conceptual topography theory's account of semantic memory tries to reconcile modular semantics, i.e., distinct stores of features, with unitary semantics. It attempts to do so without introducing a unitary store of features, only by introducing a store of associations over the features[6].

### 2.6.2 Model details

Based on the description of the conceptual topography theory, it appears as if some of the required functionality can be captured using self-organising maps (SOMs). According to the proponents of this theory, a convergence zone is a core component of how the semantic system derives (pre-)semantic features, concepts, categories and ultimately domains from feature maps, which the authors use to mean raw modal input, e.g., from the retina. Stacks of convergence zones are used to gain more and more abstract and high-level representations of modality-specific input and at the topmost level cross-modal convergence zones — which span across as

[6]This claim is possibly a little duplicitous, since associations over features can be seen as a type of (higher-level) feature.

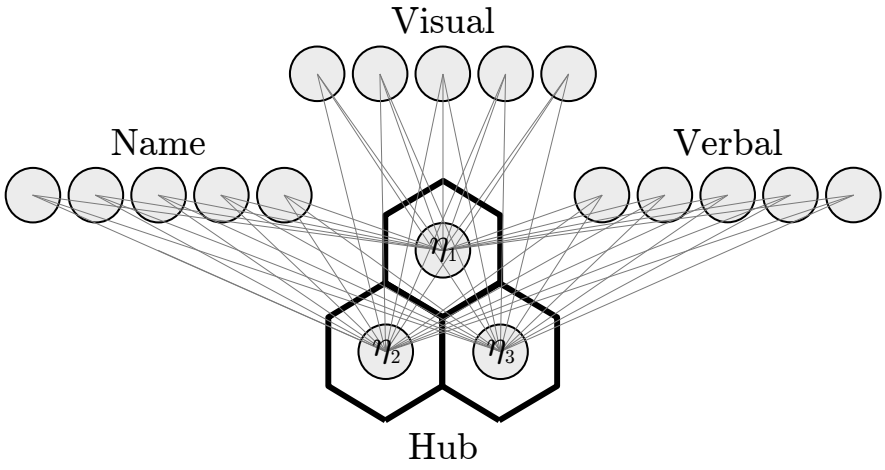


Figure 2.20: This is a toy model with the same architecture as the conceptual topography model. The three input pools are connected bidirectionally to the hub units. The SOM is trained on the whole pattern, while the three input pools are trained to link each subpattern to its activity as calculated by Equation B.10 on the surface of the hub map.

many modalities required — discover and represent the correlations between modality-specific convergence zones (see Figure 1.3).

Here the focus will be on creating a model of the cross-modal convergence zones, using SOMs, which aim to find the correlations over (pre-)semantic features. The next few sections will describe this form of network, and will explain how SOMs can be used to form a basic model of the similarity-in-topography principle; which claims that the "spatial proximity of two neurons in a convergence zone reflects the similarity of the features they conjoin[, as the] two sets of conjoined features become more similar, the conjunctive neurons that link them lie closer together in the convergence zone's spatial topography" (Simmons & Barsalou, 2003, p. 457).

In Figure 2.20, a toy version of the architecture of the conceptual topography model can be seen, in which the hub is a SOM and the inputs are supervised network layers. The SOM is trained on the full feature patterns, meaning that the topology of the map reflects the structure of whole patterns including orthogonal names, while the modal spokes are trained on their specific subpattern. In some ways, this model can be seen as a SOM-based model within the hub-and-spoke theory, because of the similarities between the two theories. A central amodal hub, in this model a SOM but in the original model a layer of classical supervised network hidden units, is used to represent semantic content and is bidirectionally connected to modal inputs. This means that the spokes contain (pre)semantic features, while the hub contains pure semantic representations in its topology (as well as the features themselves within its codebook vectors).

### 2.6.3 Implementation details

The implementation consisted of five layers of units: the three modal stores, the cross-modal layer, and the SOM cells. The convergence zone and the modal pools were trained concurrently using the equations described in appendix B. The topology of the SOM hybrid network can be seen in Figure 2.21, which shows both the emergent clusters, on the left, and on the right, an example target for the input layers. The aim of training the SOM is to create individual best-matching units (BMUs) for each input pattern, and for these unique BMUs to be linked to their respective subpatterns when they are applied over the input pools.

The SOM used to represent the cross-modal convergence zone had size $10 \times 50$ and contained hexagonal cells on a toroidal map. It was trained on the same patterns as the hub-and-spoke model using epochwise weight updates for both the supervised connections and the SOM weights. The Widrow-Hoff delta rule with the standard logistic function was used for the supervised connections, while the batch training algorithm and a Gaussian neighbourhood function was used for the cross-modal SOM. The targets for the Widrow-Hoff rule emerge on the surface of the SOM, in a similar way to those in the modality-specific model. As training of both SOM and supervised connections progressed the radii of active halos around the BMUs used as targets were decreased.

To present to the model a picture representing, e.g., $\langle tiger \rangle$, the visual input layer had the appropriate subpattern applied over its units, the input activations were propagated to the cross-modal SOM via the visual-hub connections. On the surface of the SOM units took values representing a $\langle tiger \rangle$, i.e., activation was focused around the appropriate BMU for that input pattern, which was identical to the most active unit on the supervised network layer twinned with each SOM weight. See appendix B for more details.

This implementation of the conceptual topography theory, as mentioned, could also be conceptualised as a hub-and-spoke theory, with the SOM being the amodal hub. This is not due to coincidence, the two theories have substantial overlap in their account of the semantic system.

## 2.7 Discussion

We have so far seen four different families of models of semantic cognition. Although overlap occurs, they each offer a differing take on how concepts are represented. In this section both compatibility and disagreement between these accounts will be discussed, with the aim of clarifying what overarching assumptions are included in each theory, model and therefore

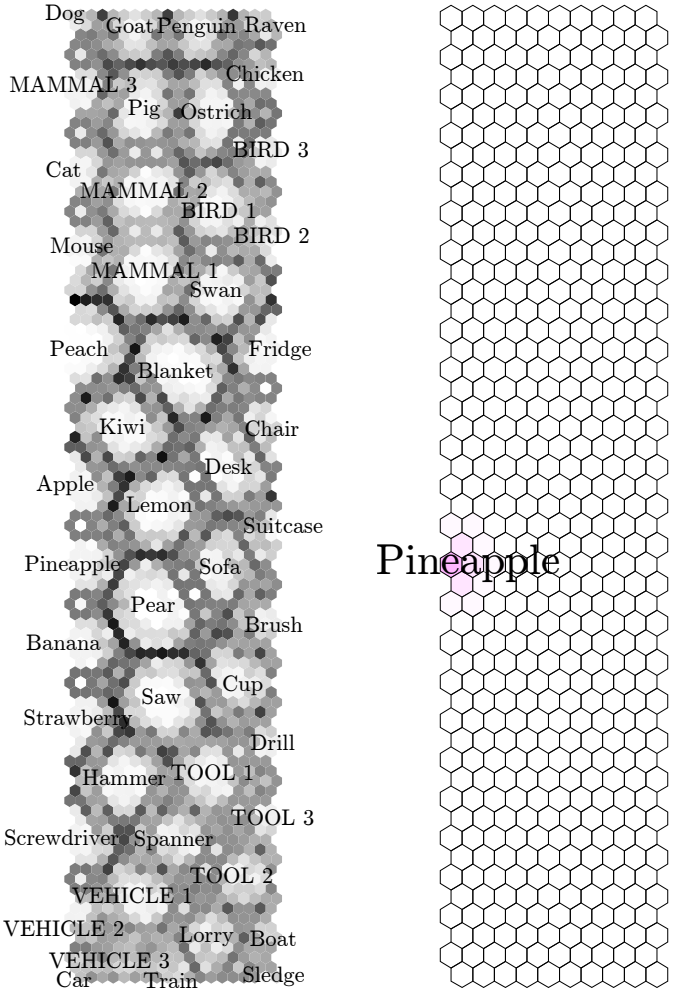


Figure 2.21: To the left is the conceptual topography model's U-matrix with BMUs labelled. The training elements cluster together based on their categories and domains. To the right is the output of the supervised units.

implementation.

The first theory and model was that of Tyler et al. (2000). Based on their interpretation of neuropsychological and neuroimaging findings (Bright, Moss, & Tyler, 2004), Tyler et al. (2000) create their model with a semantic system that does not have modular or localised stores for different types of features or indeed categories. Even so, their account can give rise to category-specific patterns of dissociation. This theoretical position is proposed as contra to the stance of Warrington and colleagues (McCarthy & Warrington, 1986, 1988; Shallice, 1988, 1993; Warrington & McCarthy, 1983; Warrington & Shallice, 1984; Warrington & McCarthy, 1987, 1994), who "argued that musical instruments and gemstones are similar to living things in that they are primarily distinguishable in terms of perceptual properties, whereas artifacts and body parts are categories of knowledge for which function is most salient" (Tyler et al., 2000, p. 196). In other words, what Warrington proposes is that features are distributed unevenly per domain

(the sensory/functional dichotomy), with animals being made up of predominantly perceptual features and inanimate objects being disproportionately made up of functional properties. On the other hand, Tyler et al. (2000) supports the view that semantic cognition comprises a unitary system, insofar as features of both kinds are stored proximally. In other words, no separate semantic stores exist for visual and verbal modalities. Features are not located in separate perceptual/functional stores (what are called spokes in the Rogers et al. (2004) hub theory, see subsection 1.3.6). Tyler et al. (2000) propose that it is correlations within and between concepts and features that cause the structural differences in concepts and categories, and not different ratios of perceptual/functional features per domain.

The Tyler et al. (2000) model is useful for understanding and computationally representing the fact that the correlations between and within concepts can organise the semantic system. The original model gives an account of this phenomenon with a minimum level of complexity in its implementation, it only includes feedforward connections, it learns using the simple feedforward variant of back-propagation, it only has three layers, and the patterns have a straight-forward structure. The main prediction of the theory is also straight-forward: feature correlation implies feature preservation. In other words, the strength of the Tyler et al. (2000) implementation of the conceptual structure theory is its explanatory power given its simplicity. The two (re)implementations show that the feature composition of concepts is sufficient to give rise to categorical structure when coupled with a feed-forward or recurrent network.

The conceptual structure theory has also been successful in giving an account for the organisation of conceptual structure in general (see: Bright et al., 2004; Randall, Moss, Rodd, Greer, & Tyler, 2004; Taylor, Moss, & Tyler, 2007; Tyler & Moss, 2001; Tyler et al., 2003), which adds credibility to the Tyler et al. (2000) model. SD patients and HSVE survivors, do indeed fare badly on items with few inter-correlations. For example, they struggle with exceptional items, e.g., tigers and zebras lose their stripes, and birds gain two legs, because they appear to normalise these items towards the average item in the category (in the case of tigers/zebras mammals) or domain (in the case of birds, animals in general) they belong to. And these patients also fare badly on domains or categories or items that are not very rich in certain features, e.g., visually poor drawings of animals are more difficult to name than visually complex depictions of inanimate objects. All of this is explained well by the Tyler et al. (2000) model, which does indeed capture this seemingly reverse-developmental trajectory in which any feature, and therefore concept (since they are defined as a set of features), that is not correlated enough with others is lost.

The second account of semantic cognition described was the hub-and-spoke theory. General

hub-and-spoke theories (and thus semantic hub models and implementations) propose that certain areas of distributed systems have concentrations of vertices around a specific node (or set of nodes), called the hub, due to its importance. This is the abstract idea behind the hub theory in Rogers et al. (2004): the anterior temporal lobe is important for the semantic system due to the fact it computes high-level amodal semantic representations as a function of low(er)-level (pre-)semantic features which themselves are derived from what is ultimately direct modal input. In some ways this part of the theory is compatible with the claims that the conceptual topography account gives, which proposes that many levels of convergence zones are stacked hierarchically to give rise to higher and higher levels of semantic content.

The generalisability of the hub theory is reflected in the modelling breadth of the hub models. This family of models provides a far-reaching account of semantic memory impairments. The model in Rogers et al. (2004) and the related model in Lambon Ralph et al. (2007) seem able to explain the breakdown of semantic memory of both SD and HSVE. Thus, this family of models appears to be versatile and informative as a general view of the semantic system both in healthy people and in patients. The family of (re)implementations presented in this and the following chapters will quantify and qualify the generalisability of the hub modelling strategy. In addition, a hybrid model using both a hub architecture and a conceptual topography style of organisation has been presented, which further tests the hub theoretic principles presented in Rogers et al. (2004).

As touched on, the hub theoretical account of semantic memory can be seen as being an intermediate position between what appear to be mutually exclusive theories of the architecture of the semantic system. On the one end of the spectrum is the account given by the conceptual structure theory (e.g., Tyler et al., 2000, as well as other OUCH accounts) which supports a unitary semantic store, in which all features, regardless of their modality, are stored cortically proximally. On the other end, are the conceptual topography and the modality-specific theories (see: Simmons & Barsalou, 2003, for comparison), which claim that every single feature is stored in a modality-based region. The hub theory has both these properties — it manages to contain an amodal unitary store for concepts (the hub), and modality-based feature stores (the spokes).

Further comparing the hub model with that of the conceptual structure theory, the former assumes that there are separate semantic areas that are modality-specific, the spokes, while the latter assumes that all features and all categories are underpinned by distributed representations within a unitary system, which also can be modality-based since both theories espouse the perceptual-functional dichotomy in some form. In light of the fact that both theories support

hierarchical processing of modal input, their significant difference must be outlined: the hub theory supports a specifically defined amodal hub, while the conceptual topography theory does not propose such an area explicitly. However, Simmons and Barsalou (2003) propose that cross-modality convergence zones exist that "create the complete representation of a category across modalities" (p. 465). These two theories seem to have extensive overlap, especially in light of the fact that they both accept and make use of the modality-specific theory as a component of the driving force for categorical and domain creation, both explicitly referring to the work of McRae and Cree (2002). So on the one hand, both models agree on the absence of category-specific loci (innate or otherwise; like in the (strong) domain-specific hypothesis of subsection 1.3.4). While on the other hand, the hub model has a series of modality-specific spokes, presumably within or directly interfacing with the semantic system, that store (semantic) features which the Tyler et al. (2000) conceptual structure theory argues against. This is an important difference as the hub theory does allow for a unitary store, as well as a set of separate modality-specific spokes, while the conceptual structure theory only allows for a unitary store. This is interesting because the literature is usually seen as falling on either side of the unitary vs modality-specific dichotomy (see Bright et al., 2004, for an overview), and choosing to include both modality-specific spokes and an amodal hub appears to be, perhaps, a compromise too far. Especially so, given one of the proponents the hub model was previously in favour of a unitary system interpretation of semantic memory (in Lambon Ralph, Graham, Patterson, & Hodges, 1999; Lambon Ralph & Howard, 2000), and against one in which features and representations are separable by modality, which is what the spokes of the hub model represent.

However, such a stark contrast between these three theories needs to be tempered with another way of interpreting their differences. The conceptual structure theory's assumption of an amodal semantics can be reconciled with the hub theory and the conceptual topography account if the hub and the cross-modal convergence zones are defined as the semantic system proper. And the spokes and the modality-based convergence zones are defined as extra- or pre-semantic. These two different ways of looking at these theories underlie some of the confusion with regards to their compatibility.

Thirdly, a modality-specific model was described. As mentioned above, the modality-specific theory was originally proposed by Warrington and McCarthy (1983, 1994) who proposed that dissociable semantic stores exist for the various modality-based features. A previous model of this theory was developed by Farah and McClelland (1991), see subsection 1.4.2, which gives a computational account of how modality- and category-specific deficits can arise. While their model appears to be a useful way of thinking about deficits, it is both modality- and

categorically-organised by design. In other words, perceptual functional features are directly associated with inanimate objects and perceptual verbal features with living things — this a priori structure can be seen clearly in Figure 1.4. This criticism has been raised previously by French and Mareschal (1998), who explain that "[t]he greatest shortcoming of the [modality-specific] model is that it fails to explain why damage should occur either (a) selectively to the perceptual features (thereby preserving knowledge of inanimate words) or (b) selectively to the functional features (thereby preserving knowledge of animate objects)." (p. 374)

To address these issues French and Mareschal (1998) propose a semantic memory model of category-specific deficits much like the one described in section 2.3, i.e., with a unitary semantic store, thus demonstrating that it is the distribution of features that determines the pattern of preservation when damage occurs. That, of course, places their model clearly within what has come to be known as the conceptual structure theory. This marks a point in the literature when the distribution of features, and therefore the internal properties of concepts, come to be seen as driving the organisation of the semantic system and not vice versa. In the late 1980s and early 1990s there was a debate centring on the issue of modality-specific vs unitary semantics (e.g., Caramazza et al., 1990; Riddoch, Humphreys, Coltheart, & Funnell, 1988; Shallice, 1993), which culminated in more complex theories of semantics (such as those discussed in chapter 1) being proposed. After the mid to late 1990s, models of semantic memory stop being organised by either modality or domain in such a direct way, with increasingly more emphasis on functional and less on structural modularity. This was in part due to the work being published at that time that demonstrated that the distribution of features can give rise to category-/modality-specific patterns of dissociation (e.g., French & Mareschal, 1998; Small et al., 1995; Thomas & de Wet, 1999; Tyler et al., 2000)[7].

Models are now created that accept patterns that bring about internal self-organisation, as opposed to it being part of the architecture. This is due to computational modelling work, as well neuropsychological, neuroimaging, and lesion studies, demonstrating that even though category- and domain-specific organisation does exist, it is not as straight-forward as dissociable stores. Structural modularity is not required for functional modularity to arise. In addition, finding the boundary between where modality-specific processing is going on (e.g., visual or motor cortices) and where such processing is semantic in nature is difficult (e.g., temporal or limbic cortices). Very few would argue against semantic features being grounded in, or derived from, one or more appropriate modalities. However, as mentioned in subsection 1.3.2, it is still

[7]Previously, two of the authors of Small et al. (1995) supported separate modal subsystems within semantic cognition (Hart & Gordon, 1992).

an open debate if features are purely semantic, pre-semantic, or purely perceptual – or, if like Simmons and Barsalou (2003) propose, that features are a function of the level of analysis of the semantic system, or the point in the semantic pipeline. They attempt to dissociate high-level properties and low-level features, but even such a differentiation will suffer due to the nature of the spectrum from perceptual/modal to semantic/conceptual knowledge. Either way, this question remains open.

In other words, modality-specific accounts have fallen out of favour for two reasons: *a*) it has been shown that structural dissociations are not required for functional dissociations, i.e., category- and modality-specific deficits can emerge in a system that is not composed of distinct regions that subserve certain divisions of semantic knowledge; *b*) in addition to not being necessary, structural modality-specific cortical regions within semantics are not able to be located in a sufficiently dissociable way. Additionally, both these reasons help to explain why, after a number of models strictly within the modality-specific theory where created, they have since been largely sidelined. Once the models are superficially shown to give rise to category- or modality-specific deficits, their explanatory power is highly limited. This being said, modality-specific theoretic approaches, albeit, perhaps, weaker than the original proposals (Shallice, 1988; Warrington & McCarthy, 1994) are still very much alive and used as the basis for constructing more complex models, e.g., the conceptual topography account.

This brings us on to the fourth model, that of the conceptual topography account. This account falls into the embodiment camp which posits that no amodal processing or representations exist in semantic memory – every semantic process is grounded in a perceptual modality. According to Simmons and Barsalou (2003) what is going on is that perceptual pathways are reactivated during semantic retrieval and that simulations or re-enactments (the process of re-activating perceptual pathways from the top down) are how semantic memories are accessed (Barsalou, Simmons, Barbey, & Wilson, 2003). While this might appear to be a novel and dramatically different approach on first glance, it shares a large number of similarities with both the hub-and-spoke and the modality-specific accounts, as already discussed. Clear parallels exist between the hub model's attractors and memories grounded in modality-specific inputs within the conceptual topography account. Perhaps the most interesting aspect of modelling the conceptual topography theory is that it is the only theory presented so far that has no computational model associated with it, as far as we know, although the authors outline how important such models would be (Barsalou et al., 2003; Barsalou, 1999; Simmons & Barsalou, 2003).

A final note should be made on the way concepts are represented because each theory has

a slightly different account. In the modality-specific and the conceptual structure accounts, a concept is a set of modality-derived features, e.g., ⟨*has two legs*⟩, ⟨*is red*⟩, and ⟨*can cut*⟩, a category is a set of concepts that share more features with each other than they do with the other concepts, and, equivalently, a domain is a grouping of categories that share many features. This means that ⌜*dog*⌝ is directly defined as the set: ⟨*has four legs*⟩, ⟨*has fur*⟩, ⟨*can bark*⟩, ⟨*can fetch*⟩, and so on; while the category ⌜*mammal*⌝ is defined as the loose family of concepts that more often than not contain features such as: ⟨*has four legs*⟩, ⟨*has fur*⟩, ⟨*is brown*⟩, ⟨*has eyes*⟩, ⟨*produces milk*⟩, and so on; and a broad domain domain such as ⌜*living things*⌝, which subsumes the previous two is composed of categories that contain (more generic) features like: ⟨*can move*⟩, ⟨*can grow*⟩, ⟨*can breathe*⟩, ⟨*has legs*⟩, and so on. In the hub theory and the conceptual topography account this is not how the semantic system internally represents concepts. For the hub theory, the amodal hub stores them as a function of these modal features, which can be seen as being pre-semantic. While for the conceptual structure account it is less clear, because the authors claim that in their account "nothing is explicitly called a concept" (Barsalou et al., 2003, p. 84). Nonetheless, conceptual processing[8] involves cross-modal convergence zones and concepts are not merely the modal input, but associations over it.

## 2.8 Summary

In this chapter, the four families of semantic memory models and their implementations were presented in detail: the conceptual structure, the hub-and-spoke, the modality-specific, and the conceptual topography models. The conceptual structure model was (re)implemented in two ways: as a feedforward network, and as a recurrent network. The hub model was (re)implemented as: a recurrent back-propagation network (with two slightly different variants), and a Boltzmann machine. The modality-specific model was implemented in a novel way using three hybrid interconnected self organising maps (SOMs). And finally, the conceptual topography model, also in a novel way, was implemented as a hybrid SOM connected to three input pools of units. In the next few chapters, the way each of these models performs on semantic tasks post-lesioning will be examined.

[8] "The study of conceptual processing will be best served by discovering and describing the relevant mechanisms, rather than arguing about the meaning of lay terms such as concept." (Barsalou et al., 2003, p. 84)

# Chapter 3

# Modelling general semantic deficits in the hub model

## 3.1 Overview

In this chapter the networks developed within the hub theory are tested on the four semantic tasks described in subsection 1.2.2. The semantic tasks are modelled in a way that allows for parallels to be drawn with participant data. More specifically, the method used to test the recurrent network, as in Rogers et al. (2004), consists of keeping the relevant input constant (be it a visual, verbal, or name pattern) while running the network. The network is then allowed to settle without any externally applied input until equilibrium is reached. Finally, the states of the output units in the required pool are compared to their targets. Presented here are the results from two families of recurrent networks, backpropagation though time and Boltzmann machine networks. Following training, the networks are able to model healthy participants, but fail to show the required effects for capturing patient behaviour. The disparity between the original hub-and-spoke model and the replications presented here is discussed as are potential reasons for why this issue has arisen.

## 3.2 Introduction

As described in section 2.4, the hub-and-spoke model comprises an amodal hub connected to modality-specific spokes. In the original implementation, as in the reimplementations discussed here, the recurrent network is separated into three input/output pools that correspond to the

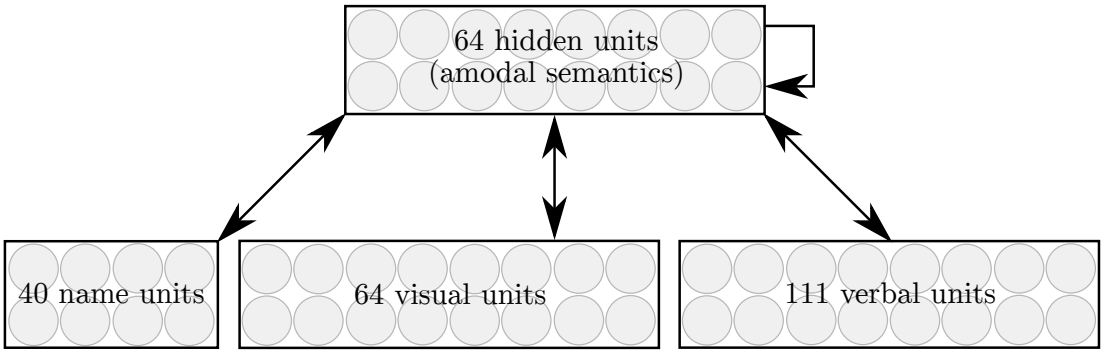


Figure 3.23: The hub model's recurrent network topology, based on figures 1 and 3 in Rogers et al. (2004). Reproduced from Figure 2.12.

properties of an item: name, verbal descriptors, and visual features, recall Figure 3.23. The connections within the semantic hub and between the input/output pools are perturbed in order to simulate SD-like lesion damage. Rogers et al. (2004) lesioned the original hub model by indiscriminately globally severing connections between units. This zeroing of the weights is claimed to be a sufficient analogue to the damage seen in the temporal lobes of SD patients. By removing randomly selected connections in increasing percentages Rogers et al. (2004) showed that their network displays neurodegeneration-like behaviour reflecting the progressive loss of semantic knowledge seen in SD.

In the implementations presented here, the same approach as Rogers et al. (2004) is used. Firstly, a randomly-selected percentage of all weights are set to zero and then the semantic tasks are run on the network. This is repeated for increasing percentages of weights, to emulate the longitudinal testing of patients as their neurodegeneration progresses. For each semantic task this is done 50 times for each percentage level of damage (e.g., from 0% to 90% of connections removed, depending on the task); paralleling 50 SD patients tested at different stages of progressive degeneration. Once the network is lesioned, settling becomes increasingly difficult and may result in dramatically different responses given the same input; the network may oscillate ad infinitum instead of settling to a stable state. Thus all the results are based on sampling the network 10 times for each of the sub-patterns it is tested on.

Since our reimplementations have pre-lesioned internal representations as found in the original hub model (recall Figure 2.16), damage can be applied to cause disruptions to the attractor basins. SD-like damage is modelled by setting increasing proportions of all connection weights to zero. This causes the network to be less adept at completing semantic tasks, as propagation of activations both within the hub and between it and its spokes is impaired. Disconnection has a pronounced effect on the semantic attractor landscape; the network can now only manage to

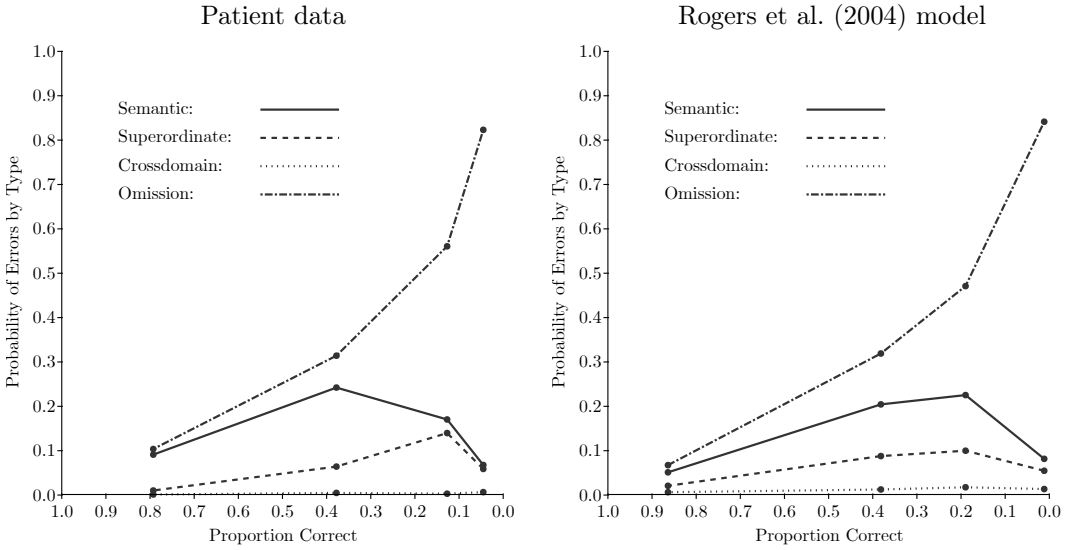


Figure 3.24: On the left are the results of the confrontation naming task from the fifteen SD patients. On the right, are the results of the original model reproduced at 10%, 20%, 25%, and 35% from Rogers et al. (2004, fig. 6)

represent a subset of the previous 48 concepts (sa discussed in detail in section 3.7). The clusters corresponding to concepts, categories, and domains are now deformed, e.g., the attractors for "cup" and "mouse", from opposing domains, are now in the same semantic cluster. This merging of conceptual representations from different domains, as opposed to categories within the same domain of knowledge, appears to signal a deviation from the hub theory's requirements. In the next few sections, the networks' performance in the semantic tasks will be presented and discussed.

## 3.3 Confrontation naming

### 3.3.1 Patient naming

Recall that confrontation naming requires subjects to generate verbal labels (names) from visual input (pictures). Rogers et al. (2004) report data on this task from fifteen SD patients, reproduced in Figure 3.24. At the earlier stages of degeneration, *omissions*, when the participant gives no answer, are relatively few but they increase dramatically as the disease progresses, until the only errors are omissions, i.e., the individual is completely anomic. *Superordinate errors*, so called because the response is not the expected name (e.g., "owl"), but something more general (e.g., "bird"), seem to follow a similar trend to omission errors. However, at the most severe stages of the disease, superordinate errors drop off due to anomia. *Semantic errors* occur when the response is from the same category as the line-drawing presented (e.g., "dog", when the

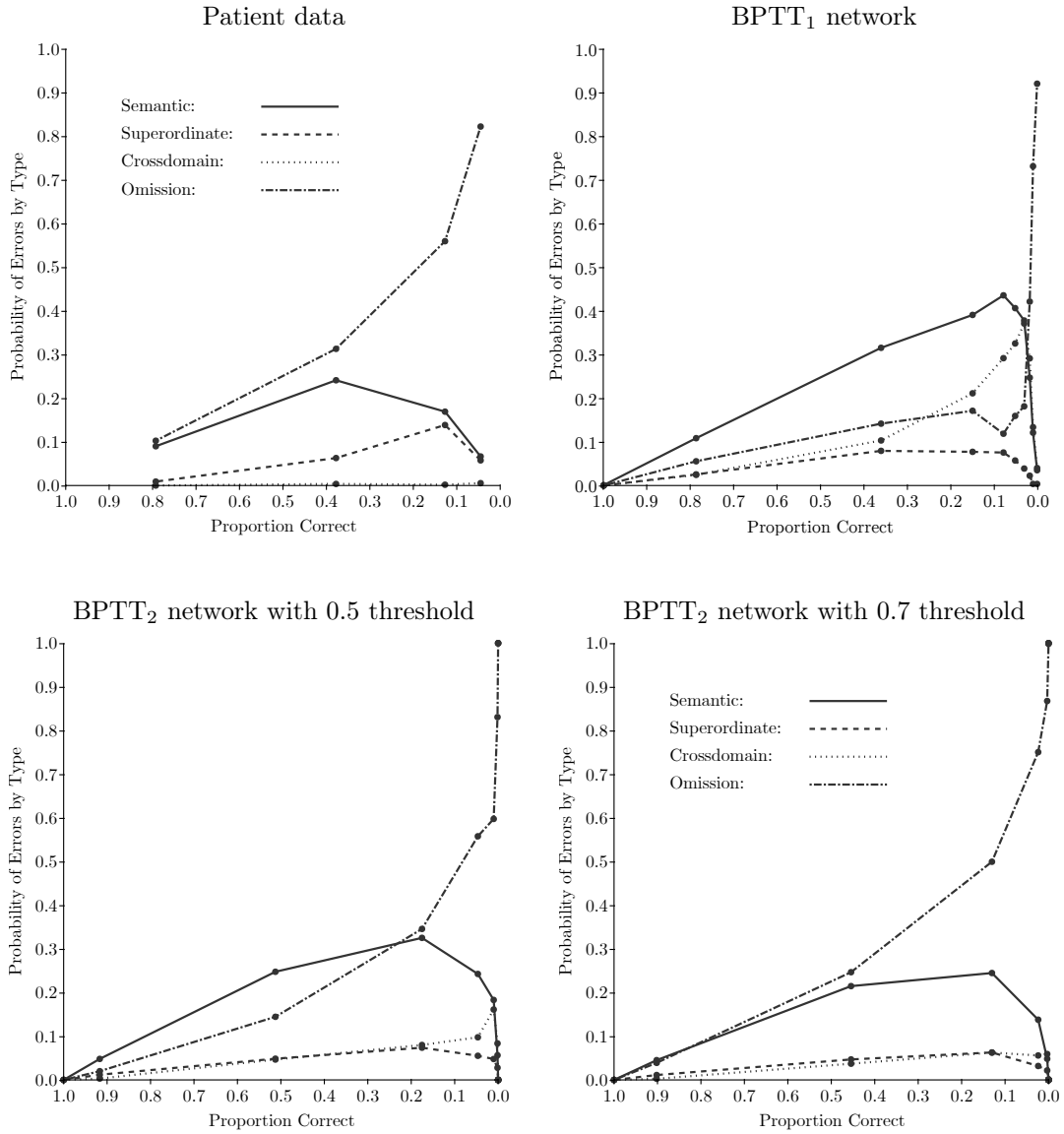


Figure 3.25: Results of the confrontation naming task of the patients from Rogers et al. (2004, fig. 6), the two reimplementations BPTT$_1$ and BPTT$_2$. The latter implementation is shown both with a naming threshold of 0.7 and of 0.5 (the one used in the original model) in order to show that the types of naming error are not overwhelmingly a function of threshold.

correct answer is "horse"); these errors are low initially, then rise, and finally return to a low level (again due to anomia). *Cross-domain errors*, where a response is given from the opposing domain to that which the stimulus belongs to (e.g., calling ⌜*horse*⌝ "car"), are almost never observed in the SD sample.

### 3.3.2 Model naming

The model implements confrontation naming by evaluating the name units' output when given a visual input, thus paralleling the visual input the patients receive (the cards depicting animals

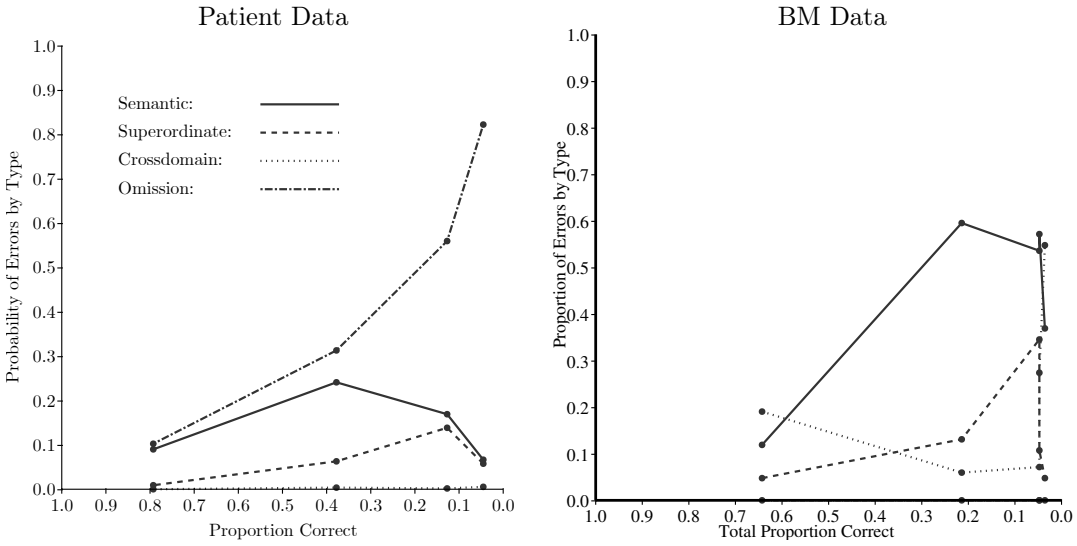


Figure 3.26: Results of the confrontation naming task of BM.

or objects using line-drawings) and the name label they produce. Following the original task design, naming involves clamping the visual units (representing the input to semantics when looking at a picture) and then allowing the network to cycle for twelve settling steps (cf. Rogers et al., 2004, p. 217). As described in Rogers et al. (2004), when real-valued activations are used (as in the BPTT implementaions), the name unit most activated above a threshold of 0.5 is considered to be the network's response. When a unit is above the threshold, the name it represents is taken to be the reply, and when all units are below 0.5 the network is considered to be anomic for the given visual input. On the other hand, when binary states are used, as in the case of the Boltzmann machine, the thresholding method of determining a response is not applicable. Instead, euclidean distance was used as the distance metric between the actual with the target output. Euclidean distance gives similar results to the original method when used as a distance metric in the BPTT networks and, unlike the original method, is applicable to the BM, which has binary-valued units activated probabilistically. So in the case of the BM, the response is derived by finding the pattern closest in Euclidean space to the name output. As described, this approach cannot produce omission errors, although they could be simulated if a threshold on the distance were added. However, the more critical aspects of the data is the rank ordering of the error types.

### 3.3.3 Results

The model reported by Rogers et al. (2004) reproduces the four important qualitative features of naming found in their fifteen SD patients. Firstly, the overwhelming proportion of errors consists

of *omissions*. Omission errors are seen to increase with the progress of neurodegeneration. Secondly, *semantic errors* initially start off low, then grow to about a quarter of responses, and finally return to a lower proportion. Thirdly, *superordinate errors* show the same pattern as semantic errors. Although at all levels of lesioning superordinate errors are lower than semantic errors, reaching only about a tenth of all responses at their highest proportion. Fourthly, *crossdomain errors* are extremely rare.

The $BPTT_1$, $BPTT_2$, and BM naming graphs in figures 3.25 and 3.26 show only a partial replication of the naming task scores as discussed above. Firstly, omissions are lower than semantic errors, but in fact they should be consistently above all other error types. Secondly, semantic errors are proportionally the highest error type. Thirdly, superordinate errors are qualitatively a good fit. Fourthly, crossdomain errors occur, when instead they should be at floor levels. This pattern of responses persists even if the value of the threshold, which determines the proportion of responses that are classified as omissions, is varied.

With respect to the models, the largest proportion of errors from 10% to 70% of weights lesioned are cross-domain errors. This means that name units corresponding, for example, to artifacts are activated when an animal is visually presented to the network and vice versa. Omission errors are defined by Rogers et al. to occur when the network fails to activate any name unit beyond a threshold of 0.5. Changing this threshold affects the relations between the error types, but does not result in a better fit to patient data. The greater the threshold the more errors are classified as omissions, and thus the remaining three kinds of naming error (semantic, cross-domain, and superordinate) are fewer; the inverse also holds. In conclusion, the reimplementation of the hub model on the naming task does not recreate the error pattern seen in the patients.

## 3.4 Sorting words and pictures

### 3.4.1 Patient sorting

The sorting task is used to determine the preservation of hierarchical conceptual knowledge in patients. The patients are given cards with line drawings or words on them and they are requested to sort these: generally into the two domains of living and inanimate objects; and specifically into each of the five categories. (For unclear reasons, fruit is exempt from the majority of patient tests in Rogers et al. (2004).) At general-level picture sorting patients score highly, and appear to remain near or at ceiling even though they proceed to lose other semantic abilities as their disease progresses. (Note that the horizontal axis on the figures reporting

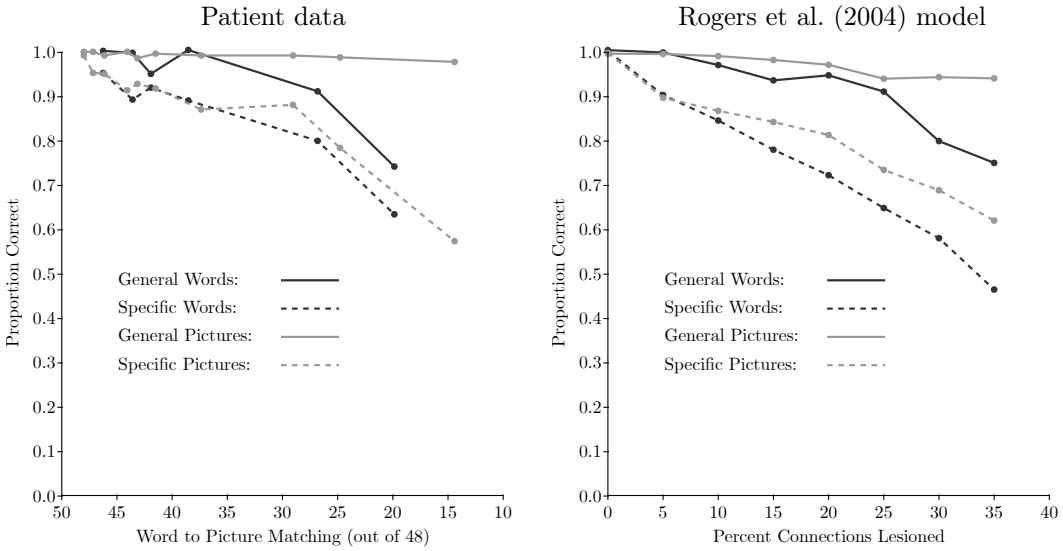


Figure 3.27: Results of the SD patients and the original model sorting task on words and on pictures. (From fig. 8., Rogers et al., 2004)

results is the patients' score on another semantic task.) For all other forms of sorting no such preservation can be detected; the patients' scores seem to be declining towards the respective chance levels for each of the sorting types.

Figure 3.27 shows the the original model behaviour next to the SD patients. The model reflects the pattern of the twelve patients tested by Rogers et al. (2004) on this task, in particular: *a*) the sorting of pictures is more preserved than that of words; *b*) sorting at a general level is retained more so than specific sorting; and *c*) the ability to classify pictures into their respective domains is largely unaffected by lesioning.

### 3.4.2 Model sorting

This task requires the network to classify name and visual sub-patterns into their respective categories and domains. This is done by clamping the input to the appropriate name, if sorting a word, or to the visual sub-pattern, if sorting a picture, and allowing the network's output to reach a final state. So either by settling, if a BPTT network, or annealing, if a BM, the network decides on the input by activating an encyclopedic unit that represents domain or category classification, meaning that once the network is in a stable state, the verbal units which represent category or domain membership are examined (cf. Rogers et al., p. 220). These encyclopedic units are pre-set category-/domain-level features for each pattern that have not explicitly been taught to the network, but have been learned as part of the whole verbal sub-pattern. The two most active of these mutually exclusive units (one for domain and one for category) are selected to be the network's response for the two levels of sorting, thus ensuring

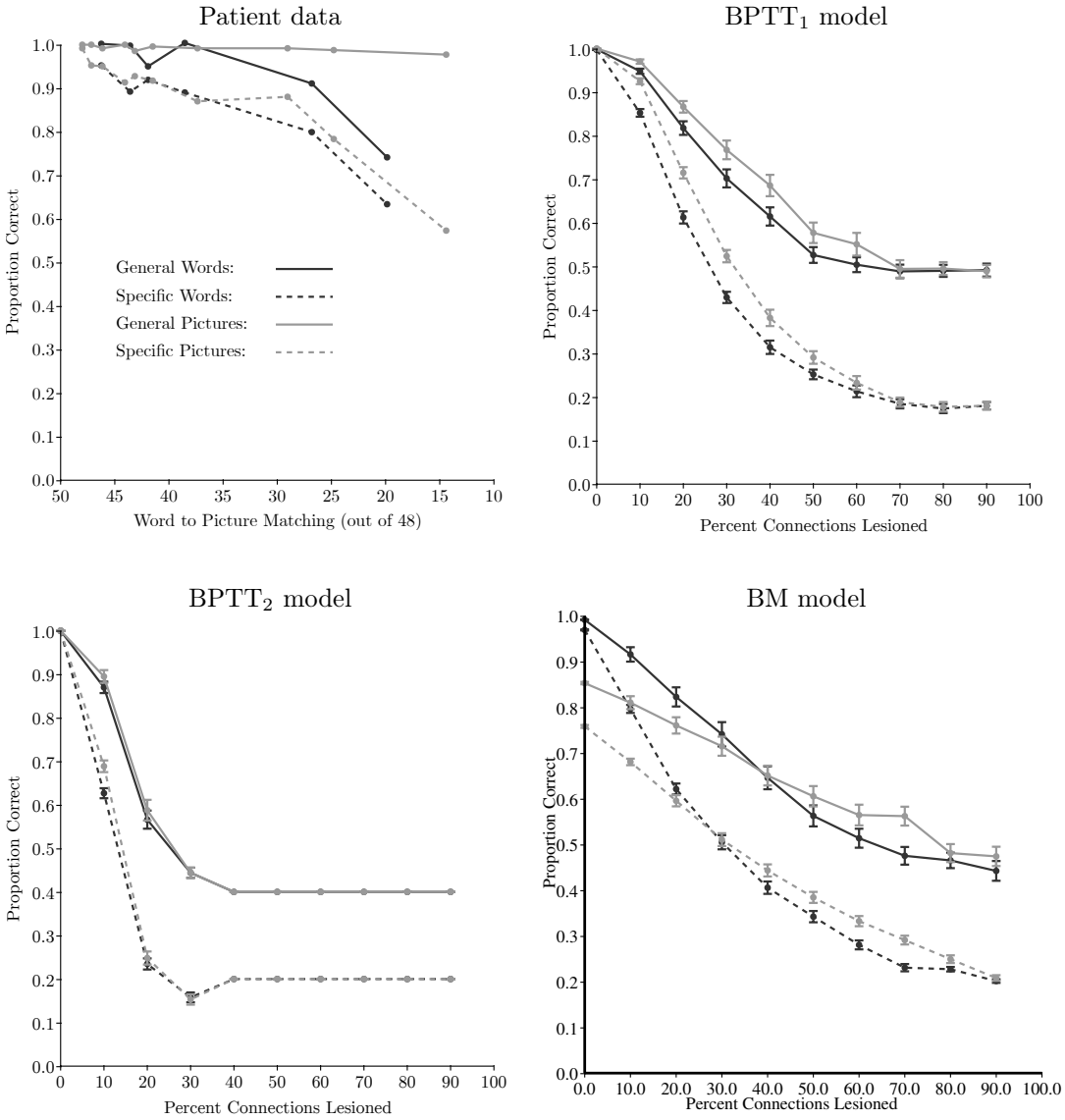


Figure 3.28: Results of the SD patients and all three models in the sorting task on words and on pictures. Errors bars represent one SE.

the task is forced-choice.

### 3.4.3 Results

In Figure 3.28, graphs of the networks' performance at sorting at increasing levels of lesioning is shown. The scores for the two general levels of sorting (represented as solid lines), for words and for pictures, follow a descent from correct to chance levels. This is expected due to the architecture of the patterns: there are two encyclopaedic units that represent the mutually exclusive facts "is an animal" and "is an artifact". In much the same way, the network's scores on the two specific sorting tasks also appear to deteriorate to chance level, this time as there are 5 categories to choose from chance is at 0.2 (as in Rogers et al., 2004, fruit is excluded in

the testing phase).

These results are relatively similar to those produced by the 12 patients tested by Rogers et al. (2004), however, there appears to be an important difference: the SD patients retain the ability to classify pictures into their respective domains well into their illness. Thus, while sorting into lower level categories is a skill that is largely lost, the two main semantic domains remain intact in SD; this also can be seen in fig. 8 of Rogers et al. While the original hub model appears to capture this dissociation, the current implementation does not. Arguably, the sorting of pictures is slightly more preserved than that of words, in Figure 3.28 (for all models), but the SD patients are all at ceiling. Again, the models are unable to fully capture this pattern of SD patient performance.

The $\text{BPTT}_1$ results, shown in the top left of Figure 3.28, indicate that the last of the properties mentioned in subsection 3.4.1 is absent: scores in general picture sorting should be near or at ceiling even after substantial (40%) lesioning. The graphs for the $\text{BPTT}_2$ and BM models do not display this property either, but nor do they consistently show the other two qualitative effects, see Figure 3.28.[1] In contrast to the original model, the scores of the three reimplementations for all types and levels of sorting tend towards baseline values. (Recall that chance for category-level sorting is 0.2 and for domain-level is 0.5 – any slight deviation from these is due to the values of the bias units.) Rogers et al. (2004) propose that their model of the sorting task is able to follow the patients' scores because "the effect of damage must be quite severe before the system begins to generate incorrect verbal information about such properties" (Rogers et al., p. 220). This does not appear to the case in the reimplemented models.

The $\text{BPTT}_1$ reimplementation manages to show a partial replication, however the $\text{BPTT}_2$ and BM do not reflect any aspect of the patient scores consistently. So while in the original model the "difference in the nature of the mapping between surface form and conceptual representations [...] underpins the difference in performance for word and picture sorting" (Rogers et al., 2004, p. 221), this does not hold as strongly for the $\text{BPTT}_2$ and BM. In addition, in the original hub model "[a]rbitrary mappings are more vulnerable to damage than are systematic mappings" (ibid, p. 221), meaning that word sorting is more fragile than picture sorting; however, this also does not generalise to all our reimplementations.

Qualitatively equivalent naming and sorting scores as those seen in the reimplementations are found over many instances of $\text{BPTT}_1$, $\text{BPTT}_2$, and BM networks (i.e., the results are not an

[1] Note that the BM does not achieve 100% accuracy on this task even when not lesioned. This is because the BM is inherently stochastic, so the probability of a single unit, e.g., the unit corresponding to $\langle mammal \rangle$, being on is a function of the co-occurrence of that unit's state and every other unit's state in the network, which will always be less than one.

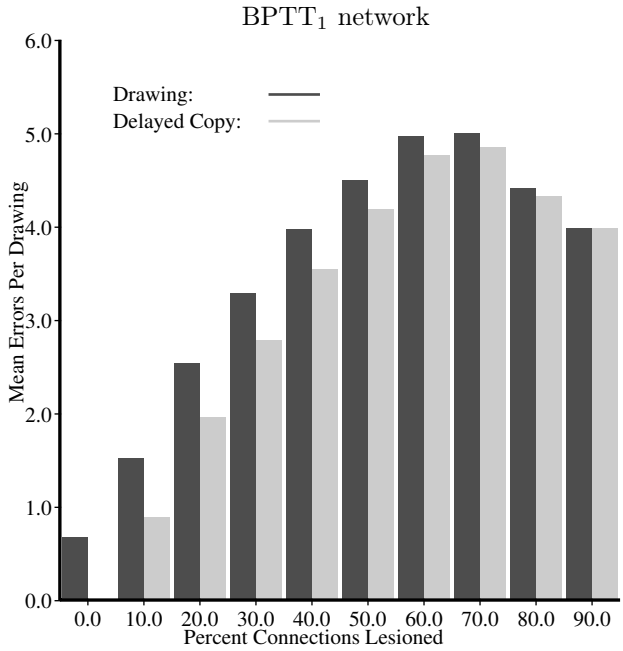


Figure 3.29: Mean overall feature errors per drawing for the drawing and delayed copying task for each lesioning level. Error bars not included because SE $< 0.002$. (Compare with Rogers et al., 2004, fig. 11.)

artifact of one set of trained weights). In addition, the training algorithm of the BPTT networks has been varied between epoch-wise, pattern-wise and sub-pattern-wise (weights updated after each name, verbal, visual sub-pattern) and it has been found to also produce qualitatively equivalent naming and sorting graphs.

A final issue concerns the way the network's performance is evaluated on this task. In all the other tasks, the network performs the same input to output mappings as those it was trained on. In contrast, this task tests the network on classification (i.e., auto-association). A more appropriate approach to the task might be to train and test the network specifically on classification as well as auto-association.

## 3.5 Drawing and delayed copying

### 3.5.1 Patient task

This semantic test has two parts, the first requires the patients to draw an object given its name, and the second part involves them creating copies of drawings after a delay. The original patients also carried out a direct (as opposed to delayed) copying task, during which they could view the original line-drawing while they drew their version. Even though only three patients are tested on this task, a qualitative trend is definitely evident: patients make more errors on

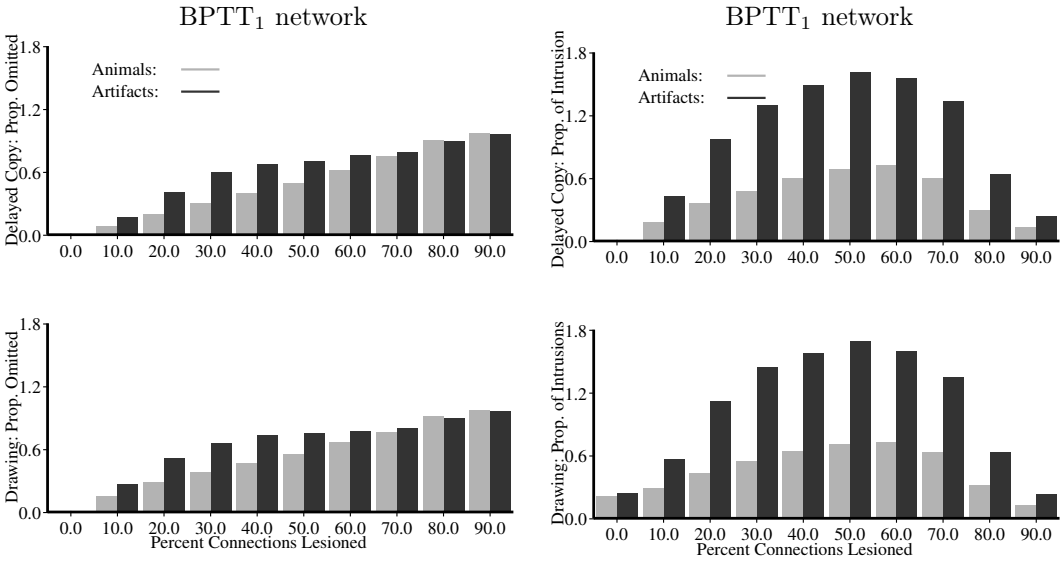


Figure 3.30: Proportion of omission (left) and intrusion (right) errors per drawing for each domain in the drawing and delayed copying task. For omissions, error bars not included because SE $< 0.003$; and for intrusions, the same but SE $< 0.105$. (Compare with Rogers et al., 2004, figs. 12-13.)

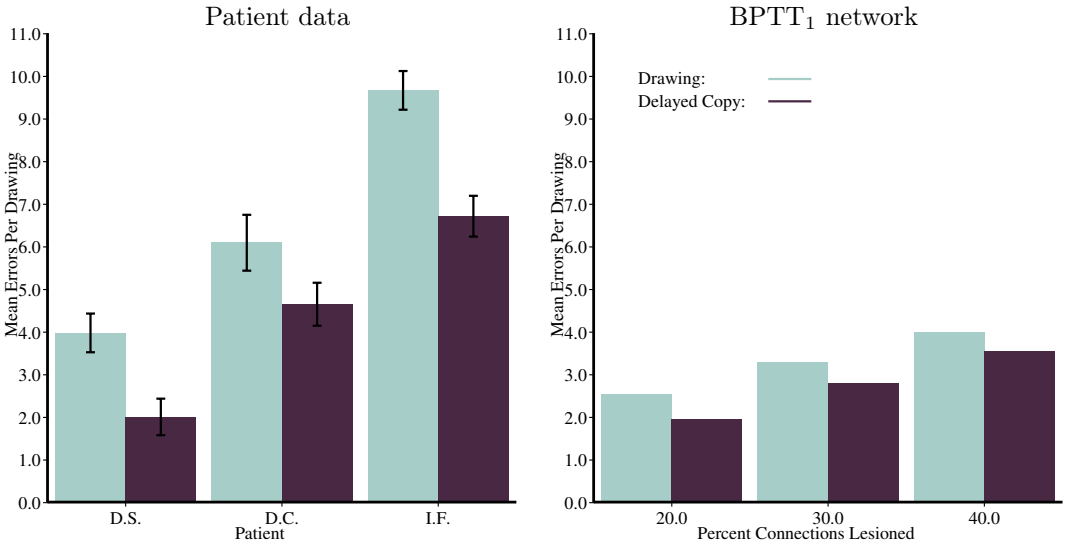


Figure 3.31: Mean overall feature errors per drawing for the drawing and delayed copying task for each lesioning level. Error bars not included because SE $< 0.002$. (Compare with Rogers et al., 2004, fig. 11.)

drawing than on copying. The networks, and the hub theory in its current state cannot perform simultaneous in/output from/to the same modality. For this reason the networks cannot model immediate drawing — only drawing and delayed copying can be modelled.

### 3.5.2 Model task

The task involves the input to be either clamped to the visual or name sub-pattern then for the specific type of clamped and unclamped settling to take place, when the network is in a stable enough state[2] the visual output is interpreted.

### 3.5.3 Results

The results obtained from running the drawing and delayed copying semantic test on the reimplementation (see Figure 3.29) appear to qualitatively match those in fig. 11 of Rogers et al. (2004). Both SD patients and the model show an increase in the errors they make when drawing and copying. Also the difference between drawing and delayed copying, that the former is more difficult than the latter per patient, is reflected in both the original model and our reimplementation.

However, when the results are further analysed, as in Figure 3.30, a different picture emerges. Rogers et al. (2004) argue that there is an underlying distinction between the scores in each domain for two kinds of error: an *omission*, a salient feature that should have been drawn but is left out by the participant (e.g., forgetting to depict a swan with wings); and an *intrusion*, a property that perhaps holds for most exemplars but is incorrectly included in the drawing (e.g., adding four legs to a swan). In the patients' drawings there are significantly more intrusions for animals than for artifacts (Rogers et al., 2004, p. 227), but no such effect for omissions. In fact, the original hub model only partially reproduces these effects, correctly showing more intrusions for animals but incorrectly showing more omissions for artifacts (see figs. 12-13 in Rogers et al., 2004). In our reimplementation of the hub model, we found that omission errors (both when copying and drawing) are higher in artifacts than animals (see Figure 3.30), though with increased lesioning severity omission errors seem to occur equally in both domains (in contrast to patients, who show no effect; while the original hub model shows the same effect as our model). The rate of intrusion errors by domain reflects neither the patient data nor that of the original hub implementation, with more intrusion errors for artifacts than animals over most of the range of lesion severity (see Figure 3.30).

## 3.6 A revised implementation of the BPTT model

The difference between our implementation and that of the original Rogers et al. (2004) model is both concerning and puzzling. It is concenring because it undermines the support of the

[2]The lesioned network does often oscillate between two or more attractors.

hub-and-spoke theory of semantic cognition. It is puzzling because the BPTT implementation sought to be faithful to the description give in Rogers et al. (2004). However, that description was lacking in detail in some respects. For example, the training algorithm is specified as "a variant of the backpropagation learning algorithm suited to learning in a recurrent network (Rumelhart, Hinton, & Williams, 1986)" (Rogers et al., 2004, p. 208) and at no point do the authors provide further information other than epochs and learning rate. This is problematic because the algorithm of Rumelhart et al. (1986) is not suitable for recurrent networks.

Moreover, there are some internal inconsistencies in Rogers et al. (2004) in their description of the original model, e.g.: the number of input units (the width of the patterns) differs between their figure 1, which shows the input pools for the model, and their figure 3, which shows the pattern prototypes for creating training and testing sets; the labels associated with patterns vary between their figures 4 and 5; and the learning rate and training epochs are dramatically different to those reported in Lambon Ralph et al. (2007), which is described as an implementation of the same model. Specifically, "[t]he model was trained with a learning rate of 0.005, without momentum, and with a decay parameter set to 0.001 to prevent individual weights from growing disproportionately large. Training proceeded for 400 epochs, at which point the model had learned to generate a steady state for all inputs in which all verbal and visual units were within 0.05 of their target states" (Rogers et al., 2004, p. 215). While in the other implementation of the hub-and-spoke model "[t]he model was trained with backpropagation over time for 10 000 epochs, using a learning rate of 0.005, at which point it was able to activate all visual and verbal units to within 0.2 of their target states for all patterns" (Lambon Ralph et al., 2007, p. 1130).

Due to the problems encountered during replication, clarifications to the original Rogers et al. (2004) paper were requested from the authors. Tim Rogers provided us with the original learning rate that was: 0.005; and the original weight decay: 0.999. Previously, a learning rate of 0.001 was used (other values were explored), however the decay rate is identical. Another point that differs is the manner in which the weight adjustments are applied, in the original model "weights were updated online after every pattern presentation [and] not in a batch following every epoch" (T. Rogers, personal communication, August 21, 2012). Perhaps most importantly of all, the original pattern set has a distinct structure from that of the the training sets that can be generated using the probabilistic prototype reported in Rogers et al. (2004). To address these differences, a further set of models was trained that takes these details into account.

Another difference, as a direct consequence of the pattern set provided by T. Rogers (personal communication, August 21, 2012), is that the input units need to subtly change in quantity

for some of the input pools. The name units remain unchanged as they are a function of the number of input patterns (48 patterns with 40 unique names, thus 40 name units), the visual units also are unchanged in quantity (64). However, the verbal units are require modification to accommodate their respective sub-patterns. The verbal units are now divided into 64 perceptual, 32 functional, and 16 encyclopaedic units, giving a total of 112 verbal units.

This new distribution of units amongst the different pools is in line with the high-level diagram of the hub model seen in Figure 1.2, but not with the number of units described in the probabilistic prototype for creating patterns in Figure A.86, nor does it match figure 4 in Rogers et al. (2004). Nonetheless, it is not expected that a significant difference in the network's behaviour will be reported due to the changes in input vector length, since the previous and new numbers of units are very close: perceptual units change from 61 to 64 and encyclopaedic from 18 to 16, meaning that the total number of verbal units has only increased by a single unit. On the other hand, what might cause a change in the processing of the network is the inter-correlations between the pattern bit vectors. It remains to be seen how much difference exists between the patterns, as acquired from T. Rogers, and those derivable from the prototype, as documented in Rogers et al. (2004).

The training algorithm used here is the same as that used in Method 4, for more details see section A.7, with the learning rate is set to 0.005. In order to explore what the effect, if any, of changing the frequency of weight adjustments would be to the network, three different networks were created, keeping the pattern set and learning rate constant: $\text{BPTT}_{\text{sub-pattern-wise}}$ weights are adjusted after the presentation of each sub-pattern, so per epoch the weights are changed $3 \times 48$ times; $\text{BPTT}_{\text{pattern-wise}}$ weights updated once per pattern per epoch; and $\text{BPTT}_{\text{epoch-wise}}$ at the end of the epoch the weights are updated, as previously in Figure 2.4.3 with $\text{BPTT}_1$ and $\text{BPTT}_2$.

Unfortunately the pattern set provided by T. Rogers (personal communication, August 21, 2012) does not conform with the requirements for performing the sorting tasks, which are that domain and category bits need to be set per pattern; see section A.2 for previous training set details. This pattern set also, as mentioned above in section 3.6, has a slightly different structure both in regards to the number of units per sub-pattern and in regards to the distributions of activated features. For details on the internal structure of all pattern sets used in this chapter see section 3.7, which analyses and discusses both the patterns and the attractors that emerge in the networks trained on those patterns. In line with the previous two names, this pattern set will be called set $C$.

The naming task, as carried out previously in section 3.3, was run on the three versions

of the hub model trained on set $C$, and resulted in similar results as those obtained before; specifically crossdomain errors were still present in substantial quantities. The results for the drawing task on set $C$ show the same qualitative effects in the models as in the patients, as before in section 3.5. In other words, the models exhibit a general increase in errors as lesioning progresses, with the deterioration more pronounced in drawing than in delayed copying. However, the network reaches the baselines for each drawing, i.e., so it falls to the level of an untrained network's score on this test, before reaching the approximately 6/5.5 and 9.5/6.5 mean errors for drawing/copying seen in patients D.C. and I.F. What this means is that while the trend of increasing errors is present in the models, all models presented here are unable to capture the patients' scores in the quantitative way the original Rogers et al. (2004) results do. This is presumably due to the fact that the drawing errors are either omissions or intrusions and the network cannot continue to activate units when lesioned.

As noted above, the sorting task cannot be performed due to the structure of pattern set $C$. It is a requirement (see section 3.4) that there exist orthogonally defined units that indicate category membership, however these were not present in the training set sent to us by T. Rogers (personal communication, August 21, 2012).

## 3.7 Attractors in the model

Recall the important role attractors play in the hub-and-spoke model, as mentioned previously in subsection 2.4.3. The process of formation and the stages of breakdown are purported to parallel the equivalent behaviour seen in semantically impaired patients. Specifically, Rogers et al. (2004) propose the attractors for animals generalise with damage whereas those for inanimate objects retain some individuality when damage occurs. This difference in the breakdown of attractors is argued to be a consequence of the differences in distribution of features in the input to the network. As such, this section aims to examine how attractors relate to the pattern set, and what effects damage has on them.

As can be seen in Figure 3.32 both pattern set $B$ (created based on the prototype in Rogers et al. (2004)) and pattern set $C$ (provided by T. Rogers, personal communication, August 21, 2012) share a similar overarching hierarchical structure. For set $B$ fruit, vehicles, tools, mammals and birds are dissociable categories; for set $C$ this is the case for tools, birds, and mammals.

As shown previously in section 2.4, Figure 3.33 depicts the hierarchical cluster dendrograms of the hidden unit states after training and settling for the two types of networks previously

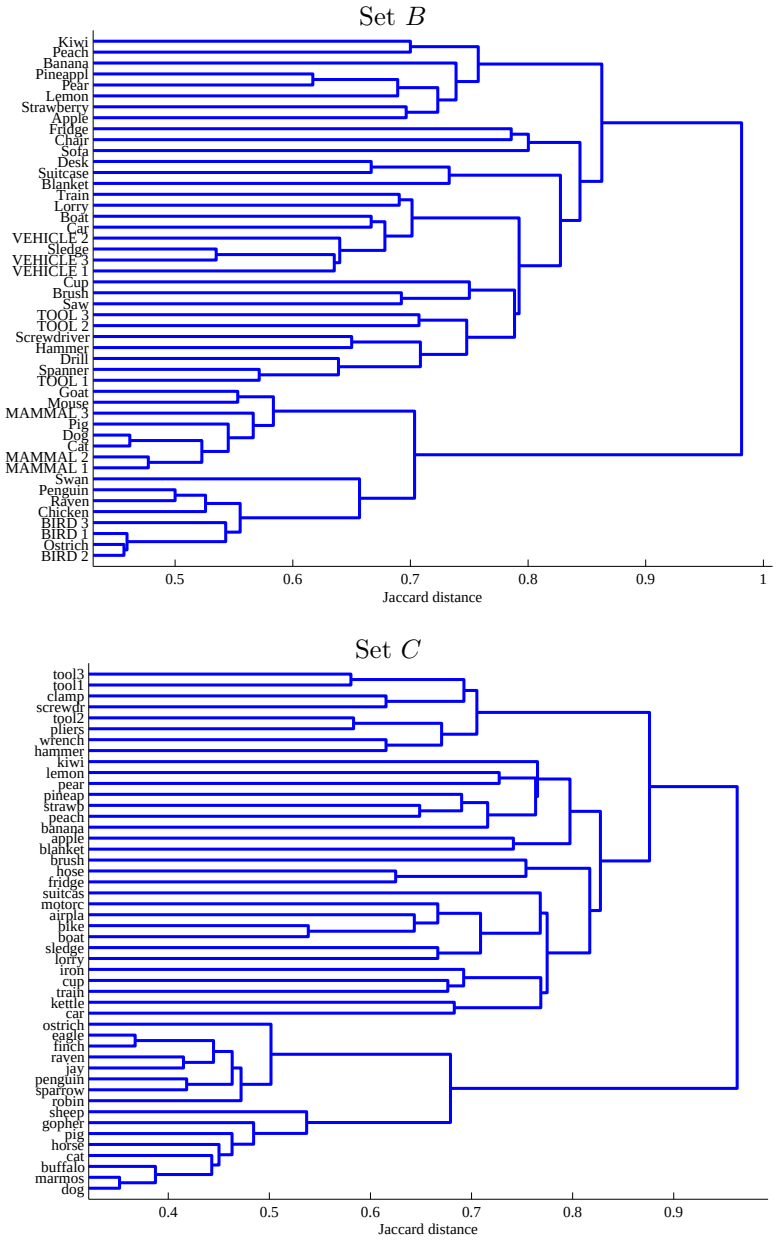


Figure 3.32: Dendrograms for whole input patterns of set $B$ and $C$.

discussed, $\text{BPTT}_1$ and BM. The BM shows a similar qualitative structure to the two BPTT networks in that all three models classify fruit as an inanimate category, although, importantly the BM does not show the same micro-structure in terms of the similartity within the two domains. In bottom half of 3.33, the animals are more different to each other than the inanimate objects, while in the other two networks and the original Rogers et al. (2004) model the opposite is the case, as required by the theory. In other words, even though the patterns have a certain macro- and micro-structure with the two domains differentiated and the animals more similar to each other than the artifacts, it is not sufficient to ensure that networks trained on them

also have this structure.

After zeroing damage the networks we have implemented compared to the original commit crossdomain errors, meaning that the two domains are not preserving their differences. As can be seen in Figure 3.34, which shows the similarity structure of hidden unit representations in the damaged network, certain concepts have now lost their correct domain membership, e.g., ⌜*cup*⌝ is not categorised within the inanimate domain by the network. This is important because crossdomain errors must be avoided all the way through zeroing damaging in order for the model to successful simulate the SD patient data. Recall that as shown in Figure 3.24, neither the original model nor the patients make naming errors that involve producing an output that is in the other opposing domain. This is the main issue with respect to replication in all variations of BPTT networks and BM networks we have attempted.

## 3.8 Discussion

Rogers et al. (2004) presented a model of the semantic system which they argued could account, when lesioned, for many of the deficits associated with semantic dementia. In support of this argument they report a number of simulations. We have attempted to replicate these simulations, but with mixed success. Thus, while we were able to recreate the basic learning performance of the intact model, we were unable to fully reproduce the patterns of behaviour seen in the lesion studies.

Rogers et al. (2004) parallel the emergence of attractors with the learning of concepts, and propose that such knowledge is amodal: the somato-sensory input from the various modality-specific pathways is encapsulated by the hidden units, which thus form semantic representations. This basic theoretical notion is successfully captured by the hub model. For the case of the deficits seen in their SD patients, Rogers et al. appeal to the attractor basins' properties post-lesioning (zeroing of connection weights). They claim that animals form a tight cluster of similar concepts, thus consisting of many neighbouring attractors, while attractors for artifacts are distal (to the average central point of their domain), which means they form distinct conceptual loci in semantic space, and therefore their attractors are further apart. When connections are zeroed the attractor basins for living creatures are held to decay to form a larger super-attractor, which has a combined attractive power; meaning categorisation of input as an animal is possible, but access to individual features might be lost. Conversely, the attractor basins of non-living things do not merge; instead they maintain their individual attractors, albeit with distorted basins, allowing slightly better performance in this domain. The evidence put forward for this

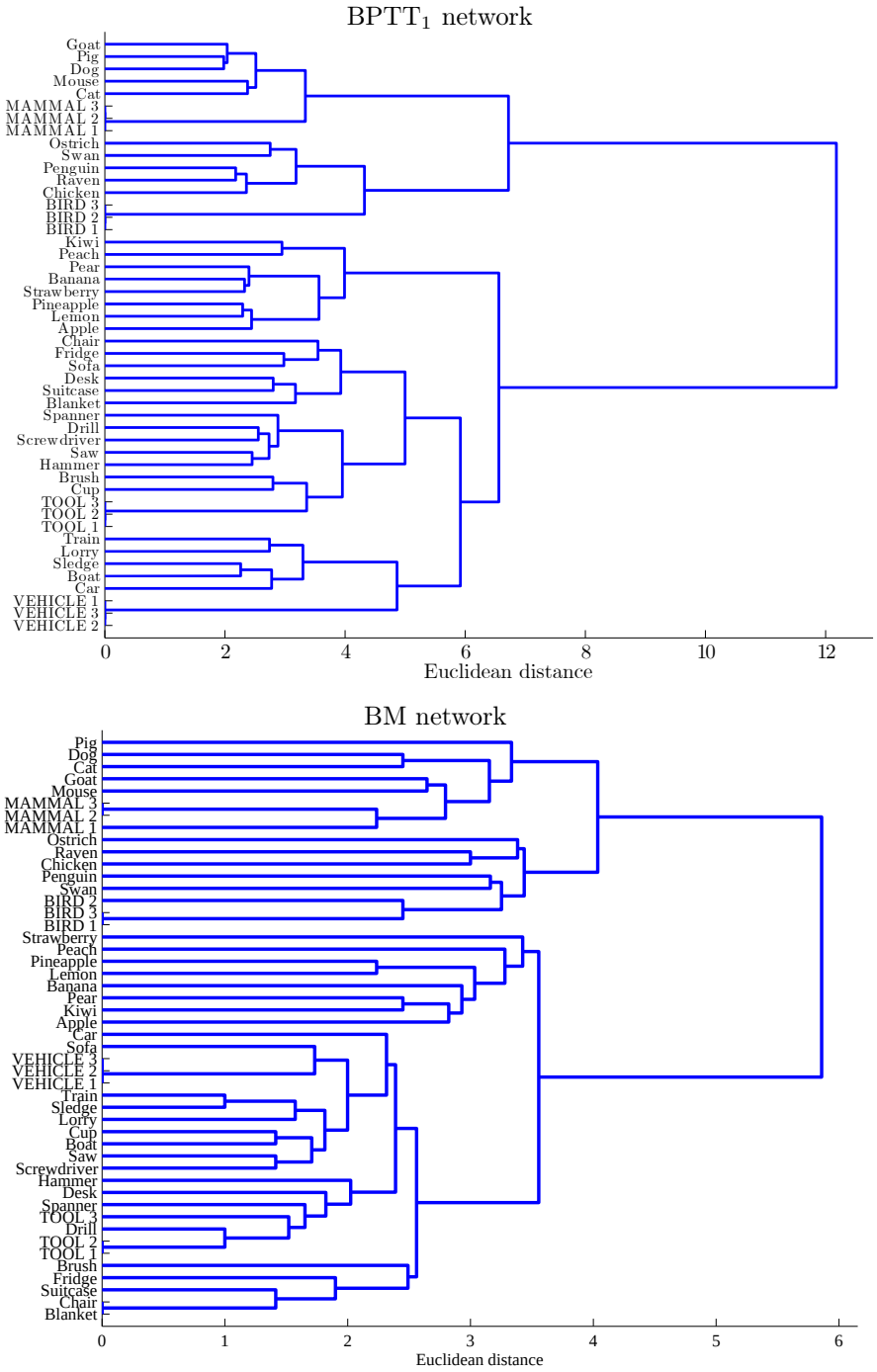


Figure 3.33: Dendrograms for the BPTT$_1$ and the BM networks' internal states given name trained on set $B$.

phenomenon is the series of graphs generated from testing the Rogers et al. model. Yet the behaviour reported in the original hub model is not found in the network trained here. Why might this be so?

One possibility is that there is an error in our replication. We do not believe this to be the case, particularly given that we have simulated the basic learning performance of the network.

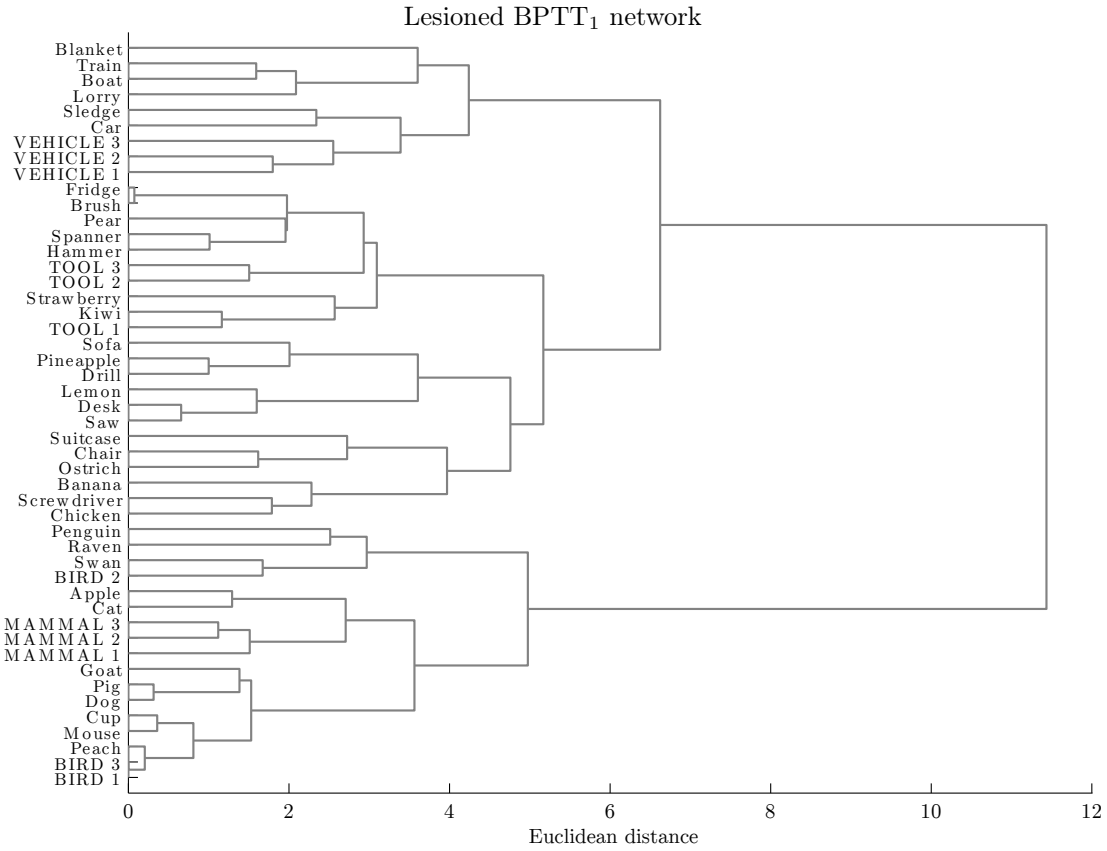


Figure 3.34: Dendrogram for the BPTT$_1$ network's internal states after 30% of weights set to zero. Shown in grey to differentiate from the intact states in the previous figures.

A second is that the difference in results relates to some difference between, for example, the learning algorithm as implemented here and as implemented by Rogers et al. (2004). This is certainly possible, given that the algorithm is not fully described in the original publication. A third is that the attractors formed by the model are dependent upon the initial random weights of connections prior to learning or the order of exemplars in the training set. However, if either of these latter two situations is the case then it calls into question the theoretical explanation offered by Rogers et al. for their results.

An important aspect of this modelling strategy, that is related to the formation of attractors, is the claimed distribution of pre-semantic (perceptual and functional) features: animals and plants are closely perceptually related to each other (due to the fact they have evolved from a common ancestor and thus are composed of generally similar body parts); whereas tools, vehicles, and other inanimate objects are not perceptually similar to each other (as they have been created by humans to solve different problems, so by definition artifacts are distinct from both living things and from each other). Without training sets that encode patterns in this specific way, no connectionist model would be capable of producing a good fit to patient data.

On this argument, the features, whose extraction from the environment itself is not modelled, play a pivotal role in giving rise to the semantic system's structure, and this is the case regardless of the network topology (be it recurrent or feedforward) or the learning algorithm. This is to say that, to a large extent, input to the semantic system should drive its organisation and also dictate the way semantic knowledge will decay. Despite this fact, the patterns used here are unable to affect the internal structure of the reimplemented hub model in the way needed when the network is damaged. This means that the qualitative and consistent effects required post-lesioning are in fact *not* guaranteed merely by the structure of the training set; even though qualitative similarity existed of the attractor structure in our reimplementations and the original model of Rogers et al. (2004). It appears that lesioning the recurrent network model by severing connections does not necessarily result in the kind of well-behaved breakdown and generalisation of attractors as supposed by Rogers et al.

To summarise, the differences between the models appear to be due to the results obtained in Rogers et al. (2004) depending on some unarticulated implementation detail. If this is so, then the required behaviour is not a necessary consequence of the model − the original model is underspecified. It seems our implementation of the BPTT algorithm yields attractors with different properties to the implementation of Rogers et al. (2004). Alternatively, it may be that the behaviour of the network when damaged depends upon, for example, some apparently irrelevant factor such as the random initialisation of the connection weights. Whatever the underlying cause of the discrepancy, further investigation is needed to discover exactly why the results obtained here differ from most of those detailed in Rogers et al. If their results are in fact reproducible, but require a very specific set-up, this suggests that the model as previously reported is insufficiently specified. Conversely, if the success of the original model is due to an artefact or randomly occurring noise then this indicates that in models of this type it is critical to present results from multiple trained models, rather than from just one, to establish whether behaviours are a necessary consequence of the model or merely one of several possible outcomes.

The results of the confrontation naming task run on the three different implementations show that internal representations do not decay in a way that replicates the patients' behaviour. So while intact naming is possible, at least within the BPTT reimplementations, the predictions made by Rogers et al. (2004) are not met. Specifically, they claim that "[w]ith increasing damage, the model becomes unable to generate any information that individuates items from the same broad domain, and representations within a given domain collapse into a single general attractor from which the model produces only those properties common to the majority of items in the

domain. [That is to say, t]he model never names an object with a completely unrelated label, because such names apply only to objects with very distal internal representations" (Rogers et al., p. 218). However, we can see from both the damaged semantic representations $BPTT_1$ has, discussed in section 3.7, and from the models' naming behaviour, that concepts from opposing domains can become much closer to each other than (what should be) neighbouring concepts. This is why a larger proportion of crossdomain errors are produced: attractor dynamics do not necessarily follow the predictions set out by the hub theory.

## 3.9 Summary

This chapter presents and discusses both direct and conceptual replication attempts of the hub-and-spoke model, using two broad families of recurrent networks and training algorithms (backpropagation though time and Boltzmann machine networks). The reimplementations are evaluated using the same tasks as those used in Rogers et al. (2004), namely computational equivalents to naming, sorting and drawing semantic tasks. During the course of reimplementation, we found a disparity between the Rogers et al. (2004) results and ours. On the one hand, we were successful in reproducing the healthy semantic system, but on the other hand the behaviour of semantically impaired patients was not fully captured by our reimplementation. These differences between ours and Rogers et al. may be due to implementation details not mentioned in the original paper. The repercussions of non-replicability on the hub-and-spoke model are discussed, as are potential causes of this problem.

# Chapter 4

# Modelling general semantic deficits in the SOM-based models

## 4.1 Overview

In this chapter the two models that use Self Organising Maps (SOMs) and their implementations are discussed with respect to modelling generalised semantic deficits such as those found in semantic dementia. The first is the modality specific model, as described in section 2.5, and the second is the conceptual structure model, previously discussed in section 2.6. While they are theoretically distinct, they share some model- and implementation-level details. The two models are run on equivalents to the naming and sorting tasks that the hub-and-spoke model and the SD patients carry out, with results comparable to those discussed previously in chapter 3. The failures and successes in capturing the patient data, which mirror the hub-and-spoke BPPT and BM networks, shed some light on the failures in both the previous chapter and the current.

## 4.2 Introduction

Recall the modality specific model, depicted in Figure 4.36, and the conceptual topography model, seen in Figure 4.37. These two accounts have been implemented using SOMs (Self Organising Maps) augmented with classical feedforward connections. Both of the models discussed here have been trained on the Rogers et al. (2004) patterns, and therefore they shall be tested in analogous ways on the same semantic tasks, as applicable. Due to the shared training set, comparison of the hub model with the two models in this chapter is facilitated. In

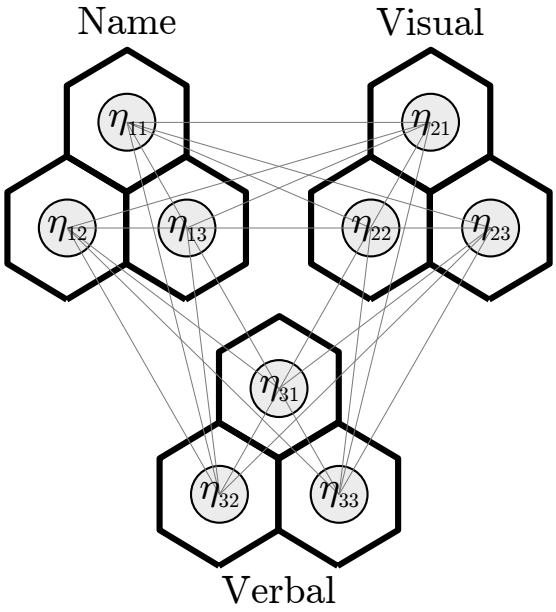


Figure 4.36: This is a schematic of the modality-specific model, simplified to only contain a few units. Each of the modal SOMs is bidirectionally connected to the other two allowing for input and output from each layer to the rest.

addition, the hub model and the conceptual structure model's architecture share many topological features. This means that the conceptual topography model, which is a SOM-based model, can be seen as a hub-like model and evaluated using the same criteria. Furthermore, the hub model's patterns can be evaluated themselves under radically different architectures of the hub-like SOM-based model and a DISLEX-like SOM-based implementation, which respectively underpin the conceptual structure and the the modality-specific models.

The modality-specific theory offers some predictions and modelling guidelines and implementation constraints. Firstly, modality-specific models must allow for different modality-specific stores, see Figure 4.36. This requirement has already been accomplished by the architecture. Secondly, the model must localise the different types of (modality-based) features into different stores. As with the previous point, this is enforced by the training regime and architecture. Thirdly, damage to an individual store must not dramatically affect the remaining semantic feature stores from functioning, thus giving rise to modality-specific deficits, like those seen in HSVE patients. Finally, damaging these stores globally, in a way that parallels SD patients' lesions, must result in what appears to be a panmodal or amodal deficit, in accordance with the reported behaviour of SD patients. These modelling requirements will be the main focus of investigation in the next few sections.

The conceptual topography theory is not fully captured by the conceptual structure model proposed here — this is because the theory is an almost complete account of cognition from the

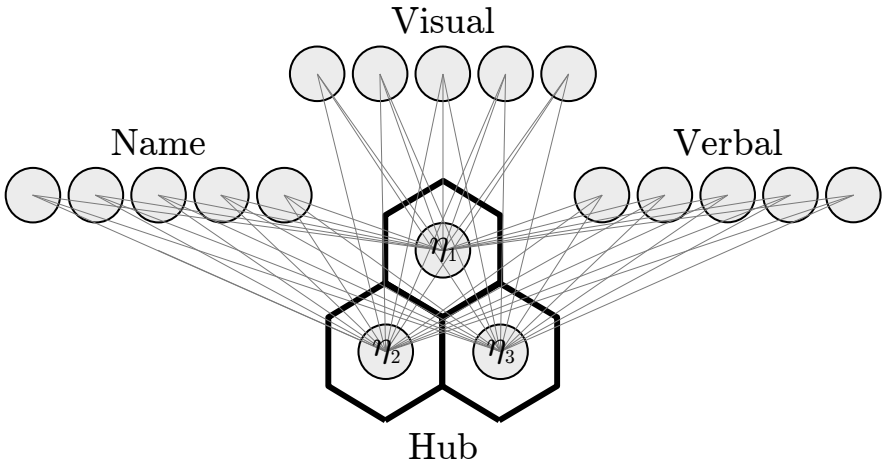


Figure 4.37: This is a toy model with the same architecture as the conceptual topography model. The three input pools are connected bidirectionally to the hub units. Th SOM is trained on the whole pattern, while the three input pools are trained to link each subpattern to its activity as calculated by Equation B.10 on the surface of the hub map.

lowest to the highest levels of analysis — but two of the key features of the conceptual topography account are, namely that (1) sensory information is stored in the separate modalities, and (2) cross-modality areas exist that store information that store information in topologically-based ways (the SOM), see Figure 4.37. What is missing are the layers below the high-level model features which form the substrate out of which they emerge, due to the fact that patterns that capture retinal cell information are largely outside the scope of this thesis[1]. Conceptual topography accounts, as well as embodied accounts in general, do not offer specific behavioural, e.g., semantic task, predictions (see section 2.6), over and above those offered by the modality-specific account. Notwithstanding, the conceptual topography theory (and embodied cognition in general) does indeed offer predictions regarding cortical organisation, as do most semantic system theories to a certain extent.

## 4.3 Modality-specific model

### 4.3.1 Confrontation naming task

The naming task in the modality-specific model consists of first lesioning the appropriate connection weights, so only zeroing percentages of weights that are taking part in this task, i.e., the unidirectional visual to name connections. This is unlike the Rogers et al. (2004) method of modelling this task which lesions increasing percentages of all the weights. This is because settling in the Rogers et al. (2004) model makes use of the majority of connections because the

[1]Recall that the Rogers et al. (2004) patterns are (pre-)semantic, and not lower-level perceptual features.

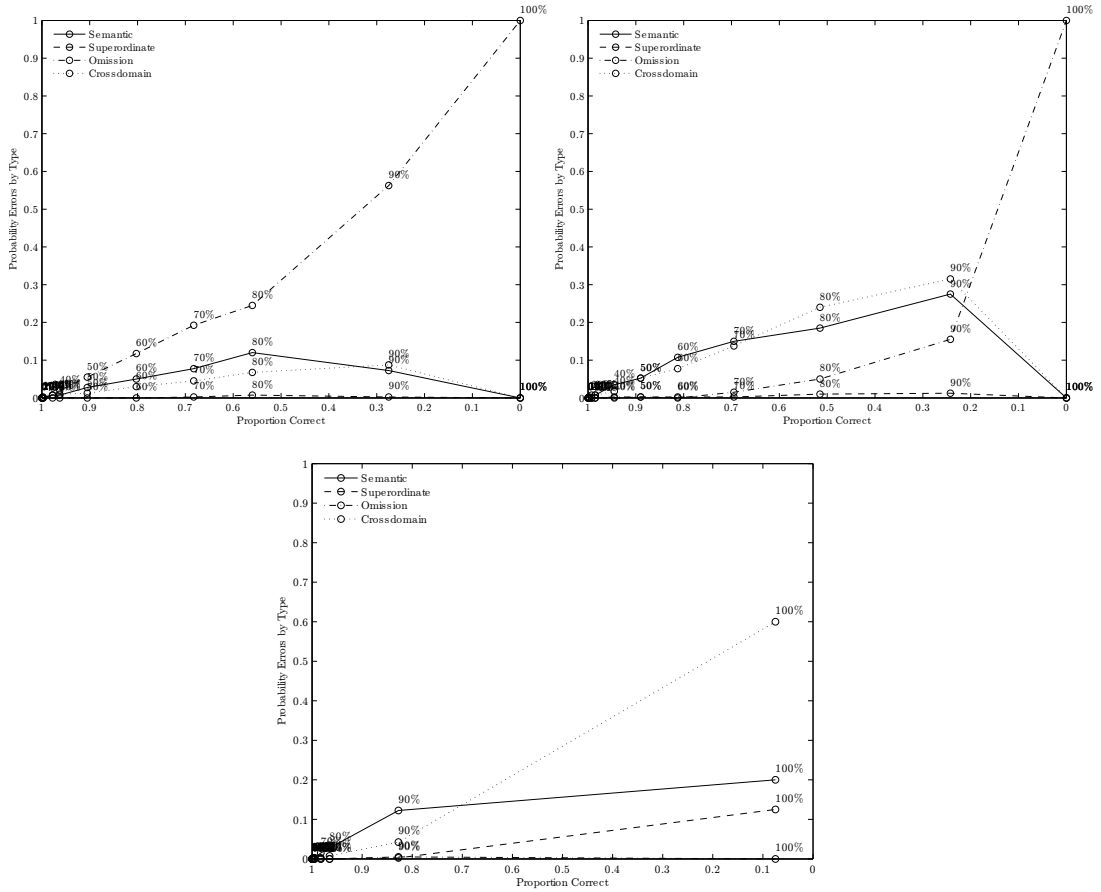


Figure 4.38: On the top left, are the results of the naming task for the modality-specific model with a radius of 10 SOM units using the nearest best matching unit (BMU) method to determine the pattern, given visual input. On the top right, the same task is carried out but with a radius encompassing the whole map. On the bottom, the naming results are shown when the SOM units are denormalised (see appendix B), which is not a function of the output radius. Due to the nature of the graph (being a two-dimensional projection of a three-dimensional graph), data points have been labelled to indicate the percentage of connections that have been lesioned.

neural network is recurrent. However, in this modality-specific implementation of the task this form of global lesioning is not necessary, due to the architecture of the model. In other words, because the SOM connectivity is feedforward, only the connection weights from the visual map to the name map take part in naming so damaging any other pathway in the modality-specific model will not affect performance in this case.

After lesioning the required percentage of weights, the inputs for each visual subpattern as defined by Equation B.10 are applied over the visual SOM and propagated to obtain activation on the name SOM. The units on the surface of the name map are in turn sorted into either a correct response or one of the possible naming errors or, if no answer can be determined, they are classified as an omission. The model's response is interpreted using the two methods described in subsubsection B.6.3.3. If the answer is exactly the same pattern as the visual

input then it is *correct*; if it is a general category name, such as one pertaining to ⌜*tool*⌝, but the input was a specific visual pattern within the category that the general name denotes, e.g., ⌜*hammer*⌝, it is a *superordinate* error; if it is from the same domain but from the wrong category, e.g., the input is ⌜*robin*⌝ but the output is ⌜*hamster*⌝, then it is classed as a *semantic* error; and if the response is from the opposing domain, e.g., the input is an animal but the output is an inanimate object, then it is a *crossdomain* error. If none of these comparisons can be made because the model has not managed to produce a clear response, because none or all units are on or off, or the search radius is not wide enough to encompass a best matching unit (BMU), then it is an *omission*.

The results of the naming task, shown on the left of Figure 4.38 are comparable to those we obtained from the hub reimplementations. In other words, the number of omissions and semantic errors does parallel those of patients but the cross-domain and superordinate errors do not capture what is going on with the Rogers et al. (2004) SD patients. The former are too high and the latter are too low, indicating that the model is unable to reproduce the exact patterns within semantic errors. The patterns of naming errors are not very dissimilar to those found in the reimplementations of the hub model.

### 4.3.2 Word and picture sorting task

Lesioning in this semantic task is carried out in an equivalent way to the previous task, but instead of lesioning the visual-to-name pathway, the visual-to-verbal and the name-to-verbal connections are zeroed. However, because of the structure of the Rogers et al. (2004) patterns, i.e., the requirement for unique output verbal units that represent category and domain membership certain problems arise. Firstly, as previously discussed an issue with regards to the original patterns we obtained from T. Rogers (personal communication, August 21, 2012) exists because they do not contain these orthogonal category and domain membership verbal units – although the pattern set created by us, described in section A.2, does contain the required verbal features.

Secondly, a slight issue arises with the modality-specific model itself. Given the architecture of the hybrid network, access to the membership units would involve denormalising the SOM weight $m_i$ that is selected as a response and then inspecting the specific verbal units within the MAU's codebook vector. (See appendix B for full details.) In other words, only one of the two methods of interpreting output work in this task. Alternatively, the three modal SOMs could be connected to their own input pool of classical neural network units, as seen in subsection 2.4.3 for a single pair of map and input/output layer. Each of the input layers would be trained

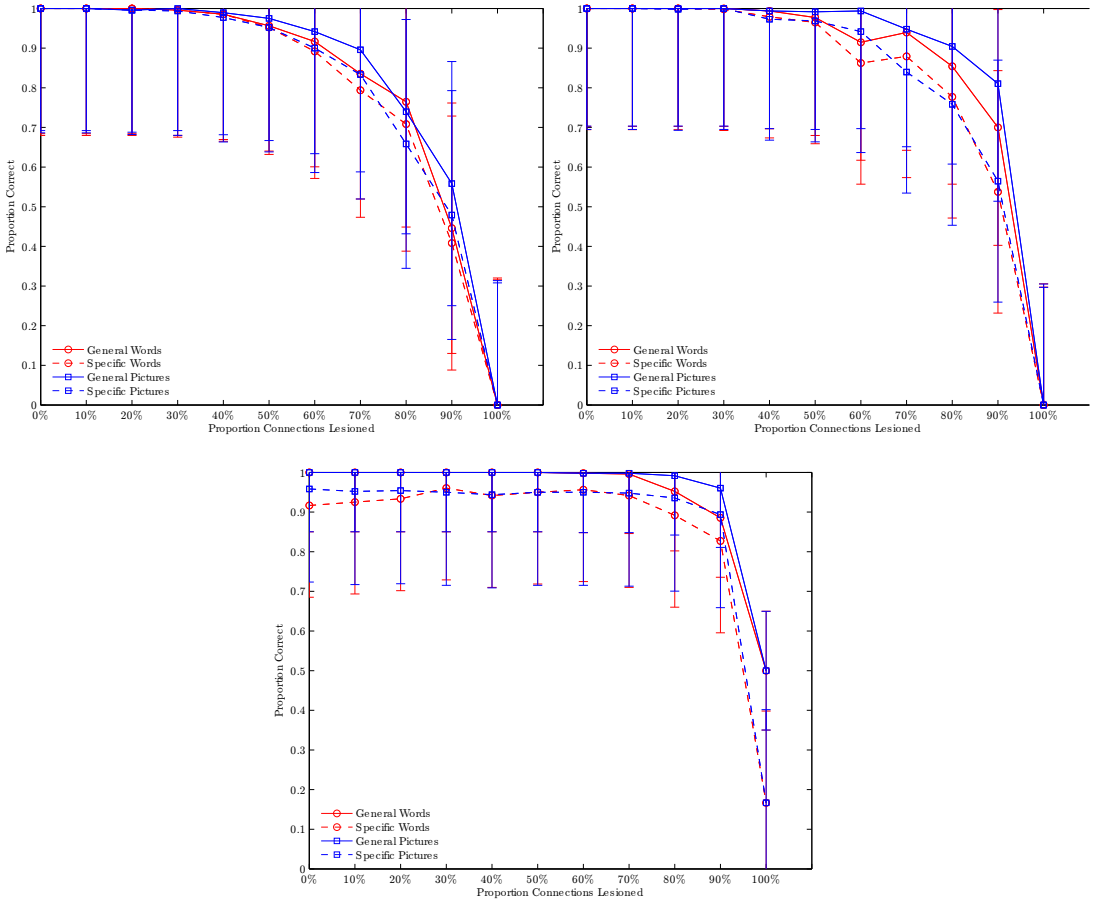


Figure 4.39: On the top left, are the results of the sorting task for the modality-specific model with a radius of 10 SOM units using the nearest BMU method to determine the pattern, given name input. On the top right, the same task is carried out but with a radius encompassing the whole map. On the bottom, the sorting results are shown when the SOM units are denormalised. Error bars represent one standard deviation.

to bidirectionally associate a subpattern onto the surface of the appropriate SOM, e.g., visual pattern input to visual SOM and vice versa; thus making the task involve only inspecting the relevant membership units in the input/output pools and not on the surface of the relevant SOM. This latter alternative is perhaps a more faithful parallel to the original task as described in Rogers et al. (2004), which involves the direct inspection of single units.

In contrast to the two methods mentioned above for modelling word and picture sorting, which would allow for direct access to category and domain membership units, the results presented here do not allow for such access. The way this task has been implemented is by comparing the output pattern's domain with the target's domain in the case of general sorting, and comparing their categories in that of specific sorting, meaning that general and specific word sorting consists of applying inputs on the name map and interpreting the output on the surface of the verbal map (as per the nearest-BMU method described in subsection B.6.1) and

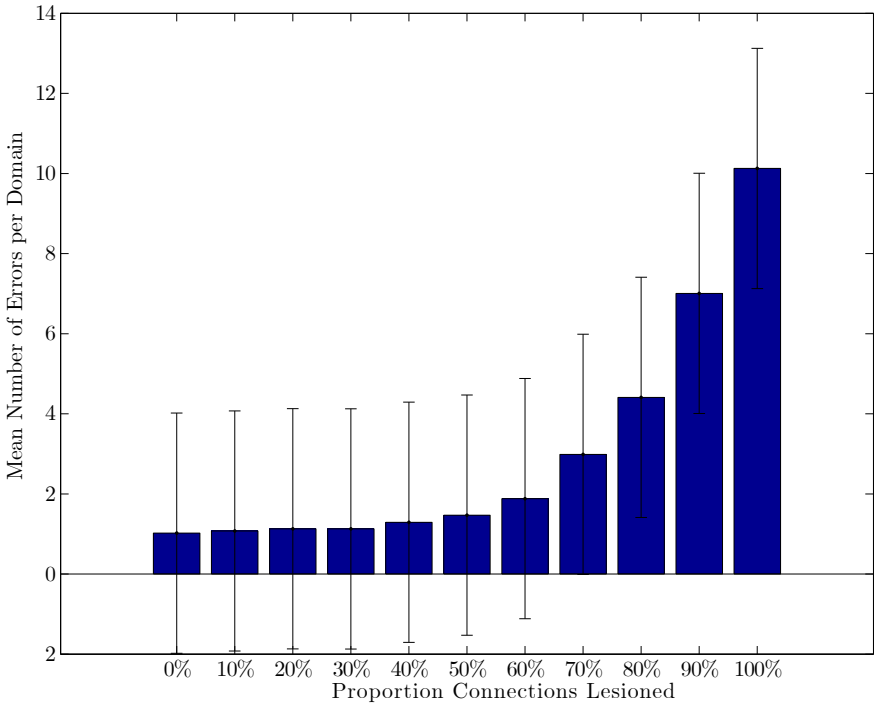


Figure 4.40: Results of drawing task in the modality-specific model. Error bars represent one standard deviation.

comparing domains or categories (without using the individual units). For general and specific picture sorting the same procedure is carried out with the input being the visual map and the destination map being the verbal SOM.

This task does not seem to fully relate back to the scores of the SD patients reported in Rogers et al. (2004). As mentioned in section 3.4, the modelling requirements are that domain-level, i.e., general, sorting is dramatically more preserved than category-level, i.e., specific, sorting. While the order of the correct scores is largely the same as in patients the preservation patterns are not.

### 4.3.3 Drawing task

The direct and delayed copying tasks cannot be carried out due to the architecture of the network. Specifically, zeroing of weights can only be applied on the name-to-visual connections, as there is no visual-to-visual pathway in this model. What this implies is that drawing can only be carried out and only by inspecting the internal representation within the BMU at that. This is because of the limitations forced upon the model by the way the tasks in Rogers et al. (2004) are carried out. Specifically, individual featured need to be accessible in the output in order for a drawing to have a meaningful evaluation. This means that the output of the SOM might not perfectly match the binary features in the input, because SOMs do not have

this requirement; their training is unsupervised. What SOMs aims for is to represent the input distribution, leaving open the possibility that the way BMUs represent their input is not as a direct copy, unlike auto-associator networks used in the hub-and-spoke and conceptual structure models. The results of running the drawing task in this way is shown in Figure 4.40. Not surprisingly, errors increase as the proportion of connections lesioned increases, though even without lesioning the model makes occasional errors.

## 4.4 Conceptual topography model

### 4.4.1 Confrontation naming task

Firstly, in Figure 4.41, the effect lesioning the input to cross-modal SOM connections and inspecting the nearest BMU to the activations on the cross-modal SOM is shown. Importantly, this method of running the network does not technically constitute a semantic task, as the interpretation of the "output" is just a function of the activations and the topology of the map, and does not involve directly accessing any name output. Most model errors are *semantic* errors meaning that the activations for one pattern become deformed so that units that belong to the neighbourhoods of other pattern's BMUs are more active. Then there are *superordinate* errors, meaning that units on the surface of the SOM are active in neighbourhoods that correspond to more general patterns (e.g., input concept: ⌜*horse*⌝, output: ⌜*mammal*⌝). The remaining two possible error classifications do not generally occur here, *omissions* because there are none as this is not an actual naming task, while *crossdomain* errors (activations are stronger in neighbourhoods from the opposing domain to the input) only seem to occur after 90% of connections have been lesioned. This allows a glimpse into the breakdown so far, but the model has not yet been used to produce a name, only to convey visual input onto the cross-modal SOM. It can be seen as an analogue to the dendrograms of the internal state of the recurrent network in Rogers et al. (2004), with the addition of allowing for a view of the model's "internal state" after damage.

Secondly, another possible implementation of this task involves analysing the most active unit on the surface of the cross-modal SOM map, using the MAU-inspection method described in subsubsection B.6.3.3. In Figure 4.42, the results of running the naming task in this way is shown, on the left with a threshold that omits responses that are below six standard deviations above the mean, and on the right with a very low threshold (to remove most omissions). It shows that, unless 90% or more connections are lesioned, no crossdomain errors ever occur. Initially, the network is run to create the same activations on the cross-modal SOM as before, and then

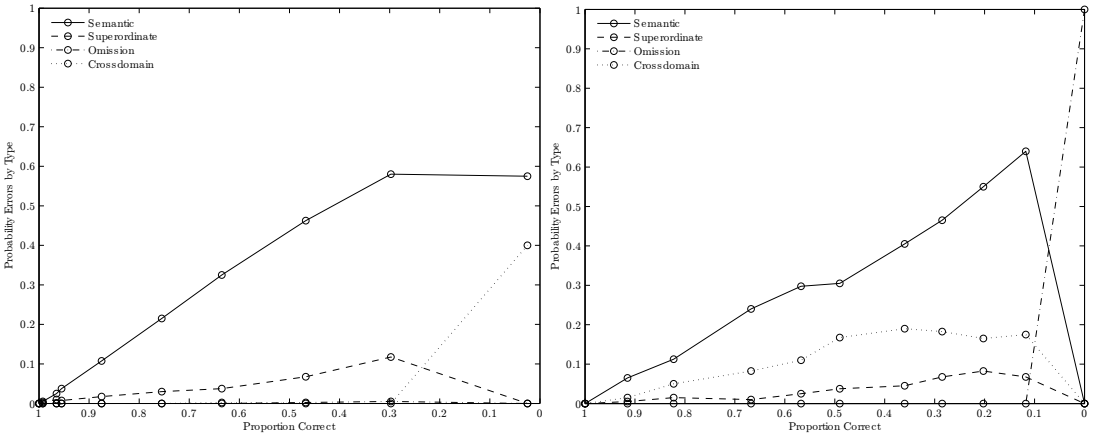


Figure 4.41: Depicted here are the naming task results for the lesioned conceptual topography model using the nearest BMU method to determine the pattern, given visual input. On the left is a network trained with the logistic function, and on the right a network trained with the hyperbolic tangent function. This method is not strictly speaking a confrontation naming task as no specific naming output is accessed, although the same input pathway is used. This does however provide a glimpse into what the activations from input to SOM have learned, and how they break down.

the most active unit's codebook vector is inspected (via denormalisation) to see which name unit is most active. This is sensible because each cell of the SOM represents an approximation to the full data set (i.e., thus containing SOM weights that look both like the original input patterns, as well as like unseen patterns that have features that resemble the distribution in the training set). Using this method a threshold can be set both at the name level (which unit within the name subpattern is most on) and at the level of activations on the surface of the map. This way of testing reflects the scores of Figure 4.41, meaning that the breakdown of the network during this version of the task reflects that of the breakdown of the activations on the map. The patient scores also show a similar pattern, but note that the scores here are largely clustered around the lowest left corner of the graph – each point represents an increment of 10% more connections zeroed, starting at 0% and finishing at 100%. The first non-zero value for semantic errors is at 30% (not very visible on the left graph), making this a more robust model with regards to lesioning damage than the original Rogers et al. (2004) version.

Thirdly, using a more typical network approach the classical artificial neural network output layer can be queried. This is trained to reproduce the patterns given as input, such as that shown on the left hand side of Figure 2.21. This method produces errors much like those reported in the reimplementations of Rogers et al. (2004) in chapter 3, as can be seen in Figure 4.43. The origin of each type of error can be seen more clearly by comparing the results of lesioning only the input to cross-modal SOM with lesioning only the cross-modal SOM to output weights – see Figure 4.44. Importantly, even though the difference between the results in Figure 4.42

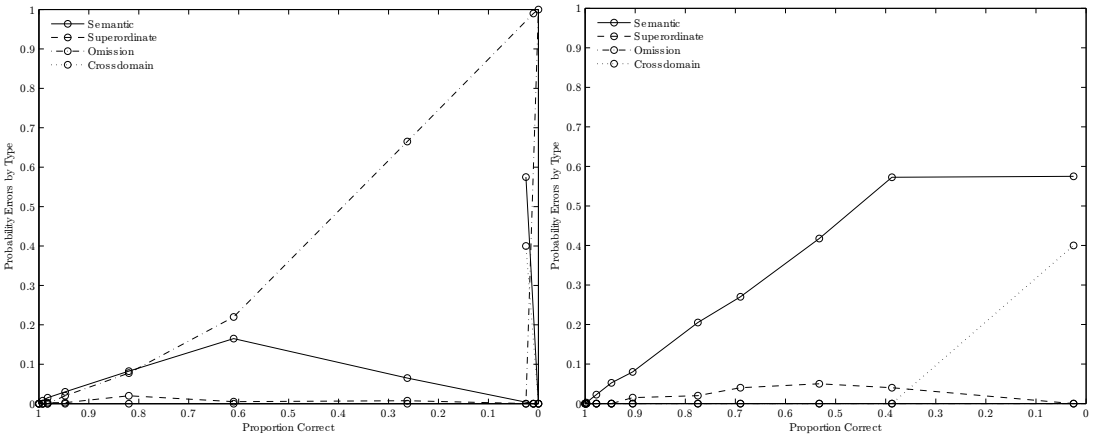


Figure 4.42: Results obtained by finding the most active unit on the SOM of the lesioned conceptual topography model as a result of visual input applied to the input units, that unit is twinned with a map cell which can be interpreted after denormalisation. On the left, are the scores with a threshold of six and a half standard deviations, meaning that if the maximum unit on the surface of the SOM exceeds this threshold it is classed as an error or correct, otherwise it is an omission. On the right side, the same scores are shown when the threshold is essentially removed, i.e., the threshold is set to the mean, which results in no omissions as there always exists a unit with activation greater than the mean during breakdown.

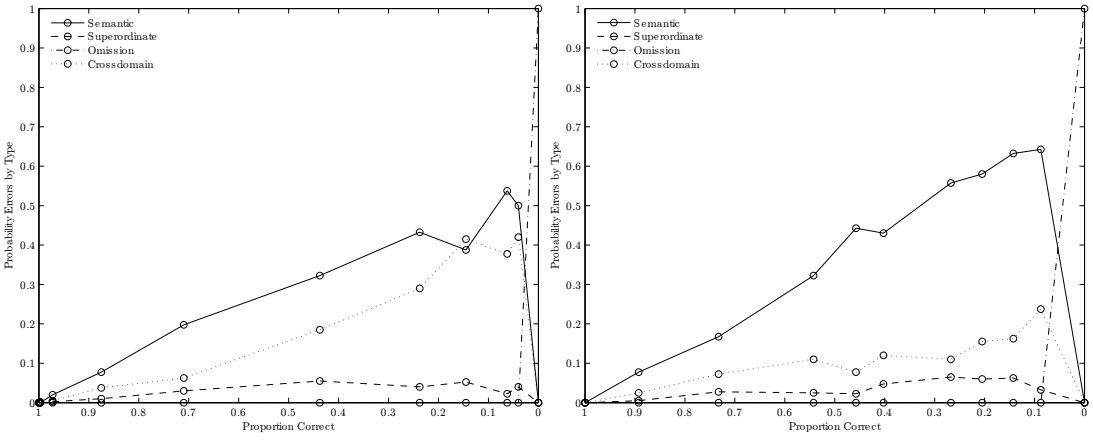


Figure 4.43: In this figure the naming scores obtained when interpreting the feedforward network output layer. On the left, the standard logistic transfer function; on the right, the hyperbolic tangent function.

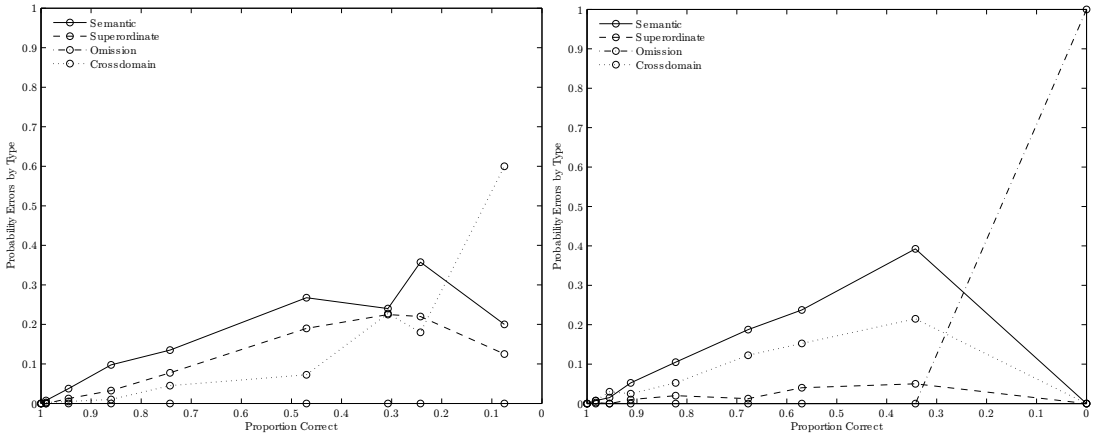


Figure 4.44: Results obtained by interpreting the output layer of the conceptual topography model, except now only certain connections have been zeroed. On the left, only the input connections are increasingly removed, while on the right only the output weights are damaged. Input connections (from the visual input layer to the SOM units) when removed do not give rise to as many crossdomain errors as is the case when removing only output connections (from SOM units to name output units).

and Figure 4.43 indicates that cross-domain errors arise only with lesioning output connections within the conceptual topography model, this is not the case in the reimplementations of the model of Rogers et al. (2004).[2]

### 4.4.2 Sorting task

The sorting task is carried out within the conceptual topography model in much the same way as the naming task above but with the appropriate modifications. And as before, with the other models of the sorting task, the patterns are required to be sorted into their respective category or domain using a set of mutually exclusive output units. Firstly, the activations on the surface of the cross-modal SOM are interpreted in a way that merely checked if the activation is in the same neighbourhood as the BMU for a category or domain that matches that of the input. The results of this method can be seen in Figure 4.45. This depicts the first part of the networks' activation being sent to the hub-like cross-modal SOM. Secondly, Figure 4.46 shows the results of sorting when responses are inferred by denormalising the most active unit on the surface of the cross-modal SOM map (see subsubsection B.6.3.3). This method results in chance performance both in the unlesioned and lesioned model, indicating that it is an inappropriate way of approaching this task. Thus, while in previous tasks this method was useful in analysing the model's internal functioning, this is not the case for this task. Thirdly, Figure 4.47 shows the interpretation of the input/output modality units, in the same way as Rogers et al. (2004).

[2]In work not reported here, the effects of lesioning only the input-to-hidden, or hidden-to-hidden, or hidden-to-output weights of the hub-and-spoke mode have been investigated and found not to eliminate the problems encountered with regards to crossdomain errors.

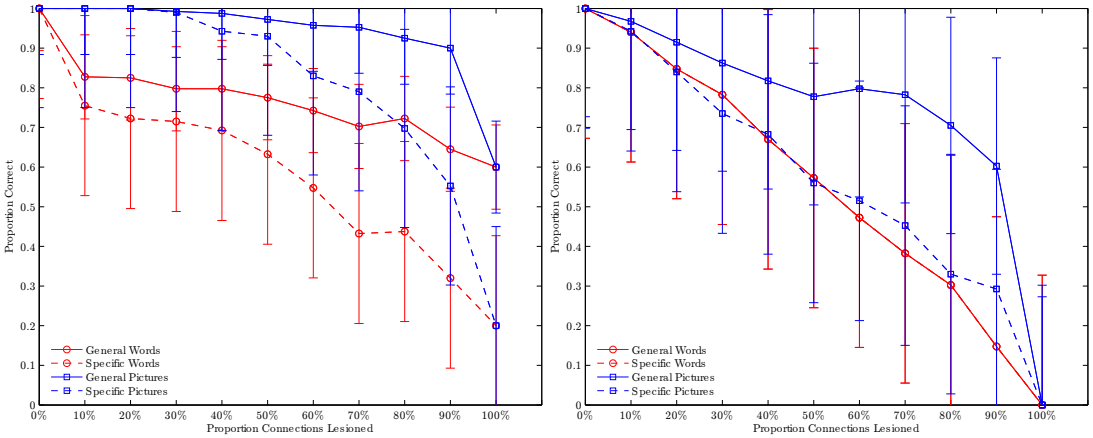


Figure 4.45: Results obtained by interpreting the activations on the surface of the cross-modal SOM: on the right, the model is trained using the logistic transfer function; on the left, using the hyberbolic tangent. Error bars represent one standard deviation.

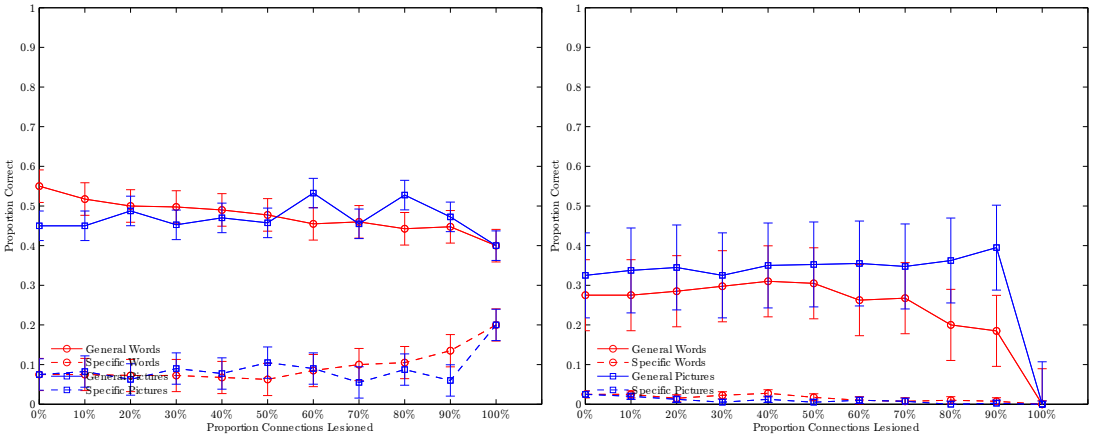


Figure 4.46: Results obtained by interpreting the denormalised weights: on the right, the model is trained using the logistic transfer function; on the left, using the hyberbolic tangent.

As previously in the hub and the modality-specific models, only a single unit that represents a category or domain is being inspected in the output units. The pattern does not seem to capture what the patients do as the required order of preservation (see subsection 3.4.1) is not present in Figure 4.47.

## 4.5 Discussion

The results presented here allow for direct and indirect comparisons to the hub model in the previous chapter. In fact many of the same problems with regards to the Rogers et al. (2004) patient data are present in the two SOM-based models as they were in the hub family of models. For both models, the results from this chapter follow the trend of the hub-and-spoke models previously, and show a failure to replicate Rogers et al. (2004) findings. In addition, the

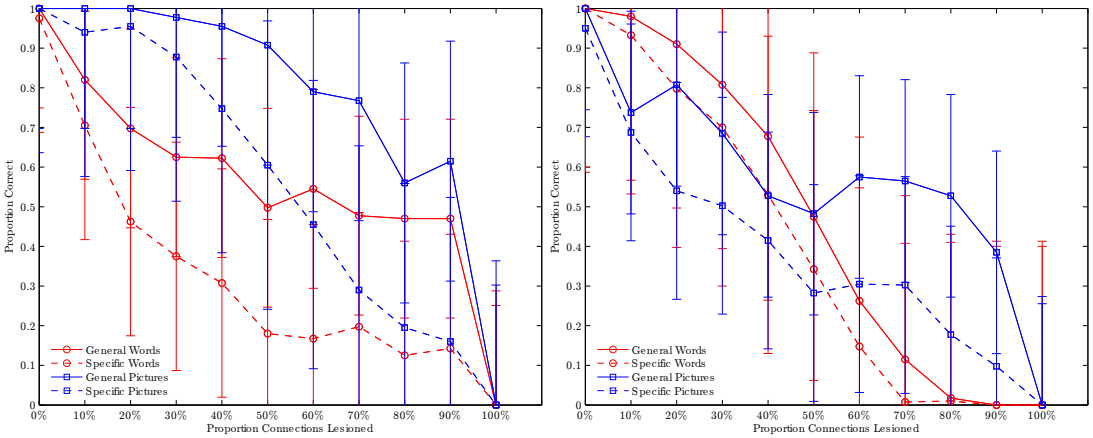


Figure 4.47: Results obtained by interpreting the output layer: on the right, the model is trained using the logistic transfer function; on the left, using the hyberbolic tangent. Error bars represent one standard deviation.

semantic task scores exhibit the same general properties as the tasks modelled in the previous chapter indicating that the inherent properties of the pattern set drive the organisation of all three models.

Looking first at the results of the modality-specific model we can see the inability to produce similar scores to patients closely matches the failures in replication in chapter 3. This indicates that the theory, which is dramatically different, and the architecture, which is driven by the theory, are largely independent of the patterns. In other words, given the Rogers et al. (2004) patterns, the theory and architecture, be they a modality-specific model with a hybrid-SOM architecture or a hub model with a recurrent network (implemented as a Boltzmann machine or a back propagation through time network), will still show very similar patterns of behaviour in semantic tasks. This is an interesting observation, given that models of semantic memory often if not always describe their architecture as contributing to their explanatory power.

Relating back to the modality-specific theory's requirements (recall section 2.6) we can see that this model is partially successful. The modality-specific model is trying to capture, using separate semantic scores, the impression of a generalised semantic deficit. This much it accomplishes, and in a way that is similar to the semantic scores produced in a model with a centralised semantic store. In other words, it appears as if topologically the semantic system is not affected, be it a hub-and-spoke or a modality-specific architecture, when it comes to modelling lesioning damage and semantic testing.

The cross-modal convergence zone here has been implemented using various map sizes and SOM three-dimensional properties (see appendix B). The hub-like SOM created a single unit (without a neighbourhood of similar units) that best represents a concept, then the hybrid units

connected this neighbourhood of units to in/output units that represent sensory information, thus forming the basis for the representation of a pattern. This structure, when damaged, decays into a structure that allows for cross-domain errors, indicating that the relative distances on the map is not large enough to stop these errors occurring. The distances on the surface of the map are a function of the patterns in much the same way as the dendrograms and PCA in section 3.7 and indeed the recurrent networks themselves used to reimplement the hub model.

As seen in section 2.6, the conceptual topography model stores features in the hub-like cross-modal convergence zone in a way that is compatible with the SIT principle (which requires similar concepts to be stored topologically proximally). Regardless, this modelling property does not appear to contribute to modelling generalised semantic deficits any more of less than the modality-specific model's architecture. In other words, semantic deficits found in SD patients, according to Rogers et al. (2004), do not display the same kinds of patterns found in this model.

To summarise, the naming scores of both models do not parallel those of the Rogers et al. (2004) SD patients. Importantly cross-domain errors seem to be the most problematic, with respect to modelling, as they occur in every single model so far, but purportedly never in the SD patients. Similarly, for the sorting task, neither model can achieve a near ceiling score like the SD patients do in general sorting. This again has been a common problem for all three families of models regardless of implementation details. This indicates that the structure of the pattern set is driving the modality-specific model in a similar way to the hub-and-spoke model. That is, the pattern set in hub-and-spoke, conceptual topography and modality-specific families of models is seen to shape the internal organisation regardless of other implementation and model details. So the attractors in a recurrent network model, and the representational structure of the two hybrid-SOM models, breakdown in seemingly very similar ways. This relative invariance of the behaviour in the semantic tasks from one set of models to the other would be very promising if both families of models showed the appropriate behaviour given the patients. But that is not the case.

It remains to be seen if this can be overcome in a specific hub topology, but it seems unlikely given the patterns and the training and testing regimens. Specifically, the patterns do not appear to represent the two domains in a separably enough way (see section 3.7). In addition the way the tasks are modelled goes against the way the training is carrier out. The models are generally trained to autoassociate but the naming and sorting tasks require classification. These are different enough problems that artificial neural networks and other types of implementations need to be specifically trained to solve them. More importantly perhaps, it is disappointing that the Rogers et al. (2004) patients (including the original ones) do not produce any cross-

domain errors but the Rogers et al. (2004) patterns consistently, over three radically different architectures, do.

## 4.6 Summary

In this chapter, two models, using a SOM network to provide the semantic stores, of the conceptual topography and the modality-specific accounts have been created to model the Rogers et al. (2004) tasks. The results here show a strong invariance to those presented in chapter 3, indicating that the patterns of the hub-and-spoke model are more responsible for driving the internal structure of the model than any higher- or lower-level implementation detail intrinsic to the learning algorithm or architecture of the models. In other words, the patterns have more importance than the models themselves in controlling behaviour on semantic tasks. This is problematic for the hub theory and will be discussed in detail in chapter 8.

# Chapter 5

# Modelling category-specific semantic deficits in the conceptual structure model

## 5.1 Overview

This chapter aims to replicate and examine the Tyler et al. (2000) model, which is an account within the conceptual structure theory of semantic memory. The tasks replicated here are the same as those presented in the original mode, however they are less easy to evaluate against patient data then those in the previous chapters. This is because the predictions are more centred around features and their preservation than a specific pattern of scores in a semantic task. The results presented here replicate and thus corroborate the account of the conceptual structure theory, indicating that it provides a plausible starting point for further modelling. Building on the successful replication, some further tasks are also carried out based on patient behaviour.

## 5.2 Introduction

Recall the conceptual structure model, discussed in subsection 1.4.3 and in more detail in section 2.3. The conceptual structure model by Tyler et al. (2000), see Figure 5.49, is used to model the two types of tasks in the original publication. The two types of testing are: looking at individual features and how their activations are affected by removal of connections; and

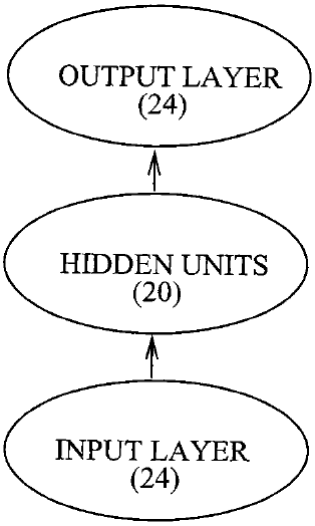


Figure 5.49: Architectural structure of conceptual structure model (Tyler et al., 2000, fig. 2).

measuring at overall output and how that is affected by the incremental zeroing of weights. The first set of tests examine the effect of distinctiveness on features both within and between the two domains. The second type of test is analogous to the word-to-picture matching task performed by the Rogers et al. (2004) hub model, which requires patients and the model to map words to their corresponding pictures. Both tests require the network to be run in the same way, with the only difference being that the mapping task is on the conceptual level (taking into account the whole output), while the other test only inspects the activations of individual units. Neither of the two types of test are able to be quantitatively matched to patient scores; only the qualitative features of the model's behaviour after damage can be discussed.

The tests have been carried out on three implementations: a faithful replica of Tyler et al. (2000), $\text{FF}_{epoch\text{-}wise}$, i.e., an epoch-wise feed-forward network trained using standard back-propagation; a pattern-wise version of the previous network, $\text{FF}_{pattern\text{-}wise}$; and a recurrent version trained using back-propagation through time (BPTT). In the results reported in this chapter, the tests have been carried by lesioining the networks 100 times. In the case of both epoch-wise and pattern-wise feed-forward reimplementations this was repeated with 300 different networks. For the BPTT, only 10 networks were used, but, because the BPTT network is non-deterministic, each pattern was sampled 100 times.

## 5.3 Individual features

### 5.3.1 Correlated properties

Properties that are highly correlated both within a concept (co-occur with other features in the same pattern) and between concepts (the same feature is present in many concepts) are predicted to be less prone to damage, while the inverse (features that are infrequently encountered) will be less likely to be preserved. Distinctive features only occur rarely in a concept and are; in the case of non-living things, uniquely correlated with another distinctive feature or, in animals, correlated to shared features, see Table 2.1.

The damage to distinctive features between domains can be seen in Figure 5.50. In the original Tyler et al. (2000) model, until the connections are 60% lesioned distinctive perceptual features for artefacts have a lower error, contra to the animals; after 60% both are too damaged to show any reliable difference (Tyler et al., 2000). In our three reimplementations, trained with feedforward epoch-wise, pattern-wise and recurrent epoch-wise learning algorithms, the results are qualitatively the same. Thus, the Tyler et al. (2000) results are replicable regardless of learning algorithm and frequency of weight updates to the extent that these options have been varied. The recurrent model, labelled BPTT in Figure 5.50, shows a move convex error curve as lesioning damage increases, but the relative ordering of the two types of error remains the same, with animals' distinctive features being more fragile than those of artifacts.

To compare and contrast the preservation of shared and distinctive features in the four models see Figure 5.51. These graphs make it clear that shared features are indeed more preserved, within the living domain. Distinctive features are more easily lost after lesioning, while shared features are relatively unscathed, remaining under 0.05 proportion error up to removal of 35% of connections.

### 5.3.2 Perceptual properties

In accordance with the conceptual structure theory, perceptual features should be preserved in the case of living things due to the structure of the patterns: animals have shared perceptual properties that are highly correlated to shared functional properties, while artefacts do not have this structure (see Table 2.1). On the contrary, living distinctive perceptual features should be prone to higher levels of error because they are not as correlated as their inanimate object counterparts. Thus, a dissociation between the domains should be seen in regards to the preservation of perceptual properties.

These predictions based on the pattern set, and general overarching theory, are supported

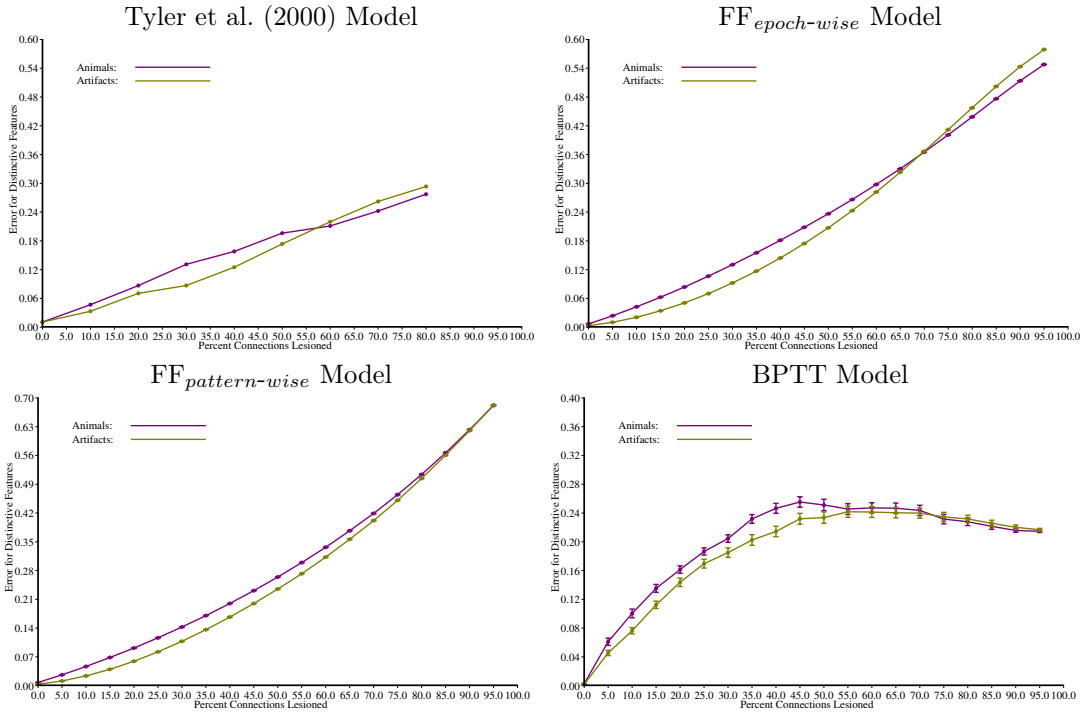


Figure 5.50: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning (Compare with Tyler et al., 2000, fig. 3.). Error bars represent one standard error.

by Figure 5.51. For a comparison of the two domains see Figure 5.52, which shows that living things' shared perceptual features are better preserved. In both graphs, correlation of any kind is shown to make individual properties more immune to damage than lack thereof.

### 5.3.3 Functional features

Functional features are distinctive for artefacts, but shared for animals, so there is a qualitative distinction between biological function and the function of artefacts. Comparing functional features for each domain in Figure 5.53 is essentially the same as comparing distinctive perceptual and distinctive functional within the artefacts domain, as seen in Figure 5.50. Yet again, the two graphs illustrate that correlation implies preservation, and that the properties of the data set can account for the category-specific dissociations seen in patients.

## 5.4 Identity mapping

In order to investigate the models' proficiency on the concept level the identity task is used. This is a task that is similar to word-to-picture matching (see subsection 1.2.2) with respect to the semantic abilities it measures. The input is presented to the network and it must reproduce

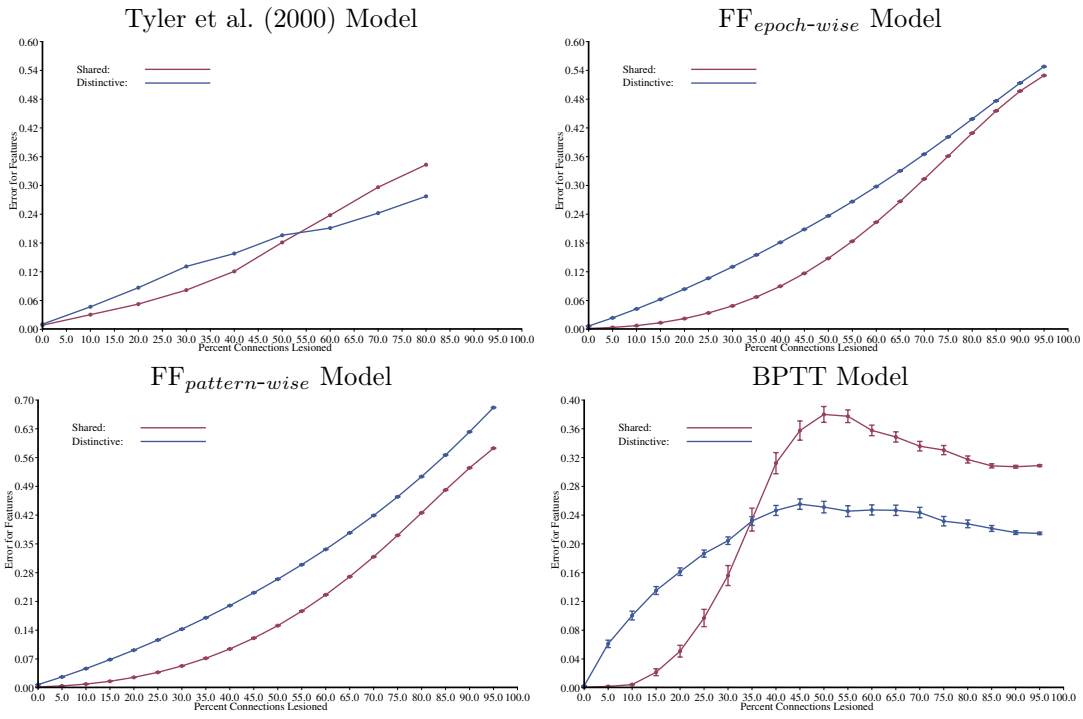


Figure 5.51: Mean absolute error for shared and distinctive perceptual features for living things (Compare with Tyler et al., 2000, fig. 4.). Error bars represent one standard error.

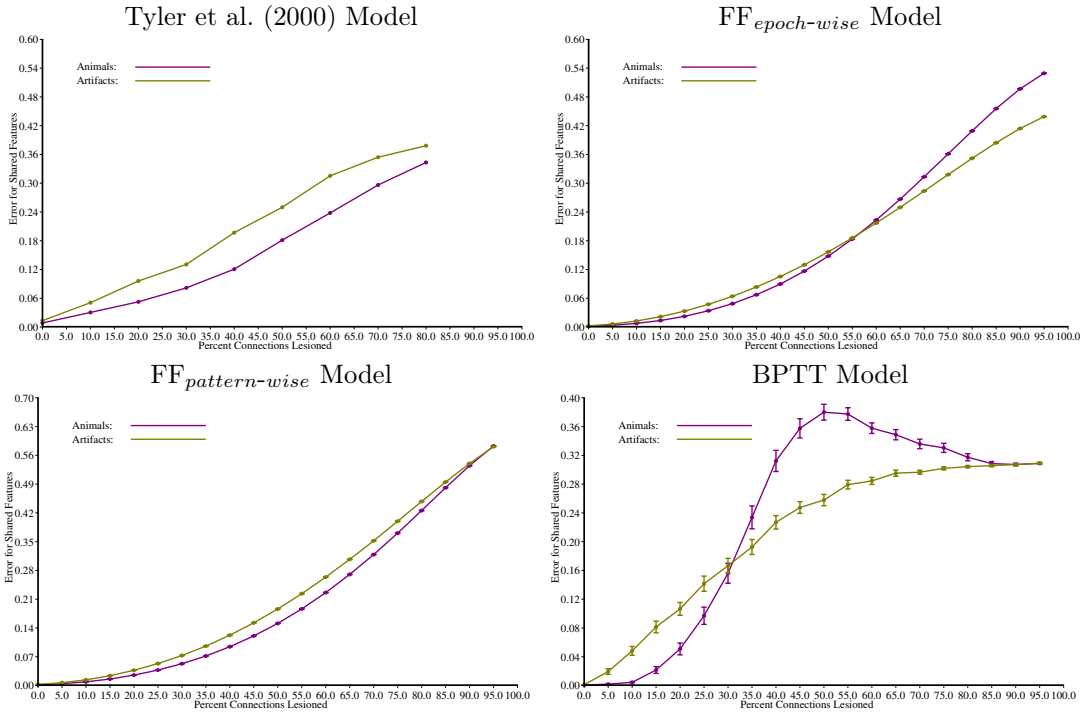


Figure 5.52: Mean absolute error for shared perceptual features for artefacts and living things (Compare with Tyler et al., 2000, fig. 5.). Error bars represent one standard error.

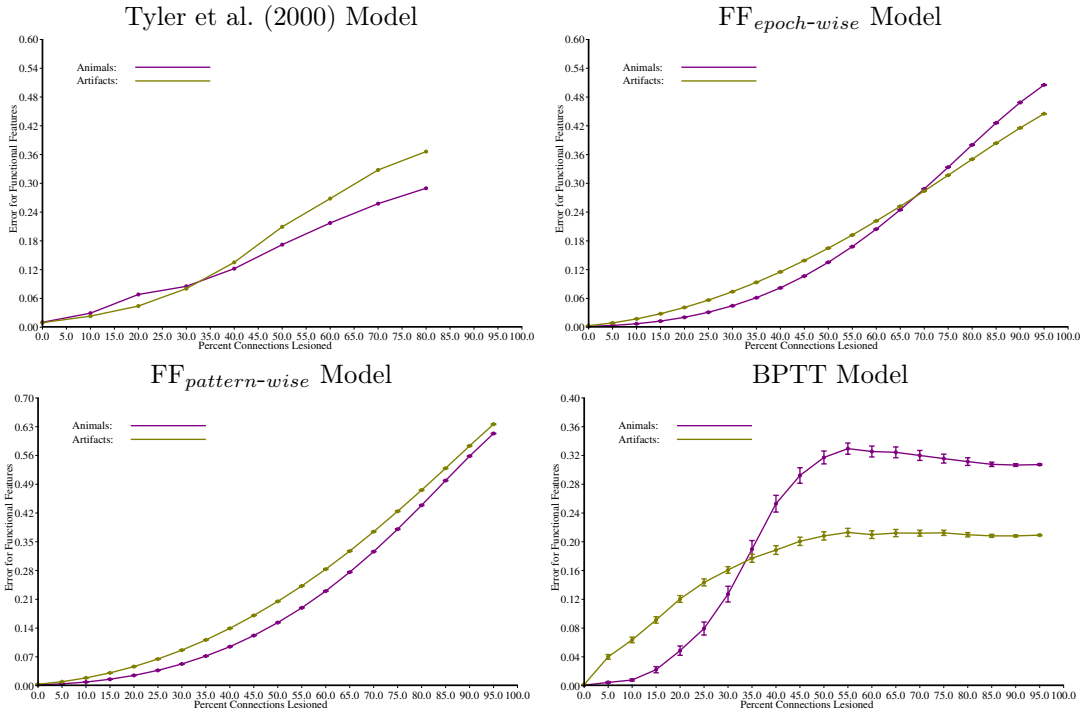


Figure 5.53: Mean absolute error for functional features for artefacts and living things (Compare with Tyler et al., 2000, fig. 6.). Error bars represent one standard error.

the same activations in the output units. Euclidean distance is used to measure the difference between the output and the target; if the output is closer to another pattern then that pattern chosen as the model's response. The output is interpreted as: *a*) *correct match*, the network has produced the same output as input, *b*) *within-category error*, the output is from the same category as the target, *c*) *between-category error*, the network's response is from the same domain but a different category as the target, *d*) *between-domain error*, the output activations are closer to a pattern from the opposing domain.

In Figure 5.54 it can be seen that there are more correct responses for artefacts than for living things. However, in the original model at approximately 60% connections lesioned, a crossover effect occurs. This doubly dissociates living things from artefacts, and allows the original Tyler et al. (2000) model to account for the same patterns seen in patients. Because this effect is not seen in the reproduced model, it might be the case that the original model is using a slightly different learning algorithm, although it seems to be the case that the sampling in the original model is biasing the result towards appearing as if the crossover is more pronounced. The good news, is that with a small adjustment to the learning rate the FF$_{epoch\text{-}wise}$ (the direct reimplementation) is also able to have an equally pronounced crossover effect at the same level of lesioning damage. The recurrent BPTT model is unable to show a crossover effect here.

For a more detailed look at what is going in this task, see Figure 5.55, which shows the

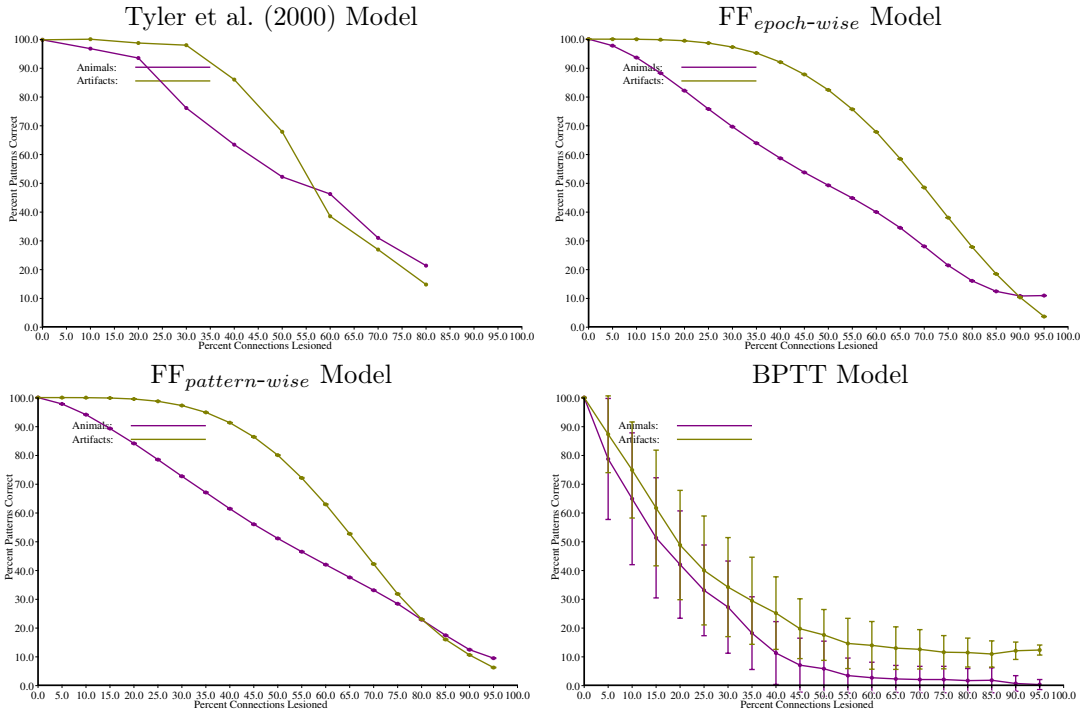


Figure 5.54: Percentage of patterns correctly mapped for the two domains (Compare with Tyler et al., 2000, fig. 7.). Error bars represent one standard error.

percentage of between- and within-category errors for each domain.

## 5.5 Modelling pre-morbid organisation of semantic cognition

### 5.5.1 Introduction

The previous sections explain and demonstrate how the Tyler et al. (2000) model is a useful and replicable way of modelling the category-specific patterns of behaviour seen in patients. The success of the model is based on the specific distribution of features per domain causing differentiation in the stored representation so that after damage each domain is affected qualitatively and quantitatively differently. This shows that inherent properties of concepts make them more or less vulnerable to lesioning damage, and thus that category-specific patterns of behaviour should be a common finding in patients with semantic deficits, this is a contentious issue as such patterns are not found in most patients (see chapter 8; Caramazza & Shelton, 1998; Gaffan & Heywood, 1993; Sheridan & Humphreys, 1993; Stewart, Parkin, & Hunkin, 1992).

The original Tyler et al. (2000) model does not account for the fact that psycholinguistic

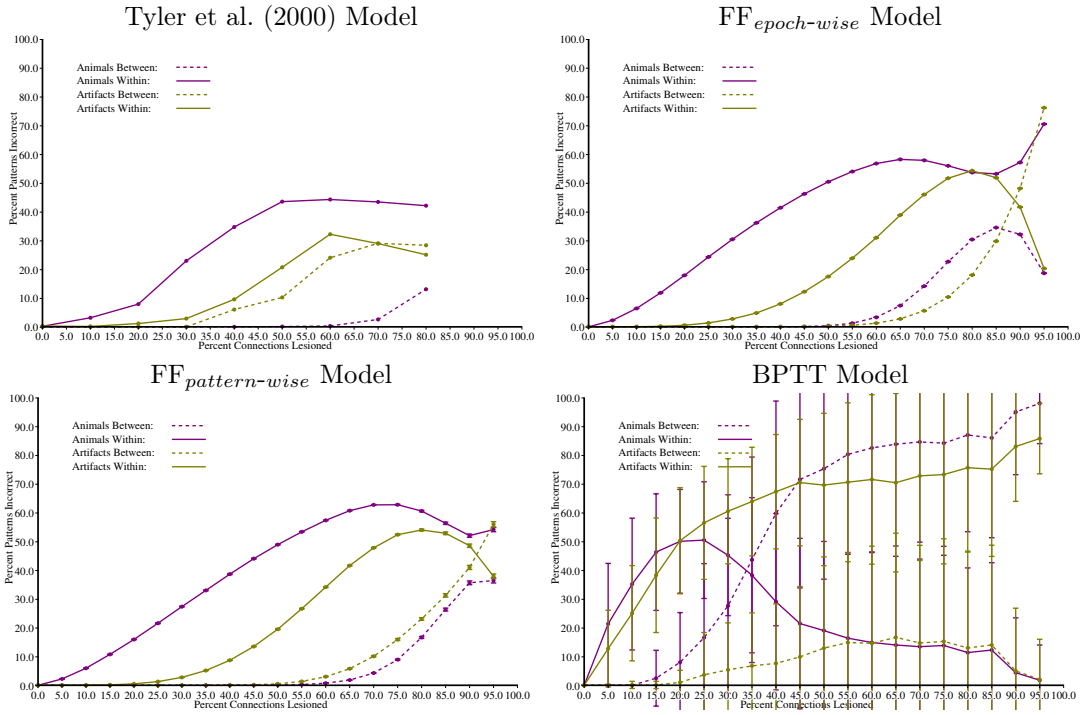


Figure 5.55: Percentage of between- and within-category errors per domain (Compare with Tyler et al., 2000, fig. 8.). Error bars represent one standard error.

variables such as the frequency, familiarity, objective age of acquisition, imageability, phoneme length, and a patient's expertise or experience contribute to the pre-morbid organisation of the semantic system (Jefferies, Rogers, & Ralph, 2011; Lambon Ralph, Patterson, Garrard, & Hodges, 2003). This pre-morbid organisation may affect the post-morbid patient behaviour, potentially contributing to, or direct causing, category-specific effects. Therefore, given that the state the semantic system is in before damage might be a good predictor of the patterns seen in some patients, e.g., category-specific preservation, this section will explore this possibility within the conceptual structure account. Specifically, the relatively simple Tyler et al. (2000) patterns are altered in order to create imbalances between the two domains, to reflect a qualitative different in age of acquisition, familiarity, and frequency of concepts, and other factors which contribute to how concepts are represented and accessed. These factors are predicted to affect the robustness of a pattern post-lesioning in the model.

| **Artefacts** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Distinctive Perceptual | | | | | | | | Shared Perceptual | | | | | | | | Functional | | | | | | | |
| **0.77** | 0.24 | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | **1.00** | 0.06 | **0.99** | 0.07 | 0.07 | 0.19 | 0.23 | 0.25 | **1.00** | 0.00 | 0.20 | 0.00 | 0.07 | 0.00 | 0.18 | 0.00 |
| 0.00 | **1.00** | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | **0.78** | 0.00 | **1.00** | 0.00 | 0.20 | 0.18 | 0.00 | 0.10 | 0.00 | **1.00** | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | **1.00** | 0.25 | 0.00 | 0.13 | 0.00 | 0.00 | **1.00** | 0.02 | **0.78** | 0.15 | 0.05 | 0.00 | 0.24 | 0.08 | 0.05 | 0.06 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 |
| 0.00 | 0.04 | 0.00 | **1.00** | 0.20 | 0.00 | 0.00 | 0.00 | **0.85** | 0.00 | **0.87** | 0.00 | 0.00 | 0.00 | 0.09 | 0.12 | 0.21 | 0.12 | 0.02 | **1.00** | 0.18 | 0.01 | 0.00 | 0.00 |
| 0.00 | 0.19 | 0.23 | **1.00** | 0.00 | 0.00 | 0.02 | 0.00 | 0.13 | **1.00** | 0.00 | **1.00** | 0.15 | 0.04 | 0.20 | 0.24 | 0.19 | 0.00 | 0.00 | **1.00** | 0.02 | 0.21 | 0.09 | 0.00 |
| 0.08 | 0.00 | **1.00** | 0.00 | 0.11 | 0.00 | 0.24 | 0.17 | 0.00 | **1.00** | 0.00 | **0.80** | 0.00 | 0.24 | 0.00 | 0.19 | 0.05 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.22 |
| 0.04 | **1.00** | 0.19 | 0.00 | 0.02 | 0.03 | 0.05 | 0.00 | 0.13 | **1.00** | 0.11 | **0.99** | 0.13 | 0.10 | 0.00 | 0.16 | 0.07 | **1.00** | 0.21 | 0.14 | 0.12 | 0.00 | 0.08 | 0.00 |
| **1.00** | 0.00 | 0.12 | 0.12 | 0.07 | 0.00 | 0.09 | 0.00 | 0.12 | **1.00** | 0.00 | **0.89** | 0.00 | 0.22 | 0.00 | 0.19 | **1.00** | 0.00 | 0.00 | 0.00 | 0.21 | 0.09 | 0.00 | 0.04 |

| **Animals** | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Distinctive Perceptual | | | | | | | | Shared Perceptual | | | | | | | | Functional | | | | | | | |
| 0.00 | 0.00 | 0.00 | 0.00 | **0.88** | 0.00 | 0.01 | 0.23 | 0.00 | 0.00 | 0.10 | 0.18 | **0.93** | 0.00 | **0.80** | 0.00 | 0.23 | 0.00 | 0.00 | 0.00 | **0.94** | 0.19 | **1.00** | 0.21 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.24 | **0.93** | 0.05 | 0.00 | 0.00 | 0.23 | 0.05 | 0.08 | **1.00** | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | **1.00** | 0.00 | **1.00** | 0.07 |
| 0.05 | 0.00 | 0.10 | 0.25 | 0.00 | 0.08 | **1.00** | 0.00 | 0.24 | 0.00 | 0.00 | 0.23 | **1.00** | 0.18 | **0.94** | 0.20 | 0.16 | 0.24 | 0.03 | 0.15 | **0.79** | 0.00 | **0.75** | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | **1.00** | 0.00 | 0.01 | 0.17 | 0.00 | **0.84** | 0.13 | **1.00** | 0.00 | 0.21 | 0.18 | 0.00 | 0.00 | **1.00** | 0.00 | **0.91** | 0.03 |
| 0.00 | 0.18 | 0.00 | 0.02 | 0.00 | 0.00 | 0.25 | **1.00** | 0.00 | 0.00 | 0.08 | 0.10 | 0.00 | **0.98** | 0.02 | **0.98** | 0.24 | 0.00 | 0.17 | 0.00 | 0.00 | **0.97** | 0.00 | **0.79** |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.21 | 0.15 | **1.00** | 0.00 | 0.09 | 0.22 | 0.21 | 0.10 | 0.00 | **1.00** | 0.00 | **1.00** | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | **1.00** | 0.14 | **0.90** |
| 0.00 | 0.06 | 0.00 | 0.02 | 0.00 | **1.00** | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | **1.00** | 0.00 | **1.00** | 0.19 | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.21 | **0.98** |
| 0.00 | 0.00 | 0.14 | 0.08 | **0.93** | 0.04 | 0.23 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **1.00** | 0.05 | **0.96** | 0.25 | 0.24 | 0.03 | 0.20 | 0.21 | **0.78** | 0.00 | **1.00** |

Table 5.5: Table showing, $e_1$, one of four different sets of 16 exemplar patterns used to train the models. These have been generated based on the original patterns, shown in Table 2.1. The features that should be considered as present, or on, are presented in bold.

### 5.5.2 Experiment 1: Exemplars versus prototypes

Using the reimplementation of the original Tyler et al. (2000) model, but training on exemplar patterns, as opposed to prototypes, might be a useful way to model familiarity. Exemplars here are defined as patterns that have Gaussian noise applied over them causing them to take on real values in the range $[0, 1]$, to parallel the variation found within a category. They are derived from the original Tyler et al. (2000) patterns, which can be seen as prototypes or ideals, by applying. The following networks are tested on a certain set of exemplars, $e_1$, after having been trained on exemplars, $e_{1,2,3,4}$. In other words, the training of the network comprises auto-associating 4 different sets of patterns, which represent exemplars. One of the four sets of real-valued exemplars created from the original prototypes can be seen in Table 5.5.

After damage to the network, the qualitative pattern of the overall error, as shown in Figure 5.56, is largely unaffected,[1] although it seems like the lesioning has less of a detrimental effect on the network in the case where it is trained on exemplars. This is expected since the noise within the patterns during training makes the network a little more robust to damage from the noise introduced by lesioning.

More interestingly, but somewhat worryingly, the effect that training with exemplars has on the network is detrimental to modelling category-specific effects in terms of the conceptual structure theory – the theory underpinning the Tyler et al. (2000) model. Specifically, looking at Figure 5.57, it is clear that the effect required to model the category-specific patients is absent. Both domains have equal error for their distinctive perceptual features even though the perceptual features are still distributed in the same way (save for the bounded Gaussian noise, recall Table 5.5). Figures 5.58 to 5.62 show familiarity does not affect the other differences in feature preservation between and within domains required to model category-specific deficits.

### 5.5.3 Experiment 2: Frequency

A second variable that has been argued to affect pre-morbid organisation of the semantic system is the frequency with which concepts are encountered. In a further set of simulations, the first $2,000$ epochs were identical to those of Tyler et al. (2000), meaning that each pattern was presented the same number of times to the network. The remaining $2,000$ epochs represent the process of specialisation of the model towards expertise in a domain. Expertise is modelled by presenting the frequent domain items twice as often, thus replacing the opposing domain's

[1] Note that the Figure 5.56 is not a graph that can be compared to human data as it does not represent anything that can be tested using a semantic task. Nor can Figure 5.56 be compared to the original Tyler et al. (2000) model since they do not present such a graph, which is not surprising since there is no interpretation with regard to human data for this graph.

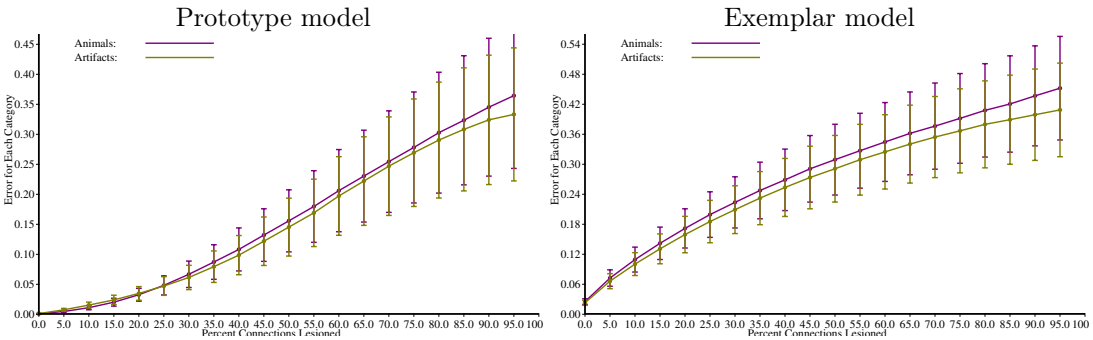


Figure 5.56: Overall error of models trained with the two different pattern sets: left, using the original Tyler et al. (2000) patterns; right, using the patterns with bounded Gaussian noise applied over them to simulate exemplars.

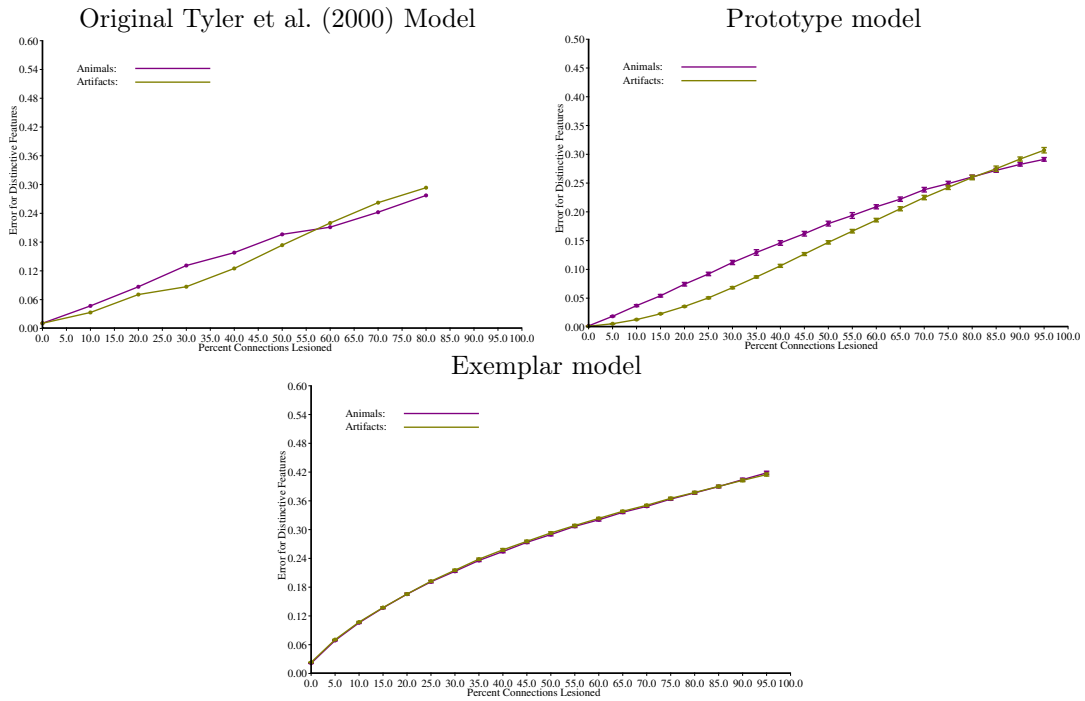


Figure 5.57: Mean absolute error for distinctive perceptual features of artefacts and living things trained with differing expertise.(Compare with Tyler et al., 2000, fig. 3.)

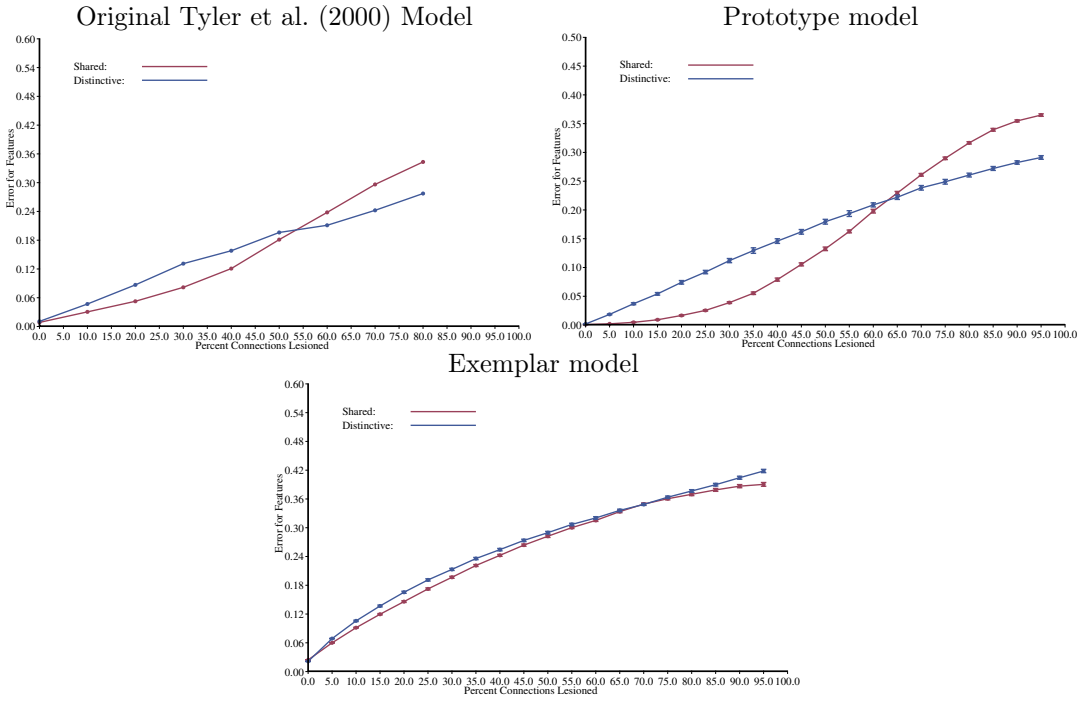


Figure 5.58: Mean absolute error for shared and distinctive perceptual features for living things (Compare with Tyler et al., 2000, fig. 4.)

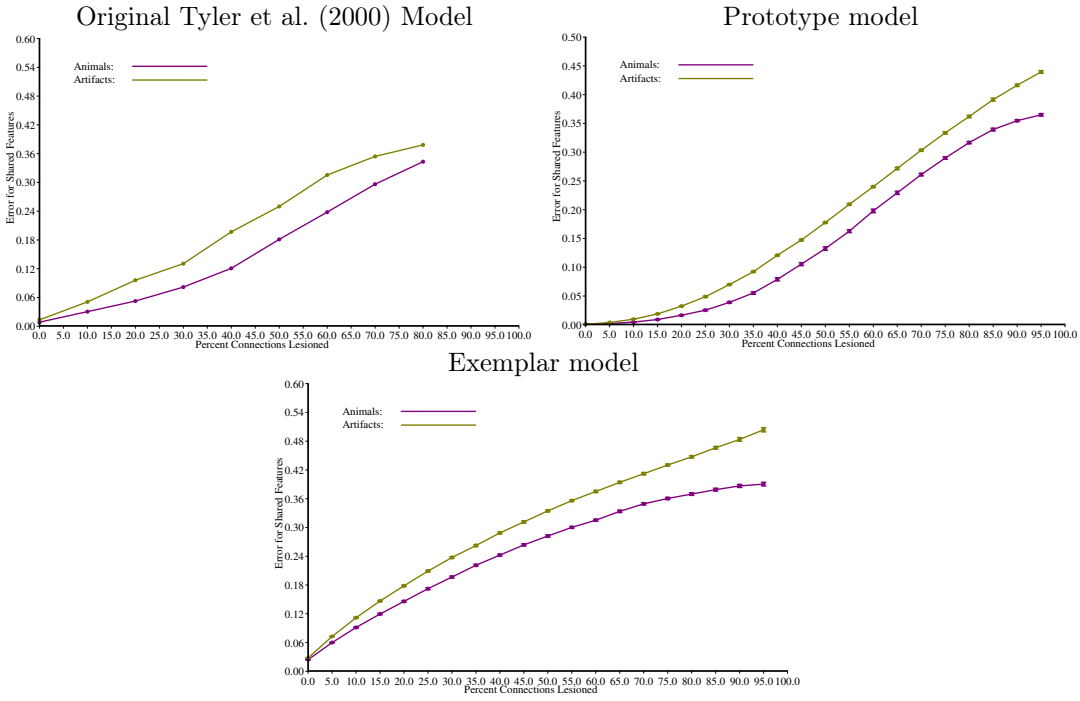


Figure 5.59: Mean absolute error for shared perceptual features for artefacts and living things (Compare with Tyler et al., 2000, fig. 5.)

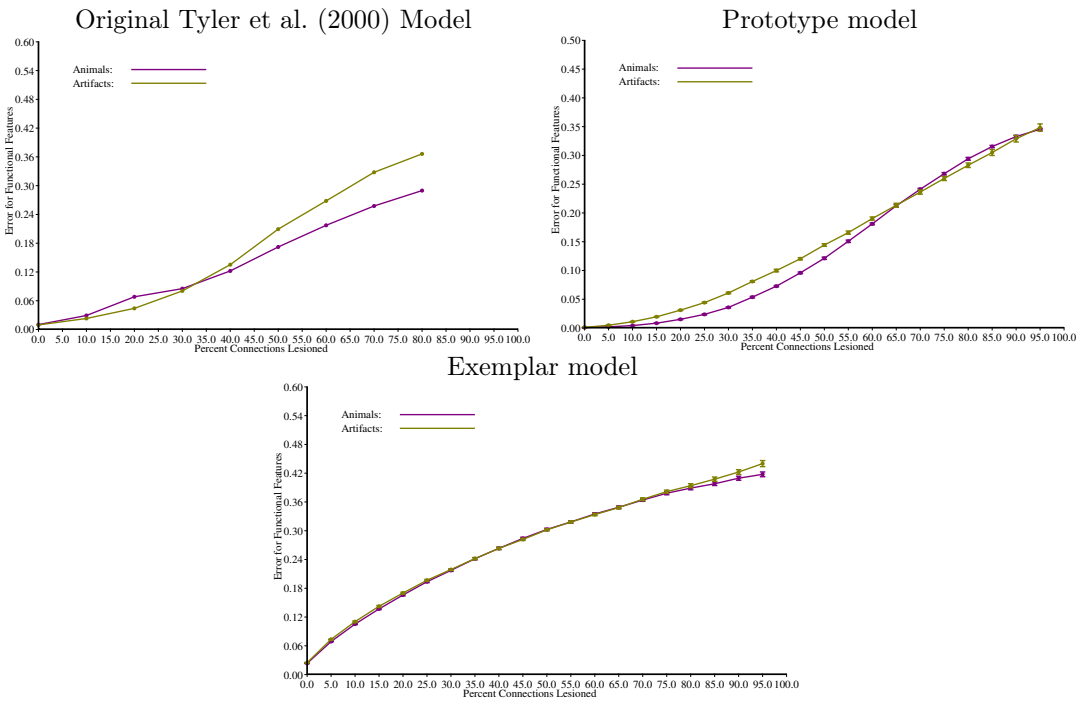


Figure 5.60: Mean absolute error for functional features for artefacts and living things (Compare with Tyler et al., 2000, fig. 6.)

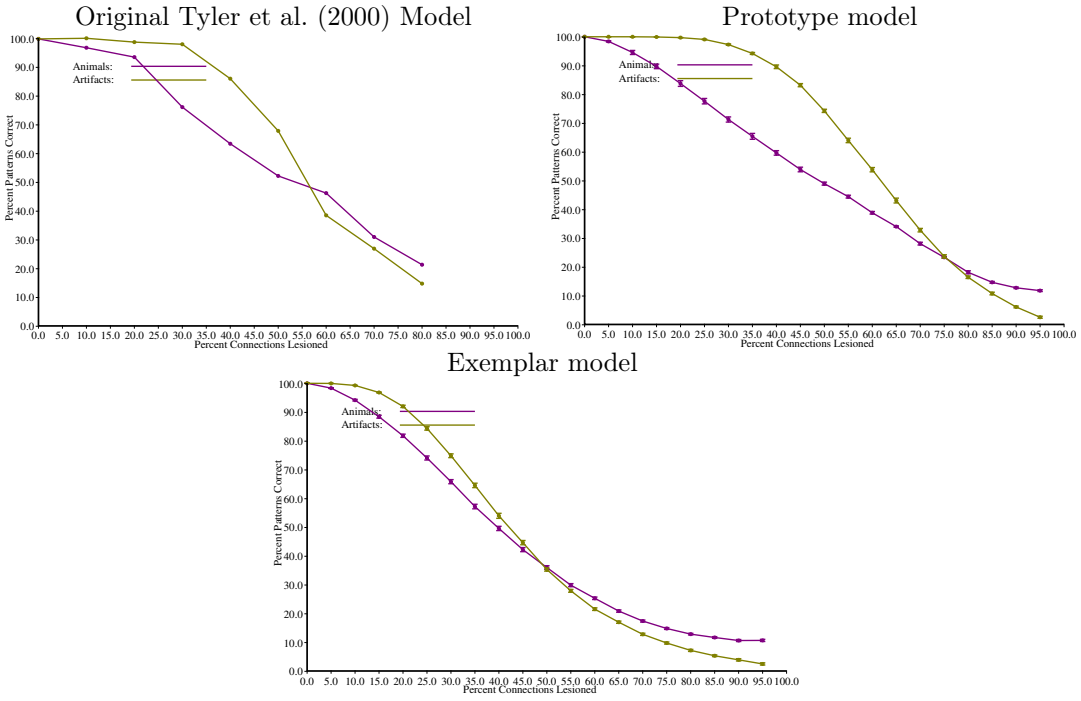


Figure 5.61: Percentage of patterns correctly mapped for the two domains (Compare with Tyler et al., 2000, fig. 7.)

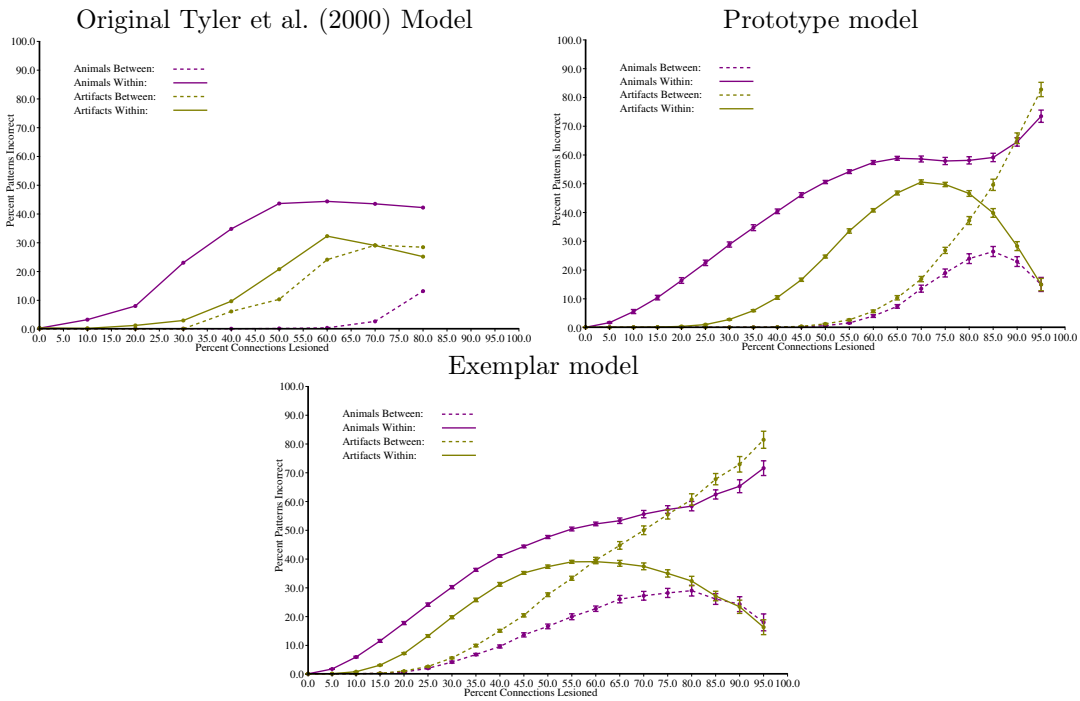


Figure 5.62: Percentage of between- and within-category errors per domain (Compare with Tyler et al., 2000, fig. 8.)

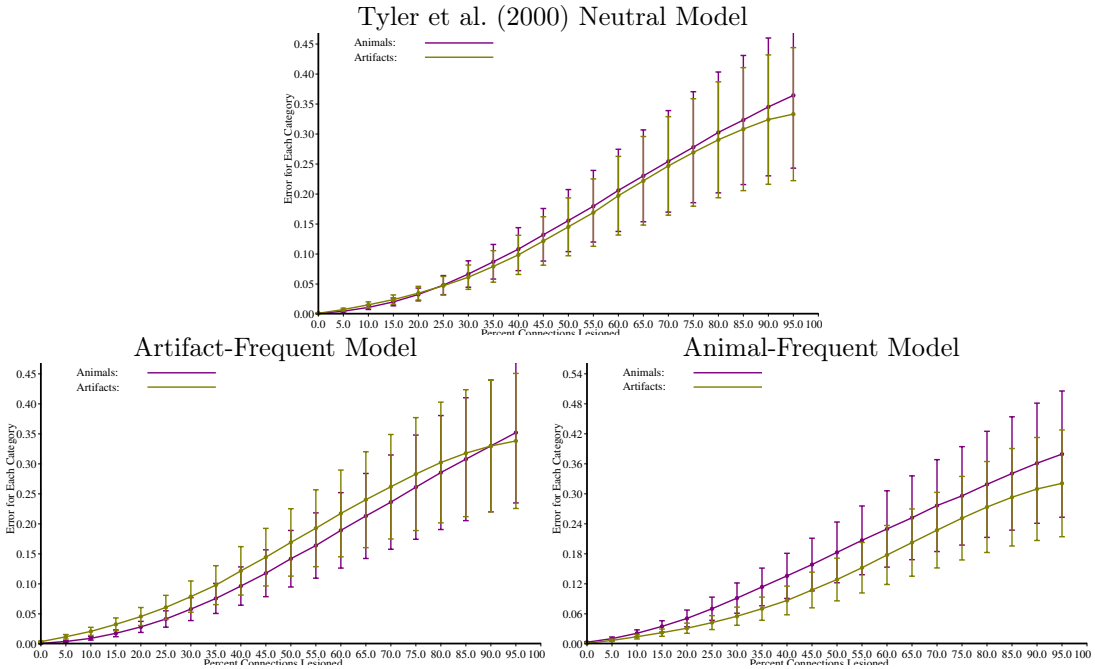


Figure 5.63: Overall error of models trained with differing expertise.

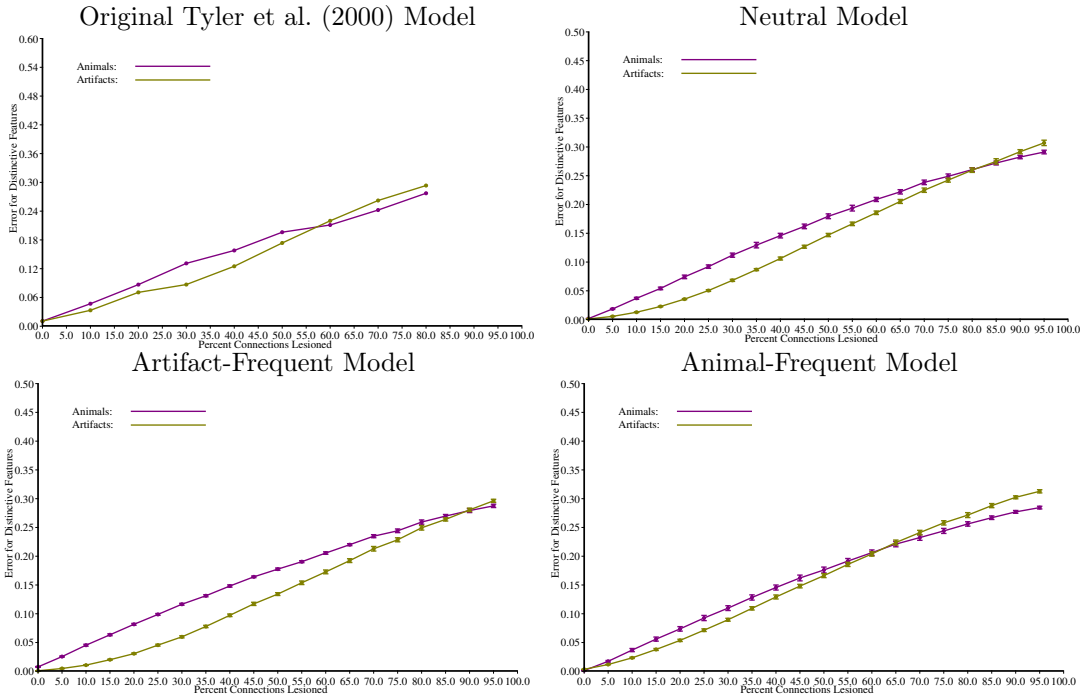


Figure 5.64: Mean absolute error for distinctive perceptual features of artefacts and living things trained with differing expertise.(Compare with Tyler et al., 2000, fig. 3.)

items and resulting in the same number of presentations. The graphs in Figure 5.63 depict the overall output error per category/domain, showing how the state of the network after damage is a function of both the patterns and the training regimen. It is important to note that these three graphs do not represent a semantic task, but merely serve to show the full output error of the models. The first sub-graph is taken from the reimplementation of the original work by Tyler et al. (2000), and shows the classic dissociation between the preservation of animals versus artefacts. The remaining two sub-graphs display the percentage correct of the model when training with one category as a specialisation.

The three different types of expertise, neutral, artifact, and animal, have a small effect on the preservation of distinctive features between the two domains, as can be seen in Figure 5.64. In the case of animals being more frequently encountered during training a subtle cross-over effect can be seen. Figure 5.65 is included to show that the internal feature-structure of living things is the same regardless of expertise. However, overall, i.e., for the animal domain as a whole, there can be seen an effect of preservation relative to the two other conditions, see figures 5.64 to 5.69.

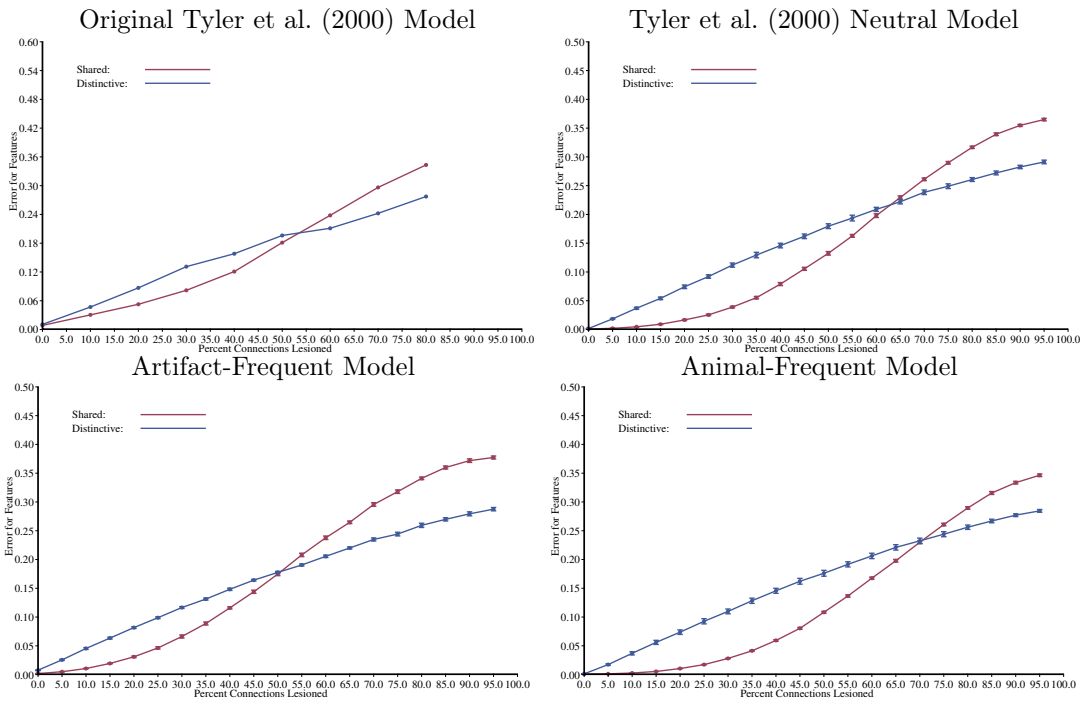


Figure 5.65: Mean absolute error for shared and distinctive perceptual features for living things (Compare with Tyler et al., 2000, fig. 4.)

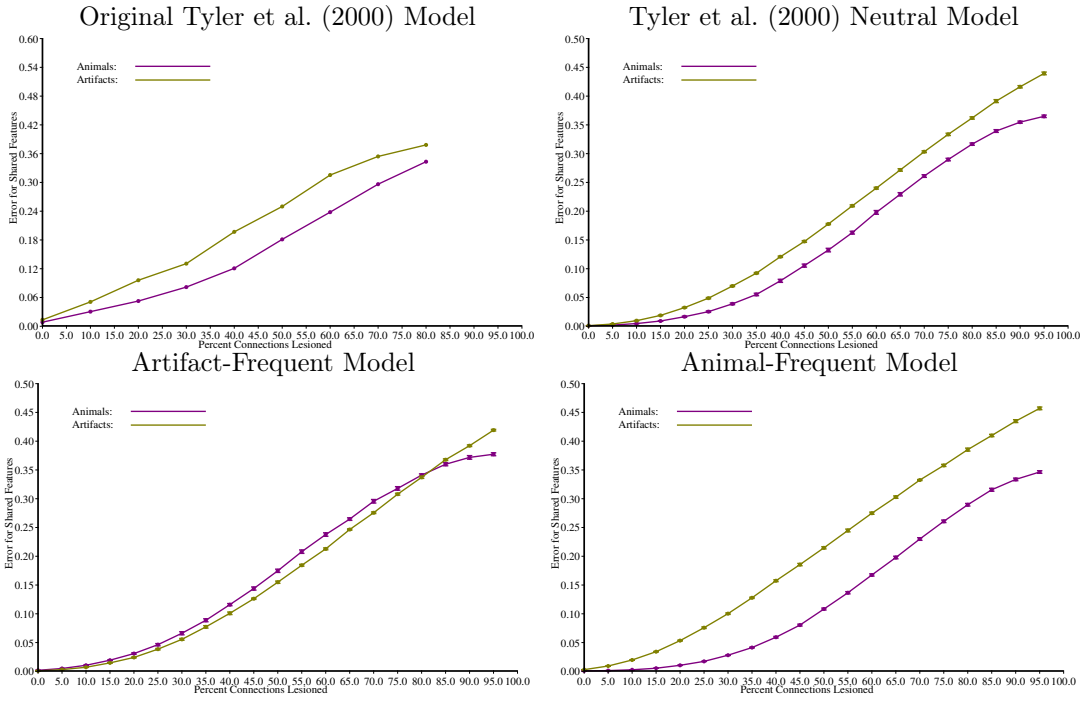


Figure 5.66: Mean absolute error for shared perceptual features for artefacts and living things (Compare with Tyler et al., 2000, fig. 5.)

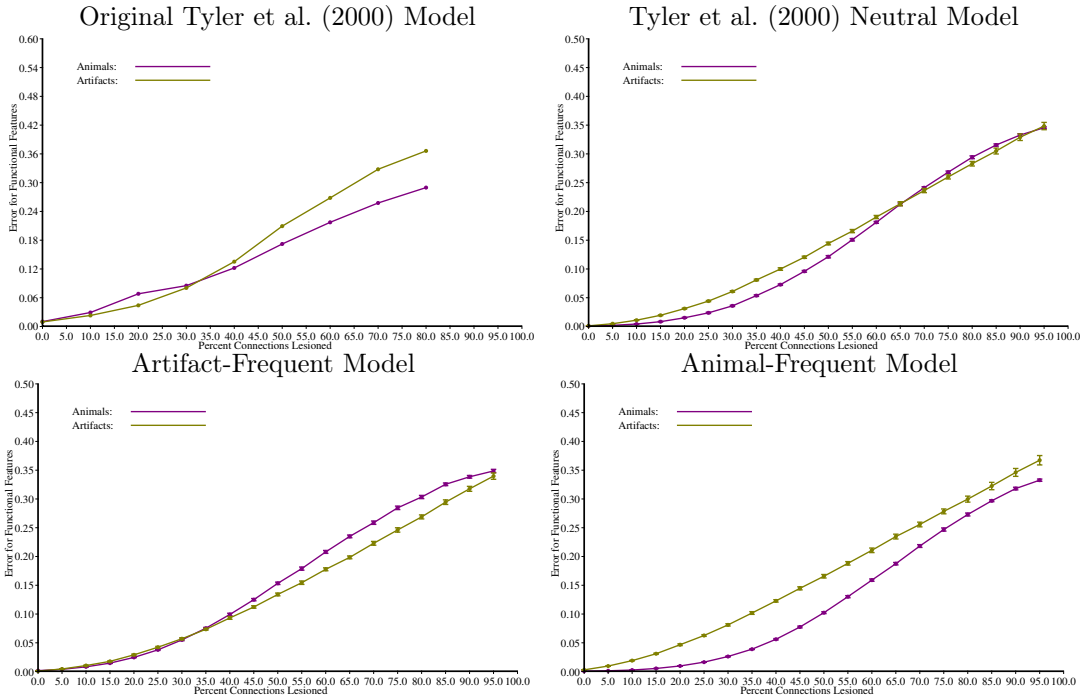


Figure 5.67: Mean absolute error for functional features for artefacts and living things (Compare with Tyler et al., 2000, fig. 6.)

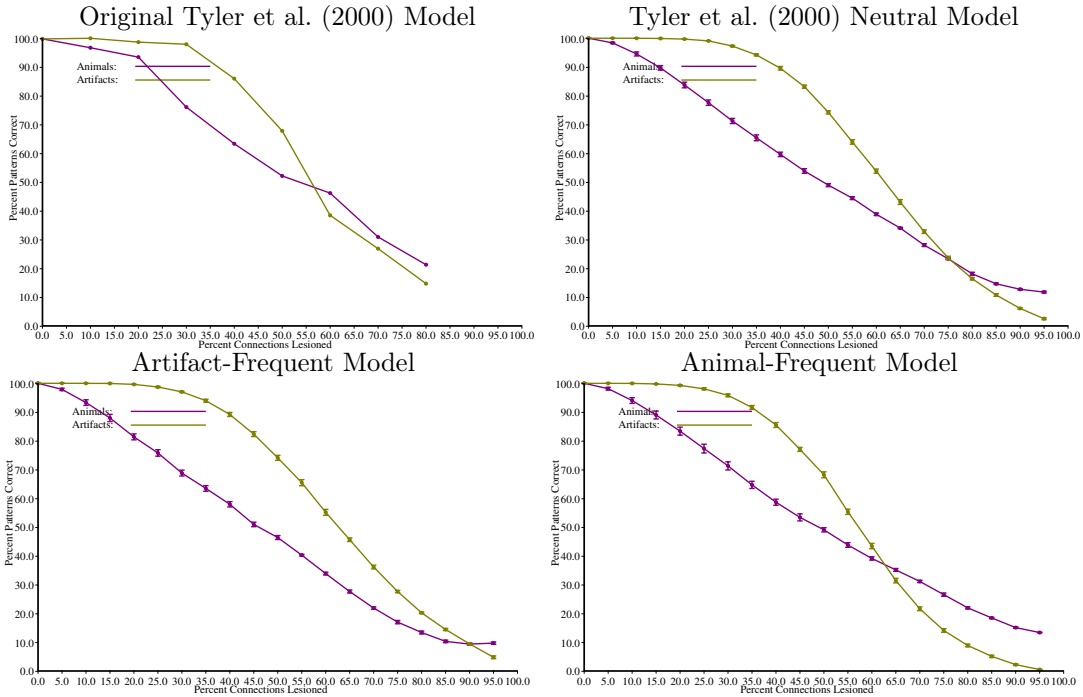


Figure 5.68: Percentage of patterns correctly mapped for the two domains (Compare with Tyler et al., 2000, fig. 7.)

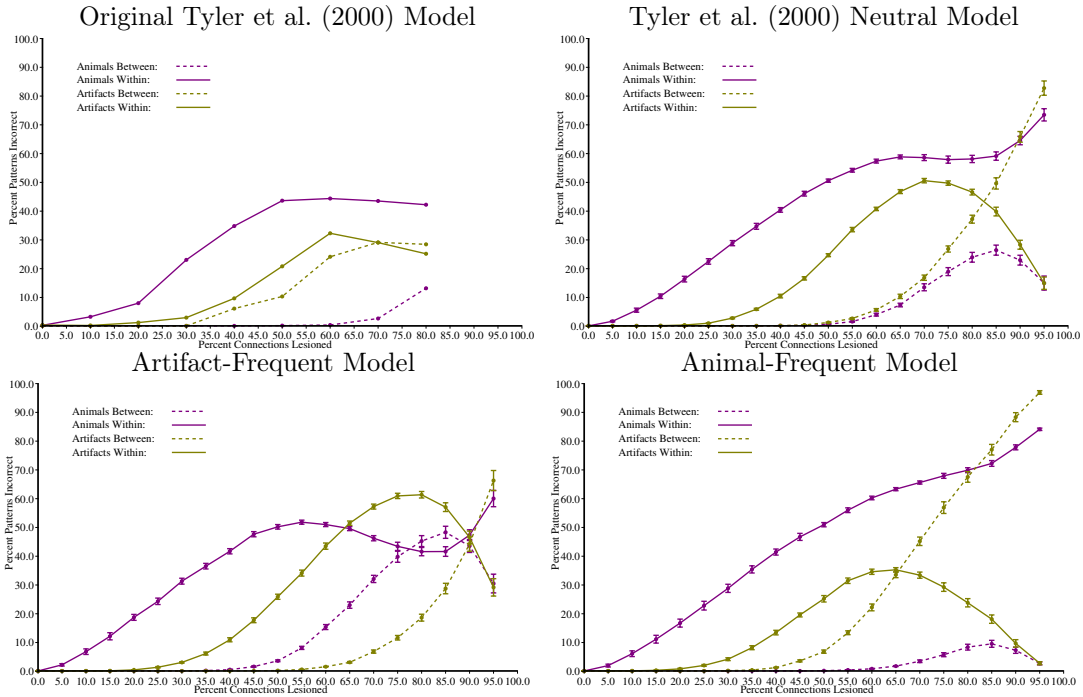


Figure 5.69: Percentage of between- and within-category errors per domain (Compare with Tyler et al., 2000, fig. 8.)

## 5.6 Discussion

This chapter presents a successful replication of the results in Tyler et al. (2000), adding credibility to their conceptual structure account. As everything was based on their article, but otherwise independently programmed, it also confirms that their theory and implementation details suffice to create a complete replication. Their results and these show very strong support for the notion that the distribution of features, both within and between concepts, can provide a mechanism for driving category-specific deficits. In other words, while the simple patterns are created based on attempting to code for their theory, the results support their hypothesis that if features are unevenly distributed across categories then that can ground a mechanistic account for category-specific deficits.

Replication is almost complete in terms of the original model ($FF_{epoch\text{-}wise}$), with a superficial difference: because the sampling here is significantly higher then in Tyler et al. (2000) the graphs presented in this chapter are smoother and better approximations for an average model, as are those presented in appendix C. This of course hides the fact that many models fall on various points of the spectrum with respect to showing pronounced category-specific biases one way or another, as initialisation noise can cause small differences between models. However, a larger sample of models allows for a better overall picture of their behaviour post-lesioning.

Another issue with replication, albeit arguably a small one, is that in Tyler et al. (2000),

the authors expected their results for the difference between functional features per domain (in their figure 6) to be like the actual results in Figure 5.53. In other words, they predicted the results obtained here, but their own results do not match that prediction. This leads to the belief that there must have been either a bug in their code for running the task and/or for generating the graph, or a sampling error due to their implementation being run fewer times. The original Tyler et al. (2000) predictions for this graph does indeed match our results. (See Cooper & Guest, 2014, for discussion of these issues.)

The other types of models ($FF_{pattern\text{-}wise}$ and BPTT) also show almost identical patterns of behaviour post-lesioning as the epoch-wise, even though BPTT has recurrent connectivity. This invariance over feedforward and recurrent connectivity lends even further support to the Tyler et al. (2000) and, thus the slightly extended model in Taylor et al. (2007), accounts of semantic cognition. So while the three different implementations of this model are relatively simple, they all capture the predictions set out in Tyler et al. (2000). Moreover, while there is some variance in the space of models created based on the original Tyler et al. (2000) specification, the patterns are extremely robust over a large range of implementation variables (see appendix C).

Given the results and replications discussed so far, the conceptual structure account as given in the Tyler et al. (2000) model is very robust with respect to modelling methodology and implementation, especially in comparison to the incarnations of the hub-and-spoke models in chapter 3. The conceptual structure model shows that features that are correlated are indeed preserved from damage when compared to sparsely correlated features, both within and between patterns and categories. This mechanism is also used in the hub-and-spoke model, since both models espouse similar theoretical positions in terms of feature preservation. In the results presented here it appears that the mechanism, i.e., that correlation implies preservation, is sufficient to model the kinds of category-specific deficits seen in HSVE and other types of patients. This is done without recurrent connectivity and appealing to attractor states; notwithstanding, the hub-and-spoke model was originally presented as an SD model, as opposed to one for category-specific deficits. However, as shall be seen in chapter 6 the hub-and-spoke model also provides an account for category-specific deficits.

The original Tyler et al. (2000) model's results are further supported by the results of an augmented model presented in Greer et al. (2001). This model uses the same architecture but with more complex, and psychologically-plausible patterns based on participants descriptions of features. In other words, the same kind of process of feature extraction that the Rogers et al. (2004) model used to create patterns is used in the feed-forward Tyler et al. (2000) architecture

to create the Greer et al. (2001) model. Since their results show the same patterns of damage as the original model this adds strength to the claim that correlation implies preservation and that distinctive and shared features are distributed unevenly across domains.

More support for the scientific usefulness of the conceptual structure account, and by extension the models based on it (Tyler et al., 2000; Greer et al., 2001) is that their predictions about preserved features have been used to predict patient behaviour. A patient, RC, performed as predicted when tested on specific features (Taylor et al., 2007). He was tested on specific items and asked if they posses certain properties, for example: "a butterfly – does it have legs?", and "ambulance – does it have wheels?" to test shared features for animals and vehicles; and "zebra – does it have black and white stripes", and "drum – is it round and hollow?" to check RC's abilities on distinctive features. His performance supports the predictions that shared features are preserved, while distinctive features are lost.

The modelling results presented in this chapter also support the notion that familiarity and frequency of stimuli can affect semantic memory organisation within the conceptual structure model in a way that appears to parallel patients. Moreover the results presented here provide a way of combining psycholinguistic knowledge for concepts storage with that of HSVE patients who show a category-specific effect or otherwise. This provides support to similar models, such as French and Mareschal (1998), which proposes that psycholinguistic properties of concepts affects reaction times for feature verification for both domains of knowledge in much the same way as the modelling work presented here.

## 5.7 Summary

This chapter presents a successful replication of the original conceptual structure model by Tyler et al. (2000), as well as implementations with recurrent connectivity. The implementations show that feature distribution in and of itself provides a mechanism to explain the differences in concept preservation after lesioning damage, as predicted. The more a feature is shared within a domain and the more other features it correlates with within a pattern and between patterns the more likely it is to be preserved post-lesioning. This is the case as a general rule of neural networks and will occur regardless of architecture. However, it is a useful principle to explore since it seems to offer some parallels and perhaps an explanation of how patients can present with category-specific deficits. The second part of the chapter demonstrates that neuropsychological and neurolinguistic properties of semantic cognition, such as familiarity and frequency of stimuli and concepts, affect differences between the two domains. This means

that in addition to (or maybe regardless of) the inherent domain structure in terms of the distribution of features, it is possible for domains to differ because of life decisions (such as specialising in a specific domain of knowledge).

# Chapter 6

# Modelling category-specific deficits in the hub model

## 6.1 Overview

In this chapter, the hub-and-spoke model (see section 2.4) is used to model category-specific semantic impairments. Modelling such intra-semantics deficits was done both in the original Rogers et al. (2004) paper and subsequently in Lambon Ralph et al. (2007). As in the original work, these deficits are modelled for SD by removing connections and for HSVE by adding noise to connection weights. The lesioned networks are tested on the confrontation naming, sorting words and pictures, and drawing and delayed copying tasks. The results are not straightforward. On the one hand, the only case in which the category-specific effects are qualitatively comparable to that of patients is when modelling SD patients on the drawing and delayed copying tasks. On the other hand, in all other cases the models failed to replicate either the behaviour of the original models or the observed category-specfic deficits of patients. More worryingly, closer examination of the patient data brings into question whether SD patients actually show category-specific semantic deficits.

## 6.2 Introduction

Building on the work in Rogers et al. (2004), Lambon Ralph et al. (2007) created an extended version of the hub-and-spoke model, which provides a model for category-specific deficits in herpes simplex encephalitis (HSVE) patients. This is confusing in light of the fact that the

Rogers et al. (2004) version of the hub-and-spoke is also able to model category-specific deficits, although in that case in SD patients. This chapter will examine how the two proposed types of deficits are modelled in the hub-and-spoke model and attempt to tease apart if and how both can be modelled.

In patients, the canonical histopathological aetiologies underlying category-specific deficit are herpes simplex virus encephalitis (HSVE), a viral infection, and less often brain injury due to head trauma, or stroke. The general consensus in the literature is that SD only very very rarely is the cause of category-specific deficits. more on this in section 8.1, and more on the specifics of HSVE in section 8.3.

Despite the clinopathological inconsistencies with respect to SD and category-specificity, Rogers et al. (2004) use a similar mechanism and explanation as Tyler et al. (2000) to model category-specific deficits. In other words Rogers et al. (2004) depend on the imbalance of features within patterns, in order to appealing to both the principles behind the conceptual structure theory (subsection 1.3.5) that correlation of features implies preservation from damage, and by extension the sensory/functional dichotomy (subsection 1.3.3). In the original description of the model the authors claim that "error types should vary depending on the density of the semantic neighborhood. Specifically, domains with a high degree of similarity structure offer more opportunities for the semantic system to be "captured" by incorrect attractors and, hence, more opportunities to make errors of commission. Unstructured domains offer fewer such opportunities, and consequently we would expect to see a greater proportion of omission errors in such domains." (Rogers et al., 2004, p. 219) This implies that the authors believe that at some point the semantic deficits resulting from SD will manifest as category-specific. Note that their patient data hints at such an effect, although it is not statistically significant (see section 8.1).

## 6.3 Category-specific lesioning damage

For Rogers et al. (2004), the dynamics of the hub network itself inherently give rise to category-specific deficits by lesioning of weights by setting them to zero, previously described section 3.2. Although, the category-specific deficits in Rogers et al. (2004) are not those of HSVE, but a type purported to arise within SD, more on this in section 8.3.

On the other hand, the same authors model HSVE-like damage using noise applied over the connections in increasing percentages. Lambon Ralph et al. (2007) propose a hypothesis about the kind of damage Gaussian noise might cause — that attractor states when damaged by noise

are damaged in qualitatively different ways, thus causing category-specific dissociations. They base the use of noise damage, to compliment disconnection for SD-like deficits, on neuroimaging research. Specifically, Noppeney et al. (2007) report largely overlapping (but not completely) loci of damage for both SD and HSVE patients, which Lambon Ralph et al. (2007) believe supports the notion that it is the quality of the damage that causes the category-/domain-specificity as opposed to any neuroanatomical localisation. A central theoretical claim is that category-specific and general semantic impairments are seen as functionally but not structurally different. They also state that this "particular form of disruption employed to simulate HSVE was, however, selected post hoc precisely because, given an adequate understanding of how the model works, it seemed likely to produce an HSVE-like pattern of behaviour." (Lambon Ralph et al., 2007, pp. 1132-1133). In their own words:

> The second form of damage [i.e., noise added to connections] was motivated by considering how processing in the intermediating layer might be disrupted so as to produce the category-specific pattern typically observed in HSVE. Specifically, we damaged the model by disrupting the values of the weights projecting in or out of the semantic layer with increasing amounts of random noise. This manipulation distorts the signals passing between layers without attenuating them: with increasing damage, inputs can still strongly drive the semantic units, albeit in increasingly random directions. This in turn means that the model will tend to confuse items with similar internal representations. Since animals tend to have somewhat more similar internal representations than do artefacts, we reasoned that this form of disruption would tend to produce a category-specific impairment—and indeed, subject to this form of disruption, the model consistently produced a category-specific naming deficit of a magnitude comparable to that observed in the patients.
>
> (Lambon Ralph et al., 2007, p. 1132)

Lambon Ralph et al. (2007) only show individual lesion-levels as required. They do not present the full longitudinal scale of test scores post-noise damage. This makes comparison slightly difficult when replicating, but Lambon Ralph et al. (2007) do include the percentage of damage or amount of noise for their results. All the tasks are run as before in chapter 3, with a different form of lesioning to conform to the specifications of Lambon Ralph et al. (2007). Specifically, for network $\text{BPTT}_1$ with noise added to each connection with a maximum standard deviation increasing till it reaches a value of 1.45 on the final data point. It is not clear why they pick that specific range of noise to add to their model. In order to investigate how well their

model parallels the patients they match their model's behaviour with that of patients' in the word-to-picture matching task and then use the same configuration to model the other semantic tasks. The model's behaviour is reported at all levels of lesioning to allow for a broader picture of its scores in the relevant semantic tasks.

## 6.4 Confrontation naming

This task is run as before, in section 3.3, with the exception of noise being used to perturb connection weights. The expected result based on patient data here is that both lesioning types should look similar in Figure 6.71 as this is an overview of naming regardless of domain. Based on Lambon Ralph et al. (2007) one would expect the noise damage to produce fewer errors, however this does not seem to be the case. The two forms of damage seem to produce similar results, with the order of errors roughly the same. Noise with standard deviation of 0.87 can be directly compared to lesioning 60% of connections, as both result in damaged models that achieve 0.54 proportion correct on the naming task. Zeroing damage to the $\text{BPTT}_1$ network shows slightly higher probability of cross-domain errors the higher the percentage of lesioning compared to when noise is applied to connections. Additionally, at complete disconnection, when all weights are set to zero, as expected all naming responses are omissions, with noise of course such a result is not possible since noise will always allow the network to produce a response. There are no HSVE patients to which this task can be compared to, since no longitudinal data is presented in Lambon Ralph et al. (2007). However, based on comparing the two lesioning types overall on naming it does not seem that they differ in terms of the effect they have on the behaviour of the network.

To examine further the potential category-specific differences the naming data are analysed by domain in Figure 6.72. When looking at each domain separately, it can be seen that the two types of lesioning continue to show very similar category-specific patters across all types of error. For the disconnection type of damage, the prediction in Rogers et al. (2004) "is that error types should vary depending on the density of the semantic neighborhood. Specifically, domains with a high degree of similarity structure offer more opportunities for the semantic system to be "captured" by incorrect attractors and, hence, more opportunities to make errors of commission. Unstructured domains offer fewer such opportunities, and consequently we would expect to see a greater proportion of omission errors in such domains." (p. 219) The less structured domain referred to here is that of inanimate objects; they propose that artifacts have less structure because they lack the self-similarity found within $\ulcorner animals \urcorner$.

In short, it is proposed that "[a]s [disconnection] damage increases, errors of omission are more likely to occur in the domain of artifacts at all levels of severity, whereas errors of commission occur relatively more frequently for animal items." As can be seen in both the patients and the model in Rogers et al. (2004) (see Figure 6.72) it is indeed the case that animals are marginally more likely to produce a semantic or superordinate error — and that inanimate objects are slightly more likely to elicit an omission. However, this does not replicate for either forms of damage in the models discussed here. What happens instead is that all types of error seem roughly equally likely, with the exception of cross-domain errors which seem to be more likely if the target is an artifact.

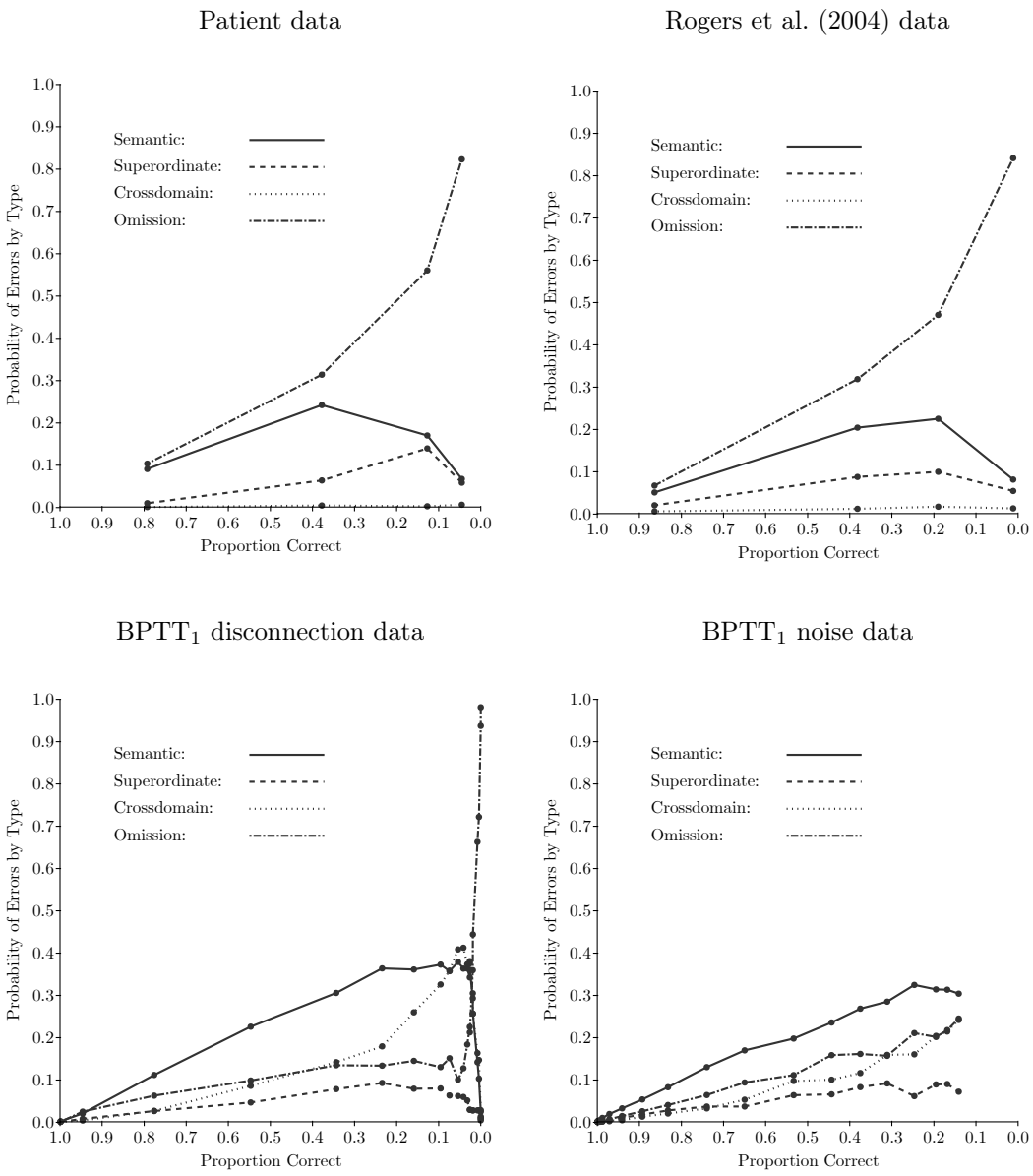


Figure 6.71: Four naming graphs representing, going clockwise, the patient data, the original model, BPTT$_1$ with noise damage, and BPTT$_1$ with disconnection damage.

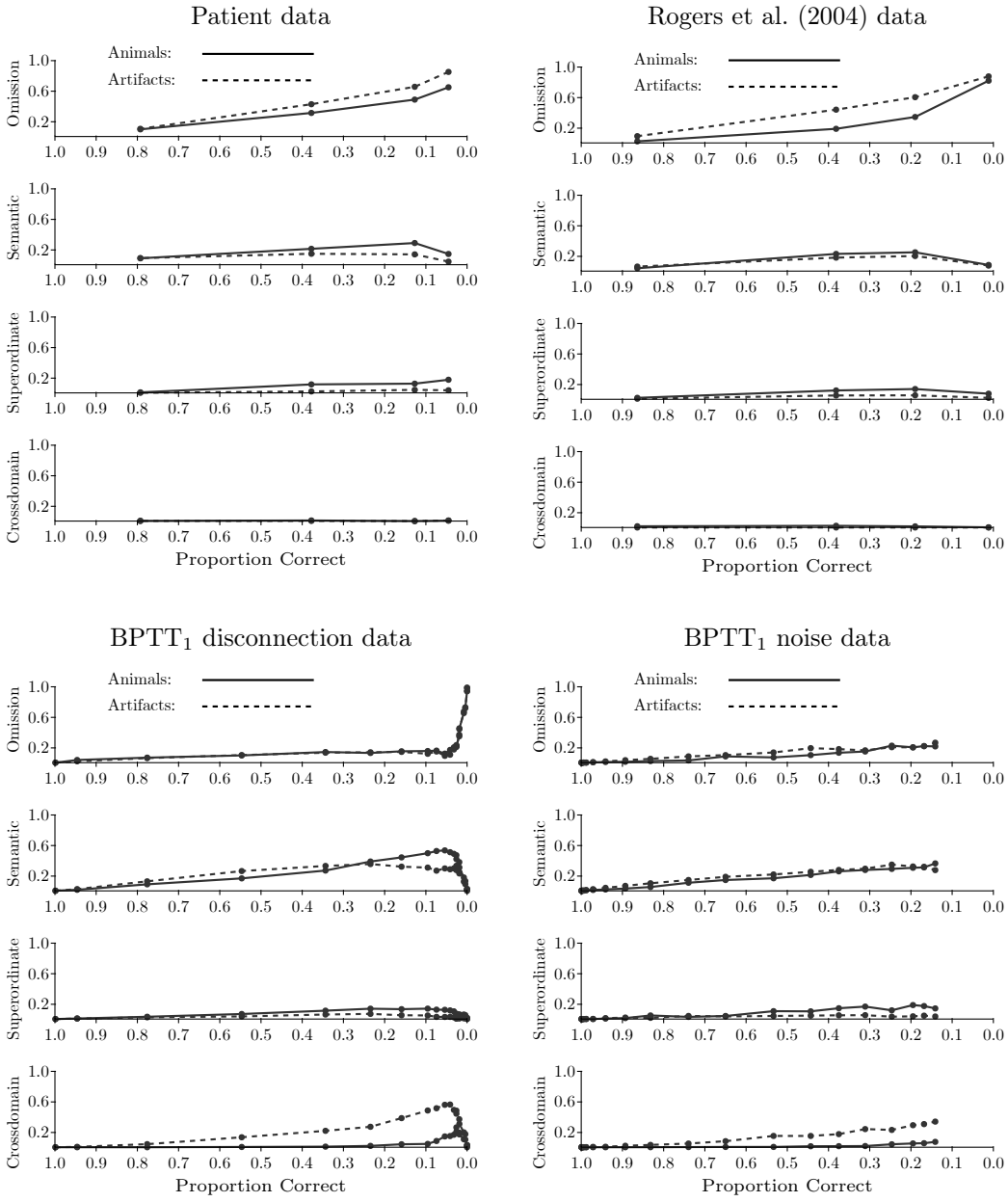


Figure 6.72: The naming results for each domain for the patient data, the original model, $BPTT_1$ with noise damage, and $BPTT_1$ with disconnection damage.

## 6.5 Sorting words and pictures

The sorting task is carried out as in section 3.4, but again weights are now damaged with noise as well as using direct disconnection. The prediction here is (even though this task is not carried out in Lambon Ralph et al., 2007) that sorting will be easier in the noise condition, because as stated in the original two publications HSVE patients are more capable at carrying out semantic tasks than SD patients. This effect was also found in Lambon Ralph et al. (2007).

In Figure 6.73 all four graphs are shown, depicting both the original Rogers et al. (2004)

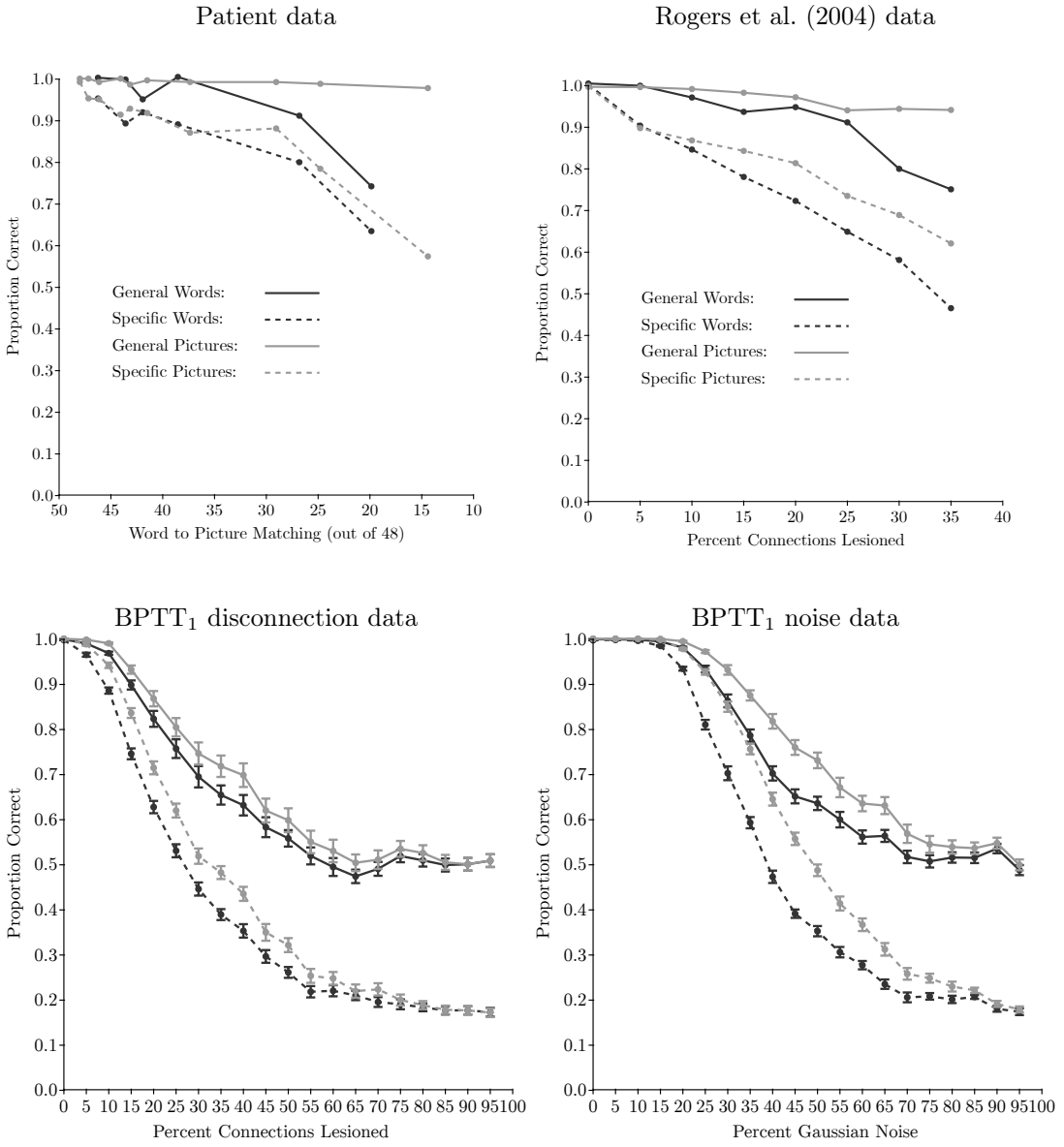


Figure 6.73: Error bars, where present, represent the standard error of the mean, in the case of the original Rogers et al. (2004) model none are provided. In the case of noise, 100% damage corresponds to Gaussian noise with standard deviation of 1.45.

patients and model, and the noise and disconnection damage to BPTT$_1$. It can be seen that apart from not replicating the required ceiling effects discussed in section 3.4, the noise and the disconnection show no difference in how they affect the network's behaviour. Regardless of the type of damage, sorting performance invariably tends towards chance levels for each domain and category. This indicates that it is not the case that noise damage is a more mild form of damage.

## 6.6 Drawing and delayed copying

The drawing and the delayed copying tasks are again as carried out in section 3.5 with both types of damage on the $BPTT_1$ network. As Lambon Ralph et al. (2007) only gives naming task results, comparison is not possible for the noise condition, but the disconnection damage can be compared to the category-specific predictions in Rogers et al. (2004).

In Figure 6.74, the overall drawing and copying scores are shown for noise and disconnection. The difference between the two types or damage is almost non-existent, with the only difference being a function of the severity of the respective form of damage. The same can be said for Figure 6.75, where the difference between the two types of damage is even less pronounced. The problem here is that while all three hub-and-spoke models show the same trend, animals are more likely to produce omission errors, the three SD patients show no clear category-specific effect.

For the intrusion errors, shown in Figure 6.76, given they are so rare in patients and the original model, it cannot be said that either of the two forms of lesioning are effective at capturing the behaviour of patients. This indicates that no level of lesioning can match both the intrusion and the omission errors of the patients nor those of the original model.

## 6.7 Discussion

This chapter sought to investigate the potential for HSVE-like and/or category-specific-like patterns of dissociation in the hub-and-spoke model. This type of semantic impairment is seen in HSVE patients, although, in Rogers et al. (2004) they do indeed model category-specific dissociations, basing this on the small sample of SD patients. This is proposed as part of the hub account by both Rogers et al. (2004) and Lambon Ralph et al. (2007). Notwithstanding, the two papers model different (proposed) category-specific patient groups in different ways. This creates some confusion (see section 8.1), which is further complicated by the proposed type of lesioning in Lambon Ralph et al. (2007), disconnection, failing to replicate and produce a category-specific effect.

Rogers et al. (2004) propose that category-specific deficits arise in the hub-and-spoke model because of the intrinsic nature of the attractors that emerge, which they believe bear a close relation in form and function to representations in the anterior temporal lobe. As mentioned, Tyler et al. (2000) propose a similar explanatory mechanism, without requiring attractors, in their model of category-specific deficits. Appealing to the dynamics of attractors requires exploring their behaviour carefully before and after damage. However, this part of the evidence

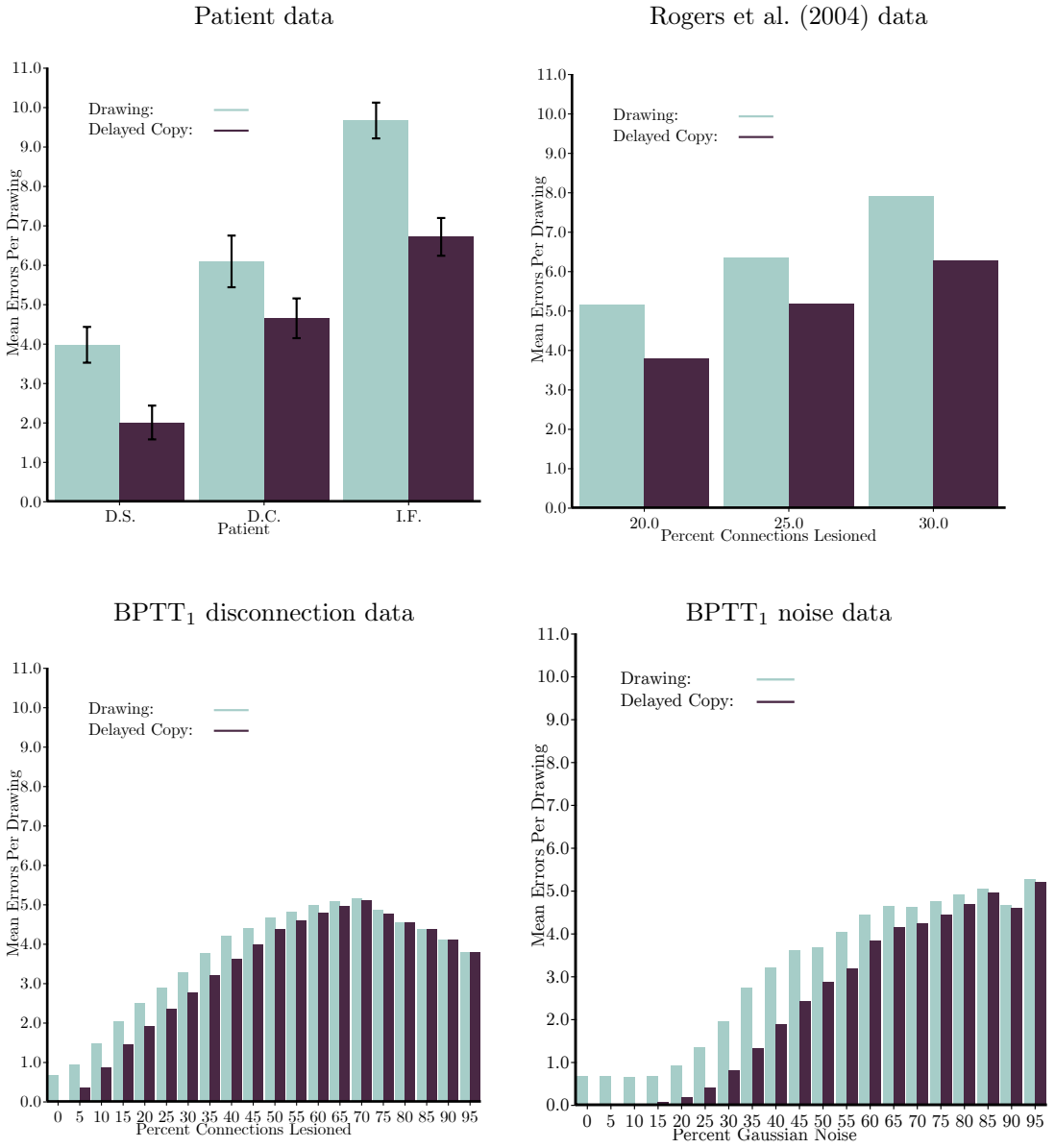


Figure 6.74: Error bars, where present, represent the standard error of the mean, in the case of the original Rogers et al. (2004) model none are provided. Again, in the case of noise, 100% damage corresponds to Gaussian noise with standard deviation of 1.45.

is not presented by Rogers et al. (2004). Instead, indirect evidence from semantic task results is. The problem with this is that semantic task behaviour could be explained perhaps more simply without appealing to attractors.

Moreover, both Rogers et al. (2004) and Lambon Ralph et al. (2007) make explicit appeals to attractor states and internal representations in their respective implementations of the hub model, they appear to contradict each other. For example, with respect to the internal states of the hub model Rogers et al. (2004) claims that "[a]s [disconnection] damage increases, errors of omission are more likely to occur in the domain of artifacts at all levels of severity, whereas

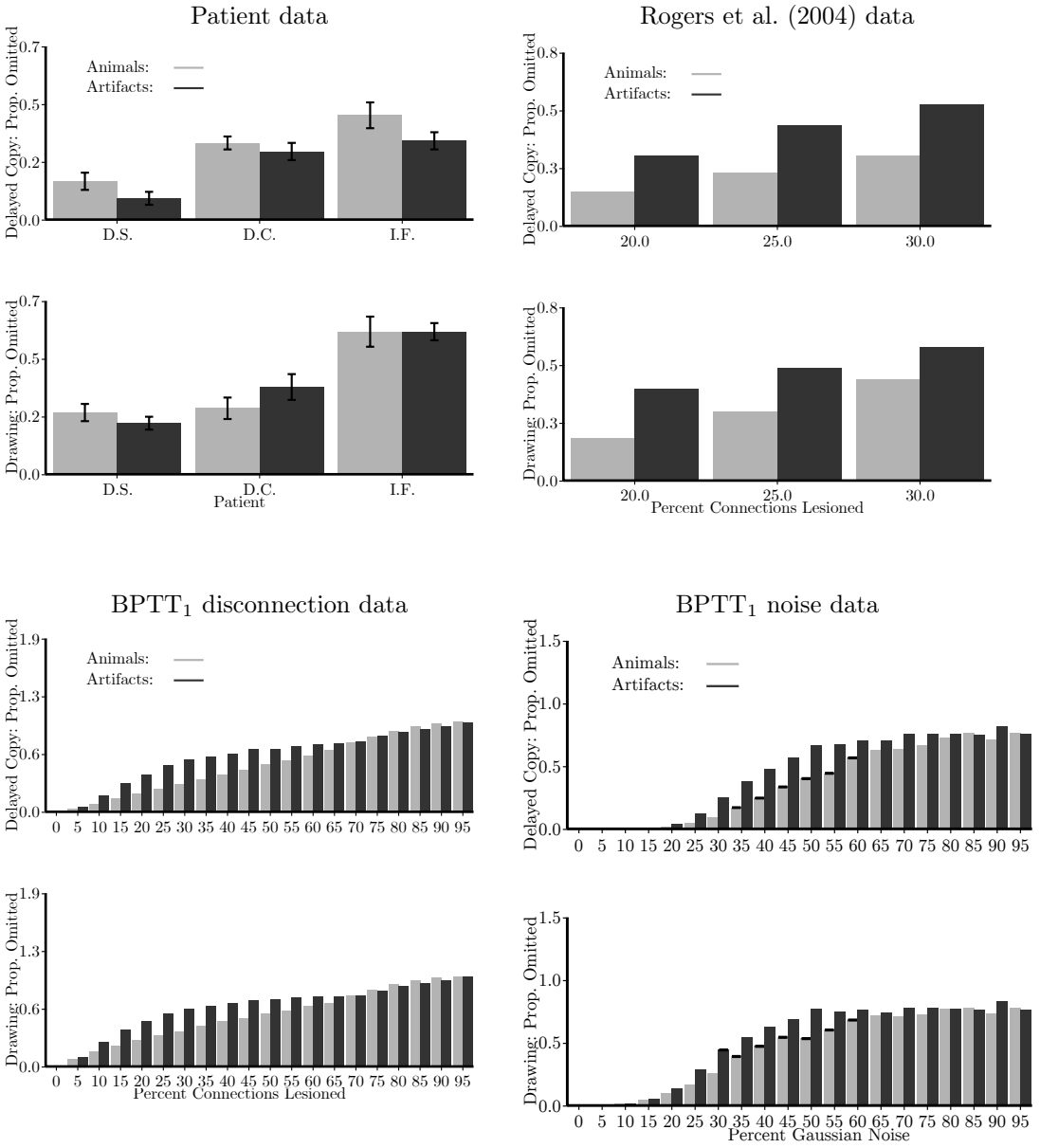


Figure 6.75: Error bars, where present, represent the standard error of the mean, in the case of the original Rogers et al. (2004) model none are provided.

errors of commission occur relatively more frequently for animal items" (Rogers et al., 2004, p. 219). On the other hand, Lambon Ralph et al. (2007) propose that disconnection damage "does not exhibit a category-specific impairment: naming is equally poor for both domains and the overall degree of naming impairment is comparable to that observed in the patients" (p. 1132).

The results obtained here imply that noise is not qualitatively different to disconnection in terms of naming. If any difference does exist it is in the quantity of damage as a factor of the amount of noise/disconnection. In other words, if a connection is completely lost, or merely impaired, the number of errors in a task will be a function of how many connections are

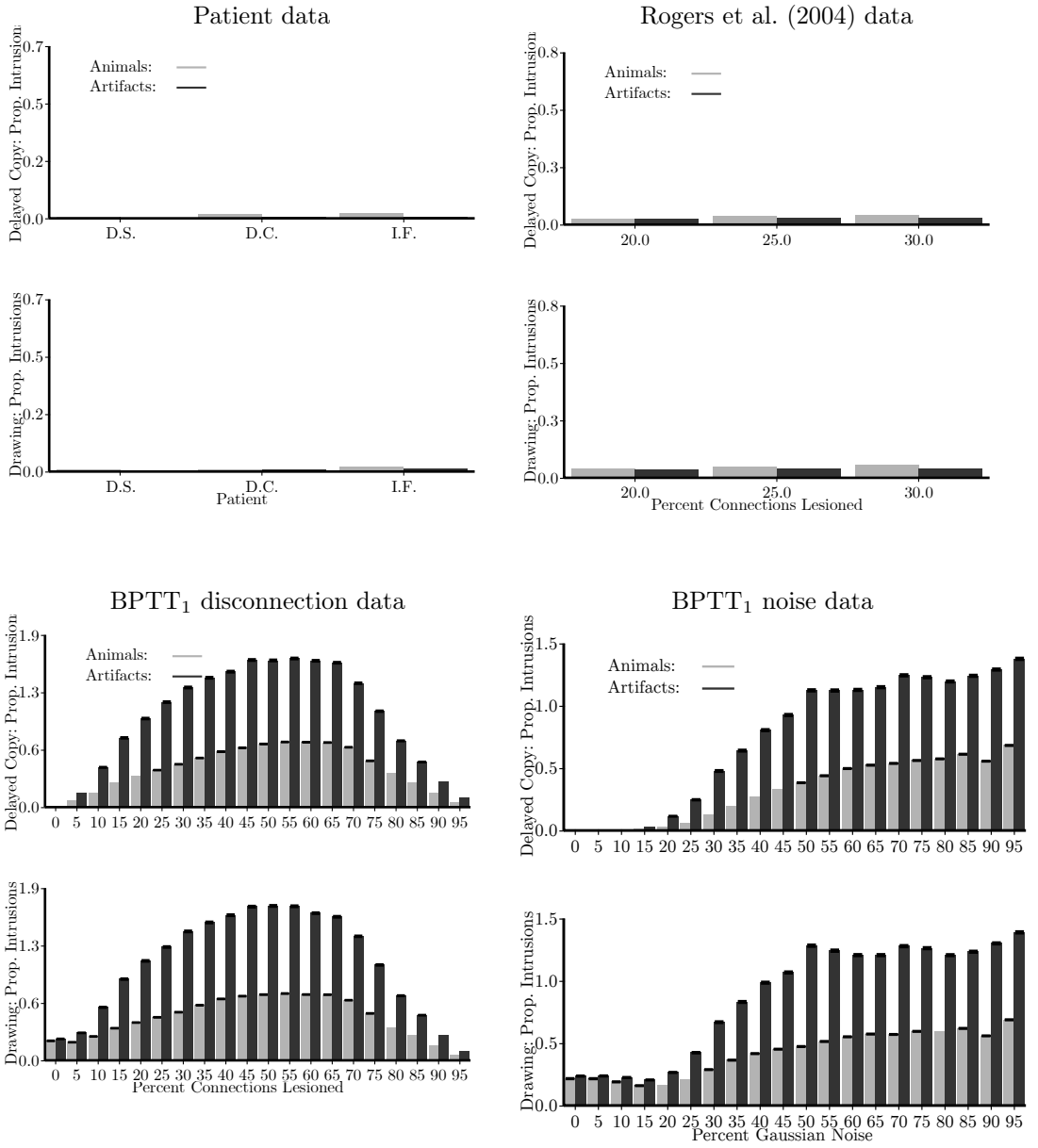


Figure 6.76: Error bars, where present, represent the standard error of the mean, in the case of the original Rogers et al. (2004) model none are provided.

thus affected and not the type of damage. But because disconnection damage produces a more pronounced effect than noise in the case where a network is run with most connections missing (because this essentially means no activations are being propagated at all) it might superficially seem like there is a qualitative difference in such extreme cases. In most other cases, with less noise/disconnection damage the effect of lesioning on task behaviour is much the same and hub-and-spoke models can be created that match each other in post-lesioning behaviour using either method. It remains to be seen if this is an issue with the specific reimplementation considered here or with the hub-and-spoke architecture and recurrent networks in general.

## 6.8 Summary

In this chapter, the two types of lesioning damage, both proposed to have the ability to cause category-specific effects in the hub-and-spoke model, have been investigated. The first type, which was also examined in chapter 3, is used in Rogers et al. (2004) to model SD patients (whom they claim can show category-specific effects). The second type of network damage, adding Gaussian noise to connection weights, is used in Lambon Ralph et al. (2007) to model the category-specific patterns of dissociation seen in HSVE patients. Neither of these two forms of damage are shown to recreate the original model results, nor do they capture the dissociations seen in patients. In fact, noise damage as well as removing connections appear to be qualitatively similar forms of damage in terms of the behaviour of the damaged networks on the semantic tasks. This finding has repercussions for the generalisability of the original hub model's results, and should caution against using noise as a distinctly different form of damage to disconnection.

# Chapter 7

# Modelling category-specific semantic deficits in the modality-specific model

## 7.1 Overview

In this chapter, the modality-specific SOM-based model is damaged with the intention of simulating category-specific patterns of deficits, building upon the work carried out in chapter 4. This model can, in theory, account for category-specific affects in the same way as similar models, e.g., Farah and McClelland (1991). In other words, by making use of the sensory/functional dichotomy one can model category-specific deficits, provided a mechanism exists that links sensory properties predominantly to animals and that associates functional features with artifacts. Specifically, this chapter will examine what effect noise — the type of damage proposed in Lambon Ralph et al. (2007), as being able to give rise to category-specific impairments — has on a dramatically different modelling architecture which is nonetheless trained on the same pattern set.

## 7.2 Introduction

The modality-specific account of semantic cognition has inspired various models based on its basic principle that separate semantic stores exist per modality. The sensory-functional hypothesis has been used in other modality-specific models (see subsection 1.4.2 and section 2.5).

For example, Farah and McClelland (1991), which uses a very different architecture to the model described here, models the category-specific patterns of dissociation seen in patients as a function of the modality-specific differences in both their features distributions and storage locations in the network. In fact, the hub-and-spoke model also makes use of some of the basic principles of the modality-specific theory, however, the presence of a hub (which is amodal) goes against the assertion that all semantic knowledge is inherently modality-specific. The model described here (and previously in chapter 4) does not provide an amodal store. Every part of semantic cognition is modality-specific, and each modality is connected to the rest to allow for cross-modal intergration (e.g., a visual representaion of a concept to be associated with an auditory representation).

Notwithstanding their differences, the Farah and McClelland (1991) model and that presented here share the same set of theoretical assumptions. The modality-specific model here uses the Rogers et al. (2004) patterns, which are designed based on participant data and have a built-in imbalance between the distribution of features both between two domains and between verbal and visual features. As such, this should provide a way to determine if this imbalance (as also proposed in the original paper: Rogers et al., 2004), varies across architectures; and additionally, to see if indeed the Rogers et al. (2004) hub-and-spoke patterns can give rise to category-specific deficits. Furthermore, the association of noise, in the hub-and-spoke account, with category-specific patterns of dissociation is a partially theoretical position as it proposes that purportedly qualitatively different forms of damage can account for category-specific effects, thus precluding or diminishing the possibility of category-specific effects being caused by a different lesion locations.

In order to simulate the HSVE category-specific pattern of dissociation (i.e., with animal concepts showing a more dramatic loss when compared to inanimate things), the same type of lesioning damage will be used as in chapter 6 and Lambon Ralph et al. (2007). Specifically, Gaussian noise will be applied in increasing amounts to all connection weights. Comparison between this form of damage and zeroing will be carried out to see if the differences are qualitatively different as opposed to merely quantitative.

## 7.3 Confrontation naming task

In the naming task, the model was run as before, using the third method described in appendix B. So the network is run using classical network propagation from the appropriate input SOM to the appropriate output SOM. The results of the overall network error are displayed in Fig-

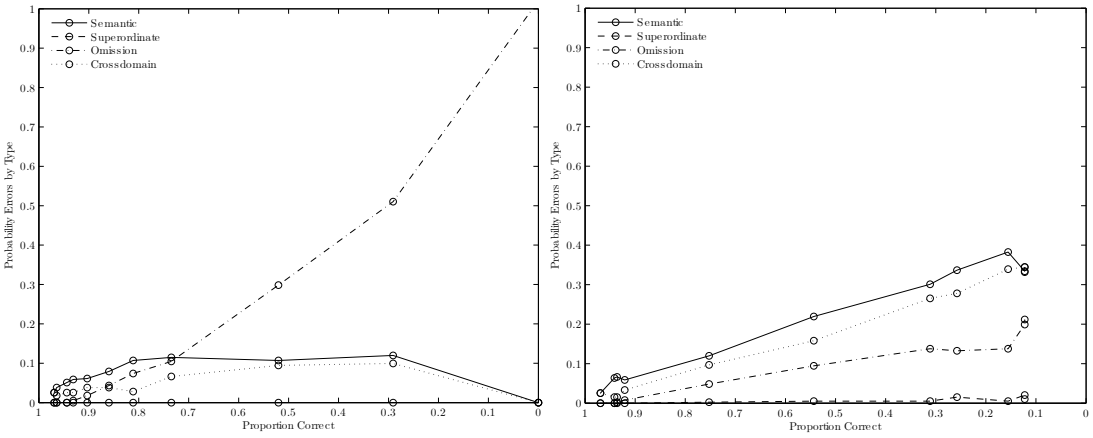


Figure 7.78: The naming scores of the modality-specific model for zeroing damage, on the left, and noise lesioning, on the right.

ure 7.78. Comparing, the zeroing damage and the noise damage, it can be seen that differences exist. For example, the most common error type for the zeroing damage is omissions, whereas for noise it is semantic errors that are most common. However, the distribution of errors is very similar in terms of the quality of damage if the noise damage is taken to represent a milder form of lesioning, because it affects the network to a much lower level of severity, resulting in a graph that represents a zoomed-in version of the first few data points in the zeroing graph. Specifically, between 1 and 0.85 in the zeroing graph the ordering of the probability of errors is the same as for the whole of the noise graph.

Breaking down the scores of naming by domain, reveals some finer differences between removing connection weights completely and introducing noise. In Figure 7.79, there is a clear difference in omissions with noise producing a category-specific effect. When noise damage is used, animals produce slightly more omissions. This is not what would be expected given the original model and patient data from Rogers et al. (2004), where the artifact domain is shown to produce more omissions – see Figure 6.72 – although it is consistent with the type of category-specific deficits found in HSVE and other patients in (Lambon Ralph et al., 2007). The lesioning by zeroing does produce the same pattern as the hub-and-spoke replication, in Figure 6.72. In others words, no discernible difference can be seen between the two domains with respect to omissions. Below omissions are the semantic errors, here the expected pattern is for a slight benefit to artifacts in the case of zeroing. For noise it is not as clear, as Lambon Ralph et al. (2007) do not break down naming errors in the same way. Either way the pattern of semantic errors appears to be similar for both types of lesioning in terms of the preservation of animals being slightly more pronounced. Gaussian noise does seem to produce a lot more semantic errors as it increases than disconnecting lesions does. Superordinate errors, are extremely low

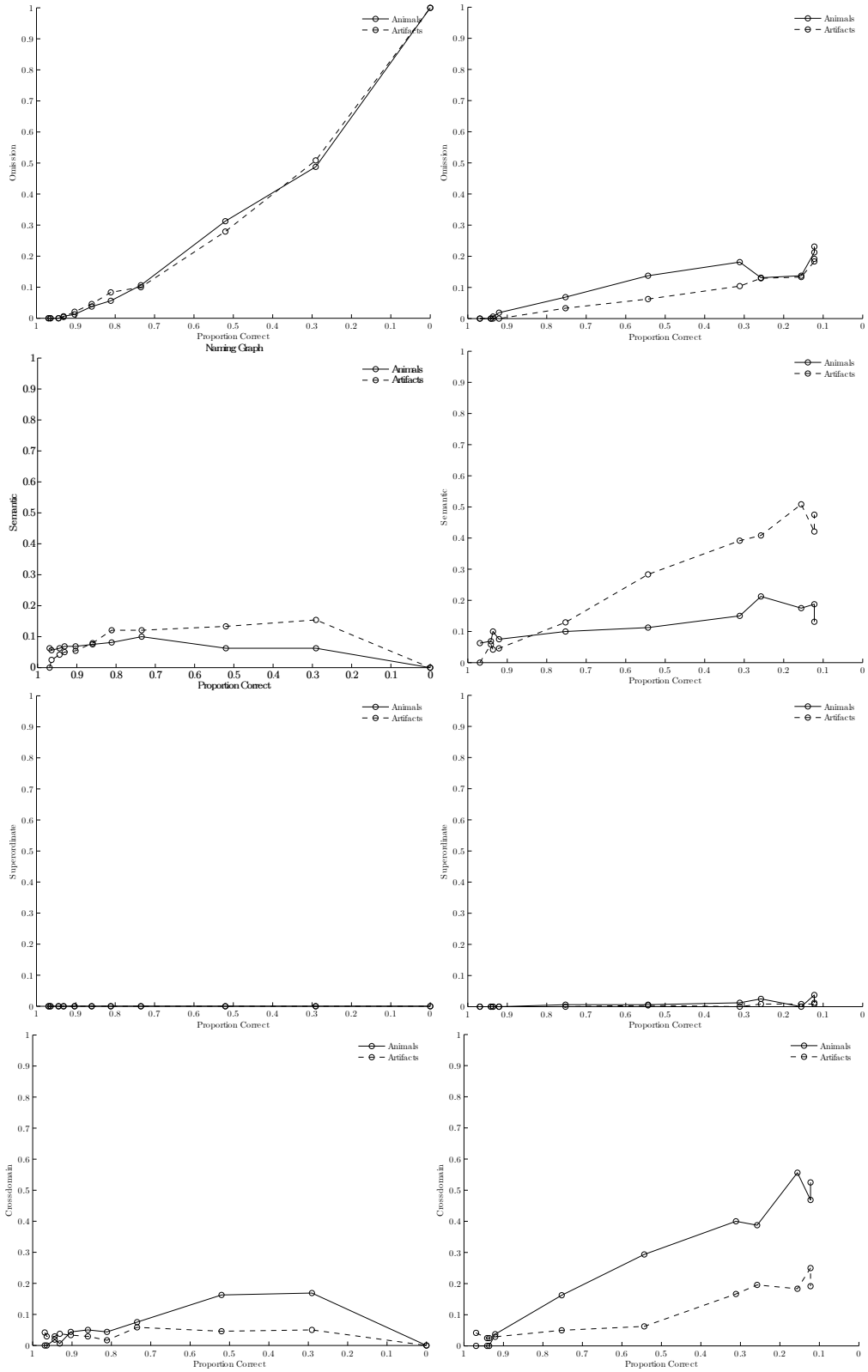


Figure 7.79: Probability of types of error by total proportion correct for each domain. The effect of removing connections is shown on the left and adding noise is shown on the right.

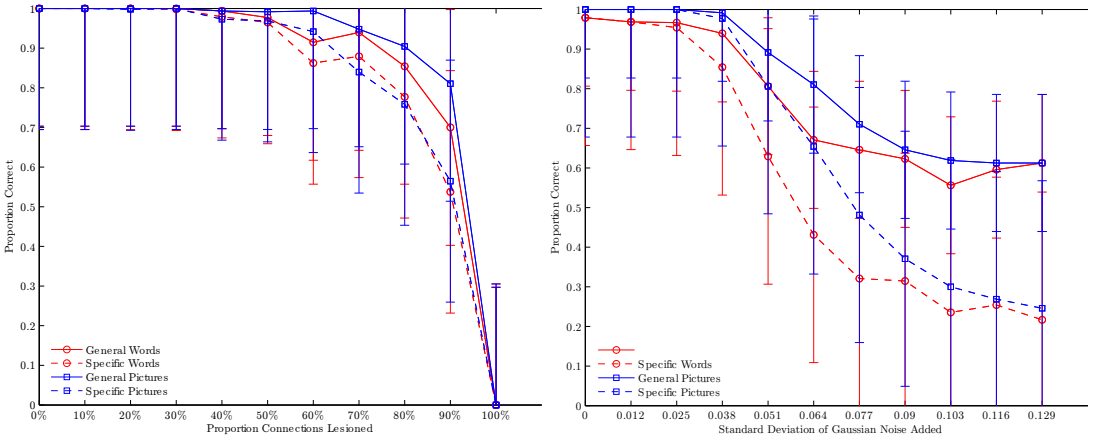


Figure 7.80: Sorting scores for both removing connections (left) and adding noise (right) with SOM radius covering the whole map. Error bars represent a SD of 1.

in this model regardless of the type of damage. The final panel shows crossdomain errors. These are the ones that should be extremely low, as in the patients and Rogers et al. (2004) model they are at floor levels. In this model though, and with both kinds of damage, crossdomain errors happen fairly frequently even at low levels of damage. Disconnection produces about a maximum of 0.2 probability of an error being a superordinate error, which adding noise is able to make almost half the errors fall in this category. This is very far from the pattern of errors the Rogers et al. (2004) SD patients show.

## 7.4 Word and picture sorting task

This task is again run as before in Figure 4.3.1, by comparing the domains/categories of the output with the input patterns using the method of subsection B.6.1. Adding noise causes a different pattern to that of disconnection in this case. After Gaussian noise is applied to connections, general sorting is more preserved than specific sorting, much like noise and zeroing of connections did in the hub model replications. The damage caused by zeroing, which was previously presented in , does not separate the general and specific sorting as clearly as the noise does Figure 7.80 in the modality-specific model. However, most importantly, none of the models here show a pattern that can parallel the Rogers et al. (2004) patients in full. So while the Gaussian noise added to connections qualitatively parallels SD patients behaviour, the proportion correct of general picture sorting should be at ceiling, if modelling sorting in SD. Lambon Ralph et al. (2007) do not present data for sorting in HSVE that can used here. Notwithstanding, it is still useful to compare noise damage with connection weight disconnection.

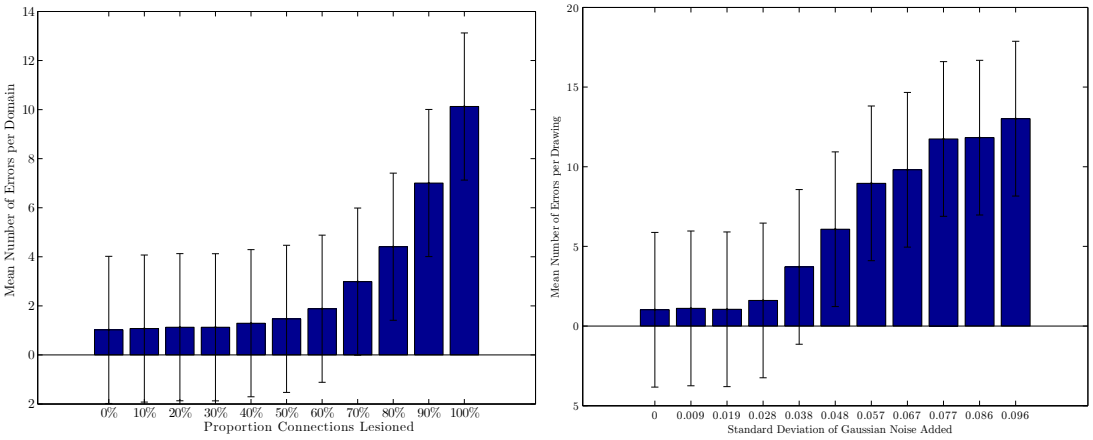


Figure 7.81: Drawing scores for both removing connections (left) and adding noise (right) with a SOM radius of 1. Error bars represent a SD of 1.

## 7.5 Drawing task

The drawing task is carried out as before, although the architecture of the model imposes certain limitations on how extensively the task can be carried out. Specifically, the implementation of the task is less able to closely follow Rogers et al. (2004) specification, as individual features of a drawing are cannot be detected in the response without denormalising the SOM's activations. This makes the task significantly different in implementation to the previous two. The differences between noise and disconnection damage appear to be minor – more a function of the severity of damage than type of lesioning (see Figure 7.81).

## 7.6 Discussion

The results presented here from the modality specific-model closely parallel the results shown in the previous chapter, 6. However, in the naming tasks, there is an important difference with the results from the hub-and-spoke model: the modality-specific model does not produce as many superordinate errors. This is because superordinate errors are problematic, in SOM-based models. The reason for that is that it is not as easy as in the hub network model, to define them. When interpreting the output states of the SOMs, i.e., looking at their relevant BMU, the classical network connections do not favour BMUs that represent superordinate domains. This is something that must be addressed, but is nonetheless a small issue since many models do not use superordinate category-levels. Ironing out these problems will shed even more light on the similarities that seem to exist between radically different models, and modelling architectures.

Notwithstanding, the fact that the results presented in this chapter qualitatively match those produced in chapter 6 is indicative of a common property of the Rogers et al. (2004) training

set. In fact, not only is this property common over implementations of the hub-and-spoke model (e.g., $BPTT_{1,\ 2}$ and the BM versions) but it is also common over different models within other theories (e.g., the modality specific and conceptual topography models). Furthermore, it seems that the results specific to this chapter, concerning noise damage to the SOM-based model, do not produce results dramatically different to the equivalent in the other models mentioned. This invariance is very interesting, and indicates that extreme caution should be taken when ascribing causal roles to model properties or implementation details. Without adequate research into alternatives and variation of "implementation" details, claiming a causal role for a hub-like topology, or any other theoretical position, is problematic.

## 7.7 Summary

In this chapter, the modality-specific model is damaged using Gaussian noise in order to model category-specific patterns of behaviour like in Lambon Ralph et al. (2007), as modelled in the previous chapter. The noise damage produces no qualitative differences between removing connections in this model, and continuing to show the same patterns of behaviour with the hub-and-spoke model. The results here replicate those found in chapter 6, showing that regardless of network topology, learning algorithm, lesioning type, and overarching theory, the Rogers et al. (2004) patterns result is very similar results post-lesioning.

# Chapter 8

# General discussion

## 8.1 Introduction

The time-line of theoretical and empirical understanding of semantic cognition, unsurprisingly perhaps, closely follows the milestones in understanding cognition and the brain general. In other words, going from largely symbolic (e.g., Fodor, 1975; Pylyshyn, 1984) to more perceptual-based theories (e.g., Newell, 1990), and connectionist approaches such as the network-based models and theories discussed in this thesis (chapter 1). In light of the history, slightly more interesting is perhaps how early on in the general understanding of semantic cognition computational modelling plays a role.

The first person to use the phrase "semantic memory" was Quillian (1966), conceptualised as the opposite or complement to episodic memory (memory for specific autobiographical events). Shortly after that, Collins and Quillian (1969, 1972) proposed a semantic cognition model based on a computer system for storing semantic information, designed in such a way as to minimise redundant features being stored. Their computational model stored features and concepts as nodes on a tree. To use an example from Collins and Quillian (1969), a ⌜*canary*⌝ is stored as a node with connections to feature-nodes, e.g., ⟨*yellow*⟩ and ⟨*can sing*⟩, as well as connections to nodes that represent other concepts, and are hierarchical in nature, e.g., ⌜*bird*⌝. ⌜*Bird*⌝ itself points to ⟨*has feathers*⟩, ⟨*can fly*⟩, and so on. Exceptions to archetypal concepts, e.g., ⌜*ostrich*⌝ point to negative features such as ⟨*cannot fly*⟩, and so on. They found that their model provided an explanation, and predictions, for human reaction times in various semantic tasks Collins and Quillian (1969, 1972).

After the Collins and Quillian (1969) model, which was almost entirely devoid of modality input — although it does contain modality-derived (pre)semantic features, see subsection 1.3.2

— other models and theories emerged that had more modality-based input. Some theories, such as embodied semantic memory (see subsection 1.3.7), are at the other end of the spectrum completely, such that no amodal knowledge exists in their derived models. In the last 30 or so years the theories for semantic cognition have overwhelmingly found a balance between the two extremes of the amodal-perceptual spectrum (see: Meteyard et al., 2012, for their version of the spectrum).

The models and theories described and replicated in this thesis all derive some of their foundations from the late 60s and early 70s in terms of features and concept structure. Given that computational accounts have been at the centre of research into semantic memory, and into semantic cognition more generally, the interdisciplinary investigation into semantic cognition owes a lot to computational modelling. This being said, some aspects (both models and theories) of the computational investigation of semantic cognition have certain drawbacks, e.g., lack of reproducibility and/or not enough grounding in empirical findings.

What this means is that while there has been a great shift from the time of Collins and Quillian (1969, 1972) away from thinking of semantic cognition as amodal, completely devoid of perceptual complications, and uniform, completely consistent regardless of the way memories are accessed, this shift has perhaps not gone far enough. While ways of understanding semantic cognition such as the modality-specific, the hub-and-spoke, and the conceptual structure theories and models, must on some level simplify the complexities of patient data, they must also attempt to capture it as closely as possible. So while a shift towards including more perceptually-derived and modality-based aspects of semantic knowledge has occurred, a related shift of suing patient and healthy participant results to guide models has not gone far enough. Many theories appear to rely more on model- and implementation-based properties, e.g., shared features (Tyler et al., 2000) and attractor basins (Rogers et al., 2004), than on human data, for support.

## 8.2 Re-examining the assumptions of the hub-and-spoke model

An example of the semantic abilities of SD patients can be found in Rogers et al. (2004) and Rogers, Lambon-Ralph, Patterson, McClelland, and Hodges (1999), who also provide their hub theory and model to account for their behaviour during semantic testing. On the one hand, their model (as originally reported) seems to produce almost exact matches for the their patients' scores in many cases (although, there are exceptions in which their model shows

different effects, e.g., in the drawing and copying task, Rogers et al., 2004). On the other, it is relevant to note that their sample is small and thus runs the risk of being uninformative, perhaps even misleading, if taken too generally. In other words, if their model as originally presented closely matches their patient data, but that data appears to be too narrow a view of the symptoms of SD, then their model is similarly too narrow.

The Rogers et al. (2004) patient data seems to suggest that SD patients perform in a certain way, and as such they put forward an expected pattern of performance on semantic tasks that the model after damage must conform to (e.g., never committing crossdomain errors when performing the confrontation naming task, because according to their model the superordinate details of items are preserved until the very final stages of the disease). It is thus important to investigate whether the assumptions behind the creation and evaluation of the original hub model are compatible with SD patients as they are described in the broader literature.

These are the central assumptions behind the original hub model about SD patients' semantic abilities:

1. General knowledge is better preserved [than specific knowledge].
2. Typical properties are over-extended to inappropriate but related objects.
3. Performance is better with pictures than with words.
4. Different patterns of errors are observed for living and non-living things.

Rogers et al. (1999, p. 1, punctuation added)

Applying the above "rules" to the confrontation naming task, it can be inferred that: rule 1. excludes the incidence of crossdomain errors, rendering them rare or non-existent; 2. allows semantic errors (i.e., patients confuse the names of things within the same category); 3. does not apply to this task – recall that for naming the input is a word and the output is (pointing to) a picture; and 4. means that in SD patients, like in HSVE patients, a category-specific effect can be seen dissociating the two domains of knowledge.

For the sorting task: rule 1. indicates that general-level sorting should be closer to healthy participant scores and better than sorting into more specific categories; 2. also adds to the expectation that specific category sorting should be more difficult for the SD patients than sorting into domains; 3. means that picture sorting will be more preserved, presumably regardless of the level of sorting, than word sorting; and 4. would apply just as in the naming task, indicating that there should be a difference between the patient scores in living versus inanimate sorting, again presumably regardless of the level of sorting.

This is indeed the case for patients' scores on both these tasks in Rogers et al. (2004), which is as expected since they created these rules based on their experience with SD patient scores. But are these general rules really applicable across the full spectrum of SD patients? Do these rules really capture the full picture? Or are they simplistic to the point of being misleading? The following sections aim to be answer these questions specifically for each axiom-like rule above, and aim to provide a view of what SD patient behaviour looks like in general.

### 8.2.1 Re-examining semantic dementia

In order to see the Rogers et al. (2004) patients in a wider context, i.e., situated on the spectrum of possible SD patients, this and the following sections will describe other samples of SD patients, with the aim of making the case that some of the hard and fast rules used to underpin the modelling success of the Rogers et al. (2004) hub-and-spoke model might not necessarily be based on representative (interpretations of) patient behaviour.

Firstly, it is important to note that the amount of cortical atrophy causes qualitative as well as quantitative differences in cognition. At the onset of SD symptoms, patients show relatively preserved cognition outside the realm of semantics. Meaning their episodic memory, and perceptual, spatial, praxic, and non-verbal executive functions are intact, as are their drawing skills, orientation in time, and simple calculation (Gorno-Tempini et al., 2011; Hodges & Patterson, 2007; Hodges et al., 1992; Warrington, 1975). In other words, in early SD, patients show a classic dissociation – semantic cognition is impaired while the rest of their abilities are largely intact. In typical SD, there is a deterioration in expressive and receptive vocabulary; the former manifesting itself as a "loss of memory for words". Words that cannot be remembered are replaced with generic terms, e.g., "thing" instead of "kettle"(Hodges & Patterson, 2007). Nevertheless, their ability to repeat even complicated words is unimpaired meaning they can use a low frequency word in a sentence, e.g., "hippopotamus", without any knowledge remaining of what that word means (Hodges & Patterson, 2007).

As the disease progresses, the symptoms not only get worse (Davies, Graham, Xuereb, Williams, & Hodges, 2004; Galton et al., 2001; G. B. Williams, Nestor, & Hodges, 2005), i.e., they become increasingly anomic, but the patients also start to display behavioural and personality disturbances in line with the other FTLD dementias (meaning that both frontal and temporal lesions have developed). These include mental rigidity, clock-watching, obsessive puzzle solving and game playing, irritability, disinhibition, altered eating behaviour, diminished awareness of emotions, and so on (Fletcher & Warren, 2011). At such an advanced stage of neurodegeneration the patients are usually also mute, not just anomic; and can suffer from

emotional withdrawal/coldness, depression, apathy, and lack of empathy (Hodges & Patterson, 2007).

In addition to the heterogeneity introduced by the stage of neurodegeneration, there is also another factor that might exist on a case-by-case level, such as the patient's experience or expertise. The effect means that if a patient is an expert in a particular domain, i.e., types of trees, that area of semantics will be relatively shielded from damage in comparison to other conceptual categories. So pre-morbid familiarity might actually affect the organisation of the semantic cognitive system, bringing about deficit patterns that appear to be category-specific (Jefferies et al., 2011).

### 8.2.2 The "classic" semantic dementia patient

Assuming that such "typical" SD patients exist, the SD patients in Hodges et al. (1992) are widely cited as being the likely candidates. Usefully, they perform very similar semantic tasks[1] and use the exact same stimuli[2] (known as the Cambridge semantic memory battery, Hodges & Patterson, 2007) as in Rogers et al. (2004). These five patients are all anomic, but have relatively spared cognition outside the semantic system. In other words they show a classic range of SD symptoms:

**P.P.** has completely lost all semantic abilities (e.g., experimenter: "Have you been to America?"; P.P.'s answer: "What's America?"), but has preserved episodic memory, drawing/copying, and visuospatial skills.

**F.M.** cannot name relatives, food, and so on, but can carry out other important tasks — thus, she too has impaired naming, but preserved visuospatial skills, episodic memory, and comprehension.

**M.C.** has naming and comprehension problems, surface dyslexia, but intact episodic memory — however, over the course of testing, she deteriorates into severe SD with the obsessive behaviours, disinhibition, mental rigidity, irritability, indiscriminate eating, etc., that go with it.

[1] "[T]he five sub-tests consist of: (i) category fluency for each of the six main categories plus two lower order categories (breeds of dog and types of boat); (ii) naming of all 48 line-drawings; (iii) picture sorting at superordinate, category and subordinate levels; (iv) picture pointing to spoken word using within-category arrays; (v) generation of verbal definitions in response to the spoken name of the item." (Hodges et al., 1992, p. 1795)

[2] Their set of stimuli "contains 48 items chosen to represent three categories of animals (land animals, sea creatures and birds) and three categories of man-made items (household items, vehicles and musical instruments) matched for category prototypically." (Hodges et al., 1992, p. 1795) This is the same set that is used by Rogers et al. (2004), see their Patient Method section.

**J.L.** has lost his naming abilities altogether — he does not understand the meanings of words, presents with severe word-finding difficulties, and shows signs of surface dyslexia, but has preserved autobiographical memories.

**E.P.** has equivalent problems with the meanings of words — she can perform, but not name, folk dances, can follow complex instructions, and her episodic memories are, as with the other patients, generally preserved.

The profile of these patients is very similar, they all show severe loss of vocabulary, and a dramatic inability to use and understand words, they have difficulties reading low frequency irregular words, but can perform at normal levels on word repetition tasks, have normal grammar, and intact autobiographical memory. They cannot provide exemplars given a category, specifically so in the case of narrowly-defined categorisation. Even if they are presented with a line-drawing of an item they struggle to provide a name. They also have difficulty categorising items into orthogonal divisions to the classical domain and categories, e.g., electrical versus non-electrical objects, although, in more general domains, i.e., animals versus objects, they fare better at dissociating between the two. They also have trouble giving sensible definitions for items (e.g., M.C. defines ⟨lion⟩ as: "Is it an animal? ...it has little legs and big ears, they sleep alot [sic], see them in shops.", Hodges et al., 1992, Appendix). Additionally, chimeric pseudo-items, made from the halves of two real stimuli, are very hard for them to dissociate from real ones. Over the course of testing, some of the patients deteriorate into much more severe dementia with the behavioural symptoms and full mutism mentioned previously. In short, they show typical SD deterioration in their memory for concepts, as defined by the clinical diagnostic criteria of Neary et al. (1998).

### 8.2.3 Rogers at al. (1999, 2004) assumptions

As touched on previously, it is important to evaluate the general applicability of the axiom-like assumptions for SD patient behaviour presented in Rogers et al. (1999, 2004). These serve in the original hub model and in general as both concise descriptions of how the SD patients' damaged semantic cognition functions and as benchmarks for evaluating models' fit to patient behaviour, and by extension their goodness as models of semantic memory. The following sections will each tackle an assumption, its evidence, its repercussions, its predictions, and its usefulness as a "rule".

#### 8.2.3.1 Is general knowledge better preserved?

More general concepts are more likely to be remembered, as evidenced by the patients in tasks such as category fluency, and sorting into general, intermediate, and specific levels of classification (Rogers et al., 2004). For the former task patients are unable to generate words when asked to list breed of dogs, or types of boat, faring better when asked to generate animals or vehicles (Hodges & Patterson, 2007, 1996). When looking at their behaviour in the latter task, it becomes apparent that their ability to sort under the first two more general conditions is still well above chance. As noted, "[s]uperordinate knowledge was better preserved than subordinate knowledge, although it is noteworthy that the normal control subjects showed the same pattern, albeit at a superior level of performance." (Hodges et al., 1992, pp. 1796-1797), meaning that this is the case for normal as well as impaired semantic cognition, and thus should be part of any semantic memory model. Superordinate feature knowledge, as opposed to superordinate classification, paints a similar picture, see Table 8.9. Superordinate features in patients are the least likely features to be given in a verbal definition, but the percentage of superordinate features is almost the same for both patients and neurologically healthy controls, meaning that proportionally the superordinate knowledge is preserved, somehow being shielded from damage.

With regards to the sorting task "[a]ll the patients even the most impaired, were perfectly capable of sorting the 48 cards at the highest order (living versus man-made), demonstrating preservation of this broad concept[.] At the category level, [f]our of the five showed no impairment, their scores falling within two standard deviations of normal." (Hodges et al., 1992, p. 1795) PP, who was the patient that struggled the most, is still above chance – her score is 72% proportion correct, chance is at 33%. These patients were subjected to a third level of sorting, which involved distinguishing between, e.g., fierce versus tame animals, and kitchen versus non-kitchen items. Scores for all of the SD patients on this task are more than three standard deviations below healthy levels attained by normal controls. This is a result of the fact they experience a loss of infrequent and irregular knowledge, as is their surface dyslexia, which causes exceptional words such as "pint" to be read like "mint". In other words, the patients show a non-verbal visual equivalent of surface dyslexia by making these typicalisation/generalisation errors, e.g., identifying a peacock as merely a bird or animal, or drawing it without its distinctive tail (Fletcher & Warren, 2011; Patterson, 2007).

Rate and Type of Attributes in Verbal Definition

| Attribute Type | AN | CS | MA | AT | SL | KH | Patient Mean | Control Mean | Worst Control |
|---|---|---|---|---|---|---|---|---|---|
| Number of attributes given (Percentage of own total production) | | | | | | | | | |
| Sensory | 158 (40%) | 39 (21%) | 53 (37%) | 91 (35%) | 59 (34%) | 103 (38%) | **83.8 (35%)** | **512.5 (57%)** | 351 (55%) |
| Functional | 154 (39%) | 81 (45%) | 62 (44%) | 113 (43%) | 75 (43%) | 106 (39%) | **98.5 (41%)** | **196.8 (22%)** | 152 (24%) |
| Encyclopaedic | 68 (17%) | 39 (21%) | 26 (18%) | 35 (13%) | 20 (12%) | 36 (13%) | **37.3 (15%)** | **125 (14%)** | 91 (14%) |
| Superordinate | 44 (11%) | 23 (13%) | 1 (1%) | 18 (7%) | 19 (11%) | 13 (5%) | **19.7 (8%)** | **62.7 (7%)** | 44 (7%) |
| Errors | 0 (0%) | 0 (0%) | 0 (0%) | 5 (2%) | 0 (0%) | 11 (4%) | 2.7 (1%) | 0 (0%) | 0 (0%) |
| Total | 424 | 182 | 142 | 262 | 173 | 269 | 242 | 897 | 638 |
| Percentage of control mean performance | | | | | | | | | |
| Sensory | 31% | 8% | 10% | 18% | 12% | 20% | 16% | | |
| Functional | 78% | 41% | 32% | 57% | 38% | 54% | 50% | | |
| Encyclopaedic | 54% | 31% | 21% | 28% | 16% | 29% | 30% | | |
| Superordinate | 70% | 38% | 20% | 29% | 30% | 21% | 20% | | |
| Total | 47% | 20% | 16% | 29% | 19% | 30% | 27% | | |

Table 8.9: "Analysis of the rate and type of attributes produced in verbal definition" (Lambon Ralph et al., 2003, table 5)

#### 8.2.3.2 Are typical properties over-extended to inappropriate but related objects?

The claim that "[t]ypical properties are over-extended to inappropriate but related objects" (Rogers et al., 1999, p. 1) provides a slightly incomplete picture of what really occurs. The reason for this is that while it is true that many SD patients appear to generalise or typicalise concepts, e.g., draw a peacock as a quadruped (because presumably most animals have four legs; Hodges & Patterson, 2007), it appears to hold only for some modalities/tasks and for some stages of the disease.

So, typical properties for a category can be applied to concepts within that category that nonetheless do not posses this typical characteristic (a visual parallel to surface dyslexia), but the real story is not so simple. What happens is that at first, *a*) a semantically related item replaces the target, not a more general one (e.g., a picture of a tiger is named as "lion"). Then, *b*) a semantic relation that is very high in familiarity and frequency replaces the original (e.g., ⌜*tiger*⌝ becomes ⌜*cat*⌝). After that, *c*) a very vague superordinate name is given instead (e.g., ⌜*tiger*⌝ becomes ⌜*animal*⌝). And finally, *d*) anomia will set in (e.g., the patient says "I do not know"), although circumlocutions (e.g., "It comes from Asia") and mutism are also possible responses (Hodges & Patterson, 2007; Hodges, Graham, & Patterson, 1995).

This "highly characteristic pattern" (Hodges & Patterson, 2007, p. 1006) is not exactly reflected in the scores of patients in Rogers et al. (2004) — see Figure 8.83 for a reproduction of their patient naming data. Their scores show an overriding tendency to make omissions (equivalent to either anomia, mutism, or a circumlocution[3]), which, given the above description of how semantic errors change over time, should not initially be the most common naming response in typical SD patients. In fact, SD patients should only be more anomic/mute than not in the late stages of the disease — unless, of course, the patients are more advanced in their disease when tested and thus have passed some of the stages mentioned in citeAhodges07. Unfortunately, details of these patients' disease progression are not given.[4]

Rogers et al. (2004) state that "[a]s the degree of impairment increases, so do the observed proportions of omission errors and to a lesser degree superordinate errors. By contrast, semantic errors initially rise with severity but then decline, with patients in the fourth quartile making fewer semantic errors on average than patients in the third. Relatively few cross-domain errors are observed at all." (p. 217) Typical features are *not* always over-extended to neighbouring concepts. For example, in table 6 of Hodges et al. (1992, p. 1797), the patients are at the earlier stages of SD and produce semantic relations to the target: the description of "guitar" by E.P. and F.M. matches that of a violin, the description given of a ⌜*violin*⌝ is "made of metal" by J.L., and that of ⌜*eagle*⌝ by F.M. is of a nocturnal bird with big eyes that "stands up", presumably an owl. These are not more general items – arguably, an owl, for example, is not a bird that has been generalised (see, McRae & Cree, 2002, for what constitutes a general bird). The same cannot so strictly be said for the naming responses of a single SD patient, KL, also tested on the Cambridge semantic memory battery, which both Hodges et al. (1992), and Rogers et al. (2004), use for their patients. KL, given a line drawing of a sock, replies "boot", and when shown a squirrel he calls it a "chicken" (see Table 8.10). Are these instances of over-extended features or of something else, e.g., semantic interference? See Table 8.10 for some of his naming responses.

It has just been shown that the "typical" element of this assumption might not always hold, i.e., that properties might be over-extended but that they are not necessarily typical properties. The "related" part of this statement also is called into question in light of SD patient data. That is to say that crossdomain errors are quite possible, contra to what Rogers et al. (2004) claim

[3]"The vast majority of omission errors were cases in which the patient was unable to provide any name for the objects (although sometimes they would attempt to describe it). A very small number of visual errors were also grouped into this category." (Rogers et al., 2004, p. 217)

[4]For the patients who took part in the naming task the details given are: "Each patient was tested at least once, and in most cases patients were tested several times over a span of several years. On average, each patient participated in 3.8 testing sessions; the greatest number of testing sessions with a single patient was 10. The total number of sessions across patients was 57." (Rogers et al., 2004, p. 217)

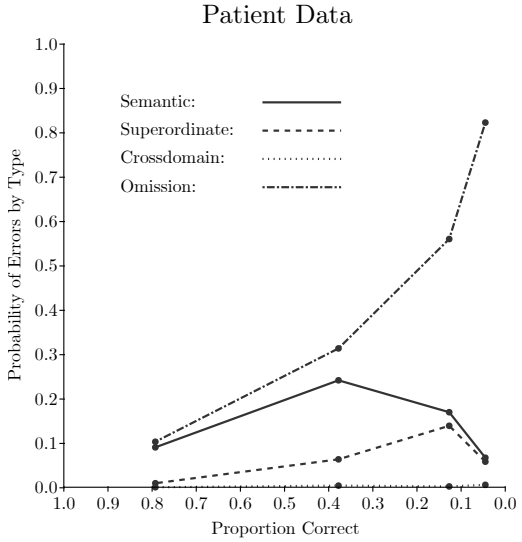


Figure 8.83: Scores of the Rogers et al. (2004) SD patients in the confrontation naming task.

in their patient and model naming graphs (see Figure 8.83). For example, patient JL, contrary to what would be expected, does make cross-domain errors: he calls pictures of a pumpkin, a fish, a deer, and a peacock "vehicle"; replies with "animal" (or part of one) when shown a telephone, a banana, a skirt, a piano, a helmet, a roller-skate and other toys; he also labels different pieces of clothing as "a man" and other related body parts, e.g., glove becomes "hand" (perhaps for obvious reasons); asparagus is "a machine" and subsequently he replies with "cut holes" to the same picture; a potato is "building" and "for the wet"; and so on (Hodges et al., 1995, appendix). It is important to note that most but not all of these errors occur at the later of the four reported stages of testing. Also important is the fact that JL seems to be a very typical SD patient, scoring on the other tasks very similarly to the Rogers et al. (2004) patients – meaning there is no evidence to suggest that he is at a different level of SD than the other patients.

In conclusion, this rule holds for a specified range on the spectrum of SD neurodegeneration, somewhere near the middle. Outside that range it is not the case that "[t]ypical properties are over-extended to inappropriate but related objects" (Rogers et al., 1999, p. 1). At each stage of conceptual loss mentioned before, it is superficially true that the general knowledge remains intact until the patient is anomic. Although, that does not mean that everything is generalised at all stages in the progression of SD.

#### 8.2.3.3 Is performance better with pictures than with words?

It is important to bear in mind that the diagnostic criteria for SD state that the patient must display multi-modal semantic deficits, that affect both verbal and non-verbal faculties of

Naming Line Drawings Test

| | **Round 1** | **Round 2** | **Round 3** | **Round 4** |
|---|---|---|---|---|
| **Mammals** | | | | |
| Pig | *Pig* | On farms | Dog | Dog |
| Elephant | *Elephant* | Horse | Horse | Animal |
| Squirrel | Cat | Chicken | Cat | Dog |
| **Birds** | | | | |
| Chicken | *Chicken* | *Chicken* | Bird | Animal |
| Ostrich | Swan | Bird | Cat | Animal |
| **Insects** | | | | |
| Ant | Bird | Bird | Cat | Animal |
| Bee | Bird | Animal | Cat | Don't know |
| **Water Creatures** | | | | |
| Alligator | Small dog | Fish | Cat | Animal |
| Lobster | For eating | Don't know | Don't know | Don't know |
| **Fruit** | | | | |
| Orange | Apple | Apple | Don't know | For food |
| Pineapple | Food | Food | Growing | Don't know |
| **Body parts** | | | | |
| Hand | *Hand* | *Hand* | *Hand* | *Hand* |
| Lips | *Lips* | *Lips* | To eat | A hole |
| **Household** | | | | |
| Chair | *Chair* | *Chair* | Table to sit on | To sit on |
| Cooker | Radio | Radio | Box | Don't know |
| Envelope | *Envelope* | Letter | Book | Don't know |
| **Musical Instruments** | | | | |
| Violin | Music | Music | Music | Don't know |
| Trumpet | Music | Blow it | Music | Don't know |
| **Clothing** | | | | |
| Sock | Boot | Boot | Shoe | Foot |
| Waistcoat | Shirt | Jacket | Jacket | Coat |
| **Tools** | | | | |
| Scissors | *Scissors* | *Scissors* | To cut things | A machine |
| Screwdriver | Unscrew things | Knife | Knife | Something |

Table 8.10: "Four successive test rounds over about 18 months in a single patient [KL] with SD."(Hodges & Patterson, 2007, table 2) Italics indicate a correct response.

Direct and Delayed Copying

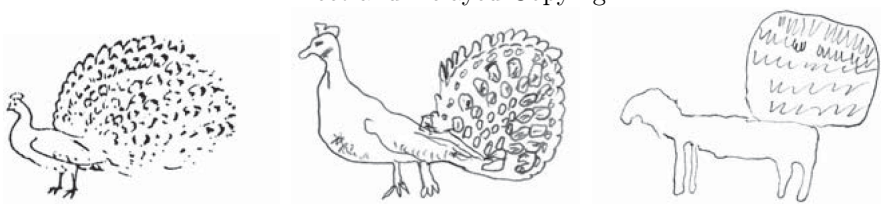

Figure 8.84: The rightmost image is a line-drawing of a peacock shown to SD patient to first copy, and then reproduce without looking. The second image shows the direct copy the patient drew whilst looking at the target. The third drawing was created approximately 15 seconds after the target was removed (prior to being studied for 5 seconds). The delayed copy clearly shows an example of a stimulus that has been modified beyond recognition, resembling a mammal more than a bird. While the direct copy shows the patient is definitely able to draw a peacock under the right conditions; meaning that drawing skills are not compromised in and of themselves as the peacock can be draw given a stimulus to copy from, but the semantic content required to produce them independently is not available any more to SD patients (drawings reproduced from Hodges & Patterson, 2007, fig. 2, E). This can be seen as the visual/drawing modality analogue of surface dyslexia. "Other drawings where patients with SD must wait briefly before reproducing the target reveal camels lacking their humps, rhinos lacking their horns, and seals with legs rather than flippers." (Hodges & Patterson, 2007, p. 1009)

memory (Hodges & Patterson, 2007; Neary et al., 1998). Visual stimuli do seem to be more easily identified than words by the SD patients tested in the picture and word sorting task in Rogers et al. (2004), but no statistical test is carried out to determine whether or not this effect reaches significance. While Hodges et al. (1992) do not carry out both a picture sorting and a word sorting task like Rogers et al. (2004) do, their patients do perform a picture sorting task at three levels and various word-based tasks, and picture pointing to spoken word (also known as word-to-picture matching). Their results indicate that indeed picture-based tasks are easier for the SD patients. It is important to bear in mind here, as will be discussed later on, that this may be the effect of a confounding variable, as it is not made clear if care was taken to match the stimuli for complexity.

The Hodges et al. (1992) patients score significantly better than the Rogers et al. (2004) patients: $t(13) = 2.3413$, $p = 0.0358$; at domain-level sorting (greater in the picture-sorting task than the word-sorting task, presumably because the former patients are all at ceiling for this task). At category-level sorting, they perform the same: $t(13) = 0.3777$, $p = 0.7117$. The first t test indicates that perhaps the Rogers et al. (2004) patients are further along the degeneration spectrum of SD then those of Hodges et al. (1992), which ties in with their naming scores being overwhelmingly dominated by omissions even in the first data points.

Category Fluency Test

| | Apr 1991 | Sep 1991 | Mar 1992 | Sep 1992 | Controls Mean (SD) |
|---|---|---|---|---|---|
| **Animals** | 10 | 6 | 3 | 3 | 17.5 (4.0) |
| **Birds** | 6 | 4 | 0 | 0 | 15.9 (4.6) |
| **Sea Creatures** | 4 | 3 | 1 | 0 | 14.0 (4.5) |
| **Dogs** | 1 | 0 | 0 | 0 | 10.9 (3.5) |
| **Household Items** | 9 | 7 | 1 | 0 | 17.5 (3.4) |
| **Vehicles** | 5 | 3 | 2 | 0 | 11.8 (2.6) |
| **Musical Instruments** | 0 | 0 | 0 | 0 | 15.3 (3.4) |
| **Boats** | 2 | 2 | 0 | 0 | 10.9 (3.4) |

Table 8.11: In the table above JL's performance in a category fluency task is compared with 25 controls of similar ages and a range of educational levels (Hodges et al., 1995, table 2). "Considering performance as a whole, JL was well outside the normal range even on initial testing, and his ability to generate instances of these categories declined dramatically over the subsequent 1.5 years. The most dramatic difference between JL and controls, however, was on the more specific categories; for the two broadest categories (animals and household items), on initial testing JL was actually just within the normal range (taken as 2 SDs from the mean). This, then, is the first indication in JL's data of a performance discrepancy between broader and more detailed information. Note that JL's greater success when asked to generate instances of the broader categories cannot be explained, or at least not entirely, on the basis that this is an easier task. Normal subjects produced more animals than types of dog, too, but only about half-again as many (roughly 18 vs 11) whereas for JL's initial session this ratio was 10:1. In another example, normal subjects produced hardly any more household items than musical instruments, whereas for JL this contrast was a dramatic 9:0. [...] Naming on the semantic battery was impaired to a very severe degree even on initial testing (JL 17/48 = 0.35, controls 0.91 f0.05) and then deteriorated until, in September 1992, he was able to name only 5 of the 48 line drawings. Naming was in fact administered a fifth time, in March 1993, when his score was 2/48. An analysis of the pattern of responses over five test sessions showed a striking item-by-item consistency (Table 4); there were only three instances (out of a total possible of 157) where a naming error on a particular item was followed by a correct response; in all other instances a naming error was followed by an error on all subsequent occasions." (Hodges et al., 1995, p. 471).

Animal Fluency Test

| Apr 1991 | Sep 1991 | Mar 1992 | Sep 1992 | Mar 1993 |
|---|---|---|---|---|
| cow | horse | horse | dogs | |
| bullock | cows | cows | cats | |
| sheep | bullock | birds | horses | |
| lamb | duck | | | |
| pig | cats | | | |
| dog | dog | | | |
| horse | | | | |
| cat | | | | |
| birds | | | | |
| geese | | | | |

Table 8.12: JL's performance on the animal fluency test. He is able to retain "bullock" and "cows" from April to September 1991, and "bird" is preserved in March 1992 (even though it was not there in September 1991). By the penultimate round of testing though, he only could list three, stereotypical perhaps, animals. At the final testing session he was completely unable to list any animals (reproduced from Hodges et al., 1995, table 3)

| | Apr 1991 | Sep 1991 | Mar 1992 | Sep 1992 | Mar 1993 |
|---|---|---|---|---|---|
| lorry | ☑ | ☑ | ☑ | ☑ | ☑ |
| bike | ☑ | ☑ | ☑ | ☑ | ☑ |
| telephone | ☑ | ☑ | ☑ | ☑ | ☐ |
| aeroplane | ☑ | ☑ | ☑ | ☑ | ☐ |
| motorcycle | ☑ | ☐ | ☑ | ☑ | ☐ |
| bus | ☑ | ☑ | ☑ | ☐ | ☐ |
| fish | ☑ | ☑ | ☑ | ☐ | ☐ |
| monkey | ☑ | ☑ | ☐ | ☐ | ☐ |
| duck | ☑ | ☑ | ☐ | ☐ | ☐ |
| cooker | ☑ | ☐ | ☐ | ☐ | ☐ |
| toaster | ☑ | ☐ | ☐ | ☐ | ☐ |
| helicopter | ☑ | ☐ | ☐ | ☐ | ☐ |
| deer | ☑ | ☐ | ☐ | ☐ | ☐ |
| rabbit | ☑ | ☐ | ☐ | ☐ | ☐ |
| mouse | ☑ | ☐ | ☐ | ☐ | ☐ |
| tiger | ☑ | ☐ | ☐ | ☐ | ☐ |
| chicken | ☑ | ☐ | ☐ | ☐ | ☐ |
| kettle | ☐ | ☑ | ☐ | ☐ | ☐ |
| Total Correct | 17 | 9 | 7 | 5 | 2 |

Table 8.13: JL's naming consistency for the 18 line drawings, from a total of 48. He never named any of the 30 remaining items correctly in any of the five testing sessions (reproduced from Hodges et al., 1995, table 4).

#### 8.2.3.4 Are different patterns of errors observed for living and non-living things?

There is no indication that the SD patients described in Hodges et al. (1992) have a dissociable difference between their inability to name different kinds of boats (inanimate domain) and their inability to name various breeds of dog (animate domain), nor is there any significant statistical evidence of any category-specificity in the naming task. These and many other points regarding the category-specific dissociation in SD (or indeed lack thereof) are tackled in subsection 8.2.4.

Patient JL (see tables 8.11, 8.11, and 8.13), also shows no such pattern when completing a category fluency task — JL does not produce significantly more items between domains: $t(15) = 1.4135,\ p = 0.1728$. Perhaps, given that certain scores are apparently arithmetically very different within the two domains in JL's responses, e.g. compare musical instruments to household items, it would be more appropriate to consider within-domain differences too. Although again, the difference might be down to familiarity and frequency as opposed to category- or domain-specificity as there is no indication these factors are controlled for.

"Naming was impaired to a very severe degree in all five patients. This is a relatively easy naming test as evidenced by the near ceiling performance of the normal controls. There was no clear category-specificity, in that all patients were severely impaired at naming line-drawings of

both living and man-made items." (Hodges et al., 1992, p. 1795) Thus contradicting rule 4.

The naming behaviour of the Hodges et al. (1992) patients is inline with Rogers et al. (2004) — there is no indication that patients make crossdomain errors. "Apart from omissions, naming errors were semantic in type, being either within-category (e.g. hippopotamus for rhinoceros; violin for accordion), superordinate (e.g. animal, musical instrument) or circumlocations [sic]." (Hodges et al., 1992, p. 1793).

Due to the nature of this claim and how central Rogers et al. (2004) posit that category-specific deficits are in SD, the next section, deals with it in much more detail.

### 8.2.4 Category-specific deficits in semantic dementia

The issue of category-specific deficits in SD appears to be a relatively long-contested one in the general literature. In addition, category specific deficits, as mentioned above, are one of the assumptions that the original Rogers et al. (2004) model makes. We believe that this debate has been largely resolved, given the following evidence. Furthermore, we posit that such category-specific deficits do not form part of a canonical definition of SD, nor do they appear in any FTLD patients, other than in exceptional cases. True category-specific effects (ones that are not a product of familiarity, frequency, stimulus complexity, and other confounds) are nonetheless highly valuable and worthy of study. However, we argue, these effects do not have to necessarily be included in SD theories and models because they are outlier cases; although of course an ideal theory should explain outlier cases in some way.

Rule 4 states that "different patterns of errors are observed for living and non-living things" (Rogers et al., 1999, p. 1). This means that damage to the semantic system in a way compatible with SD is expected to produce a category-specific effect[5], i.e., the relative preservation of manmade over biological kinds. In Rogers et al. (2004), this is what is examined and found to be the case both for the original hub model and for the patient data presented therein. However, the authors themselves dispute the existence of these effects in most, if not almost all, SD patients in their own published work (both before and after 2004):

> There is only one convincing report [other than the one presented herein] of a category-specific deficit in a case of semantic dementia (patient MF: Barbarotto et al., 1995).
>
> (Lambon Ralph et al., 1998, p. 315)

[5]The phrase category-specific in the semantic memory literature usually (and perhaps slightly confusingly) implies a domain-specific difference between behaviours. However, that does not mean that other category-specific differences, e.g., between number and non-number words in SD patients (Jefferies, Patterson, Jones, Bateman, & Lambon Ralph, 2004), do not exist.

Semantic dementia [...] accounts for only a fraction of the cases with category-specific impairment that have been described in the literature to date, and among the 40 or so published cases of SD, only six have been reported as showing a category-specific pattern (Barbarotto, Capitani, Spinnler, & Trivelli, 1995; Basso et al., 1988; Breedin, et al., 1994; Cardebat, Demonet, Celsis, & Puel, 1996; Lambon Ralph et al., 1999; McCarthy & Warrington, 1988)[.] The absence of a category-specific difference is also apparent from studies that have focused on concept definition (Lambon Ralph et al., 1999) and non-verbal forms of assessment (Bozeat et al., 2000).

(Garrard, Lambon Ralph, & Hodges, 2002, p.155)

[C]ases of semantic dementia – a progressive syndrome in which semantic knowledge undergoes a profound degradation – typically do not show the consistent preservation of one semantic domain relative to another (Hodges et al., 1995; Lambon Ralph, Graham, Patterson, & Hodges, 1999).

(Rogers & Plaut, 2002, p.14)

[P]atients with semantic dementia [...] tend not to have a category-specific semantic impairment. The one case with a clear category effect is different (MF: Barbarotto et al., 1995). MF's temporal lobe atrophy involved medial structures including the hippocampus and parahippocampal gyrus to a much greater extent than that normally seen in semantic dementia.

(Lambon Ralph et al., 2003, p. 311)

[T]he combination of semantic dementia and category specificity is something of a rarity.

(Lambon Ralph et al., 2003, p. 319)

[A]s is typically the case (Bozeat et al., 2000; Lambon-Ralph et al., 2003; Moss et al., 2005), the SD patients studied here exhibited a non-category-specific semantic impairment affecting both living things and artifacts.

(Noppeney et al., 2007, p. 1139)

[T]he contrast between HSVE [herpes simplex virus encephalitis] and SD [is] in terms of both the severity of the semantic deficit (which is often either absent or mild in

> patients with HSVE, as opposed to progressive and ultimately profound in SD) and its pattern (which is frequently category-specific in HSVE, but very rarely so in SD)[. B]oth diseases implicate the bilateral ATL in semantic processing, [therefore] it must be the specific nature and/or distribution of the brain abnormalities in SD that produces the pervasive disruption – across all categories and all modalities – of conceptual knowledge that defines this condition.
>
> (Patterson, Nestor, & Rogers, 2007, p. 980)

> Despite their substantial semantic impairment, it is rare for SD patients to present with a category-specific pattern (for example, combining across studies, performance on different categories has been investigated, with all appropriate controls, in 30 SD patients and only one showed a living < manmade pattern: Lambon Ralph et al., 2003), so the group provides an important neuropsychological and neurological baseline against which to compare patients with category-specific semantic impairment.
>
> (Lambon Ralph et al., 2007, p. 1128)

From the above extracts it can been seen that in fact there is agreement amongst the Rogers et al. (2004) authors (including other researchers) that SD typically does not produce category-specific deficits. In fact, the final quote from Lambon Ralph et al. (2007) refers to the use of SD patients as a group of participants that show a global semantic deficit, in order to contrast their behaviour to those of category-specific patients. This in and of itself is indicative of the dependable nature of SD as a non-category-specific semantic disorder, as these two groups of patients are often compared and contrasted in the same exact tasks in the literature.

Some of the authors undertake their own investigation into SD and category-specificity (in addition to that carried out in Rogers et al., 2004) by testing patients on category fluency with a breakdown of behaviour by domain (although category-specificity was not their primary concern; Bozeat, Ralph, Patterson, Garrard, & Hodges, 2000). Given the statements in Rogers et al. (2004) about the breakdown of the two domains (based on different distributions of features: perceptual/functional), one would expect a different response based on domain. However, as can be seen in Table 8.14, the difference between the domains is not significant; neither within controls $t(17) = 0.5636, n.s.$ nor within SD patients $t(9) = 0.2489, n.s.$

The same sort of investigation was carried out in Lambon Ralph et al. (2003, by a subset of the hub-and-spoke theory authors); they found no significant differences except in certain specific patients and cases. Some of their results can be seen in Table 8.15. The patients

| Test (maximum score) | JP | WM | SL | JC | DS | AT | DC | JW | JH | IF | Control mean (SD) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Category fluency** | | | | | | | | | | | |
| living | 42 | 38 | 23 | 15 | 3 | 14 | 6 | 3 | 5 | 2 | 60.3 (12.6) |
| manmade | 37 | 29 | 20 | 21 | 10 | 18 | 4 | 4 | 7 | 5 | 54.8 (10.3) |
| **Naming** (64) | 59 | 57 | 45 | 43 | 17 | 17 | 11 | 9 | 6 | 1 | 62.3 (1.6) |
| **Word-picture matching** (64) | 64 | 63 | 60 | 58 | 58 | 57 | 36 | 23 | 18 | 18 | 63.7 (0.5) |
| **Word synonyms** | | | | | | | | | | | |
| concrete (25) | 13[a] | 21 | 15[a] | 12[a] | 12[a] | 16 | 14[a] | NT | 12[a] | 13[a] | 23.7 (1.3) |
| abstract (25) | 14[a] | 18 | 15[a] | 8[a] | 14[a] | 14[a] | 13[a] | NT | 13[a] | 13[a] | 23.0 (2.1) |
| **Pyramids and Palm Trees** | | | | | | | | | | | |
| words (52) | 48 | 48 | 46 | 44 | 46 | 45 | 25[a] | 32 | 25[a] | 28[a] | 51.1 (1.1) |
| pictures (52) | 49 | 52 | 48 | 41 | 46 | 47 | 36 | 27[a] | 37 | 22[a] | 51.2 (1.4) |

Table 8.14: Assessment of semantic memory.
NT: Not tested.
[a]Score not significantly better than expected by chance. (Bozeat et al., 2000, table 2)

show no significant difference in their production of exemplars per domain. In the naming task patients CS and KH show a significant difference between living and inanimate objects, and KH also maintains this disparity between the domains in word-to-picture matching. However, KH is the only patient that maintains the category-specific difference when psycholinguistic factors such as familiarity and frequency and other potential confounds are controlled for (using stimuli developed for this purpose, Lambon Ralph et al., 1998). Patient MA however, in the controlled naming task, displays a profound inverse category-specific effect. Notwithstanding a relative preservation of biological kinds over manmade objects.

As mentioned above, the Lambon Ralph et al. (2003) patients as a group do not show category-specific effects. Lambon Ralph et al. (2003) test their SD patients on the full Snodgrass and Vanderwart (1980) picture set (which is large and thus can allow for checking for a possible domain effect however small while controlling for confounds). They found that the factors significantly affecting word-to-picture matching are the objective age of acquisition of the concept (this and the following stimulus properties are taken from Morrison, Chappell, & Ellis, 1997), and if and only if KH is included in the analysis, the domain. For naming the significant factors are stimulus familiarity, objective age of acquisition, imageability, and phoneme length (the longer the word the better the performance; Lambon Ralph et al., 2003, see tables 3a and 3b). This further casts doubt on the interpretation of the Rogers et al. (2004) patients' semantic task scores, especially, since no statistical test is carried out to underpin the interpretation they provide.

Even within Rogers et al. (2004), it is explained in the introduction that SD "provides the clearest evidence of a relatively pure semantic impairment that affects all modalities of testing

Semantic and Naming Assessments

| | | | Patient | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Test | Subtest | Max | AN | CS | MA | AT | SL | KH | Mean |
| PPT | Pictures | 52 | NT | 41 | 41 | 47 | 44 | 42 | |
| | Words | 52 | NT | 39 | 42 | 45 | 38 | 40 | |
| Fluency | Letters (FAS) | N/A | 40 | 14 | 9 | 20 | 30 | 13 | |
| | Man-made[a] | N/A | 34 | 25 | 8 | 18 | 7 | 14 | 17.7 |
| | Living [a] | N/A | 47 | 13 | 7 | 14 | 12 | 8 | 16.8 |
| | | | | | | | | | $t = 0.23$ |
| | | | | | | | | | $p = 0.83$ |
| 64 naming | Man-made | 32 | 32 | 28 | 5 | 12 | 7 | 26 | 18.3 |
| | Living | 32 | 32 | 19 | 8 | 5 | 11 | 16 | 15.2 |
| | $\chi^2$ | | | 5.13 | 0.39 | 2.88 | 0.70 | 5.61 | $t = 1.24$ |
| | $p$ | | | **0.02** | 0.53 | 0.09 | 0.40 | **0.02** | $p = 0.27$ |
| 64 word-picture | Man-made | 32 | 32 | 29 | 30 | 30 | 24 | 31 | 29.3 |
| matching | Living | 32 | 32 | 22 | 27 | 27 | 24 | 20 | 25.3 |
| | $\chi^2$ | | | 3.48 | 0.64 | 0.64 | 0 | 9.65 | $t = 2.28$ |
| | $p$ | | | 0.06 | 0.42 | 0.42 | 1 | **0.002** | $p = 0.07$ |
| Controlled set | Man-made | 30 | 28 | 15 | 7 | 11 | 10 | 24 | 15.8 |
| naming | Living | 30 | 30 | 14 | 17 | 6 | 13 | 17 | 16.2 |
| | $\chi^2$ | | 0.52 | 0.07 | 6.94 | 2.05 | 0.64 | 3.77 | $t = 0.13$ |
| | $p$ | | n.s. | n.s. | **0.008** | 0.15 | n.s. | **0.05** | n.s. |

Table 8.15: Assessment of semantic memory. (Lambon Ralph et al., 2003, table 2) PPT: Pyramids and palm trees test (Howard & Patterson, 1992). NT: Not tested. Boldface text: significant.
[a]4 categories.

Naming to Description and Description-to-Picture Matching

| | | Patient | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | AN | CS | MA | AT | SL | KH | Mean |
| **Naming to Description** | | | | | | | | |
| Definition type | | | | | | | | |
| Functional | /64 | 48 | 21 | 37 | 18 | 24 | 37 | 30.8 |
| Sensory | /64 | 42 | 15 | 27 | 8 | 13 | 16 | 20.2 |
| $\chi^2$ | | 1.35 | 1.39 | 3.13 | 4.83 | 3.87 | 14.2 | $t = 4.75$ |
| $p$ | | n.s. | n.s. | 0.08 | 0.03 | 0.05 | <0.001 | 0.005 |
| Domain | | | | | | | | |
| Nonliving | /64 | 41 | 21 | 35 | 12 | 16 | 33 | 26.3 |
| Living | /64 | 49 | 15 | 29 | 14 | 21 | 20 | 24.6 |
| $\chi^2$ | | 2.4 | 1.4 | 1.13 | 0.19 | 0.95 | 5.4 | $t = 0.5$ |
| $p$ | | n.s. | n.s. | n.s. | n.s. | n.s. | 0.02 | n.s. |
| **Description-to-Picture Matching** | | | | | | | | |
| Definition type | | | | | | | | |
| Functional | /64 | 62 | 48 | 47 | 55 | 24 | 49 | 47.5 |
| Sensory | /64 | 61 | 41 | 38 | 53 | 13 | 43 | 41.5 |
| $\chi^2$ | | 0 | 1.81 | 2.84 | 0.24 | 4.6 | 1.39 | $t = 3.77$ |
| $p$ | | n.s. | n.s. | n.s. | n.s. | .03 | n.s. | .01 |
| Domain | | | | | | | | |
| Nonliving | /64 | 63 | 46 | 42 | 56 | 16 | 53 | 46 |
| Living | /64 | 60 | 43 | 43 | 52 | 21 | 39 | 43 |
| $\chi^2$ | | 0.83 | 0.33 | 0.04 | 0.95 | 0.95 | 7.58 | $t = 1.16$ |
| $p$ | | n.s. | n.s. | n.s. | n.s. | n.s. | .006 | n.s. |

Table 8.16: (Lambon Ralph et al., 2003, tables 4a and 4b)

and all conceptual domains[, further clarifying that] the observed deficits are typically neither category nor modality specific" (p. 206). Although it is important to contrast the previous statement regarding modality-specific differences with this one: "IW resembles the pattern shown by the majority of patients with semantic dementia, [...] her semantic impairment seems to affect visual knowledge much more than associative–functional information (Basso et al., 1988; Breedin et al., 1994b; Moss et al., 1995; Cardebat et al., 1996; Srinivas et al., 1997; Tyler and Moss, 1998; Lambon Ralph et al., unpublished data)." (Lambon Ralph et al., 1998, p. 327) But then Hodges, Bozeat, Ralph, Patterson, and Spatt (2000) state that: "Our finding that semantic impairment was associated with poor use of common objects, while in line with the previous studies of Hamanaka and colleagues, and Hodges and colleagues (Hamanaka et al., 1996; Hodges et al., 1999), seems to contradict several reports[.] Why were the patients assessed here, but not those reported by others (Buxbaum et al., 1997; Lauro-Grotto et al., 1997) markedly impaired in using common objects? Furthermore, how can a number of the patients included in the present study still engage successfully in hobbies and sports requiring object use?" (p. 1921) Indicating that the authors are actively aware of modality-specific dissociations in SD being anything but clear-cut. Category-, but not modality-, specificity is mentioned again in the discussion section: "Patients with semantic dementia [...] typically do not show preservation of knowledge for one domain relative to another (Lambon Ralph et al., 2001)" (Rogers et al., 2004, p. 230).

So while it is clear that a dissociation between the two domains in patients is rare (so exceptional, in fact, as to define SD as a semantic impairment that indiscriminately affects all of semantic knowledge), it is not entirely clear what the original hub model is capturing when the task results are shown broken down by domain. Specifically, they claim "domains with a high degree of similarity structure offer more opportunities [...] to make errors of commission[, while in u]nstructured domains [...] we would expect to see a greater proportion of omission errors" (Rogers et al., 2004, p.219). This statement is followed by the behaviour of the original hub model and of the patients in the confrontation naming, the word and picture sorting, and the drawing and delayed copying tasks, in which a distinct category-specific effect can be seen across the two domains, and further dissociating fruit from the living/non-living classification. Although, the effect of category on the behaviour is not tested statistically, with the only exception being that a significant main effect of domain is found in the patients' drawing and delayed copying task scores: "$F(1, 415) = 56.0,\ p < .001$, with patients making a higher proportion of intrusion errors for animals than for artefacts." (Rogers et al., 2004, p. 227)

The patients in Lambon Ralph et al. (2003) showed no category-specificity (except KH, as

mentioned previously) but did show significant modality-specific differences, showing a preservation of functional over sensory features in their definitions in both naming to description and description-to-picture matching tasks, see Table 8.16. Both these tasks involve the 64 items used previously in the naming and word-to-picture tasks of both Rogers et al. (2004) and Lambon Ralph et al. (2003). For each of these, two definitions were written — one about the functional features and the other emphasising the sensory features of the item in question. This dissociation (between the preservation of sensory features in comparison to functional without a coupled preservation of animals in comparison to non-living things) further casts doubt over the proposed substrate of category-specific deficits, because the patients do show a modality-specific dissociation but do not have category-specific difficulties (Lambon Ralph et al., 2003). This means that the proposal in Rogers et al. (2004) that a category-specific dissociation in SD might be caused by a disparity in the preservation of sensory versus functional features cannot hold, at least not for the patients presented in Lambon Ralph et al. (2003).

Additionally, the Lambon Ralph et al. (2003) patients show a reverse ratio of sensory to functional features to the healthy participants, with the former's sensory to functional ratio at 35% : 41% and the latter's at 57% : 22%. This is a significant crossover interaction: 2 (subject group) $\times$ 2 (feature type: sensory vs. functional) ANOVA $F(1, 14) = 90.8,\ p < 0.001$; with post-hoc t-tests showing that the type of feature was significant for healthy controls and approaching significance for patients (in the opposite direction), SD: $t(5) = 2.2,\ p = 0.08$; controls: $t(9) = 12.1,\ p < 0.001$, full details given in Table 8.9 (Lambon Ralph et al., 2003). Further casting doubt on Rogers et al. (2004) theory, as they claim that healthy semantic systems are biased towards functional features and the functional modality.

This raises the question of whether Rogers et al. (2004) are investigating the effect that different feature-based structures have on the robustness of representations (something which is arguably more easily controlled for in the model that in the patients) or category-specificity in SD (which they previously explain does not exist except in some very rare exceptional cases). This question might not be answerable by appealing to their own writings given the contradicting statements outlined previously. Moreover the authors themselves claim "there is rather little evidence to suggest that this poor visual knowledge in SD leads to category-specific effects." (Lambon Ralph et al., 2003, p. 310) But perhaps more importantly, are these actually just two sides of the same coin? Surely, if responses indeed differ as a function of the domain being tested, then by definition the pattern is a category-specific one that should arise in SD, regardless of whether or not the reason behind the dissociation is due to the way features are distributed or the way the semantic system is organised or both or neither (provided there are

no confounds).

A third option is also available as an interpretation of what they are modelling: it could be that the original hub model is investigating the distribution of errors per domain. These errors in number are identical, but at further analysis one domain has more omissions, the other has more commissions[6], and so on. No other authors have looked into this detailed distribution so closely as to uncover domain-specific differences. It is unlikely, however, to be the case that this is indeed what they are proposing, as, for example, in the patients' scores for the drawing and delayed copying task the domain of animals consistently has more errors overall (Rogers et al., 2004, see fig. 1.2; although, the model shows the inverse pattern of omissions to the patients).

In the discussion in Rogers et al. (2004), this question (of what it is they are modelling when they present category-specific patterns of dissociation) is somewhat tangentially addressed. They stipulate that semantic domains: differ in neighbourhood density (i.e., the number of proximal representations); in the regularity of features (i.e., some features can be shared over most exemplars of a domain, making items without them, but within the domain in question become regularised, thus incorrectly obtaining the shared features); in the breadth of features (i.e., a feature shared over all living things is more robust to damage than one shared over marsupials only); as well as differing by familiarity and frequency – these latter two properties are not modelled in the original hub, nor is visual/perceptual complexity addressed in any obvious way. They also explain that:

> [A]nimals tend to share a greater number of properties with their semantic neighbors than do artifacts. In [the original hub-and-spoke] model, artifact representations are more sparsely distributed across a broader region of the space. These factors lead to different patterns of errors in animal and artifact domains[; as such], they must be added to the long list of potential confounding factors in experiments that purport to reveal true category-specific deficits.
>
> Rogers et al. (2004, p. 231)

Furthermore, it is perhaps interesting to note Rogers et al. (2004) do not consider category-specific effects that arise from the structure of domains as bona fide. This is not something we are comfortable with, especially since the structure is one derived by the semantic cognitive system and therefore part and parcel of the system itself. In addition, we believe this is not a useful distinction given that the neuropsychological data from patients does not make this

[6]Rogers et al. (2004) "observed [...] a greater proportion of omission errors for naming of artefacts and a greater proportion of commission errors for naming of animals" (p. 229)

distinction (although, psycholinguistic factors such as stimulus complexity, familiarity, and frequency are often controlled for), and given that both true- and pseudo-category-specific effects (however they may be defined) are in need of an explanation whether we accept the distinction or not.[7]

It is also of importance to recall that one of the Rogers et al. (2004) authors, Matthew Lambon Ralph, also describes an SD patient with the inverse to the category-specific deficit they model in Rogers et al. (2004) — thus doubly dissociating the biological/manmade domains within SD-compromised cognition (Lambon Ralph et al., 1998). However, not only does patient IW show better performance on biological kinds, but she also has a deficit in visual perceptual features, which completely dissociates between features and domains/categories (Lambon Ralph et al., 1998). This finding, appears to contradict the Rogers et al. (2004) assumption that the reason category specific deficits arise (ersatz or otherwise) is due to the feature structure of certain domains, since the domain most subserved, according to Rogers et al. (2004), by perceptual features is the animate, and thus that domain should be most affected. In most cases indeed that is the case but IW (along with six other studies) showed a clear dissociation between her knowledge of perceptual properties and biological kinds (Caramazza & Shelton, 1998; Funnell & De Mornay Davies, 1996; Laiacona, Barbarotto, & Capitani, 1993; Lambon Ralph et al., 2003, 1998; Moss, Tyler, Durrant-Peatfield, & Bunn, 1998; Samson, Pillon, & De Wilde, 1998; Sheridan & Humphreys, 1993). These cases serve to show that the original hub model's account for category-specific deficits in Rogers et al. (2004), and modified for HSVE in Lambon Ralph et al. (2007), cannot be the full story.

The questions raised in this section about whether or not category-specific deficits arise in SD, and if they do why, will be addressed to some extent in the next section, but it remains unclear why if category-specific deficits are so rare in SD, they are modelled as part of the syndrome in Rogers et al. (2004). To further add to the confusion, the evidence against category-specific deficits occurring in SD that has been presented so far comes from the same authors as Rogers et al. (2004) and has been published before 2004.

### 8.2.5 A more general consensus

Category-specific deficits are so rare that even within the classically-associated aetiology of such deficits, herpes simplex virus encephalitis, these effects are very rare (Lambon Ralph et al., 1998). So even though most category-specific patients have suffered from HSVE (Lambon Ralph

[7] Rogers et al. (2004) do, however, appear to recant their true vs pseudo distinction within category-specificity later on in their discussion section.

et al., 2003; Moss, Rodd, Stamatakis, Bright, & Tyler, 2005; Mummery, Patterson, Hodges, & Wise, 1996), most HSVE patients do not show the pattern of dissociation of interest, in fact most do not even show a naming effect regardless of dissociations based on category (Kapur et al., 1994). Other common aetiologies amongst category-specific patients are: head injury (Farah & McClelland, 1991; Farah et al., 1989; Laiacona et al., 1993; B. A. Wilson, 1997); stroke, also known as cerebrovascular accident (Farah & Wallace, 1992; Forde, D. Francis, Rumiati, & Humphreys, 1997; Hart, Berndt, & Caramazza, 1985; Howard, Best, Bruce, & Gatehouse, 1995; Sartori, Miozzo, & Job, 1993); and Alzheimer's dementia (Silveri et al., 1991; Mauri, Daum, Sartori, Riesch, & Birbaumer, 1994; Montanes, Goldblum, Boller, et al., 1995; Daum, Riesch, Sartori, & Birbaumer, 1996; Tippett, Grossman, & Farah, 1996; Gonnerman, Andersen, Devlin, Kempler, & Seidenberg, 1997; Garrard et al., 1998). However, as with HSVE, the patients that do show category-specific impairments are vastly in the minority compared to their respective original populations (Lambon Ralph et al., 1998).

On the one hand, what this means is that all the category-specific patients on their own indeed do show dissociations within their semantic knowledge hence they are used as single-patient or small-group case studies of category specificity. On the other hand, if these patients are construed as just another set of data points in an analysis of the cognitive symptoms of their respective aetiologies then the category-specific impairments would disappear, as Lambon Ralph et al. (1998) also notes.

SD patients are even less likely than HSVE patients to be seen displaying category-specific patterns of dissociation, so much so that only about a handful have been found in all the years of studying SD (Garrard et al., 2002). This is even more surprising given that HSVE, which is the most common aetiology for category-specific deficits, is not diagnosed based on the potential cognitive repercussions but often merely on positive virology[8], while SD is always diagnosed based on both the cortical locus of damage and semantic degradation (Noppeney et al., 2007). So if they were to both give rise to category-specific effects, it would be more likely to be seen in SD than HSVE(provided base rates are taken into account, as HSVE is much more common than SD: Garrard et al., 2002). To add to the confusion the different syndromes (e.g., Alzheimer's dementia, the other types of FTLD, stroke, head injury, etc.) that affect semantic memory, with apparently the same locus of neurodegeneration, appear to affect it qualitatively differently (Harciarek & Kertesz, 2009).

[8]Although, only using virological data to diagnose HSVE is prone to false positives. To solve this problem neuropsychological and clinical features are taken into account too, especially when the investigation itself is neuropsychological (e.g., Kapur et al., 1994; Utley, Ogden, Gibb, McGrath, & Anderson, 1997, recommend and carry out both types of testing).

The classical, most common, pattern of category-specific deficits is the relative loss of biological kinds, meaning living things, such as animals, plants, fruit, vegetables (Bunn et al., 1998; Caramazza & Shelton, 1998; De Renzi & Lucchelli, 1994; Humphreys & Riddoch, 2003; Laiacona et al., 1997; Moss et al., 2005, 1998; Sartori & Job, 1988; Silveri & Gainotti, 1988; Tyler & Moss, 2001; Warrington & Shallice, 1984; Warrington & McCarthy, 1987). As mentioned, this is most often found in HSVE patients, who are often paired with SD patients in order to show how semantic deficits can be global, affecting all semantic cognition, in SD and specific affecting only biological kinds and leaving man-made objects relatively intact (e.g., Moss et al., 2005; Noppeney et al., 2007). Something that seems to consistently correlate with the loss of animals and plants is the loss of perceptual (usually specifically visual) features, leading many to believe that perceptual features are what underpin the representation of biological kinds (Breedin et al., 1994; Basso, Capitani, & Laiacona, 1988; Cardebat, Demonet, Celsis, & Puel, 1996; Lambon Ralph et al., 1999; McCarthy & Warrington, 1988; Srinivas, Breedin, Coslett, & Saffran, 1997; Parkin, 1993).

It is important to note that even in syndromes classically associated with category-specificity patients have been discovered who do not show the effect after the two domains are matched for frequency, familiarity, and visual complexity (Caramazza & Shelton, 1998; Gaffan & Heywood, 1993; Sheridan & Humphreys, 1993; Stewart et al., 1992). Additionally, it is also of value to note that patients with the inverse category-specific difficulty have been documented, although more rarely, i.e., having more problems with manmade items — doubly dissociating the two domains (Funnell & De Mornay Davies, 1996; Lambon Ralph et al., 1998; Moss et al., 1998; Samson et al., 1998). Moreover, patients have been found that dissociate knowledge of biological kinds and of perceptual properties (Caramazza & Shelton, 1998; Funnell & De Mornay Davies, 1996; Laiacona et al., 1993; Lambon Ralph et al., 2003, 1998; Moss et al., 1998; Samson et al., 1998; Sheridan & Humphreys, 1993); so while often these two abilities are highly correlated, patients have been found that do not adhere to this principle. A memorable example of both a perceptual/animate dissociation and of perhaps failing to control for confounding factors is patient JBR. He was initially tested by Warrington and Shallice (1984) and found to show a dissociation between knowledge of perceptual features and animals. But when subsequently tested by Funnell and De Mornay Davies (1996), his category-specific deficit remained, however the dissociation between physical and functional properties was absent. This is of course not to say that category- and modality-specific dissociations do not really exist in patients, but that in some cases these effects can be explained by familiarity, frequency, and the complexity of the stimuli – in other words, factors such as the design of the semantic tasks themselves play a role

in determining the pattern of scores (McRae et al., 1997; Caramazza & Shelton, 1998).

While the theory behind the Rogers et al. (2004) proposal that different semantic divisions have different internal properties, also known as the sensory functional theory (Caramazza & Shelton, 1998), which can in turn give rise to category-specific effects (supported by many other authors too: Allport, 1985; Gainotti & Silveri, 1996; Hart & Gordon, 1992; Shallice, 1988; Silveri & Gainotti, 1988; Tyler et al., 2000; Warrington & McCarthy, 1983, 1987; Warrington & Shallice, 1984) appears sensible, it is not made clear why category-specific deficits do not arise in SD patients more frequently (some have even gone so far as to call it a paradox; Garrard et al., 2002). To begin to tackle these issues, we accept that category-specific deficits and SD are complicated, not conforming to any theory precisely (Grossman, 2010). This point serves to partially explain why Rogers et al. (2004) and other publications by the same authors are unclear, even contradictory, on the matter. The reasons behind the confusion are many. Notwithstanding, certain consistent properties of SD patient cognition can be gleaned from the literature.

Category-specificity in SD is proposed to be underpinned by the feature distribution between the two domains (by a number of authors: Allport, 1985; Farah & McClelland, 1991; Gainotti & Silveri, 1996; Hart & Gordon, 1992; Rogers et al., 2004; Shallice, 1988; Silveri & Gainotti, 1988; Tyler et al., 2000; Warrington & McCarthy, 1983, 1987; Warrington & Shallice, 1984), but that does not hold up to scrutiny when all patient data is examined. This theory regarding categories being grounded in different distributions of features (e.g., perceptual vs functional) is espoused by other researchers in this area, as well as Rogers et al. (2004) and Tyler et al. (2000, who propose the conceptual structure model), and supported by data from patients and healthy individuals.

The sensory functional theory states that the differences between categories can be boiled down to the differences between features: sensory features contribute more to the semantics of animals and other biological kinds, while functional features are the foundation for inanimate objects. However, some patients have been shown to have dissociations between modality and functional features, and category and domain (Caramazza & Shelton, 1998; Funnell & De Mornay Davies, 1996; Laiacona et al., 1993; Lambon Ralph et al., 1998; Moss et al., 1998; Samson et al., 1998). Indicating that domains of knowledge might not be subserved merely by the feature distribution, or perhaps further still that defining features as either functional or sensory might not be the full story — perhaps top-down high-level features such as ⟨mammal⟩ can be as categorically valuable as ⟨has fur⟩. This point is only tangentially addressed in Rogers et al. (2004) since they do indeed include such top-down features, called encyclopaedic,

encapsulated within the so-called verbal features. Nevertheless, they do not model patients who have compromised perceptual features, but have a difficulty with artefacts instead of animals. To further add to the confusion, SD patients often show preserved colour knowledge, which has been found to contribute more to biological kinds than artefacts, the domain they are held to be worst at (Grossman, 2010; Harciarek & Kertesz, 2009; McRae et al., 1997; Robinson & Cipolotti, 2001; Rogers, Patterson, & Graham, 2007). Colour knowledge, represented seemingly by the visual features, is not explicitly discussed as being preserved in any way in the original hub model (Rogers et al., 2004).

Dissociations occur within the two classical domains as well as over them. This implies that these large domains are not dependent on a particular kind of feature, as differences within them cannot be explained by such a simple mechanism. Either something else is going on entirely or something is missing in this description (Grossman, 2010). Animals are sometimes found to be harder to semantically process, i.e., more intact, than fruits and vegetables (Caramazza & Shelton, 1998; Hart & Gordon, 1992), and sometimes found to be the opposite (Farah & Wallace, 1992; Hart et al., 1985), between as well as within the same patients; while SD patients have been found that are better at biological kinds than artefacts (Biran, Chatterjee, & Glosser, 2002; Lambon Ralph et al., 1998), as well as patients with other aetiologies of brain damage (e.g., Bi, Han, Shu, & Caramazza, 2005). While this point is partially taken into account in Rogers et al. (2004), as they model fruits dissociating from the two domains of animal and inanimate, they do not model the reverse category-specific deficit or use musical instruments in the model, even though their patients are tested on them. This is unfortunate because musical instruments often dissociate from the inanimate domain, e.g., see Table 8.11. The fact that certain categories such as colour knowledge, gemstones, musical instruments, fruit and vegetables, body parts, etc., seem to dissociate from the classical domains of biological kinds and manmade objects appears to undermine the usefulness, if not validity, of the living/inanimate dissociation.

Not all features are created equal (Grossman, 2010). Some appear to be central to the meaning of concepts, e.g., a deer is a mammal, so it must produce milk. While other features, e.g., a deer's geographical distribution, are not as central to the concept ⟨deer⟩, although healthy semantic systems will generally contain both these features in some form or another. The former feature is the most necessary for the preservation of ⟨deer⟩, and Grossman (2010) claims it is not evident if this kind of integral feature is affected by SD.

Participants are inconsistent, showing different category-specific abilities based on the properties of the semantic tasks they carry out (Caramazza & Shelton, 1998; McRae et al., 1997; Tyler et al., 2000). This means that patients are sometimes impaired more in one domain, and

other times in another, depending on the task itself. Rogers et al. (2004) do not mention that these task-based differences exist, modelling all tasks as if the category-based differences are the same in all. This is a point that is rarely discussed in the literature when talking generally about SD and category-specific deficits.

McRae et al. (1997) use semantically healthy participants to compile a list of features for each concept. The categories they use are birds, mammals, fruit, vegetables, clothing, furniture, kitchen items, tools, vehicles, and weapons — three types of biological kinds and six types of artefacts. Participants listed as many perceptual, and functional features, as well as encyclopaedic facts, as possible and then their results were standardised. (A similar or equivalent procedure is described in more detail in McRae & Cree, 2002, which is what the patterns used in the hub-and-spoke are based on.) They then carry out a series of semantic priming experiments in order to investigate amongst other things how different prime-target combinations (e.g., eagle-hawk vs sandals-hawk) will affect the participants' ability to answer four questions: "is it animate?"; "is it an object?"; "does it grow?"; and "is it made by humans?".

McRae et al. (1997) discovered that semantic similarity when it comes to individual features predicts priming for artefacts but not living things. While the reverse is true for similarity evaluated based on correlated features (e.g., ⟨has eyes⟩ and ⟨can see⟩ are very highly correlated features). It predicts the residual variance in priming effects for biological kinds but not inanimate objects. Specifically for living things, uncorrelated features are irrelevant to predicting priming effects, while without the information on violated correlations predicting priming was not possible (instances where two features that usually are found together are found on their own, e.g., ⟨has feathers⟩ and ⟨can fly⟩ usually correlate, but it is semantically meaningful to know when they do not in the cases of, e.g., penguins, insects, or aeroplanes), making violated correlations important to domain structure when it comes to living things.

McRae et al. (1997) also ask participants, using the same prime-target pairs as above, to rate the similarity of the two words. Their results in this task differ to the previous one in an important way: individual features predict both the similarity rating of living things and of artefacts; while correlated features predict neither domain's similarity ratings. This is in contrast to before where semantic similarity between the target and the prime with regards to individual features predicts priming for artefacts only, and similarity with regards to correlated features predicts priming for biological kinds only — seemingly telling a simple story of double dissociation. McRae et al. (1997) also go back to the previous experiment and replace the semantic similarity with one calculated based on the participants' perceived similarity, instead of one based on the features of each concept. This change means that the new similarity

measure predicted priming for living things, but not inanimate objects (in line with when using individual features) — and that, as before, the similarity in terms of correlated feature pairs predicts the residual semantic priming effects for biological kinds, but not artefacts (McRae et al., 1997).

The McRae et al. (1997) results, as well as others (e.g., Caramazza & Shelton, 1998), suggest that both the structure of the task and the analysis of the results can be responsible for dissociations of the category-specific kind. In other words, the categories healthy participants (but presumably also patients) are tested on might differ based on the specific properties of the experimental manipulation itself. So care must be taken when proposing that feature-space affects semantic-space in objective ways, when in reality semantic-space might be dense when accessed using one task and sparse during another task (which directly relates to criticisms of the sensory-functional theory, and thus of the conceptual structure model too; Caramazza & Shelton, 1998).

McRae et al. (1997) explain that the differences between the two tasks are evidence that computing word meaning is dependent on the statistical regularities among properties, which is something artificial neural networks are sensitive too as well; and evidence that semantic similarity is based on overlapping features and additional information, presumably top-down higher-level knowledge. This knowledge is only used when semantic cognition has enough time to access it, otherwise fast decisions are made using feature similarity. Their results further support the idea that correlated features are indeed used for processing living things (and not for artefacts), given the semantic task is not slow enough to recruit higher-level conceptual details — what they call "other knowledge" (c.f., McRae et al., 1997, especially, p. 111).

McRae et al. (1997) claim, based on their experiments with healthy participants, that living things are more similar to each other because of their common genetic-evolutionary ancestry. Due to animals and plants conforming to their environment their phenotypes (i.e., their features) will come to be shared over a family of related species, creating dense inter-correlations of features in the semantic space of biological kinds in contrast to objects created by humans, which are not constrained by evolutionary but by societal mechanisms. In other words, artefacts tend to have fewer intercorrelated features across concepts because each one is created from a specific anthropocentric use. McRae et al. (1997) further go on to propose that colour serves as a good clue towards identifying living things, but not artefacts, since the latter are usually arbitrarily coloured and the former have relatively less variability in colouring (e.g., typical colour for bananas is yellow, turtles green, sheep white). Also when the colour of living things does change, especially with regards to food, the discolouration is semantically important (e.g., a green

banana vs a brown banana). When it comes to inanimate objects the colours usually do not carry information (e.g., furniture, clothes, tools, are not affected semantically by their colour), although there are exceptions (e.g., warning labels, road signs, uniforms) where the colour is meaningful due to socially agreed upon convention. In addition, they note that functional properties of artefacts may be used to define that object, thus placing more emphasis on one single feature and ignoring potentially other features that might correlate but are sidelined due to not being important (e.g., a hammer is defined as the implement used for hitting nails, as opposed to paying attention to the similarities between hammers, mallets, and other tools, or indeed noticing the differences and similarities between the many different subtypes of hammer, as probably is the case with the many different breeds of dog). Meaning that on the flip-side people, according to McRae et al. (1997), tend to anthropomorphise animals, thus paying attention to their agency, patterns of behaviour, and multiple functions (e.g., a dog is for many things depending on the circumstances, for companionship, for guiding the blind, for herding sheep, and so on).

Perhaps what really happens, is that such category-specific effects do exist in SD but they are so mild as to usually be undetectable, or they go both ways, within patients, and thus essentially cancel each other out. Alternatively, the exceptional patients might just be the products of random noise (e.g., pre-morbid organisational differences), and therefore cannot be seen as candidates for explaining "normal" patterns of breakdown in semantic cognition, in much the same way as a neural network model may display such non-representative outlier behaviour when lesioned. Alternatively, it could be that many category-specific SD cases are not ones in which familiarity, frequency, and stimulus complexity are controlled for, leading to dissociations that are only category-specific on the surface. So when these factors are taken into they account the dissociation between domains disappears (as seen in Caramazza & Shelton, 1998; Gaffan & Heywood, 1993; Sheridan & Humphreys, 1993; Stewart et al., 1992). In any case, there is little empirical support so far for including category-specific patterns of dissociation in any model of SD (even if soem SD patients can be seen to show mild category-specific dissociations, e.g., Garrard et al., 2002; Lambon Ralph et al., 2003).

### 8.2.6 Does a hub-and-spoke topology exist?

Another central assertion of the modelling work in Rogers et al. (2004) is the topology of the network itself, which is meant to parallel the semantic system. The account of semantic cognition given in Rogers et al. (2004), and by the hub-and-spoke semantic theory in general, is that a network of amodal and modality-specific structures exist in the brain. These are

connected in such a way that sensory information is sent to the hub from the spokes and vice versa. This is the simplest formulation of a hub-and-spoke theory for semantics. And as such, is a useful and relatively common way of thinking about the semantic system (M. F. Bonner, Peelle, Cook, & Grossman, 2013; Chiou & Rich, 2014; Campo et al., 2013; Binder & Desai, 2011; Hoffman et al., 2012; Hoffman & Lambon Ralph, 2011; Hwang, Hallquist, & Luna, 2013; Jefferies, 2013; Lambon Ralph, 2014; Lambon Ralph et al., 2010, 2009; Pascual et al., 2013; Patterson, 2007; Pobric et al., 2010; Skipper et al., 2011; Tranel, 2009; Tsapkini et al., 2011).

The main proponents of the hub-and-spoke theory add on to the these basic assumptions a neuroanatomical localisation; they claim that the hub is found in the anterior temporal lobe (ALT), and furthermore that the amodal hub is likely to be the anterior temporal pole (ATP) (Lambon Ralph et al., 2007; Rogers et al., 2004, 2006). In other words, using neuroimaging data Pobric et al. (2010) and Lambon Ralph et al. (2010) propose that the role of the hub matches the locus of the ALT.

> A central theme of the Hub Account is that the ATL is amodal, which is necessary for accessing concepts that must be retrieved based on different sensory cues, such as a sound, an image, or a word. Evidence for this view is drawn primary from studies of patients with semantic dementia, a disease characterized by progressive and rapid loss of semantic knowledge and cell loss that in its early stages, is localized to anterior aspects of the temporal lobe. Patients with this disorder have semantic deficits that are characterized by amodal receptive and expressive semantic deficits that are observed in response to pictures, words, sounds, and even olfactory information (Patterson, et al., 2007; Rogers, et al., 2006).
>
> These findings are inconsistent with evidence from neuroanatomy, reviewed earlier, suggesting that there is sensory segregation of ATL function. There are other reasons to be skeptical about claims of amodality in the ATL as well. Patients with semantic dementia have cell loss that extends into regions beyond the ATL, including prefrontal cortex, and inferior temporal lobe extending into lateral temporal cortex (Hodges, 2007). It is difficult to know whether more discrete cell loss, say to inferior aspects of the ATL, would result in semantic memory deficits limited to the visual modality, as predicted by anatomical findings (Moran, et al., 1987), since semantic dementia progresses rapidly and promiscuously. Moreover ATL resection for epilepsy rarely leads to sever, amodal semantic deficits (Drane, et al., 2008). Indeed a recent meta-analyses reported that semantic tasks using visual stimuli tended to

> show greater activations in the inferior ATL while similar tasks using verbal stimuli showed greater activations in superior ATL (Visser, et al., 2009). It is therefore possible that there are functional subdivisions within the ATLs in regards to the sensory modality of the stimulus material.
>
> (Skipper et al., 2011, pp. 3420-3421)

In short, the ATL is vaguely defined as a brain region, and moreover there seems to be a lot of evidence (as reviewed in Skipper et al., 2011) that the ATL contains sensory subdivisions. This does not sit well with the assumptions behind the hub-and-spoke account, which of course require a localisable semantic story free form modality-specificity of any kind. So while, the polar region of the ATL, the ATP, appears to integrate affective and sensory information while, the general ATL contains anatomical subdivisions that function on a modality-specific basis (Skipper et al., 2011).

Regardless of whether or not the ATL has perceptual localisations, the ATP within it remains an area whose general function is not well-understood (Olson, Plotzker, & Ezzyat, 2007). Evidence seems to suggest the ATP, which *could* be seen as the required amodal hub (or a high-level convergence zone to use Simmons & Barsalou, 2003, terminology), is more likely to be associated with socioemotional regulation than amodal semantic cognition damage, e.g., Klüver-Bucy syndrome (Lilly, Cummings, Benson, & Frankel, 1983), acquired bipolar disorder, etc. (see Olson et al., 2007, for an overview). In other words, the ATP has been found to have "a significant segregation in the processing of different sensory modalities, indicating that the ATL is not a general amodal semantic processor[, but that it is instead] more sensitive to social words than nonsocial words" (Skipper et al., 2011, p. 3431). More importantly, with respect to the hub-and-spoke account (Skipper et al., 2011) did not discover a specificity for hierarchical conceptual structure in the ATP, as a function of psycholinguistic measures (e.g., typicality, frequency, familiarity) as would be required by the hub theory (Rogers et al., 2004; Patterson, 2007; Patterson et al., 2007). However, Skipper et al. (2011) do not reject the hub's existence outright, they merely contest its proposed neuroanatomical location, indicating that the anterior fusiform gyrus has more of a chance of being a semantic hub. Notwithstanding, Skipper et al. (2011) do not find evidence that the ATL as a whole is amodal, or equally sensitive to all kinds of perceptual modalities.

### 8.2.7 Repercussions for the hub-and-spoke account

Rogers et al. (2004) uses different categories to test patients and model: the former are tested on land animals, sea creatures, birds, household objects, vehicles, and musical instruments; the latter is trained and tested on mammals, birds, fruit, vehicles, tools, and household objects. This, given the issues and dissociations mentioned previously, seems an inappropriate way to compare the two semantic systems, since the dissociations between domains depends highly on what constituent items are included in the testing. Additionally, the fact that the results indeed, in some cases, match the patient scores seems suspect given that dissociations are known to arise that would render the results different had they been done on groups of SD patients. Musical instruments for example are known to dissociated from other inanimate objects. etc.

The basis upon which Rogers et al. (2004) category-specific impairments are modelled is not well explained in the original publication. However, we think we can see a light at the end of the tunnel by appealing to the views expressed by the authors previously (e.g., Farah & McClelland, 1991; Lambon Ralph, Patterson, & Hodges, 1997) as well as within Rogers et al. (2004). This work tends to support the notion (first proposed by McCarthy & Warrington, 1986; Warrington & McCarthy, 1994) that it is sensory features that underpin the representation of living things, and functional features that do the same for inanimate objects. So the authors have, during both the course of Rogers et al. (2004) and in another model by one of the same authors (Farah & McClelland, 1991), seemingly subscribed to the notion of feature distribution being the cause, so to speak, of category specific deficits in SD. Thus, as in the Farah and McClelland (1991) model, Rogers et al. (2004) align themselves ideologically with the dissociation between sensory/functional being the "same" or the "cause of" the animate/inanimate dissociation — even though as discussed previously this is not entirely accepted by the authors themselves, in light of SD patients' data.

Perhaps the most consistent description for the category-specific effects in SD documented so far that are indeed down to SD-caused cortical atrophy and not confounds, is that they do indeed exist, but arise very rarely. This indicates, in our view that two equally plausible and non-mutually exclusive explanations exist for explaining category-specific differences: *a*) atypical spreading of the pattern of neurodegeneration (e.g., "pre-semantic visual impairment" Lambon Ralph et al., 1999, or a locus of damage more similar to that of HSVE, which does indeed give rise to category-specificity), that is still nonetheless part of the spectrum of frontotemporal lobar degeneration; and/or *b*) unusual premorbid cortical organisation, which could be due to expertise in a certain area, significant difference to normals in age of acquisition

of certain concepts, or lack of deep semantic knowledge in a category – this can lead to different categories being affected differently since both the content and the structure of the semantic system are proposed to be different from the average SD patient (Jefferies et al., 2011).

Because of the general scarcity of SD patients in general and the nascent nature of neuropsychologically investigating such finer details, these within group differences have been largely overlooked and under-researched. Nevertheless, they might prove useful to understanding both the syndrome and semantic memory in general. Either way category-specificity is not, at this stage, a part of SD neither empirically nor diagnostically — category-specific deficits cannot form part of the spectrum of symptoms that SD patients present, as only a handful have presented with such deficits, leaving the door open to both an atypical spread of SD within the frontotemporal lobes, as well as rare premorbid cortical organisation. As Lambon Ralph et al. (2003) also state, "[g]iven that patients with semantic dementia are typically homogeneous with respect to both neuroanatomical and neuropsychological profiles, the within-group comparison might provide a revealing method of testing various assumptions regarding category- specific deficits." (p. 311)

In conclusion, given current SD data, it does not seem feasible to consider category-specific effects (whether they arise from special premorbid temporal lobar organisation, or atypical patterns of neurodegeneration, or one of many possible confounds, or indeed something else entirely) to be part of typical SD. Therefore, we believe modelling SD patients' scores using a computational paradigm does not require the emergence of a category-specific deficit as part of the model.

## 8.3 Re-examining category-specific semantic deficits

### 8.3.1 Herpes simplex virus encephalitis

This section will re-examine and re-evaluate the assumptions and claims related to the most common aetiology of category-specific semantic deficits. The process by which the herpes simplex virus affects the brain when manifesting as encephalitis will be described. This is important because the spreading pattern of damage is relevant to any theory or model that makes assertions or assumptions about the cortical organisation and localisation of semantic memory. Although unfortunately, none of the models examined (and indeed none of the models which have been brought to the attention of the authors) mention how the herpes virus reaches, and spreads in, the brain.

A distinction should be made between "HSVE patients" who are currently experiencing

acute encephalitis and "HSVE patients" who have been cured of their HSVE but continue to live with the cortical (and subcortical) damage the virus has inflicted on their neuropsychology. The clinicopathological literature usually refers to the former (because brain damage by HSVE can be prevented the sooner the treatment is started, but if the neurological damage has already been allowed to occur there is little that can be done histopathologically), while the neuropsychological investigations always refer to the latter (since these patients have survived and can be subjected to testing in order to explore the long-term repercussions of their lesions). As shall be seen, and for perhaps obvious reasons, these two groups of patients (acute HSVE versus HSVE survivors) are not identical in their neurological and cognitive impairments.

Herpes simplex virus (HSV) has been known about for at least 2,000 years. Recognised and named in ancient Greece after its property of spreading when it manifests as a skin condition; "herpes" (cognate with the Latin "serpere" and modern English "serpent") literally means "to creep" or "to spread" . However, it was not dissociated into the specific herpesviridae until the 20th century: herpes simplex virus, which consists of two types, as dissociated by Nahmias and Dowdle (1968), called HSV-1 (of interest here) and HSV-2 (rarely found in adults, Berger & Houff, 2008); varicella-zoster virus; Epstein-Barr virus; cytomegalovirus; and human herpes virus-6 (see table 1, Ferrari et al., 2009, for an overview of the herpesviridae).

It is estimated that about half of the human population carries HSV, which can give rise to diseases that are mild and uncomplicated, e.g., herpes simplex labialis (cold sores), as well as ones that are seriously debilitating and even life threatening, e.g., herpes simplex virus encephalitis (Whitley & Roizman, 2001). HSV-1 is the cause of approximately 95% of herpes simplex virus encephalitis (referred to as HSVE, usually in neuropsychological literature, or HSE, in clinicopathological investigations) in adults (Tien, Felsberg, & Osumi, 1993). 1 in 250,000 to 1 in 500,000 people per year are diagnosed as having HSVE, which is the most common form of sporadic fatal encephalitis (Kennedy & Chaudhuri, 2002; Whitley, 2006; Whitley & Kimberlin, 2005; Whitley & Roizman, 2001).

HSVE was initially suspected and discussed by the Mathewson Commission (1929). About a decade later, intranuclear inclusion bodies (evidence of viral multiplication in a cell) consistent with HSVE, as well as the isolation of the virus from brain tissue, was first documented in a newborn by Smith, Lennette, and Reames (1941). A few years after that, the first case of an adult, shown using equivalent evidence, was given in Zarafonetis, Smodel, Adams, and Haymaker (1944). "The most striking pathologic findings in this patient's brain were apparent in the left temporal lobe, in which perivascular cuffs of lymphocytes and numerous small hemorrhages were found. This temporal lobe localization subsequently has been determined to be

characteristic of HSE in individuals older than 3 months of age." (Whitley, 2006, p. 142)

Generally, "[e]ncephalitis is an inflammation of the brain parenchyma [meaning the glial and nerve cells] usually caused by bacteria or viruses, and often associated with meningitis" (Ferrari et al., 2009, p. 1). The encephalitis caused by HSV usually asymmetrically targets the grey matter of the temporal lobes usually causing inflammation, congestion, and/or haemorrhage. About two weeks after the onset of these symptoms, necrosis of the affected brain areas can be detected (Ferrari et al., 2009; Whitley, 2006; Whitley & Kimberlin, 2005). Diagnosis of HSVE is established by detecting the virus in cerebrospinal fluid using polymerase chain reaction[9], these findings can further be supported by the discovery of lesions in the temporal lobes, especially the left, using an appropriate neuroimaging technique (e.g, MRI, EEG, Kennedy & Chaudhuri, 2002; Whitley & Kimberlin, 2005).

As mentioned, HSVE is the most common sporadic encephalitis in the Western world (Raschilas et al., 2002). It can clinically simulate stroke or tumour, as shall be seen in the following sections (Castillo & Rumboldt, 2012). If left untreated, approximately 70% of infected individuals will die, with only 2.5% remaining neurologically healthy (Whitley & Roizman, 2001). HSVE begins with an infection (usually HSV-1 is acquired via the inhalation of respiratory droplets) of the peripheral and central nervous systems (CNS), after that the virus infects a cranial nerve and uses it to travel to the trigeminal (also known as Gasserian) ganglia (a bundle of sensory nerve cell bodies), where it can stay dormant (Castillo & Rumboldt, 2012). From these nerves, where the virus remains, it can be reactivated (often due to stress), or alternatively a reinfection may occur. Either of these two events can result in the virus reaching the brain, specifically the limbic areas (which include the temporal lobes). Once inside the brain it causes acute encephalitis, inducing haemorrhagic and necrotic damage.

There are currently a number of pathways proposed by which herpes can reach the brain. It is likely that all of them are possible and that the disease does not necessarily follow a specific nerve by its very nature; instead, it probably spreads as a function of the site of (re)infection or that of reactivation, meaning that HSV could (and does) reach the brain by following the olfactory (first cranial), trigeminal (fifth cranial), and possibly other cranial nerves (Ferrari et al., 2009; Johnson, Olson, & Buescher, 1968; Tien et al., 1993; Whitley, 2006). This is because the virus uses nerve tracts to access different parts of the nervous system. These nerves can "guide" the virus to orbitofrontal, medial and anterior temporal structures, as well as the insulae (Dinn, 1980; Ferrari et al., 2009; Tien et al., 1993; Twomey, Barker, Robinson, & Howell, 1979;

[9]Polymerase chain reaction (also known as PCR) is a technique of amplifying small samples of DNA into thousands to millions of copies of the same DNA sequence. This is useful for a diagnosis of HSVE, although it is sadly prone to false negatives at the early stages of the disease.

Ojeda, Archer, Robertson, & Bucens, 1983; Schlitt et al., 1986).

The characteristic HSVE frontal and medial temporal lobar lesions imply that the spreading of the virus is often via the trigeminal nerve from the trigeminal ganglion (Tien et al., 1993), where HSV-1 stays dormant in the body (Khanna, Lepisto, Decman, & Hendricks, 2004; Verjans et al., 2007). In other words, the pattern of spread of the virus is a function of neural connectivity – the areas of the brain innervated by the infected ganglion are those that show evidence of infection. The route of access of HSV to the CNS is a subject of debate, but most evidence points to the olfactory nerve being the the most probable candidate (Whitley & Kimberlin, 2005). This claim is based on various forms of evidence: *a*) HSVE patients, whose loci of damage are largely within the temporal lobes – the olfactory nerve projects into the limbic system, which contains a frontotemporal portion (Whitley & Kimberlin, 2005); *b*) electron microscopy that shows that in patients this neural pathway is indeed infected, with the temporal lobes, hippocampus, amygdaloid nucleus, insula, and cingulate gyrus containing HSV antigens (Dinn, 1980; Esiri, 1982; Kapur et al., 1994; Ojeda et al., 1983; Twomey et al., 1979; Whitley et al., 1986; Yochim, Kane, Horning, & Pepin, 2010); *c*) animal models that have shown that the analogues of the human temporal lobes are infected when the olfactory nerve is the HSV access route to the CNS (Schlitt et al., 1986; Stroop & Schaefer, 1986). There is also evidence for the virus spreading from the trigeminal ganglia to the frontal and temporal cortices (Kennedy & Chaudhuri, 2002; Davis & Johnson, 1979). However, these routes are by no means mutually exclusive. Any combination of appropriate nerve tracts can be used, in theory, for the virus to access the CNS.

Further evidence that the spreading pattern is a function of neural connectivity is provided by the biological properties of HSV: "neuroinvasiveness – the ability to invade the brain; neurotoxicity – the ability to multiply and destroy the brain; and latency – the ability to remain in a non-replicating form in neurons of dorsal root ganglia and the autonomic nervous system." (Whitley & Roizman, 2001, p. 1514) Additionally, this account is compatible with the human body's defence mechanisms, or rather lack thereof in this case, because spreading intraneuronally allows the virus to avoid an immune response (immunosuppressed, e.g., HIV-positive, patients show the same incidence of HSVE, although the disease progresses differently, Ferrari et al., 2009). To summarise, HSV-1 is transported by "retrograde [abnormal in direction] flow along axons that connect the point of entry into the body to nuclei of sensory neurons" (Whitley & Roizman, 2001, p. 1514).

The damage to brain tissue discussed above is focal. However, HSVE patients have been documented as having diffuse microstructural damage to white matter contralaterally to their

highly visible lesions by using diffusion-tensor imaging, as opposed to the normal MRI which only reveals the large focal lesions (Grydeland et al., 2010). So HSVE patients appear likely to have both focal and diffuse damage, which histologically dissociates them further from stroke and head trauma patients, as well from frontotemporal lobal degeneration patients (e.g., semantic dementia patients, see subsection 1.2.1). This subtle diffuse damage probably contributes, by some small amount, to their behavioural and cognitive impairments (Grydeland et al., 2010).

### 8.3.2 Clinical and neuropsychological aspects of HSVE

HSVE, before treatment, is identified by a typical set of symptoms: a sudden onset of fever and headaches; which if left untreated lead to a deterioration in mental state, seizures, and even coma. Survivors of HSVE often have: chronic epilepsy, memory, and personality disturbances, as well as, hearing loss (Castillo & Rumboldt, 2012). In other words, HSVE patients have a variety of issues due to their illness, see Table 8.19. Although approximately 38% to 56% of patients do return to normal function after appropriate medication (acyclovir), HSVE still has an "unacceptably" high mortality and morbidity rate even with the use of acyclovir (20-30%, Kennedy & Chaudhuri, 2002). Some form of significant neurological impairment can be found in most survivors (Sköldenberg et al., 1984; Whitley, 2006; Whitley & Lakeman, 1995; Whitley et al., 1986). Even individuals who have been treated very early, effectively, and made a good recovery can show significant neuropsychological and neurobehavioural deficits (Kennedy & Chaudhuri, 2002). These impairments usually include epileptic seizures. Alternatively, patients can also present with a focal neurological/neuropsychological deficit, meaning that a dissociable part of cognition appears damaged (e.g., Raschilas et al., 2002). The behavioural effects of HSVE also comprise mood and anxiety disorders (e.g., Dewar & Gracey, 2007) — and they can be a function of genetic factors (e.g., Dickerson et al., 2008).

In other words, HSVE patients show the classical "constellation" of frontotemporal features: aphasia or mutism, personality disturbances, generalised or focal seizures. In fewer cases coma, meningism, focal motor weakness, and even brainstem encephalitis have been documented (Kennedy & Chaudhuri, 2002). Patients show damage in areas such as: the hippocampus unilaterally or asymmetrically, along with adjacent areas such as the parahippocampus, the amygdala, specific temporal lobe gyri, and the temporal poles; the insula; the temporal lobes, which are usually affected unilaterally or asymmetrically and never in isolation;[10] the anterior and inferior temporal lobe gyri, which are affected more than their posterior and superior coun-

[10]This is important to bear in mind when comparing SD and HSVE patients since HSVE patients do not just have temporal lob damage.

terparts; often the substantia innominata (a region of the basal forebrain/anterior perforated substance); the fornix; the mammillary bodies of the limbic system; sometimes the thalamic nuclei, usually unilaterally; and sometimes the frontal lobes, more medially than dorsolaterally (for more details, and information on less frequently affected areas, see Table 8.17; Kapur et al., 1994).

In McGrath, Anderson, Croxson, and Powell (1997), 27 HSVE patients are neurologically and neuropsychologically evaluated, 30% of them have normal short term memory, 40% have a mild impairment, and 22% are severely affected; their mini mental state score, which gives an evaluation of cognitive function, is less than 25 in 18% of patients (a score above 27 is considered normal: Folstein, Folstein, & McHugh, 1975). Also, 65% of the patients show uni- or bilateral anosmia (i.e., loss of smell — recall the involvement of the olfactory nerve as both a potential site of initial infection and further providing a pathway into the brain), and 41% have mild non-fluent dysphasia[11] with one patient out of 27 having global aphasia. Some of the patients also have various motor impairments in their arms and faces, e.g., anterior opercular syndrome (McGrath et al., 1997). These memory and language disturbances, and more specifically, semantic ones are why HSVE patients are of scientific interest.

HSVE patients present with agnosias and anomias for visual and verbal stimuli. In addition they struggle to remember novel events and objects due to damage to their hippocampal areas (e.g., Stewart et al., 1992). Utley et al. (1997) document an HSVE survivor with anterograde and retrograde memory loss and prosopagnosia, while Stewart et al. (1992) present a survivor with anomia and language comprehension problems as well as severe memory impairments. Many authors also present similar patients with language disturbances (Yochim et al., 2010). Visual and verbal problems are rare, indicating that the deficits are mainly within the realms of semantics, language, and memory, as opposed to any specific modal pathway, e.g., the patient in Wilson et al. (1995) is shown to have intact visuo-spatial abilities thus dissociating them from his semantic and language deficits (Yochim et al., 2010), although Hokkanen et al. (1996) found that some HSVE survivors can have significant problems with tasks that are pre-/non-semantic, like drawing and copying.

Since HSVE is seen to spread to both the temporal and the frontal lobes, it would be expected to see survivors with executive functioning being impaired. This is indeed the case as shown in Hokkanen and Launes (2007). In Utley et al. (1997) 41% of participants have mild executive functioning impairments, in addition to the memory and language problems. There

[11]"Dysphasia" is often used to mean partial or mild, while "aphasia" is used to mean total or global language impairment – frequently they are used interchangeably.

are also cases with severe executive dysfunction, which is unsurprising as the locus of damage mirrors the spectrum found in FTLD (Yochim et al., 2010).

In addition, Utley et al. (1997) show a neuropsychological evaluation of 22 adults who having had HSVE have since been treated with acyclovir, see Table 8.18. These individuals' main cognitive repercussion, despite being heterogeneous, was anterograde memory impairment. The other common cognitive impairments found in these kinds of studies are: occasionally retrograde memory is affected, sometimes to the point of retrograde amnesia (e.g., four out of the ten patients in Kapur et al., 1994, showed signs of their retrograde memory being compromised); immediate memory span is only rarely affected in HSVE (e.g., no instances of affected dgit span was found in Parkin, 1993); language abilities are occasionally compromised (e.g., word-finding difficulties, anomic apahasia also known as dysnomia; Utley et al., 1997). Utley et al. (1997) also mention of Warrington and Shallice (1984),in which two HSVE patients are presented with category-specific deficits, but they believe these claims – in light of Funnell and Sheridan (1992) — are misplaced. This of course does not mean all claims of category-specific effects are products of biased stimuli, but that care should be taken when evaluating models and theories against a small sample and/or single semantic task.

In other words, HSVE occasionally gives rise to neuropsychological syndromes on the aphasic spectrum, because it can cause focal lesions in the temporal lobe. Specifically, some HSVE patients display difficulties in manipulating and accessing concepts that are specific to a certain domain of knowledge. Care must be taken not be led astray by reports of HSVE patients who show category-specific deficits, as they are not representative of HSVE patients (a point also made by Kapur et al., 1994). This is because case studies are undertaken as a consequence of the extraordinary nature of the patients, and thus can be a source of sampling bias if this is not taken into account. However, this dissociation of semantic cognition is the aspect of HSVE that neuropsychological, cognitive neuroscientific, and computational modelling investigations of semantic memory focus on, especially because it provides a segue into dissociations within the semantic system.

Many studies of HSVE survivors usually involve focussing on individual cases that are exceptional, e.g., patient SD in Yochim et al. (2010) has recurring HSVE (i.e., twice) in addition to also suffering from hyponatremia (which can cause brain damage). As others have explained:

> There is a paucity of research on the broad cognitive outcome of HSE in unselected cases. Most of the published outcome data are from single case studies that focus on various aspects of the memory impairments that are the most common sequelae of the disease. Even the few small group studies (including 4-10 participants) that

HSVE Clinical and Neuropsychological Profiles

| Age | Duration of Memory Disorder (y) | Amnesia Severity* Score | NART estimates IQ score | Verbal IQ Subtest Scores | Performance IQ Subset Scores | Picture Naming | Card Sorting |
|---|---|---|---|---|---|---|---|
| 53 | 7 | 0 | 106 | Infor = 8<br>Arith = 12<br>Simil = 6† | Pict ar = 8<br>B des = 15<br>Dsy= 11 | Impaired 2/30 | Normal |
| 39 | 15 | 1 | 117 | Infor = 6†<br>Arith = 8<br>Simil = 7 | Pict ar = 8<br>B des = 8<br>D sy = 8 | Normal 11/30 | Pronounced impairment |
| 42 | 10 | 2 | 113 | Infor = 7<br>Arith = 12<br>Simil = 11 | Pict ar = 8<br>B des = 10<br>D sy = 14 | Normal 15/30 | Normal |
| 45 | 3 | 3 | 110 | Infor = 5†<br>Arith = 10<br>Simil = 7 | Pict ar = 8<br>B des = 9<br>D sy = 10 | Impaired 2/30 | Mild Impairment |
| 59 | 7 | 4 | 122 | Infor = 10†<br>Arith = 12<br>Simil = 13 | Pict ar = 14<br>B des = 12<br>D sy = 14 | Normal 21/30 | Normal |
| 39 | 4 | 7 | (Premorbid dyslexia) | Infor = 5†<br>Arith = 6†<br>Simil = 8 | Pict ar = 13<br>B des = 12<br>D sy = 6† | Impaired 5/30 | Normal |
| 70 | 3 | 9 | 98 | Infor = 12<br>Arith = 12<br>Simil = 11 | Pict ar = 7<br>B des = 11<br>D sy = 8 | Normal 13/30 | Mild Impairment |
| 57 | 2 | 12 | (Dysphasia affected test score) | Infor = 5†<br>Arith = 7<br>Simil = 7 | Pict ar = 12<br>B des = 12<br>D sy = 11 | Impaired 0/30 | Normal |
| 65 | 7 | 13 | 89 | Infor = 7<br>Arith = 9<br>Simil = 7 | Pict ar = 7<br>B des = 8<br>D sy = 9 | Normal 11/30 | Pronounced impairment |
| 24 | 1 | 23 | 107 | Infor = 9<br>Arith = 11<br>Simil = 10 | Pict ar = 8<br>B des = 10<br>D sy = 7 | Normal 22/30 | Normal |

Table 8.17: "Clinical and neuropsychological profiles of cases of herpes simplex encephalitis
* Severity of amnesia was based on a composite score reflecting performance on the Wechsler memory scale-revised, the recognition memory test, and the current awareness test.
† Impairment.
Infor = Information; Arith = Arithmetic; Simil = Similarities; Pict ar = Picture arrangement; B des = Block design; D sy = Digit symbol." (Kapur et al., 1994, table 1)

HSVE Clinical and Neuropsychological Profiles

| Case | Executive Function Index | Verbal Memory Outcome Index | Visual Memory Outcome Index | Nonspecific Damage Index | Overall Cognitive Outcome | Remote Memory Index | CT scan: laterality of injury | CT scan: area of injury |
|---|---|---|---|---|---|---|---|---|
| 1 | + | ++ | +++ | ++ | ++ | ++ | R | T |
| 2 | - | + | + | - | + | ++ | - | nil relevant |
| 3 | + | +++ | - | + | + | ++ | L | T |
| 4 | + | - | + | - | + | - | R | T |
| 5 | + | - | - | +++ | + | NA | - | Not done |
| 6 | - | - | + | ++ | - | - | R | T & inf F |
| 7 | + | + | + | ++ | + | NA | L | T & P |
| 8 | + | +++ | +++ | ++ | +++ | +++ | R > L | T |
| 9 | - | - | - | - | - | - | - | Normal |
| 10 | - | - | ++ | - | + | - | R | T |
| 11 | - | - | ++ | ++ | + | + | R > L | T |
| 12 | + | ++ | - | ++ | + | ++ | - | Normal |
| 13 | - | +++ | + | - | + | ++ | L | T |
| 14 | + | +++ | - | - | + | - | L | T |
| 15 | - | - | + | + | + | + | R | T |
| 16 | - | - | + | - | - | - | R | T |
| 17 | - | - | - | - | - | - | R | Not defined |
| 18 | - | ++ | - | + | + | NA | L > R | operculum and Left T |
| 19 | +++ | ++ | + | +++ | +++ | - | R | T |
| 20 | - | - | + | - | - | - | R | T |
| 21 | - | + | + | +++ | + | - | R | T |
| 22 | - | - | - | - | - | NA | R > L | operculum |

Table 8.18: "Severity of impairment on neuropsychological outcome indices, and location of brain injury according to computer tomography data for adult participants"
- no impairment, + mild deficit, ++ moderate deficit, +++ severe deficit.
T = temporal, Fr = frontal, P = parietal, inf P = inferior parietal, R = right, L = left, R > L = right more than left, L > R = left more than right, NA = not applicable.
* Premorbid disorders: Case 11 has epilepsy; Case 12 has dyslexia. As such, results should be interpreted cautiously. (Utley et al., 1997, table 2)

HSVE Patient Features

| | Number (%) of patients | |
|---|---|---|
| | Brain-positive ($n = 113$)[a] | Brain-negative ($n = 85$)[a] |
| **Historical findings** | | |
| Alteration of consciousness | 109/112 (97) | 82/84 (98) |
| CSF pleocytosis | 107/110 (97) | 71/82 (87) |
| Fever | 101/112 (90) | 68/85 (78) |
| Headache | 89/110 (81) | 56/73 (77) |
| Personality change | 62/87 (71) | 44/65 (68) |
| Seizures | 73/109 (67) | 48/81 (59) |
| Vomiting | 51/111 (46) | 38/82 (46) |
| Hemiparesis | 33/100 (33) | 19/72 (26) |
| Memory loss | 14/59 (24) | 9/47 (19) |
| **Clinical findings at presentation** | | |
| Fever | 101/110 (92) | 84/79 (81) |
| Personality change | 69/81 (85) | 43/58 (74) |
| Dysphasia | 58/76 (76) | 36/54 (67) |
| Autonomic dysfunction | 53/88 (60) | 40/71 (56) |
| Ataxia | 22/55 (40) | 18/45 (40) |
| Hemiparesis | 41/107 (38) | 24/81 (30) |
| Seizures | 43/112 (38) | 40/85 (47) |
| Focal | 28 | 13 |
| Generalized | 10 | 14 |
| Both | 5 | 13 |
| Cranial nerve defects | 34/105 (32) | 27/81 (33) |
| Visual field loss | 8/58 (14) | 4/33 (12) |
| Papilledemia | 16/111 (14) | 9/84 (11) |

Table 8.19: "Comparison of findings in "brain-positive" and "brain-negative" patients with herpes simplex encephalitis (Whitley et al., 1982a and Whitley et al., 1982b)" (Whitley, 2006)
a: Of 202 patients assessed.

> have been published tend to select participants because they have memory problems rather than because they have survived HSE (4). Additional methodologic problems in many of the previous studies include an uncertain diagnosis of HSE, often based on clinical assessment rather than laboratory methods, and a wide variation in treatment factors (e.g., antiviral drug used, delay to treatment from symptom onset, duration of the treatment). The time that has elapsed between HSE and the neuropsychological assessment, and the tests used in the assessment, also vary across studies, making direct comparisons difficult.
>
> (Utley et al., 1997, p. 180)

### 8.3.3 Repercussions for accounts of category-specific deficits

The research itself results in a skewed impression of HSVE survivors, focussing only on one end of the spectrum. Patients with more extensive damage are more likely to have non-semantic deficits as well as semantic ones, since they are selected based on the cognitive repercussions of HSVE. This implies that the survivors and patients used for drawing inferences about the semantic system are already those on the more severe end of the spectrum of impairments. Therefore, care must be taken when describing these results, and indeed when describing the patients as "HSVE patients", without clarification.

With this in mind, most models for category-specific deficits are modelling outlier cases of brain damage within the broad spectrum of HSVE. In addition, models such as Tyler et al. (2000), which intend to provide a description and theory for these semantic impairments, have to be, and in the case of Tyler et al. (2000) are, described as category-specific deficit models and not as models of HSVE. So the Lambon Ralph et al. (2007) model, a hub model that purports to be emulating "HSVE patients", in fact, is modelling a very specific subset of HSVE survivors and possibly even modelling patients with other aetiologies such as head injury and stroke. The points raised in this section about the nature of HSVE are not addressed in any of these publications.

Even more importantly, as there is doubt on the results of Warrington and Shallice (1984), by Funnell and Sheridan (1992) and Utley et al. (1997), care must be taken to ensure that the patients and models being used to investigate category-specific deficits are not affected by a poor, careless, or mistaken, decisions with respect to stimuli. In addition, neuroimaging research (as seen in, e.g., Gainotti, 2005) should also be coupled with neurospychological examination to ensure that various sources of evidence are converging on the same phenomenon.

While their underpinning theories may be indeed backed up by category-specific patient evidence, the models themselves cannot claim to be models of semantic cognition in general, nor can they claim to be models of HSVE-, stroke-, or brain trauma-affected semantic cognition since these aetiologies do not usually cause category-specific deficits. They cause a spectrum of cognitive deficits largely outside semantic cognition in general, since the locus of damage is largely down to chance, although HSVE has a preference for the frontotemporal lobe (Tien et al., 1993). A balancing act must be attempted with respect to how much of the brain must be modelled and how many neuropsychological syndromes of semantic memory must be in included in a model of semantic cognition. However, we believe the line cannot be drawn with respect with what to model without explicit reference to all related empirical findings, e.g., that HSVE very very rarely results in category-specific deficits, and that SD essentially never does.

A model of HSVE should result in 70% or 20% of instances ceasing to work altogether, depending on the administration of antiviral therapy, as this is the level of mortality of patients once they are diagnosed (McGrath et al., 1997; Sköldenberg et al., 1984; Sköldenberg, 1996; Tyler et al., 2004; Whitley, 2006) and depending on the patient, as some are particularly susceptible to very negative outcomes, regardless of medication (e.g., Yamada, Kameyama, Nagaya, Hashizume, & Yoshida, 2003). Long-term survivors of HSVE report clinical features such as headache, confusion, nausea, vomiting, fever, seizures, drowsiness, abnormal mental state, meningism (neck rigidity, photophobia, and headache) and unconsciousness — with two thirds of them having residual neurological deficits (Whitley, 2006). The damage to their brain often is to the temporal lobes, with a pattern indicating that it follows the connectivity of the lobe, and hence of the semantic system itself:

> Neuroimaging with CT and MR reflects the pathologic findings of a necrotizing encephalitis involving the temporal lobe and the orbital surfaces of the frontal lobes, which may extend to the insular cortex, cerebral convexity, and posterior occipital cortex [16] (Fig. 1). The basal ganglia tend to be spared while frontal, panietal, and bilateral involvement is frequent [17]. involvement of the cingulate gyrus can be seen in HSV-1; however, this region tends to be involved later in the course of disease [18]. The characteristic location of lesions in the medial temporal and frontal lobes indicates the probable mechanism of spread intracranially along the small meningeal branches of the trigeminal nerve from the trigeminal ganglion.
>
> (Tien et al., 1993, p. 168)

The pattern of neurodegeneration seen in HSVE might be a hint as to why it is qualitatively

distinct from SD, since HSVE is likely to be following the pathways of neural connectivity — helping to potentially explain why category-specific patterns might be seen: more connectivity/overlap in neural sub-networks that deal with similar concepts, than with ones dealing with very distinct ones. This must be tempered by the fact that HSVE does also cause haemorrhage, necrosis, and neuronophagia (the destruction of a neuron due to an immune response).

Given that HSVE patients rarely display category-specific features, as mentioned, models such as Lambon Ralph et al. (2007) are in fact models of category-specific deficits and not of HSVE in general. Models that purport to model category-specific deficits in general such as Tyler et al. (2000), offer more convincing accounts of the dissociation found between domains in some patients.

Notwithstanding, the apparent success of the conceptual structure account as an explanation for category-specific deficits, it remains to be seen why such effects are so infrequent in patients. One explanation is that HSVE, stroke, head injury, and other (non-FTLD) aetiologies of category-specific deficits involve areas that are outside the temporal lobes that are nonetheless important to maintaining a healthy semantic system — as mentioned above, HSVE always involves areas outside the temporal lobes (Kapur et al., 1994). For example, damage to the posterior and/or anterior of the ventral stream of visual processing predicts a loss of living things, but cannot dissociate within this domain; while loss of non-biological items is usually found to be caused by lesions on more dorsal structures (Gainotti, 2005). These aspects of HSVE and of category-specific aetiologies in general must be taken into account when creating models.

## 8.4 Methodological considerations for semantic memory models

As seen previously, the sorts of language, memory, and recognition deficits SD, "a disorder of the temporal neocortex of the dominant hemisphere" (Compston, 2011, p. 2446), causes patients when tested are well-documented in some areas but less so in others. A dissociation between their symptoms and locus of lesion damage and those of other semantic disorders (e.g., herpes simplex virus encephalitis patients known for semantic deficits that are more severe in animals) is important. A more clear way of knowing which syndrome a patient/model really is suffering from behaviourally is required. Especially since in the case of models, neuroimaging, for example, cannot be used to detect which parts of the brain have been damaged as can be done with suspected SD patients (to rule out if they suffer from, e.g., FD, which is mainly

frontal and not temporal). This means that in the case of patients, there exists more than one source of information in order to complete a diagnosis (Gorno-Tempini et al., 2011). In order to "diagnose" a model, to really understand what it is that is being modelled, only the output of semantic task analogues can be used and/or qualitative comparisons between the model's and the patients' behaviour.

As such, the specific patterns of errors, e.g., variability in naming abilities across domains, as proposed by Rogers et al. (2004), must be used to infer what sort of semantic disorder the model is emulating. Otherwise, we carry the risk of haphazardly lesioning models (e.g., by randomly removing connections) without having strict a priori predictions and requirements for the nature of breakdown. The creation of such a benchmark is further hindered by the paucity of the reported scores in semantic tasks (e.g., very specific breakdown of the scores or indeed the raw data is rarely provided), and by the scarcity of SD patients, as "it is an uncommon disorder [...] with estimated prevalence 1–5/100,000 between ages 45 and 64" (Fletcher & Warren, 2011, p. 629). Meaning that even though SD is the second most common syndrome associated with FTLD — it only accounts for 15% of cases of patients found to have frontotemporal lesions (Snowden et al., 2002).

Perhaps, the only plausible way of maintaining a rigorous scientific classification of models into syndromes is the method Rogers et al. (2004) and many other modellers use, i.e., to fist enumerate a list of required behaviours the model must exhibit based on a large enough. Notwithstanding, that Rogers et al. (2004) approach for this was based on too few SD patients, it did allow for the evaluation of their modelling/behavioural assumptions. The problem however, is that for these assumptions to be investigated either new patient data must be collected or the model must be replicated in order to discover if these results continue to show the desired effect in similar or identical implementations. The patterns of SD patients' behaviour described in this chapter (in section 8.2.3) can be used as a basis for explicitly evaluating and "diagnosing" models which model SD. But there nonetheless, remains a lot of research to be carried out to determine which systems (e.g., olfaction, executive functioning, etc.) are involved when semantic memory is damaged. In addition, before any such model can be incorporated back into a theory it requires appropriate scrutiny to determine which parts of a model are implementation details (and thus superficial), and which are theory-level properties that must always be included.

## 8.5 Conclusion

Four influential theories and their computational models of semantic cognition have been investigated and implemented in this thesis: *a*) the hub-and-spoke account, which proposes that an amodal centralised store connected to perceptual areas accounts for the behaviour seen in healthy and both semantic dementia and herpes simplex virus encephalitis patient behaviour (Rogers et al., 2004; Lambon Ralph et al., 2007); *b*) the modality-specific account also sometimes known as the sensory/functional dichotomy, which postulates that semantic memory is underpinned by perceptually-based brain regions that each represent (pre-)semantic features separately (Farah & McClelland, 1991; Warrington & McCarthy, 1983); *c*) the conceptual structure account, which claims that the inherent distribution of (pre-)semantic features, i.e., their correlation within and between domains, drives the organisation of the system as well and the preservation of concepts after lesioning damage, giving rise to category-specific patterns of behaviour (Greer et al., 2001; Tyler et al., 2000); *d*) the conceptual topography account, which proposes that embodied approaches to thinking about semantic memory might provide the framework for understanding various deficits (Barsalou, 2010; Simmons & Barsalou, 2003).

An account relied heavily on in this thesis is the hub-and-spoke model, a semantic memory model, which was replicated both faithfully — in terms of their patterns, their topology and learning algorithm — and conceptually, using different topologies, different architectures and learning algorithms. All three broad families of reimplementations of the original Rogers et al. (2004) model, while sometimes theoretically distinct in terms of their topology (e.g., the modality specific and the conceptual structure models), and while not replicating the original results which matched the patient scores, did show the same general patterns of behaviour when tested. This indicates that the pattern set is able to drive the organisation of the models, regardless of their higher-level architecture.

In addition, a reimplementation of the Tyler et al. (2000) conceptual structure model, a model for category-specific semantic deficits, was also created which was successful in capturing the original model's qualitative effects. An extension to this model was also created which allowed for psycholinguistic variables such as frequency and familiarity of concepts can also be modelled in the same paradigm.

While on the surface these two models and their theories might appear very distinct they do share some core features. Both models define concepts as a function of features — features are split into perceptual and functional — and both models define each domain of knowledge as being composed of different distribution of these features. Both theories, the hub-and-spoke

and the conceptual structure theory structure theory then appeal to the mechanics of their own implementations to a greater or lesser extent. For the Rogers et al. (2004) model, attractors and their patterns of decay are used as to explain how semantic memories are damaged in SD and in HSVE (as modelled in: Lambon Ralph et al., 2007). For the Tyler et al. (2000) model, they appeal to the inherent statistical properties of their pattern set, which is represented in the feedforward network. In the case of the hub-and-spoke model this is a problematic method of explaining the patient data, mainly because their results do not replicate, but also because of the confounding of implementation- or model-level concepts (e.g., attractors, the pattern set, their specific learning algorithm) with theory-level concepts (e.g., high-level descriptions of patient behaviour).

It appears the hub-and-spoke account can theoretically shuttle between explaining things one way, i.e., the features are driving the attractors, or another, i.e., the features are driving the category-specific dissociations after damage (but not before?). These two explanation are clearly not incompatible but one would err on the side of preferring simpler explanations and opt for a feedforward network as in Tyler et al. (2000) or Farah and McClelland (1991). Including a recurrent hidden layer, which is what gives rise to attractor states, does not seem to provide anything over and above what the the latter explanation. Removing the existence of attractor dynamics, does not seem to affect how the pattern set shapes the internal representations of models, it merely precludes the ability to appeal to the breakdown of attractors. Appealing to attractors fails to offer anything theoretically over and above what the feature distribution can tells us a priori. This seems to be supported by the fact that a backpropagation through time model, a Boltzmann machine model, the conceptual topography, and the modality-specific models all show the same or very similar qualitative behaviour in semantic tasks. These four models are dramatically different in their learning algorithms and yet still produce the same semantic system, as seen from the point of view of testing — the only thing they have in common is that their training set is derived from Rogers et al. (2004).

Specifically, while our reimplementations are adept at patient modelling on the word-to-picture and drawing and delayed copying tasks, they do not fare well when reproducing patient scores in the naming and sorting tasks; nor do our models exhibit the required pattern of breakdown in their internal representations (Guest & Cooper, 2012; Guest, Cooper, & Davelaar, 2014). This means that the ideas encapsulated within the hub theory can lead to models that are not fully in line with the higher level aims of the theory, i.e., to explain the effects of the neurodegeneration caused by semantic dementia on the semantic cognitive system.

In the original hub model, Rogers et al. (2004) describe the breakdown in performance

of the hub model following damage as arising because "small amounts of drift may lead the network into an inappropriate proximal attractor, [thus making the model] produce incorrect responses appropriate to a semantically related object[, meaning that the attractor space is] robust even to relatively large amounts of damage, because the system's internal representations must be severely distorted before they drift out of the region to which such properties apply" (Rogers et al., p. 229). This has been shown not to hold for our reimplementations, as errors have been documented that are not semantic relations of the target response, instead they are from the opposing domain of knowledge. This is not documented in the original model, or the patients. In the reimplementations presented here it occurs even given relatively small amounts of lesioning damage. Why might our reimplementations, when damaged, fail to reproduce the behaviour reported by Rogers et al. (2004)? One possibility is that the pattern of breakdown of attractors as required by the hub theory is not a necessary consequence of a recurrent neural network trained with the structure of the training set. The hub theory assumes that attractors drift apart and merge in certain ways, as a consequence of the underlying recurrent neural network substrate, without *requiring* this at a theoretical level. But this assumption does not always hold. Based on this disparity between models and theory, it appears that the hub theory is underspecified as different implementations behave differently. Therefore, some additional theoretical constraint is required if models that implement the hub theory are to be consistent with the patients' behaviour. In our view this constraint should concern the behaviour of attractors following lesioning.

While we maintain support that a hub-and-spoke-like topology exists somewhere in the semantic cognitive system, and propose that out of all the accounts here this is the most complete and the most convincing, there remain many unanswered questions with respect to this theory as a whole. No model of the semantic system has so far shown that a different account can capture patient behaviour in the four main types of semantic task, however the hub-and-spoke theory does appear to propose a general overarching principle (centralised a- or cross-modal zones) that all other theories largely converge on. Also the relative reliance of conceptual representations on sensory and functional features as a function of domain and more fine-grained organisation seems to be an assumption common to all accounts. The issues outlined here with the hub-and-spoke theory and especially in terms of creating a robust specification of the model, must be addressed by the original authors, in order for both theoretical and empirical work to continue to be carried out reliably.

# Appendix A

# Training the hub-and-spoke model

## A.1 Overview

The learning algorithm used by the original network is described only as "a variant of the backpropagation learning algorithm suited to learning in a recurrent network" (Rogers et al., 2004, p. 208). J. L. McClelland (personal communication, 2011) confirmed that this was a variant of backpropagation through time (BPTT). Although the exact equations pertaining to the calculation of weight adjustments and the flow of signals are not specified by Rogers et al. (2004), a seemingly appropriate set was applied to the network from R. Williams and Zipser (1995). Despite the perceived suitability of this initial attempt, it revealed itself to be relatively unfruitful, for this reason a slightly modified version of McClelland's (2011) equations was also implemented. These two families of approaches share the same basic principles vis-à-vis the back propagation of error signals through the network; however, each method expands upon the mathematical complexity of its predecessor by encompassing small, but nevertheless important, differences.

## A.2 Training set

The set of patterns used by Rogers et al. (2004) to train the hub model has some very particular properties. Specifically, it contains some patterns in which visual and verbal sub-patterns are mapped onto the same name. The sharing of name sub-patterns is held to be analogous to the way a chicken, a robin, and a sparrow can all be called birds, both individually and collectively. What this amounts to here is, for example, 3 nondescript birds sharing the superordinate level name "BIRD"; forming a unidirectional 3-to-1 mapping from the three pairs of visual and

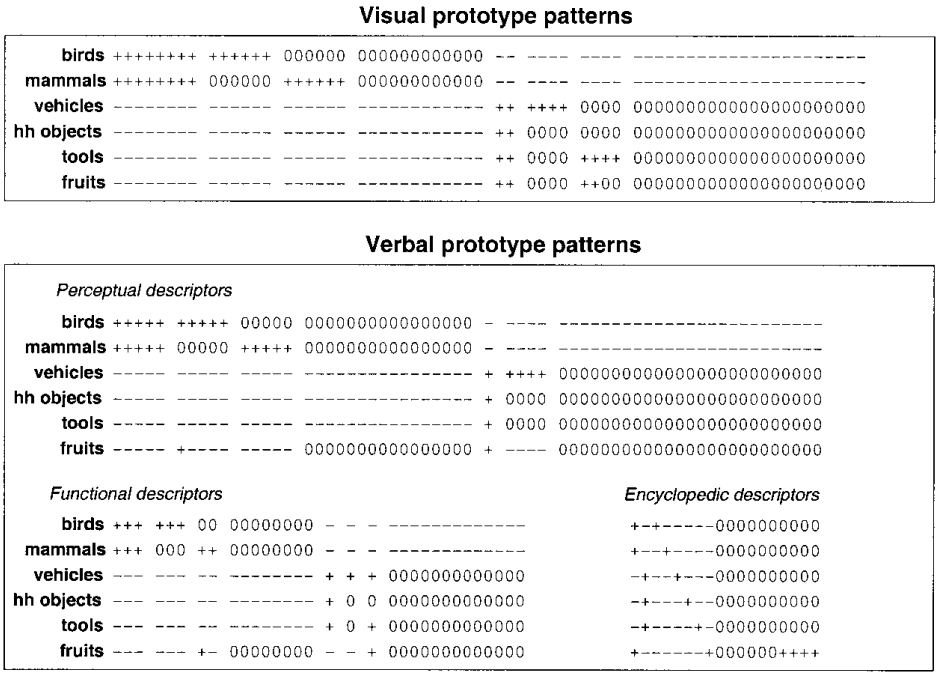


Figure A.86: Prototype feature vectors used to generate visual (top) and verbal (bottom) representation patterns for the model. Plus signs indicate units likely to be active for items in the category (turned on with $p = .8$), zeros indicate idiosyncratic units that are less likely to be active for items in the category ($p = .2$), and dashes indicate units that are never active for items in the category. hh = household. (Rogers et al., 2004, fig. 3)

verbal sub-patterns to a single name label. Conversely, if given "BIRD" their network "learned to generate visual and verbal properties common to most [birds]" (Rogers et al., 2004, p. 214). The network has to learn this slightly more complex relation, in addition to internalising the modal mappings for each pattern. These intricacies within the network's environment motivate the creation of two different training sets that partially dissociate the two forms of relations:

*Set A* contains patterns that just have a single unique name linked to them, so every item in this training set has a base name, but no names are shared over a number of objects; thus it is made up of 5 items for each of the 6 categories. These 30 patterns are created with the intention of testing whether the model can learn every possible mapping permutation between modal inputs. Set $A$ is by definition easier to learn than $B$ as it contains fewer patterns and excludes any categorical relations.

*Set B* is as faithful a replica of the training set used by Rogers et al. (2004) as possible, given the level of detail provided. This means $B$ includes a name sub-pattern that represents the label for a category in the cases of mammals, birds, vehicles, and tools. The remaining divisions within the domain of inanimate objects are not equipped with superordinate names. Rogers et al. claim that the category of household objects does not form a dis-

sociable enough, and therefore nameable, group. Although, it is less clear why fruit is not given a superordinate name; perhaps it is due to the fact fruit is not used by Rogers et al. to test patients' semantic abilities. The categories assigned their own name label (i.e., "BIRD", "MAMMAL", "VEHICLE", and "TOOL") contain the 3 sub-pattern pairs of visual and verbal descriptors that are associated with the general name and additionally 5 base-level named objects (e.g., "chicken", "dog", "train", "hammer"); resulting in each category explicitly named in the model consisting of 8 verbal-visual pairs and 6 names. The remaining 2 categories contain 8 patterns, each with a unique name. This creates a set with 48 distinct patterns overall, although there are only 40 unique orthogonal name sub-patterns.

It must also be noted that while names, in the case of set B, encode some categorical structure, there also exist domain and category bits that indicate domain/category classification for each and every pattern. This is contra to what Figure A.86 indicates, however the encyclopaedic units' prototype can be used to extrapolate what Rogers et al. (2004) must have meant by replacing the plus signs with a symbol that means "definitely on". The plus sign, which means "likely active" is not appropriate in the case of the encyclopaedic units as these domain/category units needs to be activated for all patterns to allow the network to carry out the sorting task. For details on the internal structure of all pattern sets used in this chapter see section 3.7, which analyses and discusses both the patterns and the attractors that emerge in the networks trained on those patterns.

## A.3 Back propagation through time

### A.3.1 Basic concepts

The classic back propagation through time (BPTT) learning algorithm, as defined in R. Williams and Zipser (1995), involves "unrolling" a network $N$ with $n$ units into a new feed-forward network, $N^*$, with $t$ time step layers of $n$ units. In the Rogers et al. (2004) model, $t = 28$ and $n = 279$ (divided into 215 visible and 64 hidden units). Each pair of connected units, $i$ and $j$, in $N$ is unrolled to create $t - 1$ pairs. The new unit pairs share the same weight on their connections, $w_{ij}$ (where $i$ projects onto $j$), as the original "rolled" pair with no two units being connected on the same time layer. This results in $N^*$ containing $t \times n$ units; in other words, a historical store of unit states is maintained for duration $t$.

Training is performed on the unrolled network. Initially, activations are propagated forwards

through $N^*$ from the first, $t_0$, to the last, $t_1$, time step[1]. Once all the states have been updated, the network is run backwards. During this backwards pass an error measure for each unit is calculated as a function of its state, its target, and the weights along its projections. The error signal derived from this process is subsequently used to adjust the weights. Over the course of one training epoch every sub-pattern of each training item is serially propagated through the network. This means that for each pattern, 3 different teaching stages are required in order to allow the model to internalise the mappings to and from all possible modalities.

### A.3.2 Teacher forcing & target states

On every time step, $\tau$, of $N^*$, visible unit states are clamped or have a target value set to a specific pattern bit. Teacher forcing involves clamping a unit $i$ to the state it is anticipated to have when fully trained; this amounts to ignoring input from units that project to $i$. As a result of hard clamping, subsets of units propagate a teacher signal derived from the sub-pattern currently being trained on (R. Williams & Zipser, 1989). On the other hand, a target for $i$ is set to the $i^{\text{th}}$ bit in the pattern presently clamped. Target values are equal to those used for clamping, however their function differs: a measure of unit, and by extension overall network, error can be computed by comparing actual and target states.

Rogers et al. (2004) specify that: the first 12 layers, $\tau \in [t_0, t_0 + 12)$, have the appropriate unit states clamped to the current sub-pattern, leaving the rest of the visible units free; and the following 16 layers are set the full pattern as a target, hence all visible units have a direct error in the interval $\tau \in [t_0 + 12, t_1)$. Hidden units are always unclamped, their states are free, and therefore do not directly transmit teacher signals; although, they are assigned virtual targets by the BPTT algorithm (R. Williams & Zipser, 1995).

### A.3.3 Forwards phase: propagation of activations

On the first unrolled layer, at time $t_0$, visible unit states, if applicable, are clamped to the current sub-pattern, while the rest are set to 0. Hidden units are assigned states with values selected at random from a uniform distribution of range $[0, 1]$. Non-clamped visible units can also be set to a random state in the range of $[0, 1]$ instead of 0; although, this slightly impedes

[1]The specific unrolling process used by Rogers et al. (2004) further unrolls each of the 7 time steps into 4 "ticks" each[2], resulting in $t_1 = 28$. Predefining the value of $t_1$ effectively forces the network during training to "settle"; or rather it is considered settled whether it actually has reached an inherently constant state or not, within the allotted interval of $[t_0, t_1]$. This is in opposition to allowing the reverberations to attenuate naturally over the course of as many time steps needed, as is the case with, e.g., BPTT when applied to settling networks (R. Williams & Zipser, 1995, p. 449). Forced settling does not impair learning, provided a stable enough state is reached within the predetermined time frame. Outside of the training stage, the network is allowed to reach true equilibrium.

the rate of learning (and that of setting) due to the introduction of noise to the system. For the time steps that follow $t_0$, the input to unit $i$ is calculated using:

$$\eta_i(\tau) = b_i + \sum_j s_j(\tau - 1) w_{ji} \tag{A.1}$$

where $\tau$ is an iterator variable over all the layers of $N^*$ with range $(t_0, t_1]$, $b_i$ is the bias of unit $i$ ($b_i = -2$ in Rogers et al., 2004), and $s_j$ is the state of a unit $j$ that projects onto $i$ (R. Williams & Zipser, 1995).

When unit $i$ receives input $\eta_i(\tau)$, its state changes to:

$$s_i(\tau) = \begin{cases} p_i & \text{if } i \in C_s(\tau) \\ \frac{1}{1+e^{-\eta_i(\tau)}} & \text{otherwise} \end{cases} \tag{A.2}$$

where $p_i$ is the $i^{\text{th}}$ bit of the pattern that is currently being taught to the network, and $C_s(\tau)$ is a set containing the indices of units that are clamped at time $\tau \in [t_0, t_0 + 12)$ (R. Williams & Zipser, 1995). The subscript $s$ denotes the current sub-pattern being clamped, in order to allow for the unit indices in $C$ to cycle over name, visual, and verbal, which are clamped and propagated serially through the interval $[t_0, t_1]$.

### A.3.4 Backwards phase: propagation of error signal

Starting from the last time step, $t_0$, error is calculated for each unit and sent back through the network. Each unit $i$ has an error at $\tau$, composed of a real and virtual part, that depends on its state and the errors of units in the next layer, $\tau + 1$. A unit's direct, or real, error at a given time is defined as:

$$e_i(\tau) = \begin{cases} p_i - s_i(\tau) & \text{if } i \in T(\tau) \\ 0 & \text{otherwise} \end{cases} \tag{A.3}$$

where $T(\tau)$ is the set of unit indices that are currently being evaluated against target states, i.e., $p_i$, and $\tau \in [t_0 + 12, t_1)$ (R. Williams & Zipser, 1995). These two sets, $T$, for targets and $C$, for clamped units are disjoint; a unit at any one time can be clamped or have a target, but not both. Intuitively, calculating the real error of a clamped unit will always return 0, as its state has been set to the current target.

At $t_1$ all visible states are directly compared to their targets and there is no virtual error. To calculate the virtual component of the error, direct error at $t_1$ must be percolated backwards to $t_0$ using appropriate equations. By combining the two, an overall error, $\epsilon_i(\tau)$, is derived and

based on this calculation a virtual target can be set for each free unit at $\tau$. At $t_1$, the overall error is simply:

$$\epsilon_i(t_1) = e_i(t_1) \tag{A.4}$$

For the full description of the error propagation equations see sections A.4-A.7.

### A.3.5 Weight adjustments

Once the network has been run forwards and backwards for an epoch, weight updates are calculated using the standard BPTT equations:

$$\Delta w_{ij} = \sum_{\tau=t_0+1}^{t_1} \delta_i(\tau) s_j(\tau - 1) \tag{A.5}$$

where $\delta_i(\tau)$ represents the error of the input to $i$ at $\tau$ (Rumelhart et al., 1986); for a detailed definition of $\delta$ see subsection A.4.1.

When adjustments have been collected (for all the patterns and sub-patterns), they are scaled by $\mu$, the rate of learning, which is empirically[3] set to 0.001. To ensure the pattern presentation order does not interfere with the learning process, $\Delta w_{ij}$s are applied in an epochwise manner (as opposed to after every single pattern). Weight changes are accumulated and applied if and only if the entire set of sub-patterns has been clamped and propagated exactly once.

Connection weights, despite being constrained by $\mu$, are liable to experience exponential growth; thus to counteract this tendency, $w_{ij}$s are decayed after each epoch by:

$$\phi = 1 - \frac{\mu}{3 \times |P|} \tag{A.6}$$

where $|P|$, the cardinality of the training set, is multiplied by the number of sub-patterns; the name, visual, and verbal sub-patterns amount to different input patterns in accordance with the network's view of its environment, albeit with the same targets (McClelland, 2011). For example, training set $A$ contains $|P| = 30$, therefore $\phi = 0.999988889$, and $B$ encompasses 48 patterns, making $\phi = 0.999993056$; more details on the two sets are provided in the following section.

[3]The learning rate has been found to produce the most effective learning at around 0.001 (McClelland, 2011). Values above this are detrimental to learning and will not result in minimising error; in contrast, values below 0.001 constrict the speed of learning without gaining any advantages.

## A.4 Method 1: Classic epochwise BPTT

### A.4.1 Error propagation equations

The first attempt at training involves using equations A.1-A.6 in combination with a set of equations that are based on R. Williams and Zipser (1995, p. 447, eq. 18 & 19). As usual, unit states are updated in a feedforward fashion, with activations spreading from $t_0$ to $t_1$, at which point the network is run in reverse. For the layers before $t_1$, a total error is calculated as the sum of the direct error and the virtual back propagated error:

$$\epsilon_i(\tau) = e_i(\tau) + \sum_{j \in U} w_{ji}\delta_j(\tau + 1) \tag{A.7}$$

where $U$ is the set of all unit indices in $N$, and:

$$\delta_i(\tau) = s_i(\tau)(1 - s_i(\tau))\epsilon_i(\tau) \tag{A.8}$$

The two equations presented above, A.7 and A.8, are the mechanisms employed to derive the adjustments for each $w_{ij}$. The value of $\epsilon_i(\tau)$ represents the relationship between the output error and small differences in unit $i$'s state at $\tau$. Similarly, $\delta_i(\tau)$ relates the overall network error with perturbations in the input at time $\tau$ that $i$ receives. As seen in Equation A.7, direct error, $e_i(\tau)$, is injected at every time step, this is in contrast to, e.g., the real-time version of the BPTT algorithm as presented in R. Williams and Zipser (1995, p. 445, eq. 14), which shares all other equations with this approach.

### A.4.2 Results

Over the course of training $N^*$, $N$ is sampled for output error once per epoch. This is done by applying the appropriate input values over subsets of visible units, allowing the network to settle, and then comparing the output unit states to their corresponding targets. Units are considered settled at time $t_{settled}$, which is defined as the time at which all states have changed by no more than 0.001 from the previous time step. No historical record of state values through time is needed when the network is being run for testing purposes, as error signals are not propagated and states are not unrolled.

In order to evaluate the nature of the network's internal mappings after the last training epoch, a sub-pattern is hard clamped for the interval $[t_0, t_{settled}]$; the network is sampled thus 500 times per sub-pattern. Hence, the overall network error is calculated with more accuracy

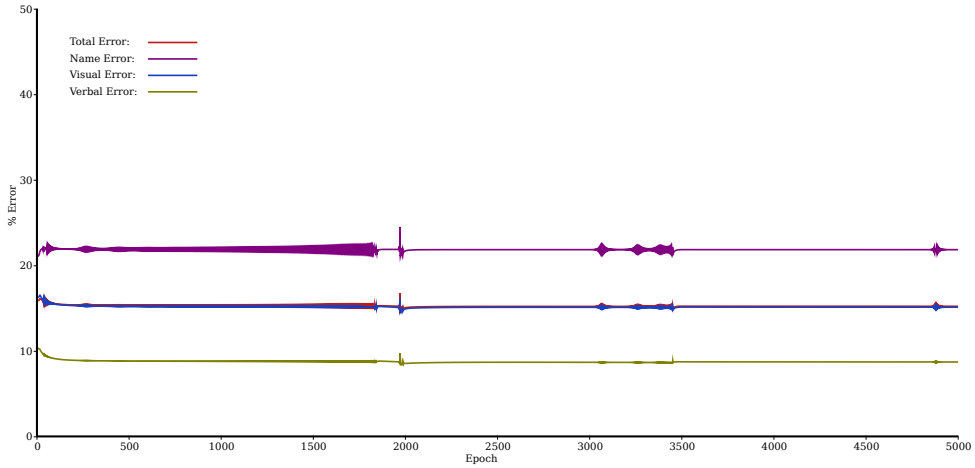


Figure A.87: Error during training on set $A$, Method 1.

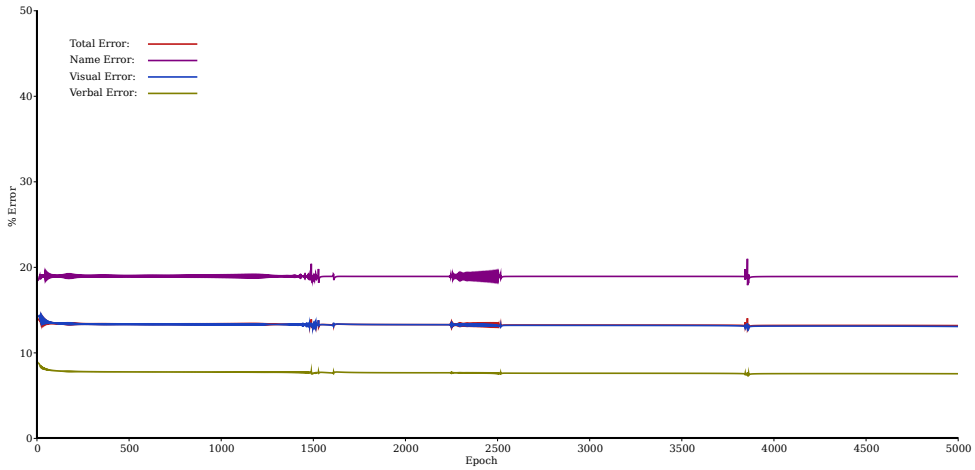


Figure A.88: Error during training on set $B$, Method 1.

then in the error graph and the coupling between each sub-pattern input and its semantic attractor state (i.e., the activations produced by the hidden units) is also investigated. The existence, or lack thereof, of attractor basins directly reflects the network's ability to map each pattern's sub-pattern to the same semantic concept, and by extension to the same output.

Figures A.87 and A.88 indicate that overall network error is unlikely to fall after the initial decline to about 15%. This level of stagnation, for both training sets $A$ and $B$, is symptomatic of the fact that the network has learned how to reach only one single attractor[4]. As a result of this, it produces static output independent of the currently clamped pattern, which corresponds to the average unit values of all members of $P$. For example, if most patterns contain a 1 at position $i$, $s_i(t_{settled}) \approx 1$; alternatively, if $p_i = 1$ in half the patterns and 0 in the rest, then $s_i(t_{settled}) = 0.5$. As is apparent from the error graphs, and from closer examination of the

[4]The two graphs are expected to differ marginally due to random variations between $A$ and $B$, this is because the two sets are generated independently and no patterns are common to both.

hidden unit activations, the network has not produced $|P|$ (i.e., one per pattern) concepts, it has instead created one single amalgamated attractor; thus rendering it unable to map input to output.

In conclusion, the weights computed using error propagation equations A.7-A.8 are very poor; they have not captured any of the inherent categorical properties of the data. Observing the error signals during training in real-time appears to indicate a large amount of noise is present, despite the injection of direct error to every time step after $(t_0 + 12)$. The extent of this noise may be partially responsible for the failure of the learning process. On the other hand, the network has demonstrated it can derive (by means of averaging) the properties present in most patterns, although Rogers et al. (2004) require this functionality over a specific category, not the whole training set. Rogers et al. do not explicitly define what output is expected in the case of category naming; whether it is binary (in line with the patterns) or real-valued. The results from this method motivate the investigation of alternative, but nonetheless similar, training techniques.

## A.5 Method 2: McClelland-based BPTT

### A.5.1 Error propagation equations

In McClelland (2011) a set of BPTT equations are described and used to train the Rogers et al. (2004) neural network architecture. This set, as in Method 1, comprises equations A.1-A.6; although, instead of using equations A.7 and A.8, error is computed using the following:

$$\epsilon_i(\tau) = \sum_{j \in U} w_{ji} \delta_j(\tau + 1) \tag{A.9}$$

$$\delta_i(\tau) = e_i(\tau) + s_i(\tau)(1 - s_i(\tau))\epsilon_i(\tau) \tag{A.10}$$

and at $t_1$, since $\epsilon_i(t_1) = 0$:

$$\delta_i(t_1) = e(t_1) \tag{A.11}$$

These equations form only a part of McClelland's reimplementation of the Rogers et al.. A significant difference between the equations presented here and those used previously in Method 1, is that the real component of error, $e_i(\tau)$, is injected directly to $\delta_i(\tau)$, as opposed to $\epsilon_i(\tau)$. This means direct error is not part of the second term of Equation A.10, as it had been previously in Equation A.7; thus making $\epsilon$ equal to virtual error only, as it does not contain any direct target comparisons.

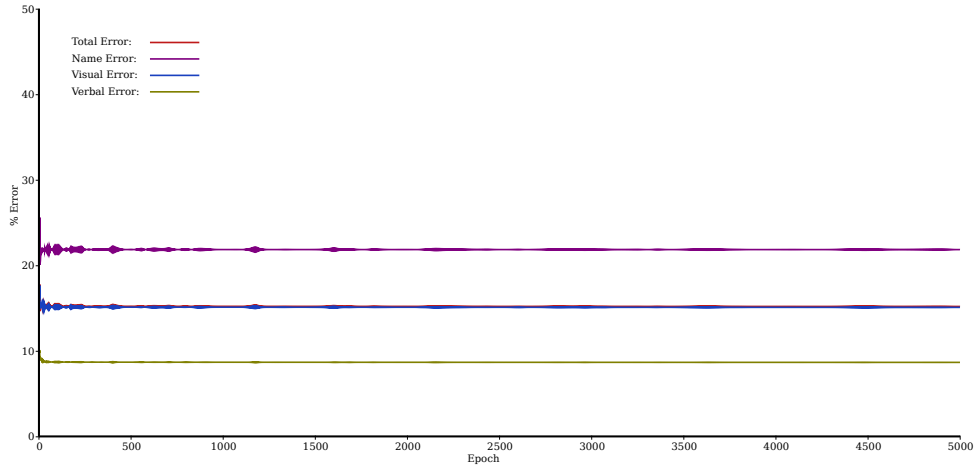


Figure A.89: Error during training on set $A$, Method 2.

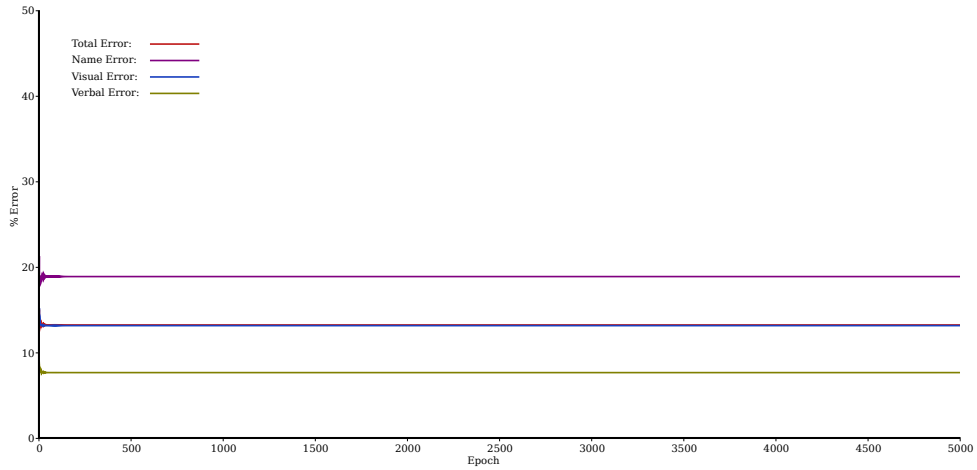


Figure A.90: Error during training on set $B$, Method 2.

### A.5.2 Results

The error graphed in figures A.89 and A.90 is sampled in the same manner as Method 1. The set of equations used in this training attempt have resulted in equivalent behaviour to that of subsection A.4.2. In other words, the network's performance is very poor; training has resulted in the creation of a single attractor, rendering the model useless at discriminating between patterns. The weights produced by both attempted training methods so far computer a single output vector given any input, specifically: the arithmetic mean of the elements of the training set. In both training methods explored so far the network will appear to be able to learn, if the training set used consists of a single pattern. To conclude, using the most basic form of McClelland (2011) BPTT equations does not further the model's capabilities. A form of enhancement, which is described in McClelland, may be required in order to increase the network's performance.

## A.6 Method 3: Time-averaged epochwise BPTT

### A.6.1 Time-averaging

The set of BPTT equations previously presented in Method 2, were originally enhanced and extended in McClelland (2011) through the use of a statistical method of noise reduction: time-averaging. Specifically, this process is applied over the variables $\eta_i(\tau)$ and $\delta_i(\tau)$ thereby enhancing the signals generated within McClelland's version of the network. This technique can be carried out on any time-varying property of a dynamic system (Hudgins & Kaspersen, 1999). Thus, by increasing the signal-to-noise ratio, the network can in theory learn over fewer epochs, whilst unrolling over fewer time steps. For example, time-averaging is appropriate for use over the states of unit $i$, since they evolve over time and results in a new variable:

$$\tilde{s}_i(\tau) = s_i(\tau)dt + \tilde{s}_i(\tau + 1)(1 - dt) \tag{A.12}$$

where $dt$ is defined as the reciprocal of the number of ticks a in single time step; thus, in this model $dt = \frac{1}{4}$ (McClelland, 2011). This new state for unit $i$, $\tilde{s}_i$, is used to replace $i$'s actual state in all equations, if time-averaging over unit states is required.

### A.6.2 Error propagation equations

In order to test whether applying time-averaging to the error signal, as calculated in Method 1, allows for a degree of noise reduction, and by extension a discernible difference in learning, the network is trained using:

$$\tilde{\delta}_i(\tau) = \delta_i(\tau)dt + \tilde{\delta}_i(\tau + 1)(1 - dt) \tag{A.13}$$

to replace every instance of $\delta_i(\tau)$ in equations A.5 and A.7. Similarly, the weights are now updated using:

$$\Delta w_{ij} = \sum_{\tau=t_0+1}^{t_1} \tilde{\delta}_i(\tau)s_j(\tau - 1) \tag{A.14}$$

reflecting the introduction of time-averaging over the variable $\delta$. Random noise in the network is expected to be minimised as a result of time-averaging; thus producing an increase in the learning abilities of the network.

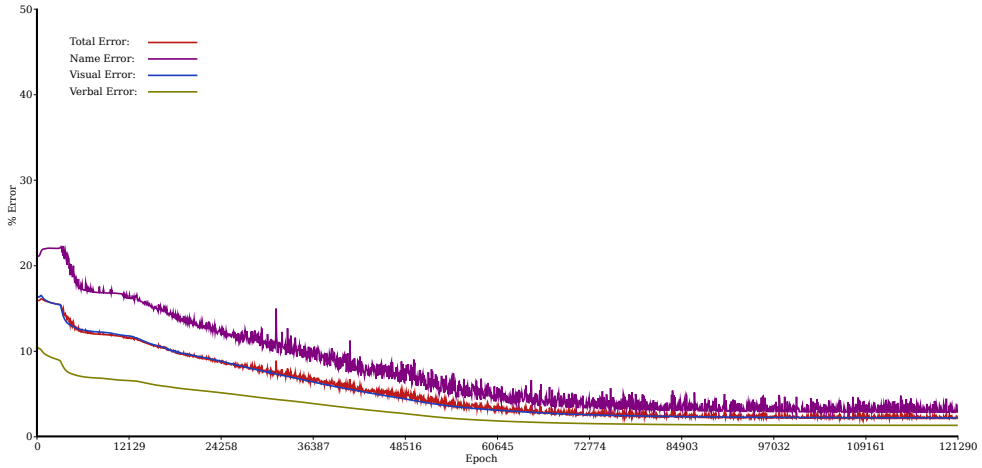


Figure A.91: Error during training on set $A$, Method 3.

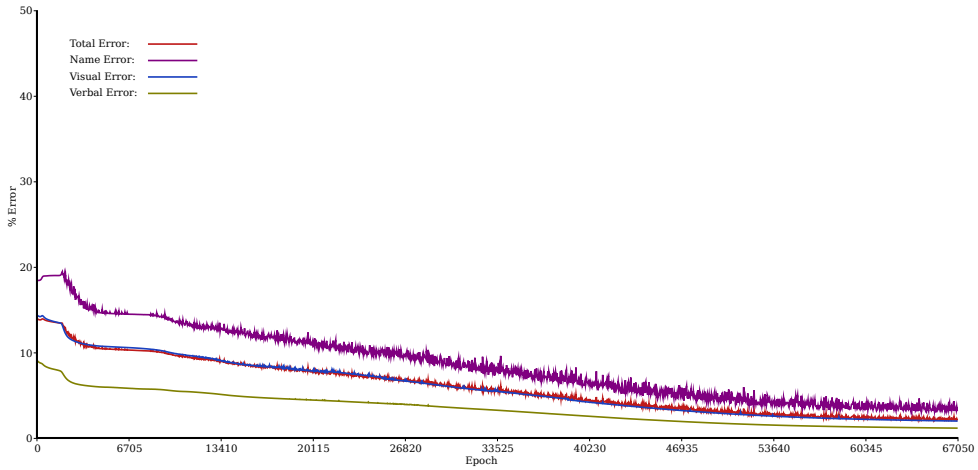


Figure A.92: Error during training on set $B$, Method 3.

### A.6.3 Results

Despite the clear indication of a decrease in error, the learning period required to attain the final level is in the range of tens (in the case of set $B$), or hundreds (for $A$), of thousands of epochs. Nonetheless, the network robustly associates a unique attractor with each pattern, in the case of set $A$, meaning it successfully can map a name, for example, to its related visual and verbal sub-patterns, along with the other two mapping permutations. The network is therefore able to derive, and arrive at, the required $|P|$ semantic attractors. On the other hand, when trained on set $B$, in the case of patterns that share a name label, the network derives a way of mapping names to visual and verbal sub-patterns that is potentially equivalent to the Rogers et al. (2004) model. Specifically, when given a superordinate name it activate visual and verbal sub-patterns that are an average of the 3 nondescript patterns that share the same name. Conversely, when provided with the visual or verbal descriptors the network activates the general-level name.

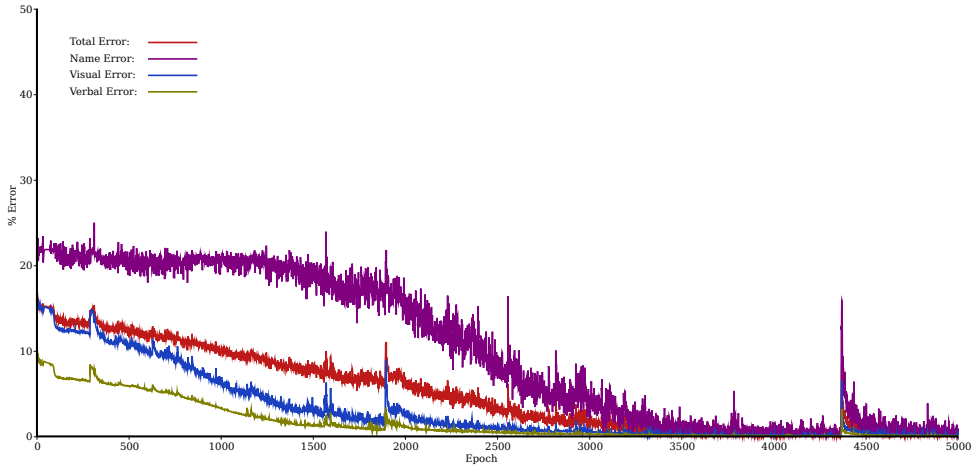


Figure A.93: Error during training on set $A$, Method 4.

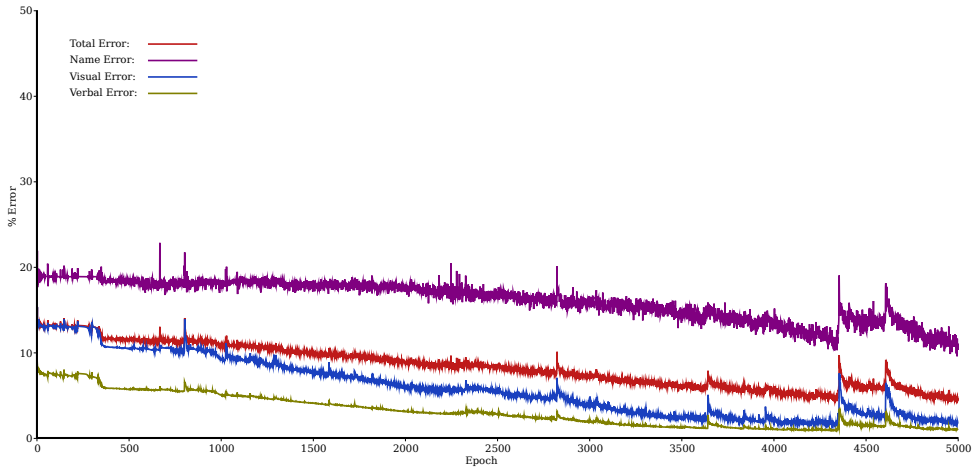


Figure A.94: Error during training on set $B$, Method 4.

Error during this training method appears to fall steadily along an appropriate learning curve, see A.91 and A.92. Although, a stagnation point for both $A$ and $B$ is reached; nonetheless, error is low enough: a unit state is on average 0.02 away from its binary target (Rogers et al. (2004) report that their unit error is 0.05 at the end of training). As time-averaging has produced a usable set of connection weights, when applied to the equations adapted from R. Williams and Zipser (1995), it seems appropriate to perform the same procedure to the equations taken from McClelland (2011).

## A.7 Method 4: Time-averaged McClelland-based BPTT

### A.7.1 Error propagation equations

In Method 2, the equations presented in McClelland (2011) were unsuccessfully used to train the network. However, the deficiency of Method 2 may be due to the omission of the process of time-averaging, which had been originally employed in McClelland. On the other hand, Method 3, which did apply time-averaging, produced acceptable results, albeit after a large number of training epochs. In-keeping with the replacement of $\delta_i(\tau)$ with $\tilde{\delta}_i(\tau)$, this training attempt combines time-averaging with the McClelland (2011)-based approach used in Method 2. Thus, Equation A.9 is changed to:

$$\epsilon_i(\tau) = \sum_j w_{ji}\tilde{\delta}_j(\tau + 1) \tag{A.15}$$

where $\tilde{\delta}_i(\tau)$ has been previously defined in Equation A.13, which at $t_1$ is simplified to:

$$\tilde{\delta}_i(t_1) = e(t_1)dt \tag{A.16}$$

The new error signal is used, as before using Equation A.14, to update the weights at the end of an epoch; the rest of the equations are identical to subsection A.5.1.

### A.7.2 Results

The results in the error graphs in figures A.91 and A.94 indicate a very fast learning curve in comparison to Method 3. For set $A$ error reaches levels below 0.1% at around $3,600$ iterations, which is significantly faster than Method 3. These results, along with a comparison of each attractor, demonstrate that the network has successfully assigned a unique attractor to each pattern, thus relating each sub-pattern to the same internal semantic state. The error appears to reach a point of stagnation in the case of set $B$; this occurs only after the network has learned every modal mapping, and is due to fact that the method of testing error does not take into account that names are shared across 3 patterns and thus, when an averaged output is given it is compared to a pattern and not to the averages of the 3.

## A.8 Conclusion of BPTT comparison

Use Method 4 as BPTT.

# Appendix B

# The self-organising map

## B.1 Overview

SOMs (self-organising maps) are a type of unsupervised neural network, also known as a Kohonen network (c.f., Kohonen, 2001). Kohonen, Hynninen, Kangas, and Laaksonen (1996, p. 4) characterise a SOM as "a "nonlinear projection" of the probability density function of the high-dimensional input data onto the two-dimensional display." In other words, SOMs are able to take as input high dimensional vectors and produce as output a lower-dimensional map composed of topologically related units that reflect the input's underlying structure. It must be made clear that the SOM is not a method "for pattern recognition; it is a clustering, visualization, and abstraction method." (Kohonen, 2001, p. XI)

SOMs are composed of a layer of input nodes, $x : x_1 \cdots x_n$, that are fully connected to a set of output nodes, $m : m_1 \cdots m_s$, known as SOM weights[1]. Every $m_i$ receives identical input to the rest of the weights (Kohonen, 1982). A simple way of visualising the matrix of weights, $m$, is by assigning each $m_i$ to a (hexagonal) cell, see Figure B.96 for an example. The input layer, $x$, represents the interface between the SOM and the patterns it is to be trained on. The inputs, which are fully connected to the weights, are a set of samples taken from a vectorial observable, $x \in \Re^n$, meaning each input vector represents a glimpse of the true distribution of $x$ that the SOM must approximate (Kohonen, 1990).

SOMs can locate clusters within a dataset much like principal components analysis (PCA), K-means, and other vector quantisation algorithms (Vesanto et al., 2000). However, within

[1]Be careful not to confuse SOM weights, which are the result of the process of classification performed by the SOM, with supervised neural network (e.g., Hebbian) weights (i.e., the connections between units).

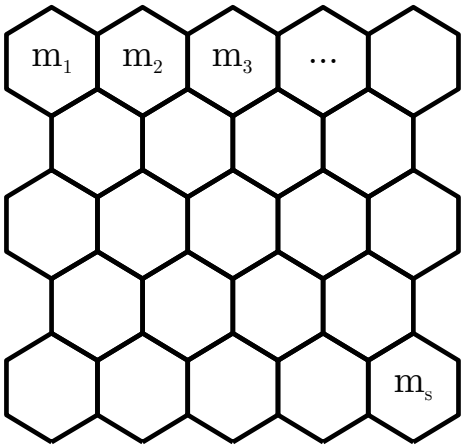


Figure B.96: Depicted is a view of a toy SOM. The matrix of weights is represented by a lattice of hexagonal cells – appropriate because each weight has been defined as having six immediate neighbours – that are arranged to reflect the hard-coded part of the topology of the map.

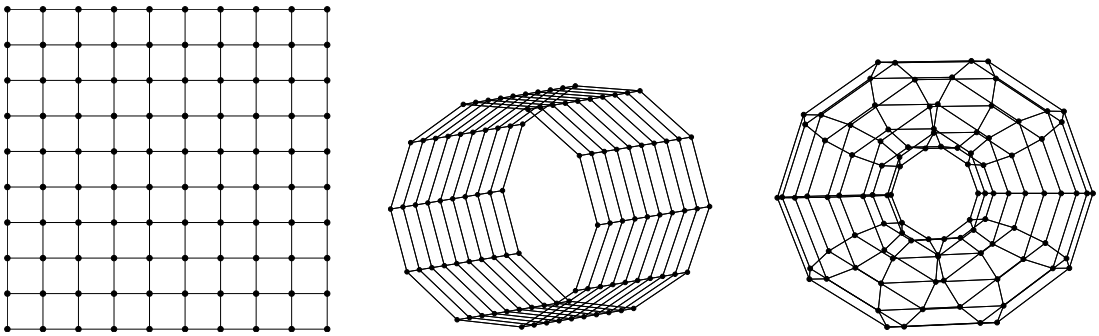

Figure B.97: In the three diagrams above each $m_i$ is depicted as a point, and a direct neighbour is represented by a vertex. On the left is the sheet map, in which the cells (which in this case are "square", so called because the have four neighbours each unless they are located on the border) are mapped onto a plane. In the middle, the SOM is mapped into a cylinder. And on the right, the SOM is mapped onto a toroid. The latter two map shapes offer a reduction in edge effects (Vesanto et al., 2000, fig. 2)

a SOM the weights are topologically related, meaning that weights that are proximal on the map are also proximal in vector space. There are two topological factors that define the map: the local lattice of individual weights, and the global map shape. The former can be seen in Figure B.96, in which hexagons are used to represent each $m_i$. Although, all three regular polygons that tessellate can be used to depict a SOM depending on if three, four, or six immediate neighbours for each cell are needed, hexagons are preferable because they are the most efficient way of tiling in two dimensions. The second property that affects topology is the global structure of the map in three dimensions. In addition to a flat map, the weights can be mapped onto the surface of a cylinder or a toroid, as seen in Figure B.97. These two forms offer an advantage over a flat map, especially the toroid, which defines every cell as having the exact same number of neighbours.

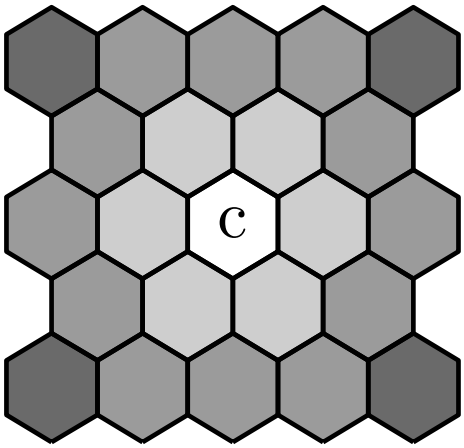


Figure B.98: This toy SOM is coloured to reflect discrete neighbourhoods of cell $c$. The neighbourhood with radius zero (comprising just $c$) is white, the one-, the two- and (part of the) three-neighbourhoods are shown in increasingly darker shades of grey (c.f., fig. 2, Vesanto et al., 2000).

Unlike other (supervised) networks, SOMs do not use their input units (directly) after they are trained in order to produce output. Instead, SOMs classify a novel or learned pattern by associating it to an element of the matrix of weights $m$, which is taken to represent it in the pattern-space. This is possible because each $m_i$ contains a codebook vector: $m_{i1} \cdots m_{in}$, where $n$ is the width of the input patterns. So each $m_i$ is compared $s$ times (the size of the map) until the best match for $x$ is located. In other words, each weight represents an approximation of an element of the population from which the training patterns have been sampled.

## B.2 Pattern-wise training

SOM weights are trained using a kind of competitive learning, meaning that each weight must compete against its neighbours to represent an input. The concept of a neighbourhood is required in order to produce topologically related subsets of weights, as opposed to independent weights (Kohonen, 1990). To ensure this happens each $m_i$ is assigned to a neighbourhood, see Figure B.98 for examples of different neighbourhood sizes.

The neighbourhood of a cell, $N_c$, is defined by its radius and the cell, $c$. For the weight $m_c$ to be selected as the centroid of the neighbourhood it must "win" the competition to represent $x$ by minimising the distance metric of choice:

$$\begin{aligned} d(x, m_c) &= \min_i d(x, m_i) \\ \Rightarrow c &= \arg\min_i d(x, m_i) \end{aligned} \tag{B.1}$$

where $i \in N_c$, and usually $d(x, m_i) = \|x - m_i\|$, although the inner product can also be used (Kohonen, 1990, eq. 4). The cardinality of $N_c$ decreases monotonically over the course of training, because at the beginning of training an initial radius for all neighbourhoods is set to a large value. In other words, at $t = 0$ (the first time step of training) neighbourhood radius is such that that it covers about half of the map, as it has been empirically discovered that this drives global organisation (Kohonen, 1990). As training progresses the neighbourhood radius should fall until it reaches a value of one. This means that at that point a single neighbourhood is made up of $c$ and its immediate neighbours, also known as the one-neighbourhood (recall Figure B.98). It is also possible to have a radius of zero, in which case each unit is directly competing with the rest, thus reducing SOM learning to simple competitive learning (c.f., Rumelhart & McClelland, 1986). These much smaller neighbourhoods are useful near the final stages of training as they enforce the teasing out of similar inputs into different neighbourhoods while preserving the previously created global topological structure.

As mentioned, SOM weights compete to represent the input. To determine which weight best represents an input pattern the weight must be closest to the input in space, as calculated in Equation B.1, at which point $c$ becomes known as the best-matching unit (BMU) for that input. The SOM algorithm then dictates that the proximity of the BMU and the other weights in $N_c$ to the input is increased. This is accomplished by bringing weights and input closer in $n$-dimensional space, but leaving all other weights intact:

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)\{x(t) - m_i(t)\} & \text{if } i \in N_c(t) \\ m_i(t) & \text{otherwise} \end{cases} \tag{B.2}$$

where $m_i(t+1)$ is the weight $i$ at the next training time step, and $\alpha(t)$ is the adaptation gain (or learning rate) which can be a linear or other function of $t$ such that it decreases over time and has domain $(0, 1)$ (Kohonen, 1990, eq. 6). This simple rule for determining the neighbourhood membership of $i$ is known as "bubble". This learning rule resembles that of the perceptron (Rosenblatt, 1957), except that there are no targets because there is no supervision (Kohonen, 1982).

A slightly more sophisticated way of updating weights, by allowing neighbourhoods to be more fuzzy by defining their boundary at the sub-cell level, is possible using:

$$m_i(t+1) = m_i(t) + h_{ci}(t)\{x(t) - m_i(t)\} \tag{B.3}$$

where the neighbourhood kernel function, $h_{ci}(t)$, can be set to $\alpha(t)$ if $i \in N_c(t)$ and zero oth-

erwise, to make this equation identical to Equation B.2 (Kohonen, 1990, eq. 7). Alternatively, to add biological plausibility (Kohonen, 1990), $h_{ci}$ can be defined as a Gaussian function of the form:

$$h_{ci}(t) = h_0(t) \ \exp\left( - \frac{d(r_i, r_c)^2}{\sigma^2(t)} \right) \tag{B.4}$$

in which the exponential function is applied to a quadratic function of the distance between the radius vectors of $i$, and the centre of the neighbourhood, $r_i \wedge r_c \in \Re^2$ (Kohonen, 1990, eq. 8). The scalars $h_0$, which defines the learning rate (equivalently to $\alpha$ in Equation B.2), and $\sigma(t)$, the width of the bell curve and hence the radius of $N_c$, decrease over time to make neighbourhoods monotonically smaller (Kohonen et al., 1996). With increasing $d(r_i, r_c)$, $h_{ci} \to 0$, meaning that cells can be left largely unaffected outside the neighbourhood, in a comparable way to Equation B.2. The main difference to the previous method is that the boundary between being an element of $N_c$ is now defined at the sub-cell level. In other words, Equation B.4 defines a three dimensional Gaussian neighbourhood centred around $c$, as opposed to previously when a step function (the bubble neighbourhood) was used (Kohonen et al., 1996). Other neighbourhood kernel functions can also be used, such as cut-Gaussian and Epanechnikov (fig., 4, Vesanto et al., 2000).

## B.3 Epochwise training

In order to train the network by presenting all patterns at once before updating the weights, the map must be treated in a slightly different way. This involved partitioning the SOM into topological regions, where each is in charge of representing one or more inputs. Before describing the simplified equations B.8 and B.9, for calculating weights and winning cells, as before, some theoretical background must be covered. Firstly, a description of the state the SOM weights must be in when trained is required. This can be achieved by setting the expectation values of $m_i(t+1)$ and of $m_i(t)$ for $t \to \infty$ to be equal, which is the case if the SOM manages to converge to a stable state (Kohonen, 2001). This means that at this stationary state $m_i(t) = m_i^*$ must satisfy the following equilibrium condition:

$$E\{h_{ci}(x - m_i^*)\} = 0 \tag{B.5}$$

as shown in Kohonen (1993, eq. 13), which can be rewritten as :

$$m_i^* = \frac{\int_{V_i} xp(x)dx}{\int_{V_i} p(x)dx} \tag{B.6}$$

where $p(x)$ is the probability density function of $x$, and $h_{ci}$ has been assumed to be the bubble neighbourhood function ($h_{ci} = 1$ if $i$ in $N_c$, otherwise $h_{ci} = 0$, see Equation B.2) and so can be omitted for simplicity (Kohonen, 2001, eq. 3.26). $V_i$ is known as the influence region of cell $i$ (Kohonen, 1993), because $V_i$ is the set of patterns that can update $m_i$ – the winning cell $c$ for each pattern $x \in V_i$ belongs to $N_i$ (Kohonen, 2001). It is now possible to to solve Equation B.6 iteratively (Kohonen, 1993). This involves classifying the samples from $x$ (chosen to be the training set) into their respective $V_i$ regions and updating $m_i^*$ (Kohonen, 2001).

In order to allow for more general neighbourhood functions the weights can be computed using:

$$m_i^* = \frac{\sum_{j=1}^{s} n_j h_{ji} \bar{x}_j}{\sum_{j=1}^{s} n_j h_{ji}} \tag{B.7}$$

where the sum is over every SOM cell ($s$ is the number of weights), $\bar{x}_j$ is the mean of the pattern(s) in the Voronoi set $V_j$ (c.f., Aurenhammer, 1991), and $n_j$ is the number of $x \in V_j$.

The above equations can be further simplified because it has now been shown that each $x_j$ (training elements are now indexed by their position in the batch, as opposed to by time of presentation) is placed into a Voronoi region of the SOM, meaning that each pattern belongs to $V_i$, where $i$ is their representation on the map (Vesanto et al., 2000). This allows the BMUs to now be calculated slightly differently to Equation B.1, to reflect the fact that the set of inputs, $x$, is now composed of patterns with width $n$ and length $p$ (the number of patterns in the batch), so the winning cell index is now:

$$c = \arg\min_i d(x_j, m_i) \tag{B.8}$$

making $c$ the best representation of sample $x_j$ on the map (NB: this must be done $p$ times to find the winner for each pattern). Weights are adjusted according to:

$$m_i(t+1) = \frac{\sum_{j=1}^{p} h_{ci}(t) x_j}{\sum_{j=1}^{p} h_{ci}(t)} \tag{B.9}$$

which defines the new weight as a weighted sum of the inputs, the value of the neighbourhood kernel function, $h_{ci}$ (Vesanto et al., 2000, eq. 4). In contrast to the method used previously to update the weights, in Equation B.3, this method contains no learning rate term and therefore converges to stabler values for $m_i$ (Kohonen, 2001). In addition, if $N_i = 1$, as is the case near

the end of training, then the epochwise training algorithm turns into the K-means clustering algorithm, if all data samples are used during training, thus guaranteeing the most accurate approximation for the input (Kohonen, 2001).

## B.4 Empirical recommendations

The procedure of training the SOM involves oscillating between Equation B.1, the competition between the SOM weights, and Equation B.3, updating the weights to reflect the new topological organisation. This might seem straightforward, however, certain precautions are advisable. While the SOM is very powerful it is still mathematically ill-posed, which means at least one of the following is true: a solution does not always exist, a solution is not always unique, the behaviour is dependant on small changes to initial conditions, i.e., the behaviour is unstable (Hadamard, 1902). Due to this, influential mathematicians (M. Cottrell, J.-C Fort, and G. Pagès) claim that "[d]espite the large use and the different implementations in multi-dimensional settings, the Kohonen algorithm is surprisingly resistant to a complete mathematical study." (Kohonen, 2001, p. XI) This translates into relative care being required when using SOMs, especially since some published models have misunderstandings incorporated into their design (Kohonen, 2001).

As explained above it is preferable to take into account certain considerations, so as to avoid training SOMs that are difficult to interpret and unstable (c.f., Kohonen et al., 1996; Kohonen, 2001). Regarding the form on the network, *i*) a hexagonal cell lattice is most appropriate because it aids visual inspection by maximising compactness as well as, more importantly, not favouring a horizontal or vertical direction over and above the others as in a rectangular lattice. Also in relation to the form, *ii*) the two sides of the SOM must not be equal, meaning that a rectangular SOM is to be chosen over a square SOM (this is because of the inherent ordering that arises in the weights, see section 3, Kohonen et al., 1996, for more details). In other words, due to the learning procedure, the $m_i$ vectors end up approximating the probability densitiy function of the input data – meaning that a circular SOM will not have a stable orientation and also that the best dimensions for a SOM are those that roughly correspond to the first two principle components of the data(Kohonen, 2001). As mentioned previously, *iii*) care must be taken to either acknowledge or prevent edge effects (also known as boundary effects), by being aware of the three-dimensional shape of the SOM (Kohonen, 1982). Also aforementioned is the requirement *iv*) to monotonically decrease neighbourhood radius over training, which at early stages organises global topology and at later stages ensures similar but nonetheless distinct

patterns are teased apart.

With regards to learning, *v*) batch training is faster compared to other methods, such as sequential training, and is more likely to converge (Vesanto et al., 2000). In the case of exceptional patterns compensation might be needed, *vi*) inputs that occur rarely in the population and hence sample, might need to be exaggerated (by means of weighting them or by forcing them to have a larger $h_{ci}$) if they are required to have their own BMU.

When in comes to initialising the SOM care must also be taken as *vii*) the initial values of the weights directly affect the average quantisation error (the mean of every $d(x, m_c)$, which is the distance between BMU and the pattern they represent, referred to as AQE from now on). One way to experimentally find a good set of value for $m(0)$ is to train many otherwise identical SOMs and compare their AQEs. Related to the AQE is the average distortion measure (ADM), a weighted measure of the distance, defined as $\sum h_{ci}\ d(x, m_c)^2$, which is also appropriate in determining[2] how good a SOM is at capturing the properties of $x$. However, empirically-based evidence suggests that *viii*) random initialisation is not the most optimum way of obtaining a useful SOM; what should be done instead is to determine the two eigenvectors of the autocorrelation matrix of $x$ with the largest eigenvalues, normalise them, multiply them by the square root of their eigenvectors, and use them to linearly initialise the map (c.f. section 3.7, Kohonen, 2001). A final point to bear in mind is that *ix*) the scaling or normalisation of the input vectors usually produces SOMs with a lower AQE and ADM.

## B.5 Visualising the SOM

### B.5.1 Component matrix

The component matrix is a way of visualising a trained SOM by using a simple lattice, ignoring any global three-dimensional shape, as seen previously in figures B.96 and B.98. However, instead of (or in addition to) labelling the cells to reflect the rank order of the weights, each cell is coloured or shaded to represent the value of a single component, $j$, of the codebook. In other words, the colour of cell, $i$, represents the value of $m_{ij}$. This results in $n$ possible views of the same SOM which depict where each input component's cluster(s) can be found on the map. By comparing different component planes it can determined if they correlate with each other or with some higher-level categorisation not explicit in the data itself.

However, because each component $j$ needs a separate component matrix in order to allow

[2]Both ADM and AQE depend on the neighbourhood kernel, so without equal $h_{ci}$ values over the SOMs the comparison is meaningless.

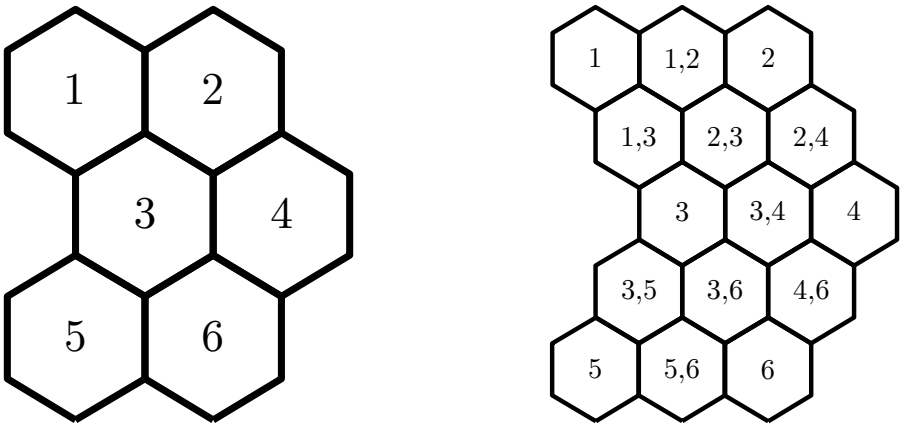


Figure B.99: On the left is a simple toy SOM with 6 weights – each cell is numbered accordingly. On the right, the U-matrix of the same SOM is shown. U-matrix cells are numbered with a single number if they present the same weights as on the left. The cells that represent the distance between two SOM weights are given a label that reflects this, e.g., the cell that corresponds to $d(m_3, m_4)$ is labelled "$3, 4$". The U-matrix allows the SOM to be shown in a way that facilitates the visualisation of clusters because the cells that represent distance will be coloured according its length.

examination of the distribution of values for that codebook element, component matrices are not ideal. If the width of $x$ is larger than about five, component matrices become an impractical way of looking at every property of the input that the SOM has uncovered. So while component matrices are useful to locate specific clusters, to determine how certain components of the input are placed on the map, and to pick up on correlations, they are ill-suited to viewing the whole map in a compact way. For a solution to this problem see below.

### B.5.2 U-matrix

The U-matrix (unified distance matrix) was designed to easily visualise both the SOM clusters and their respective distances (Ultsch & Siemon, 1990; Ultsch, 2003b, 2003a). As such, the U-matrix solves the problem of having to look through many views of the SOM in order to determine its inherent cluster structure without sacrificing a view of the topology (although other SOM visualisations are also available, c.f., Kohonen, 2001, for an overview).

The U-matrix is defined as "a collection of pairwise distances between the [codebook] vectors of neighbouring SOM [weights]" (Nikkilä et al., 2002, p. 975). This can be seen in Figure B.99, which depicts a toy SOM represented as a set of weights and as a U-matrix, which is composed of both the weights and the distance pairs between them (as before any three dimensional structure is not taken into account). The pairwise distances, Figure B.99, are coloured or shaded appropriately to display whether or not two weights form the same or a different cluster. The cells that represent the weights themselves are coloured by taking the mean of the surrounding

values (the minimum, maximum, or median are also applicable); alternatively, they can be left out of the U-matrix altogether. For an example see Figure 2.21.

## B.6 Connecting SOMs to classical layers of units

### B.6.1 Translating SOM output to activation values

In order to connect SOMs using classical neural network connections each SOM weight, $m_i$, is to be "twinned" with a classical neural network unit, also labelled $i$ (see the bottom half of Figure B.100). In other words, if a SOM has $s$ number of weights, then the twinned layer also has width $s$. An interface between the twinned classical neural network layer and the SOM's weights is achieved by defining the post-synaptic state of each twinned unit $i$ as:

$$\eta_i = \begin{cases} 1 - \dfrac{d(x, m_i) - \min_j d(x, m_j)}{\max_k d(x, m_k) - \min_j d(x, m_j)} & \text{if } i,\ j, \text{ and } k \in N_c \\ 0 & \text{otherwise} \end{cases} \tag{B.10}$$

where, as before, $x$ is the input to the SOM at $t$, the distance measure $d$ is defined appropriately (e.g., as the Euclidean norm), and $N_c$ is the neighbourhood centred around the BMU $c$ that encompasses the units $i$, $j$, and $k$ (Miikkulainen, 1997, eq. 1).

In order to use Equation B.10 during batch training, it is modified in an analogous way to before in Equation B.8. Specifically, $x$ gains a subscript because Equation B.10 must be iterated over $p$ times to find the full set of post-synaptic activations of the surface of the map for each pattern.

### B.6.2 Training connections

#### B.6.2.1 Oja learning rule

On appropriate way of training the connections between different SOMs, and between SOMs and layers of classical neural networks, is to use the Oja learning rule (E. Oja, 1982, 1989). This rule is a stable version of the classic Hebbian rule:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t) x_i \eta_j \tag{B.11}$$

where $w_{ij}$ represents the connection strength between units $i$ and $j$, $\alpha$ the learning rate, $x_i$ is the pattern value applied to the input unit $i$, and $\eta_j$ is the post-synaptic activation value of unit $j$. Classical Hebbian weights when trained iteratively will either grow unboundedly or decay to zero, which is a serious problem in most neural networks (K. D. Miller & MacKay, 1994). To overcome this issue Oja's rule can be used instead – as is the case in the models described in Miikkulainen (1997) and Sirosh and Miikkulainen (1997) that consist of SOMs connected to other SOM's using "linear Hebbian" weights (also see, Miikkulainen, 1993, eq. 7.6 & 7.7). Thus each connection is updated using:

$$w_{ij}(t+1) = \frac{w_{ij}(t) + \alpha(t)x_i\eta_j}{\|w_{ij}(t) + \alpha(t)x_i\eta_j\|} \tag{B.12}$$

where the units $i$ and $j$ can belong to two different SOMs (a destination and a source map), or one of the two can belong to a classical neural network layer that takes on pattern values or produces direct output (recall Figure B.100); $\alpha$ represents the learning rate, as usual, which can be equal to that in Equation B.2 (since these connections and the SOM weights are trained concurrently) and must, as in SOM training, decrease monotonically over time (E. Oja & Karhunen, 1985). The denominator in Equation B.12 is included in order to scale the weights; here the Euclidean norm is used: $\|x\| \equiv \sqrt{x_1^2 + \cdots + x_n^2}$ (Miikkulainen, 1997, eq. 4); meaning that the sum of squares of the weights is set to one upon every update (E. Oja, 1982). To minimise computational and time complexities, Equation B.12 can be simplified and then approximated by:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)\eta_j\{x_i - \eta_j w_{ij}(t)\} \tag{B.13}$$

if and only if $\alpha$ is sufficiently small and $x$ is limited to specific values, as shown in E. Oja (1982). In this simplified version it is easier to examine the forgetting element: $-\eta_j w_{ij}(t)$. This term is what causes each connection weight's growth to be adjusted, meaning that the larger the values that $\eta_j$ takes on, and thus the larger the Hebbian increment at the beginning of the rule, the larger the "leakage" (E. Oja, 1982).

As mentioned, if one of the $i$ or $j$ units is on the surface of a SOM and the other is on a classical input or output layer, as in Figure B.100, the latter values are taken directly from the set of patterns or targets appropriately. Alternatively, if the connection $w_{ij}$ links a unit on a map to a unit on another map, then both $x_i$ and $\eta_j$ are calculated using Equation B.10. In other words, in this latter case $x_i \equiv \eta_i$, because $\eta_i$ is being used as an input pattern.

#### B.6.2.2 Sanger learning rule

Oja's rule, when applied to most or all units in a network, is shown to extract the first principal component of the input data (Friston, Frith, & Frackowiak, 1993; E. Oja, 1982, 1989). In order to overcome this and discover more than just a single principal component (as the first might not be sufficient) the generalised Hebbian algorithm, also known as Sanger's rule and as sequential principal component analysis, can be used (Sanger, 1989; E. Oja, 1982):

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)\eta_j\{x_i - \sum_{k=1}^{j} w_{ik}(t)\eta_k\} \tag{B.14}$$

meaning that the subtractive forgetting element is only calculated based on "previous" units, i.e., ones that have had their weights updated prior to the current $j$. So when using the Sanger learning rule the first unit to have their weights updated will learn using Oja's rule – Equation B.14 when $j = 1$ is identical to Equation B.13. When $j > 1$, so for each subsequent output unit, the generalised Hebbian algorithm forces unit $j$ to learn something other than the first principal component (which is what unit 1 is leaning). In other words, each $j$ is taught to represent one eigenvector, thus obtaining a well-ordered set of principal components by decreasing eigenvalue (Sanger, 1989).

#### B.6.2.3 Widrow-Hoff learning rule

Generalised Hebbian weights have certain disadvantages regarding what they can and cannot learn. In order to overcome this issue it is useful to use the Widrow-Hoff learning rule instead, which is a simplification for the delta learning rule fro two-layered networks. This rule states that weights should be a function of the output error of the destination layer:

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)x_i(\sigma_j - y_j) \tag{B.15}$$

which is missing the factor $g'(\eta_j)$ because we only have two layers and thus can simplify the generalised delta rule, as $g'(\eta_j) = 1$

#### B.6.2.4 Lateral inhibition transfer function

The classical linear Hebbian activation function is defined as:

$$\eta_i \equiv y_i = \sum_{j=1}^{s} w_{ij}x_j \tag{B.16}$$

where $s$ is the number of input units; meaning that each $\eta_i$ is not the result of a non-linear sigmoid activation function on the pre-synaptic unit state $y_i$, but merely a sum of products. Thus the output of $i$ is identical to its input, which often results in post-synaptic states outside the standard $[0, 1]$ range. For this reason and in order to approximate lateral inhibition Miikkulainen (1997) modify this transfer function in their model. This new function remains linear but nonetheless scales the $\eta_i$ to $[0, 1]$. It will be referred to as the lateral inhibition transfer function and has the form:

$$\eta_i \equiv \frac{y_i}{\max_j(y_j)} \tag{B.17}$$

where $j$ is the index of the unit that has received the largest weighted sum (Miikkulainen, 1997, eq. 5).

### B.6.3 Self-organising maps

#### B.6.3.1 Historical details

Inspired by computational models of human memory, as well as the brain itself, Kohonen (2001) created the self-organising map (SOM), first as a rough idea in 1976, and then developing it much further starting in 1981. The SOM algorithm has a simple ultimate aim: to produce a low-dimensional output space given high-dimensional input, which places similar patterns next to each other.

The SOM was first successfully applied to the problem of speech recognition (Kohonen, Mäkisara, & Saramäki, 1984). Following this, which was at the time an optimal solution to a relatively hard problem, SOMs have been appropriated for many modelling applications (for an exhaustive bibliography of about ten thousand publications see: Kaski, Kangas, & Kohonen, 1998; M. Oja, Kaski, & Kohonen, 2003; Pöllä, Honkela, & Kohonen, 2009).

SOMs are held to reflect the same processes that organise the brain, more so than classical feedforward neural networks. To this end, Kohonen (2001) explains that "nervous systems [cannot] have a simple feedforward [...] structure, because the behaviour of even [those] species [with small neuronal networks] is already so complex and dynamic that some kind of planning functions are needed" (p. 98). In addition, the localisation of different cognitive functions in the cerebral cortex further lends support to the claim that SOMs offer a close parallel to the organising processes used by the brain. It is a well-documented fact that evidence from lesion studies and neuroimaging indicates that specific areas of the cortex are related to certain modalities, while others, e.g., the frontal lobes, are in change of integrating, associating, and controlling modality-specific brain areas. So to some degree self-organising processes do exist

in the brain, and these processes do not differ greatly in kind from the competitive learning which leads to the topologically-related output SOMs produce (Kohonen, 2001).

Furthermore, Kohonen (2001) lists three different kinds of map-like elements that can be found in the cortex: feature-selective neurons (e.g., orientation selectiovity in the primary visual cortex, Bienenstock, Cooper, & Munro, 1982), anatomical projection areas which are arranged somatotopically (e.g., the primary somatosensory cortex, which has a one-to-one mapping between a nerve region on the body and an area in the lateral postcentral gyrus, Nakamura et al., 1998), and ordered maps of abstract features (e.g., the colour map in visual area four, Zeki et al., 1991). "As no receptive surface exists for such abstract features, the spatial order of representations must be produced by some self-organising process, which occurs mainly postnatally." (Kohonen, 2001, p. 101) Moreover , these brain maps are able to plastically reorganise themselves after injury to the nervous system (e.g., the somatosensory cortex changes due to amputation, Elbert et al., 1994), after treatment (e.g., transcranial magnetic stimulation causes a regrowth of the mapping of the hand muscle in the motor cortex in stroke patients, Liepert, Bauder, Miltner, Taub, & Weiller, 2000), and after lesion damage and brain surgery (e.g., maps in the primary motor cortex for the hand and forearm can acutely reorganise themselves, Duffau, 2001).

To add to the reasons above in support of the existence of self-organising principles in the brain, three benefits to using topological organisation over and above other methods are provided: "1. By bringing mutually relevant functions close to each other, the wiring can be minimized. 2. If the responses are spatially segregated (although the underlying network may be distributed), there will be minimal "crosstalk" between the functions, and the brain architecture can be made more logical and robust. 3. It seems that for effective representation and processing of knowledge one anyway needs some kind of metric "conceptual space" to facilitate the emergence of natural concepts"(Kohonen, 2001, p. 101)

Kohonen (2001) laments the fact that even though these map-like structures, for both modalities and higher-level associative cognitive functions, are widely acknowledged and studied in neuroimaging, neuropsychology, etc., they are largely ignored by those who create cognitive models, especially those using artificial neural networks. In other words, a near universal aspect of brain form and function, "the existence of a meaningful *spatial order and organization*" (ibid, p. 101) is overlooked by a large proportion of modellers, and so by extension not incorporated in their work.

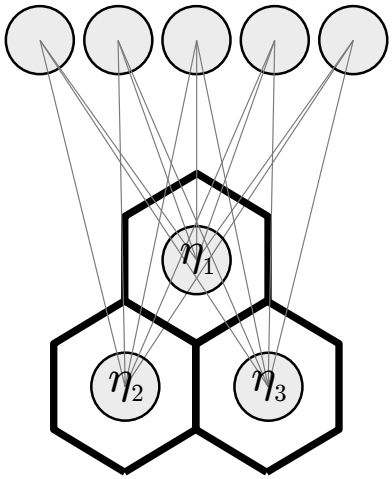


Figure B.100: In order to connect a SOM to a layer of linear units each of the SOM weights (depicted as hexagons) must, by means of Equation B.10, translate their codebook vectors into activations for their linear twins (the circles superimposed on the hexagons) to take on. Once the conversion from SOM weight to linear unit takes place, then standard connection weights can spread the activation to a linear output layer (seen at the top of the figure). Alternatively, the liner units at the top can represent a pool of input units that activate via the linear connections twinned units on the surface of the SOM, which in turn can be interpreted based on the topology of the map.

#### B.6.3.2 Classical connections between SOMs

In order for SOMs to communicate with each other or their environment (post-training) in a more traditional supervised neural network way, certain additions need to be made. Firstly, the SOMs need to be linked via some form of classical neural network connection, e.g., Hebbian. Secondly, these connections need to have a two-way interface to the cluster output the SOM provides; so SOM topological structure can both be used as input to the connections, and be used to interpret the output of these connections when they feed into the surface of the map. And finally, these connections need to trained in a way that allows them to function in the way described while being trained concurrently to the SOMs themselves.

#### B.6.3.3 Interpretation of SOM activations

To perform a theoretical inverse operation to above, i.e., to interpret the output of a linear connection on the surface of a destination SOM, two options are available, either: *a*) to discover which BMU's neighbourhood the most active unit (MAU) belongs to, named the *nearest-BMU* method; alternatively, *b*) to accept the MAU as the response of the network and access its codebook vector, which will be referred to as the *MAU-inspection* method.(It is presumed that some training has occurred prior to this, otherwise the interpretation will be largely meaningless because both the SOM weights and the classical connection weights do not represent anything

prior to training.)

In the first case the closest BMU to the MAU must be discovered. This is done by first applying the required pattern to the input units (be they on a source SOM or classical input units) and running the connections, thus obtaining a set of activations on the destination map. On the destination map, the MAU is located. Then, using a predetermined neighbourhood radius with the centre being the MAU, e.g., looking at all units in the three-neighbourhood and below, the closest BMU to the MAU is found, if any within the present radius, and returned as the model's output.

For the second method, the same process is repeating but instead of looking for the closest BMU, the codebook vector of the SOM weight that corresponds to the MAU is denormalised. This codebook vector resmbles what the SOM had determined to be an aceptable member of the real distribution of patterns, i.e., it has extrapolated based on the training set that such an element might be a variant of the dataset. And so the MAU's codebook vector can be treated as an output layer of an autoassociator and intpreted as such.

Adding this functionality allows the SOM's contribution to the topological organisation of whatever kind of hybrid network it is part of to be qualified. But this does not change the behaviour of the network as it is only a way of examining the behaviour and not of training, unlike its inverse described in subsection B.6.1, which is necessary for training the connection weights. However, it does have the potential directly dictate the behaviour of SOM-based models in semantic tasks. If the former mechanism is chosen to determine the results of, e.g., the conforntation naming task, will be partially responcible for the responses the model gives — meaning that changing the radius of search for the BMU has the potential to alter the distribution of errors, e.g., with larger radii there is a lower chance of omissions (see chapter 4).

## B.7 Note on SOM dimensions

The three SOMs used in this model have hexagonal cells (meaning that they have six neighbours), arranged in a $48 \times 40$ grid. This is due to the fact that these are the smallest non-square dimensions that allow for unique BMUs per pattern in the name-SOM (which consist of feature-poor representations of phonology, i.e., orthogonal patterns). Equivalent results can be obtained using maps with $40 \times 40$ dimensions, but this is a little unorthodox given the advice to the contrary (Kohonen et al., 1996; Kohonen, 2001, summarised in section B.4). Bearing that in mind, Miikkulainen (1997, the model that the architecture of this model is inspired by) uses square SOMs for the semantic, phonological, and orthographic maps (with dimensions $7 \times 7$, $9 \times 9$, and

$9 \times 9$ respectively). It is not clear why these proportions are used, but Miikkulainen appears to use SOMs with equal width and height in many models, e.g., Choe, Sirosh, and Miikkulainen (1996, where a $20 \times 20$ LISSOM map is used), citeA[in which $11 \times 11$ SLISSOM maps are used to model the cortex and retina]choe97, Farkas and Miikkulainen (1999, SOMs described as having $N \times N$ and $R \times R$ cells), Miikkulainen and Kiran (2009, which uses $10 \times 10$ and $12 \times 12$ maps); although there are models in which a rectangular map is used, e.g., Grasemann, Sandberg, Kiran, and Miikkulainen (2011, with $30 \times 40$ semantic and phonetic maps). As a compromise, DISSEM has an almost square shape to comply with Kohonen (2001) and DISLEX.

# Appendix C

# Effects of parameter variation on the behaviour of the conceptual structure model

The following tables of graphs depict the effect on the conceptual structure model (see chapter 5) of varying the momentum, the learning rate, the initial range of the weights, and the error function used to train the network. Changing these parameters affects both the scores in the semantic task after lesioning damage, as well as the healthy behaviour of the network itself. In the latter case, there are networks that do not learn, and so cannot be tested; when such a combination of values arises the corresponding cell will be left blank. The graphs show that, in the vast majority of cases when the parameter settings allow the model to learn the task, the model shows the hypothesised sensitivity to conceptual structure following damage.

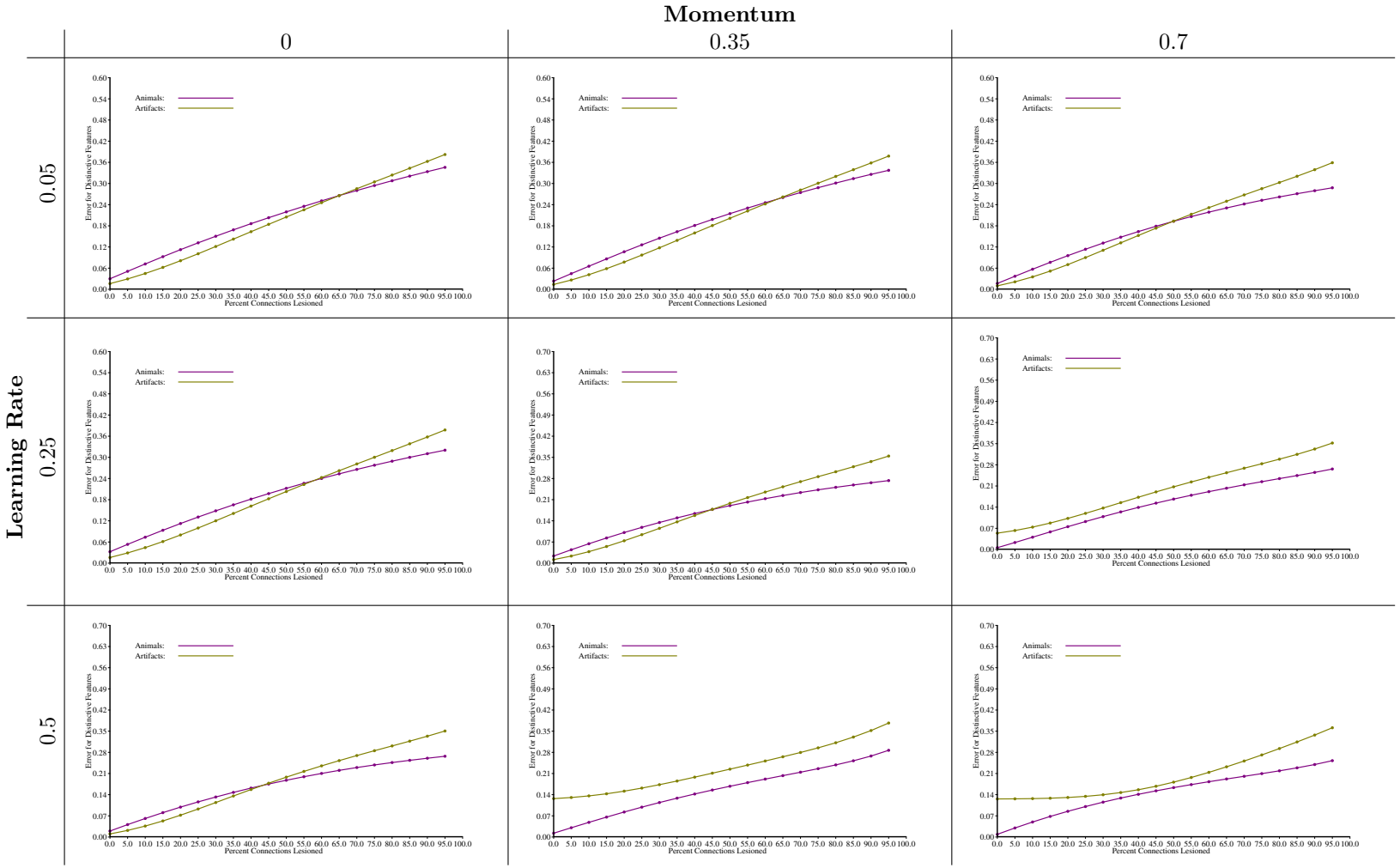


Table C.23: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (sum squared error) (compare with Tyler et al., 2000, fig. 3.).

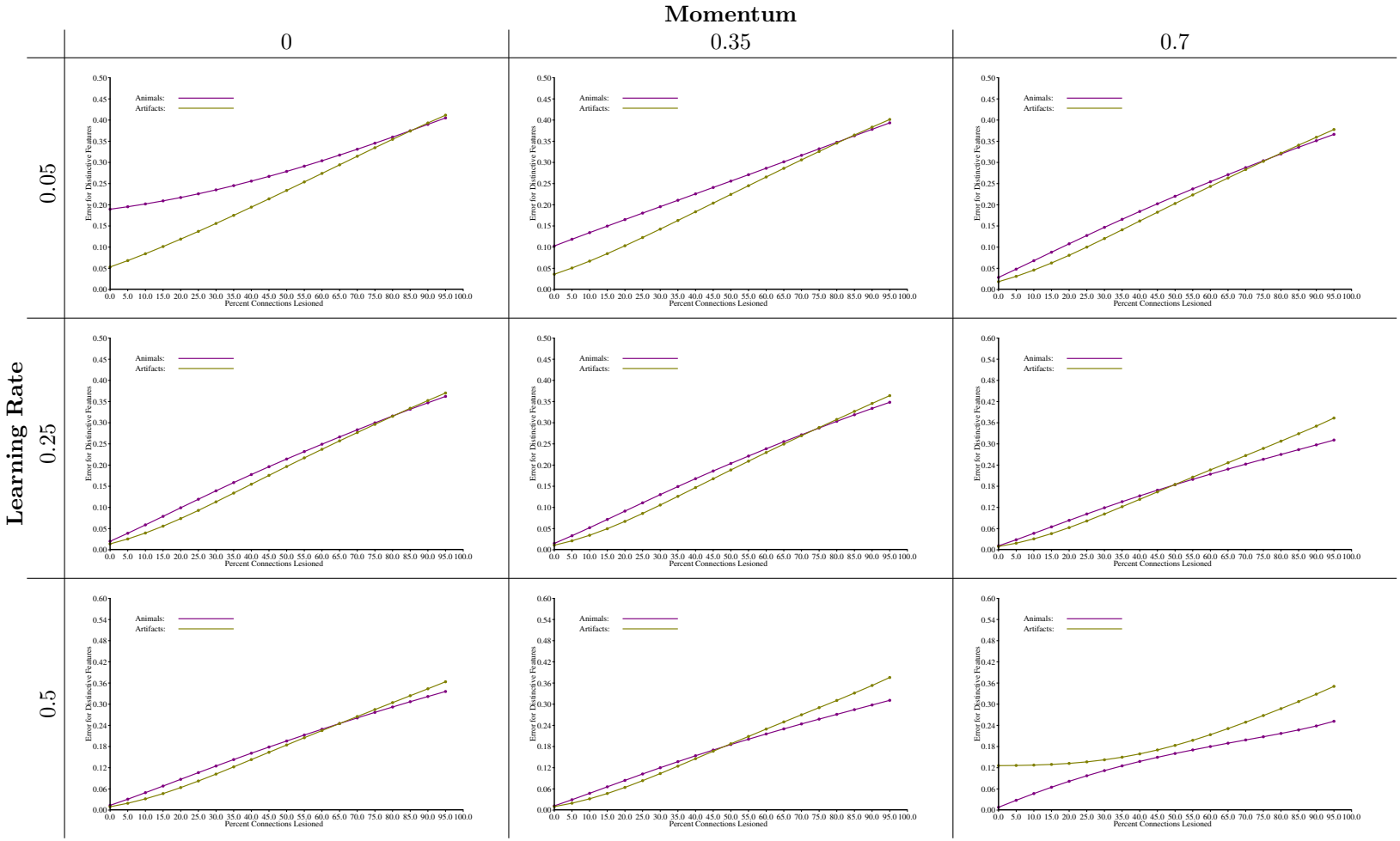


Table C.24: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 3.).

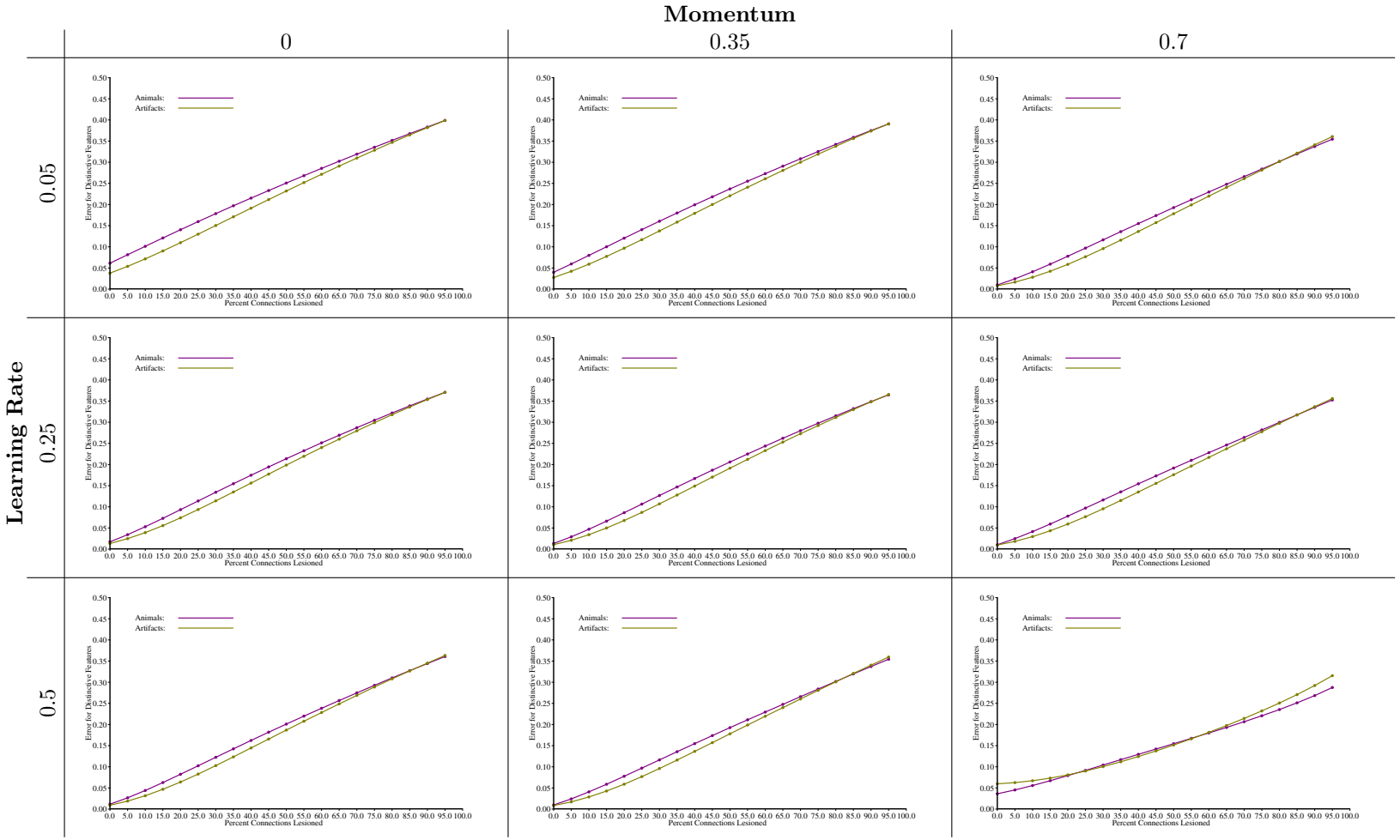


Table C.25: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 3.).

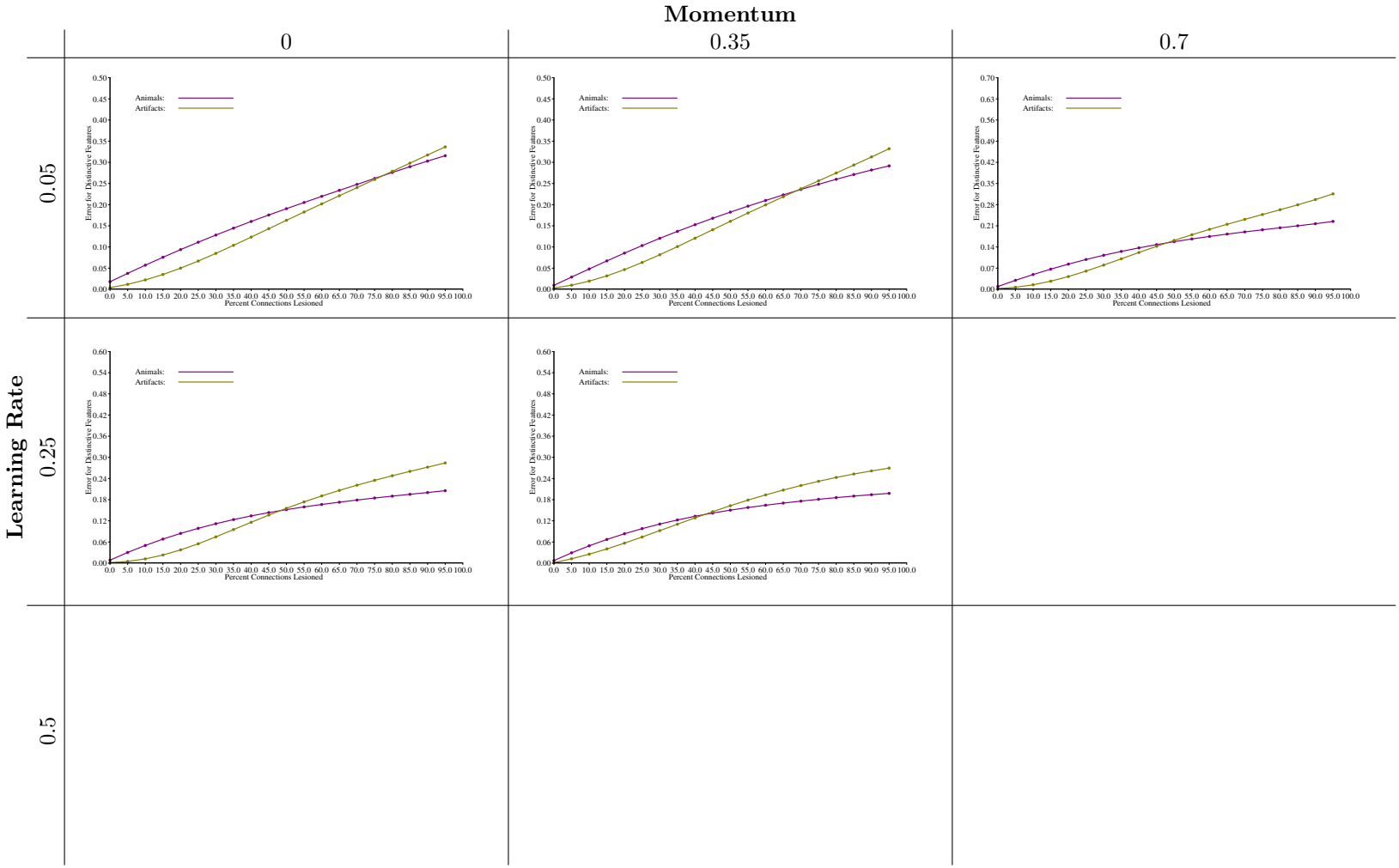


Table C.26: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (cross-entropy error) (compare with Tyler et al., 2000, fig. 3.).

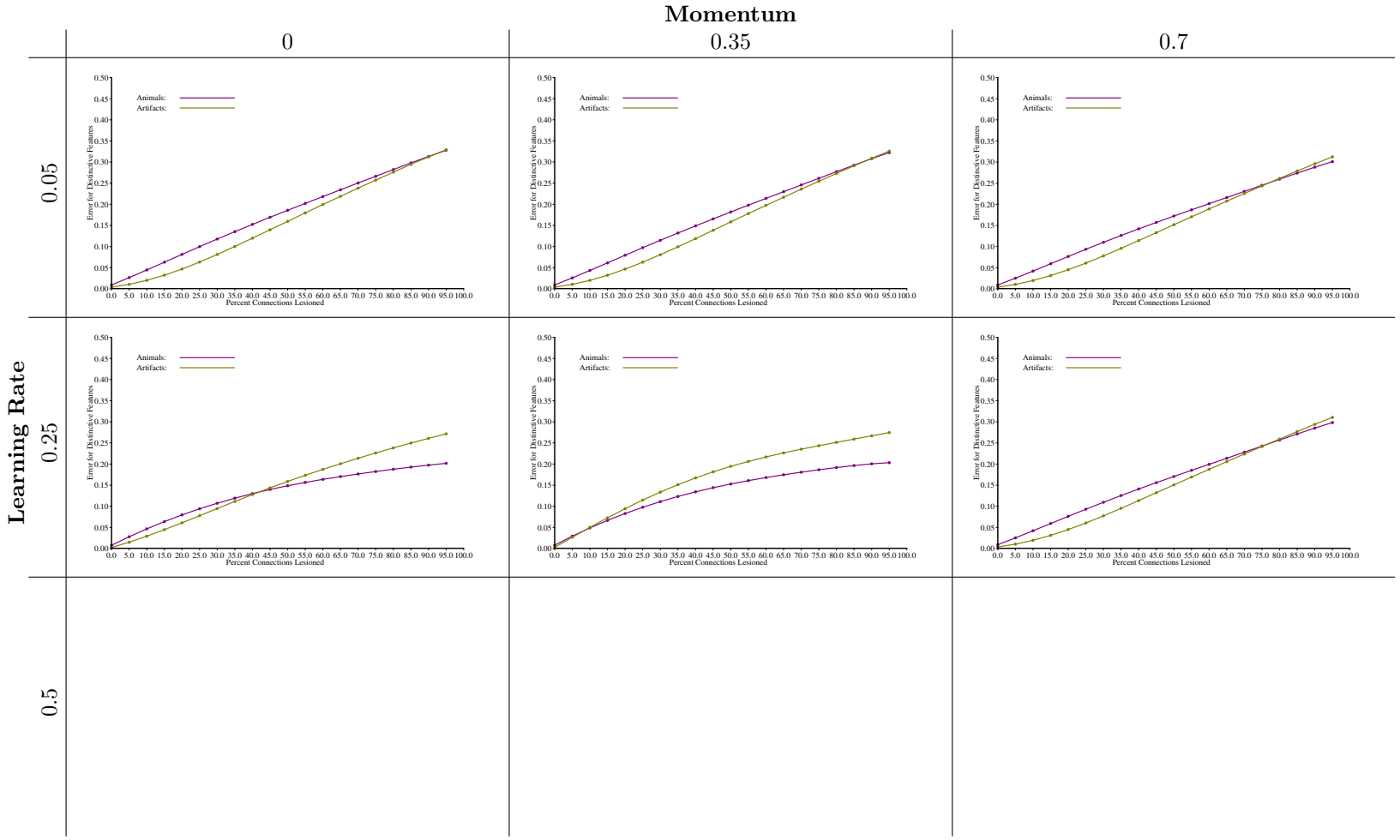


Table C.27: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 3.).

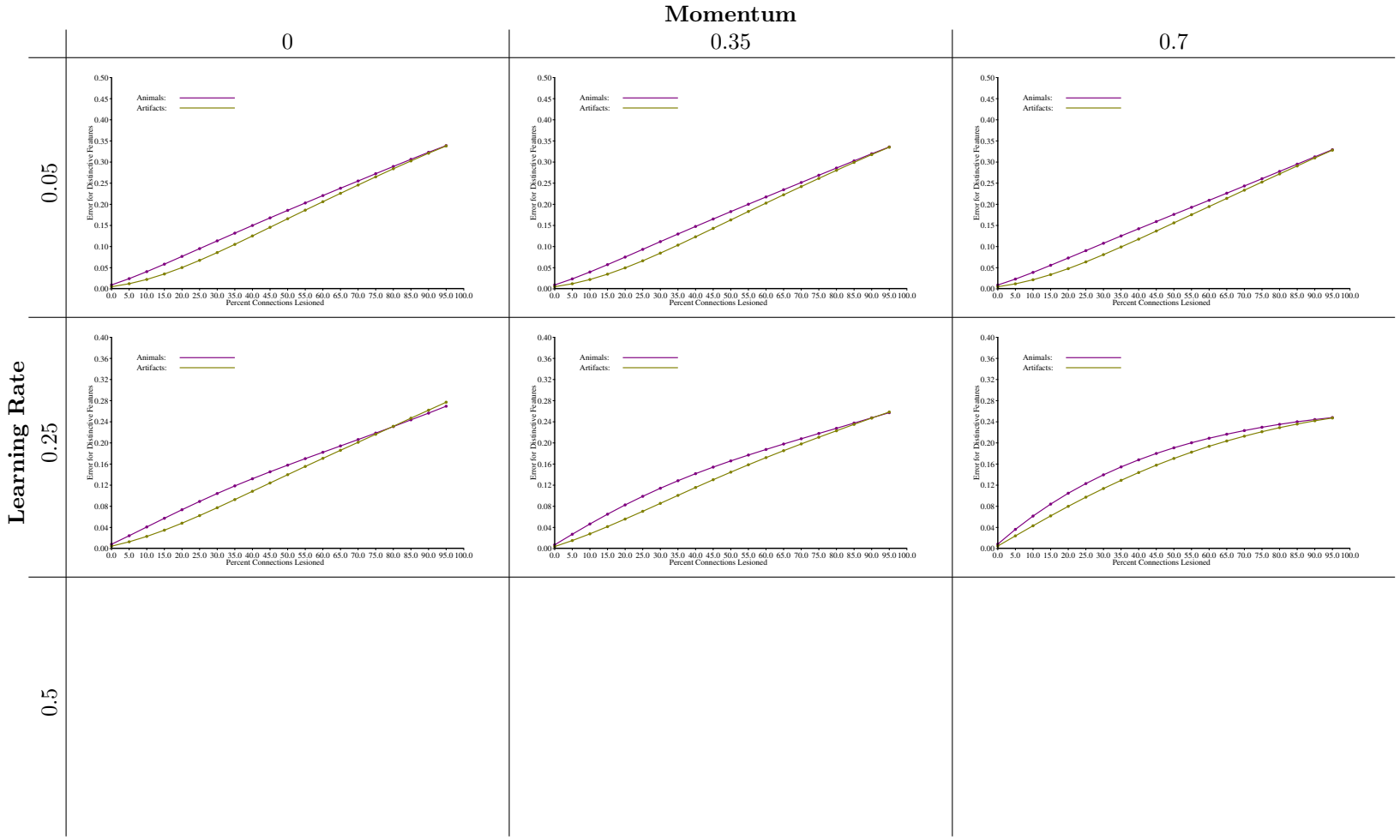


Table C.28: Mean absolute error for distinctive perceptual features of artefacts and living things at twenty levels of lesioning for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 3.).

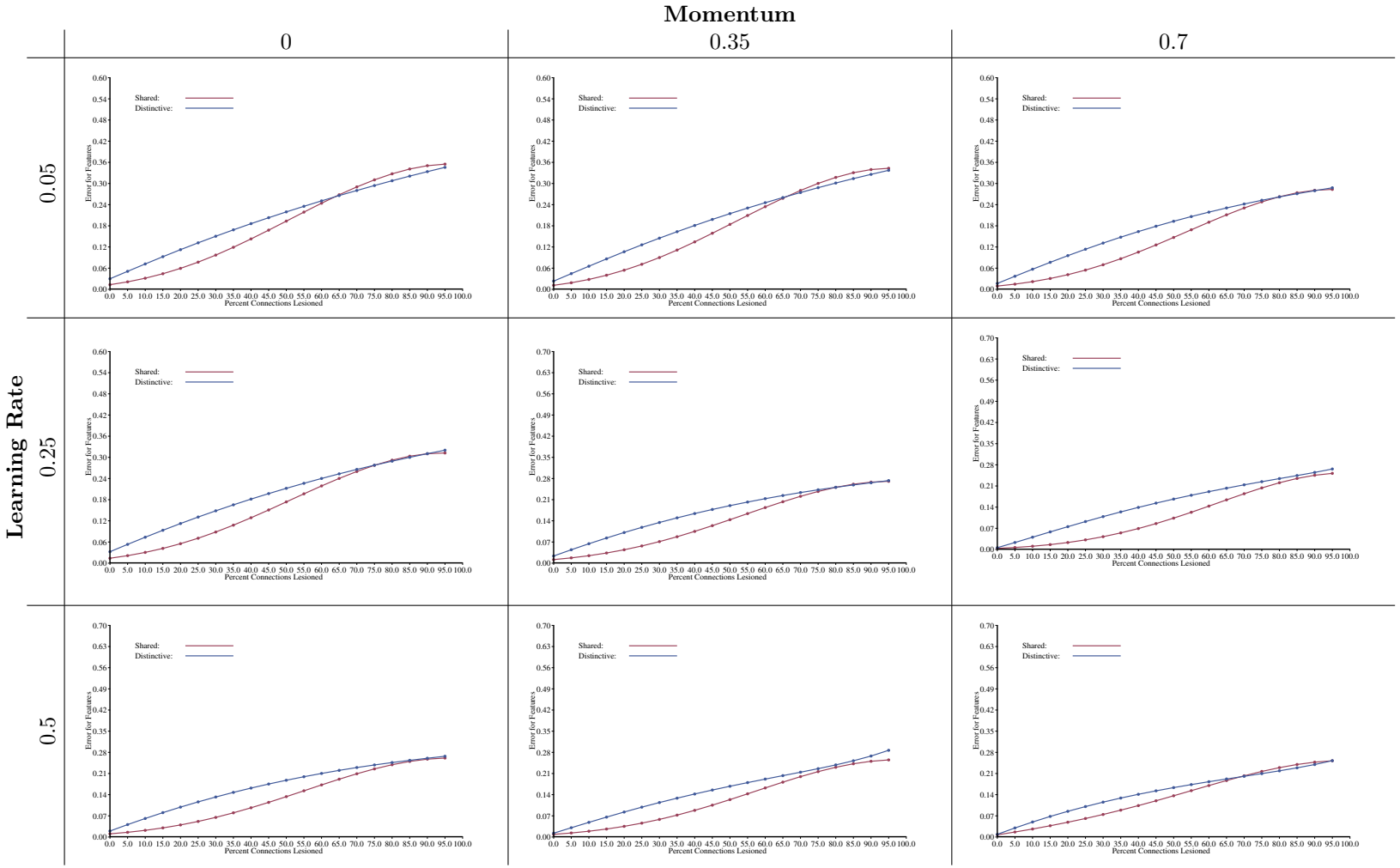


Table C.29: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 4.).

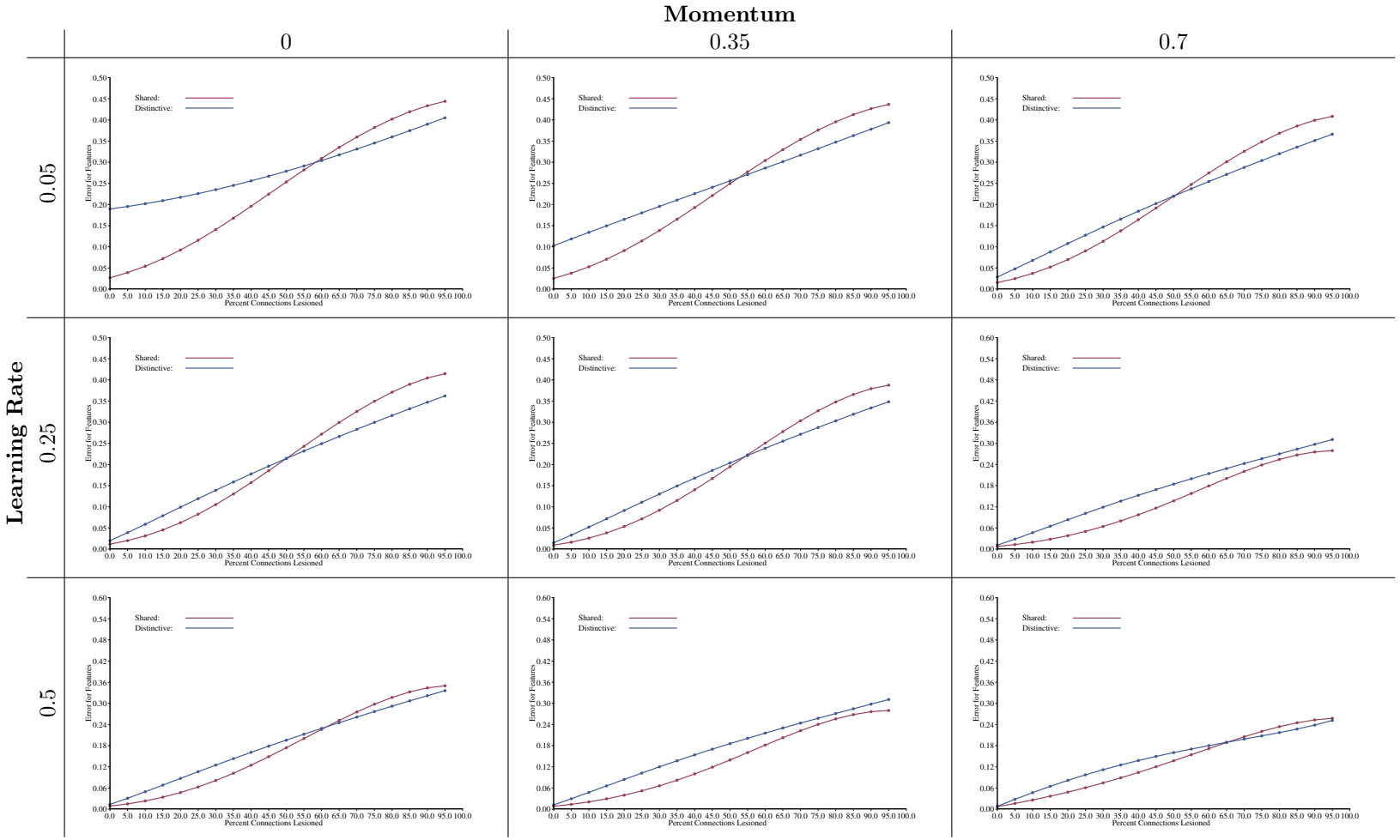


Table C.30: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 4.).

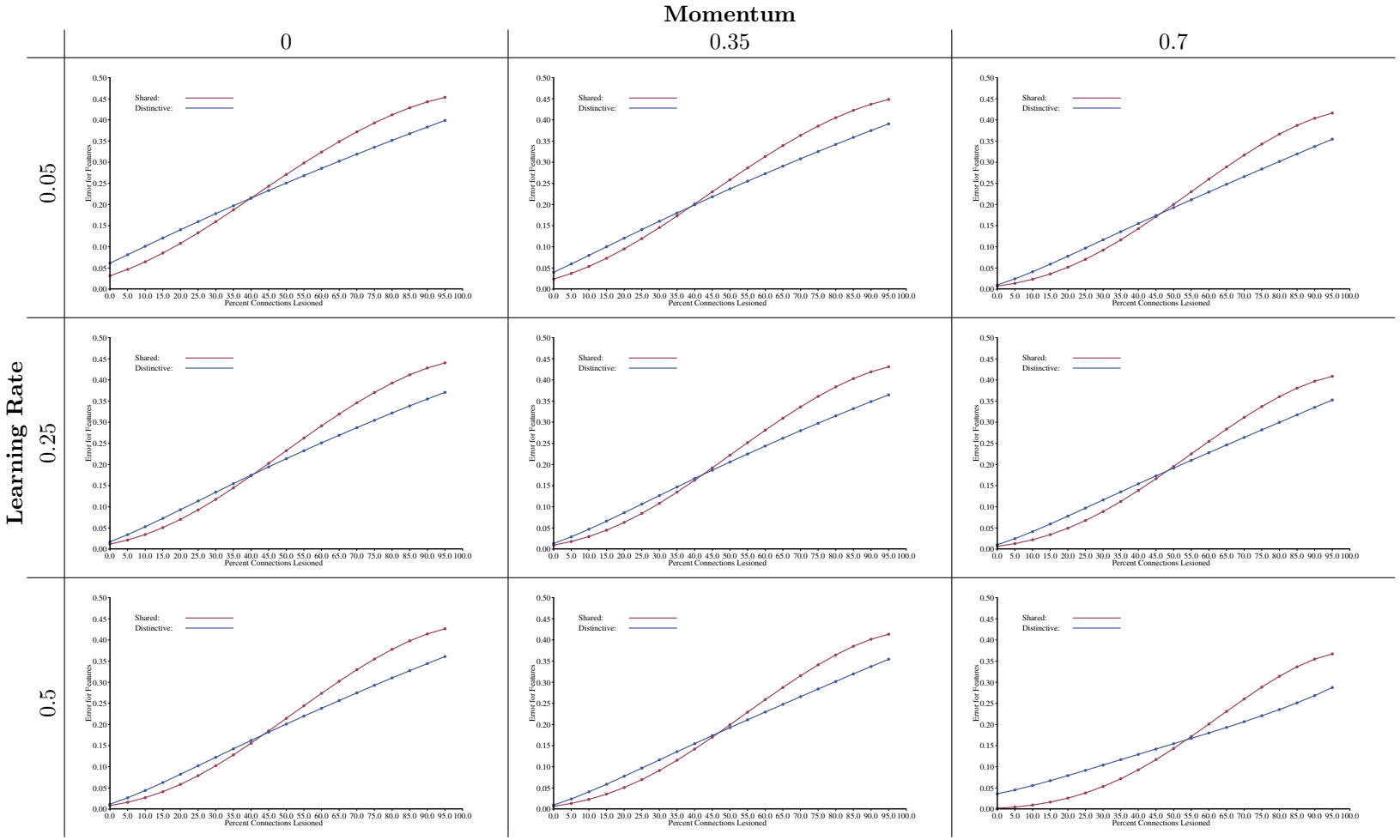


Table C.31: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 4.).

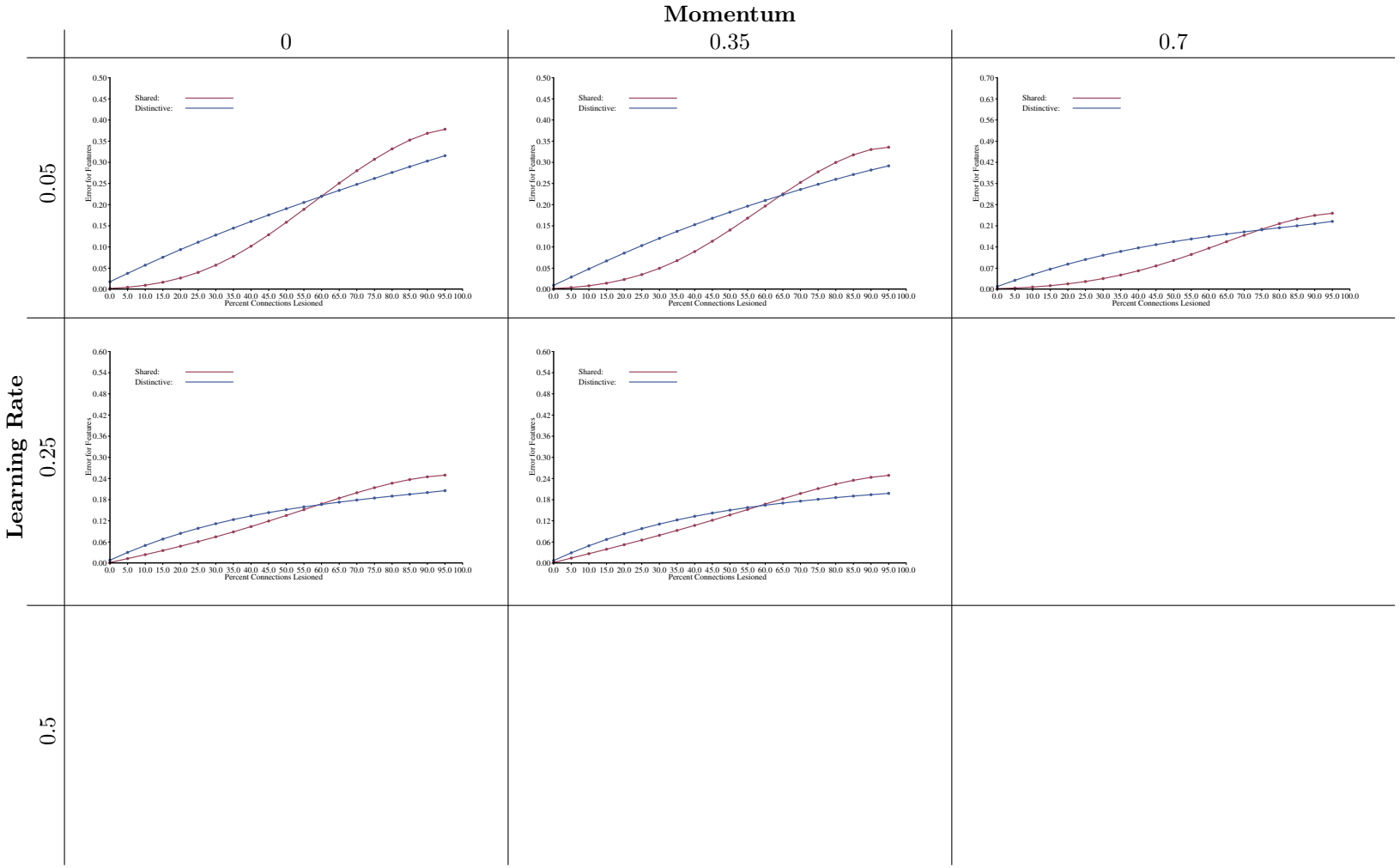


Table C.32: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 4.).

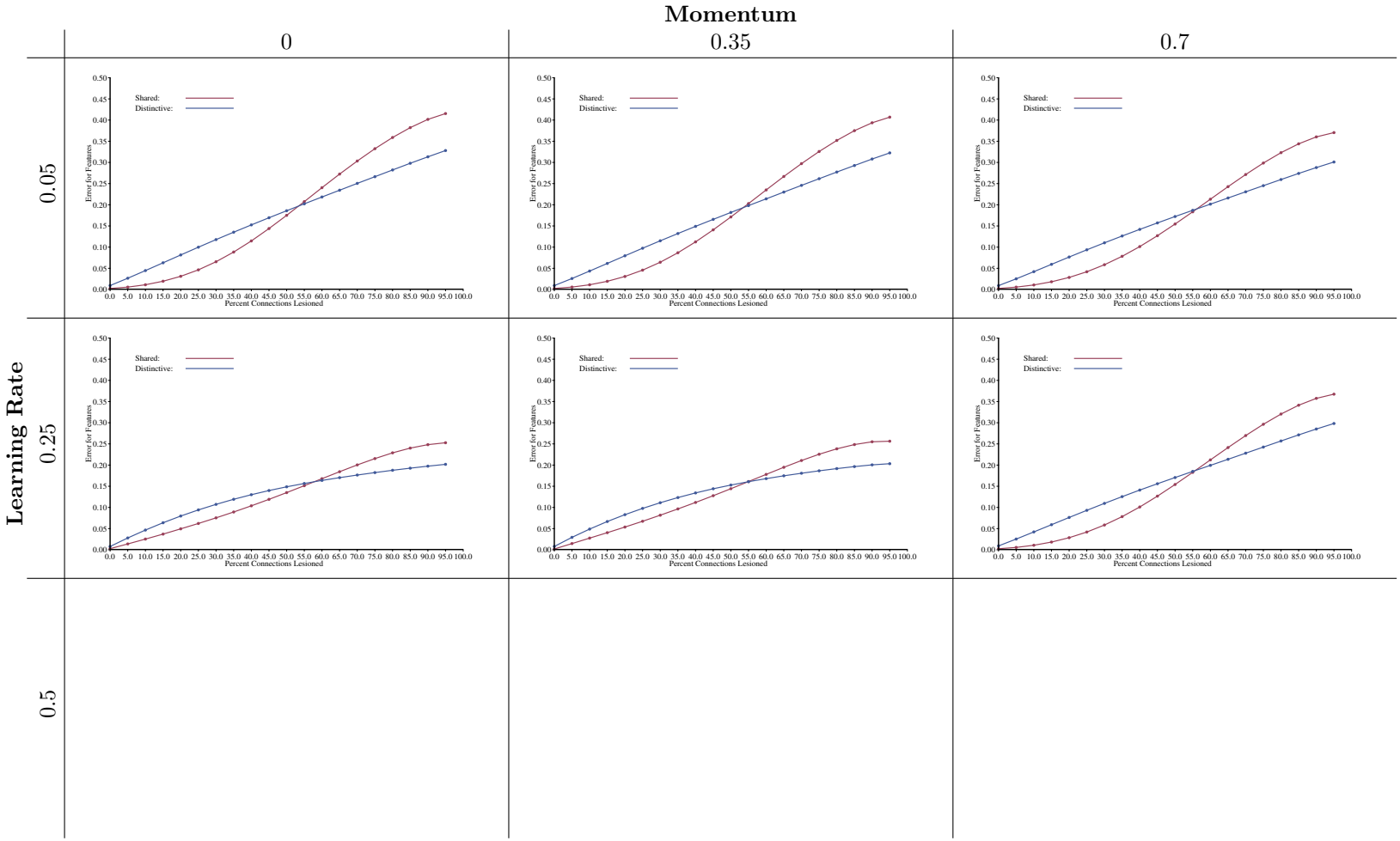


Table C.33: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 4.).

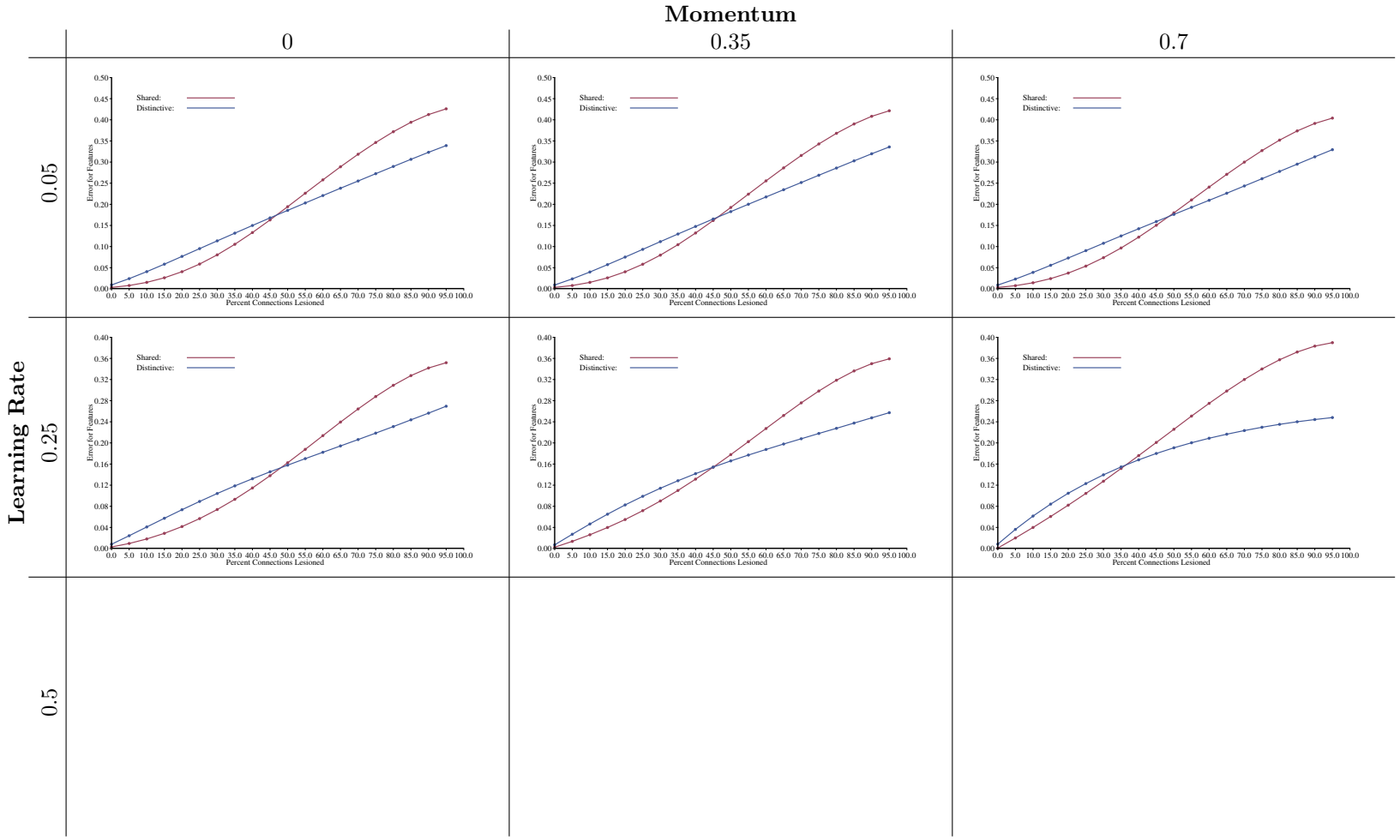


Table C.34: Mean absolute error for shared and distinctive perceptual features for living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 4.).

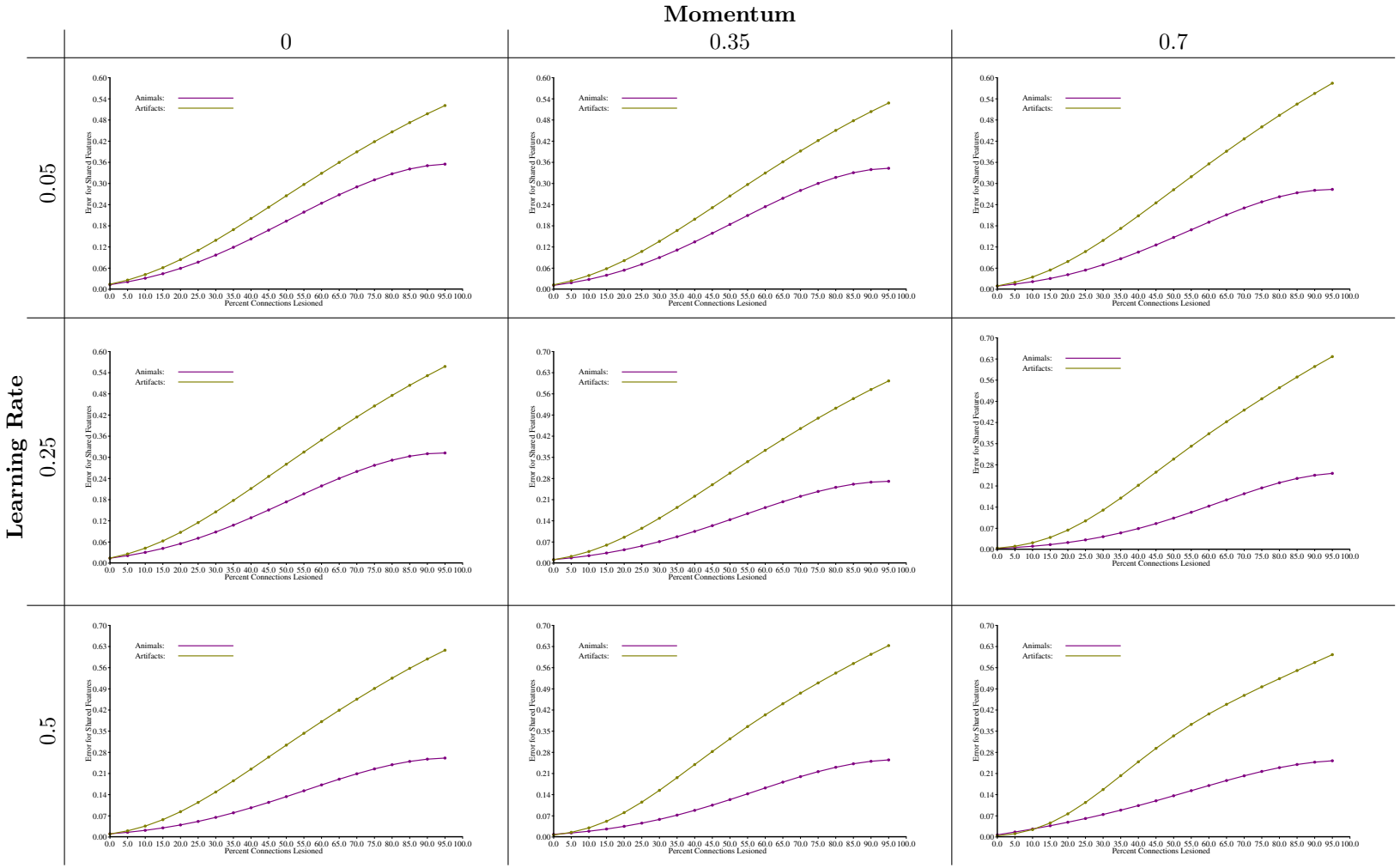


Table C.35: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 5.)

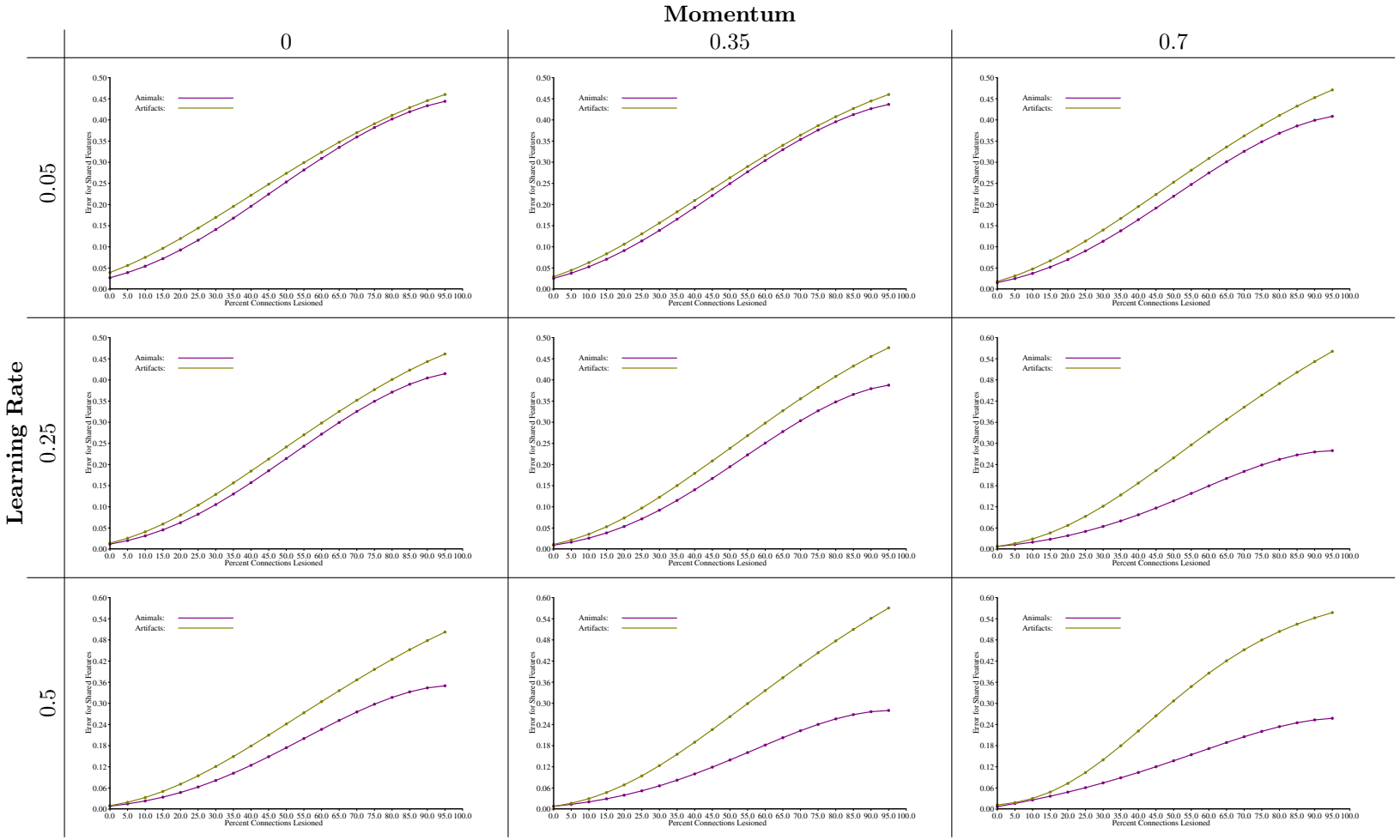


Table C.36: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 5.)

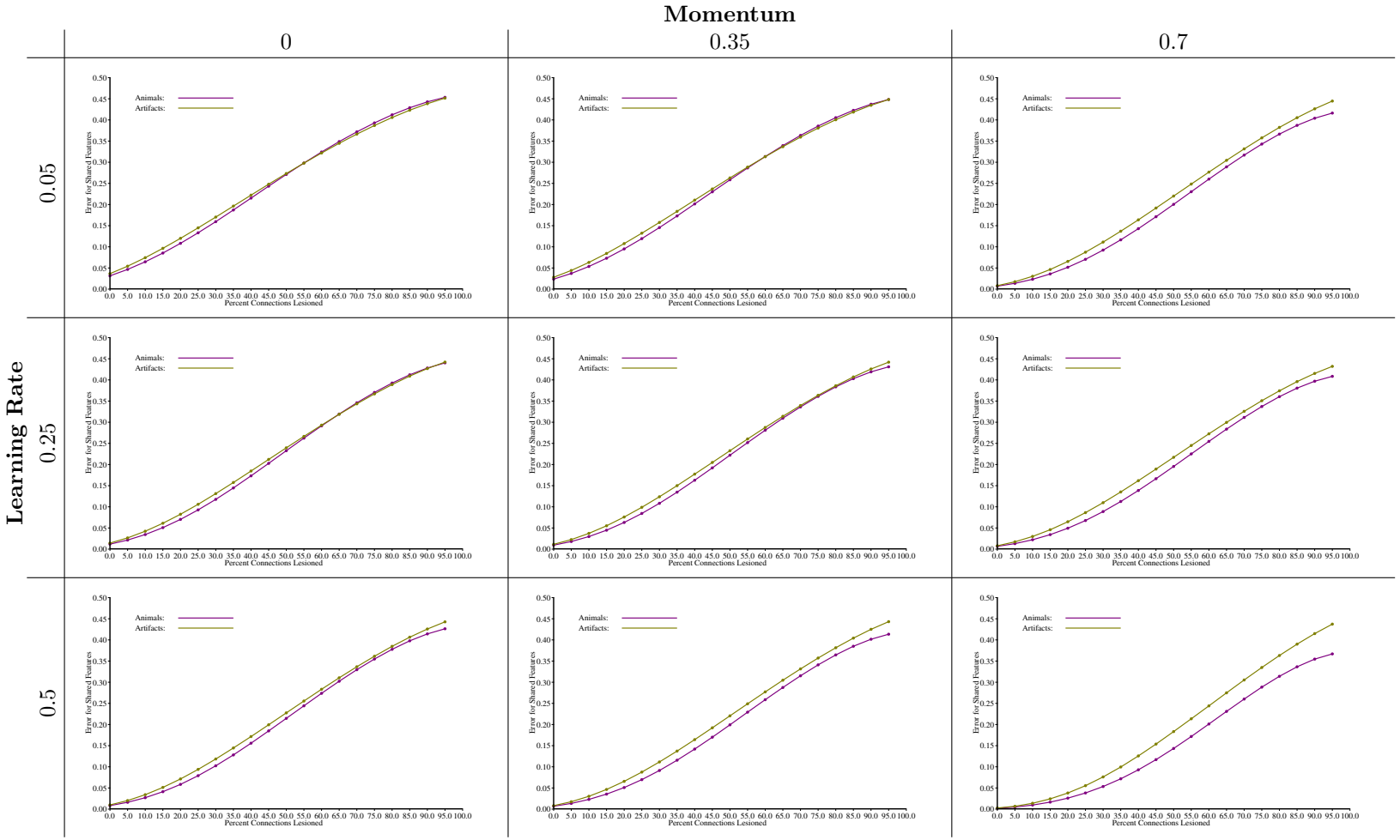


Table C.37: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 5.)

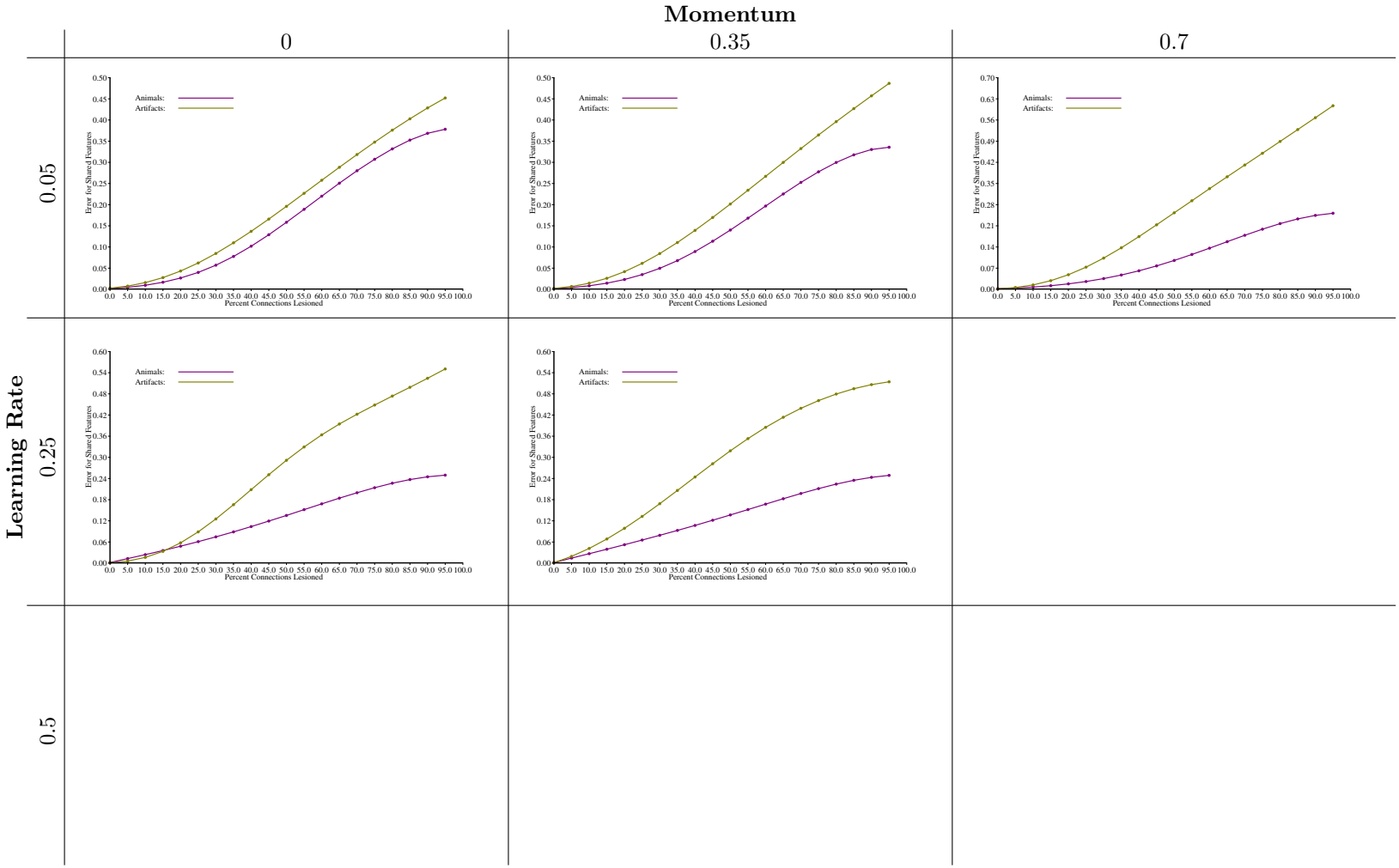


Table C.38: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 5.)

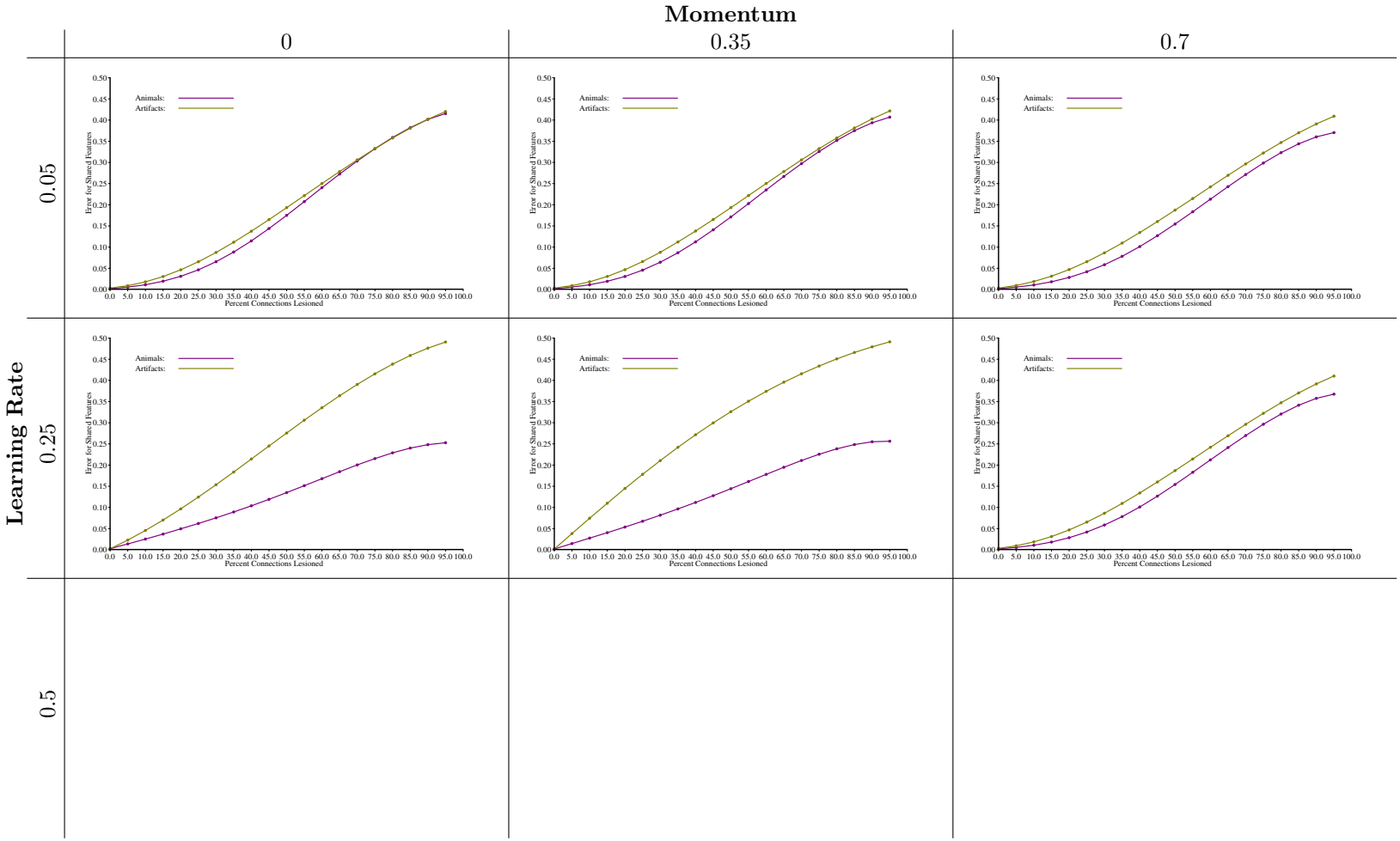


Table C.39: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 5.)

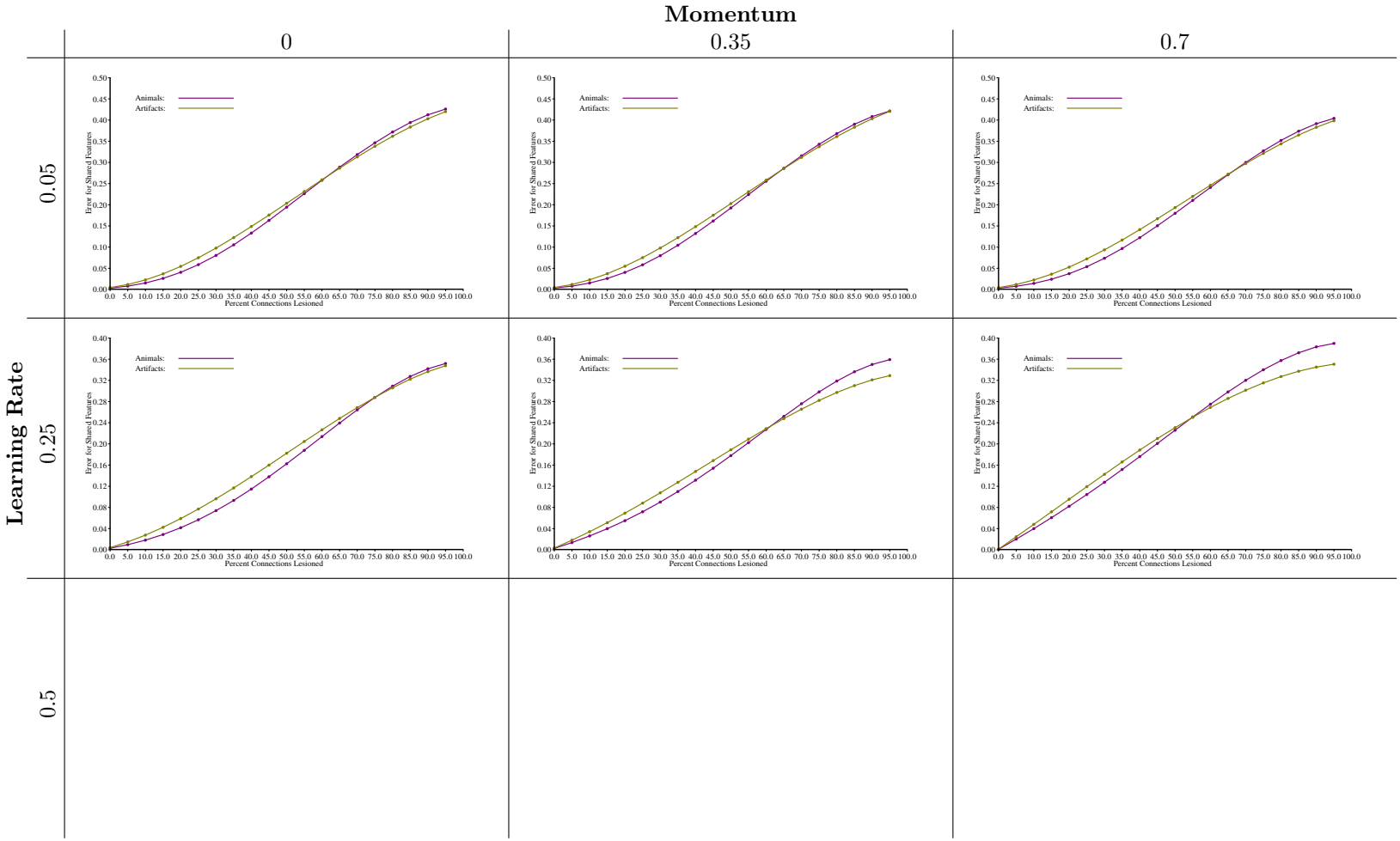


Table C.40: Mean absolute error for shared perceptual features for artefacts and living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 5.)

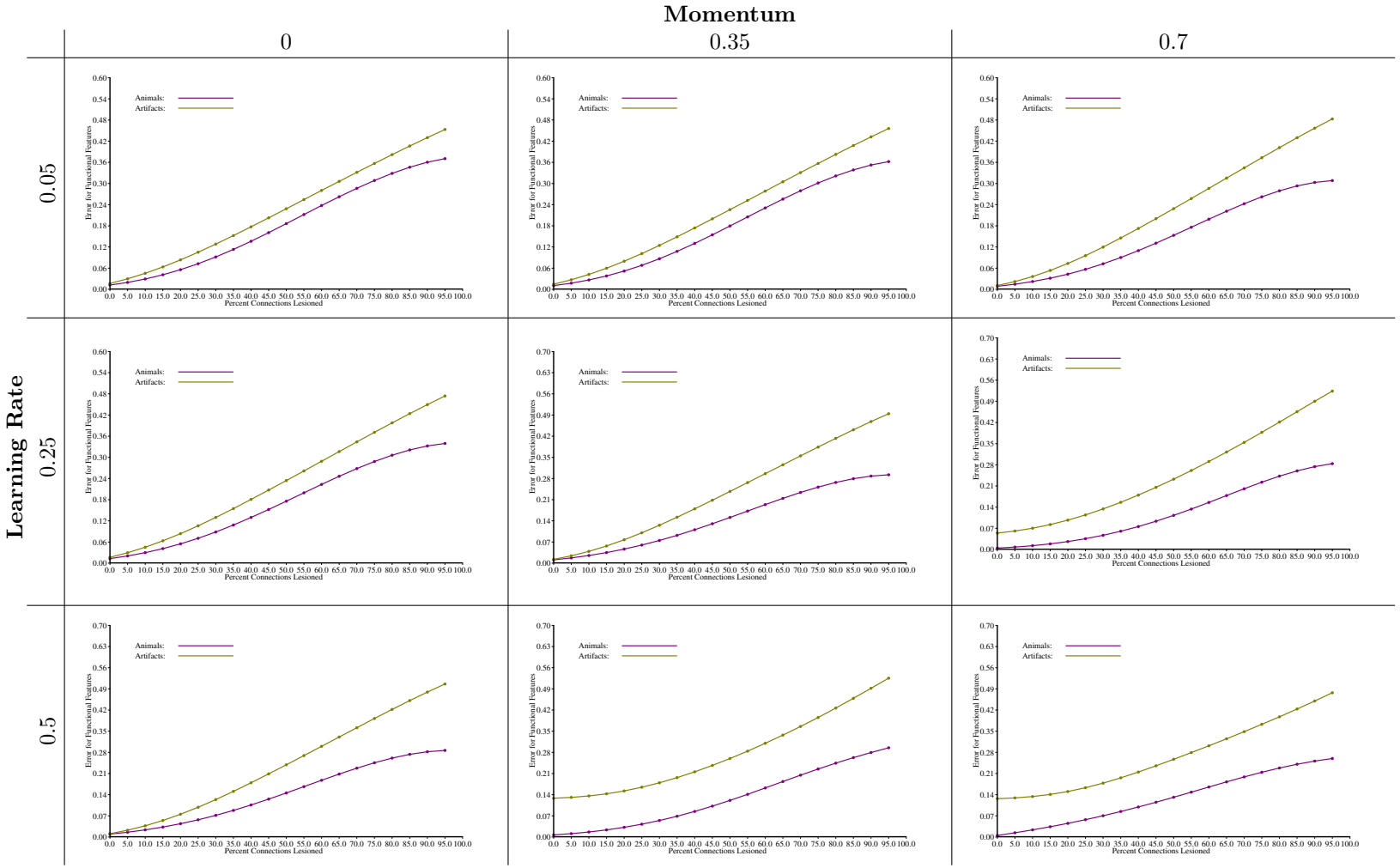


Table C.41: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 6.)

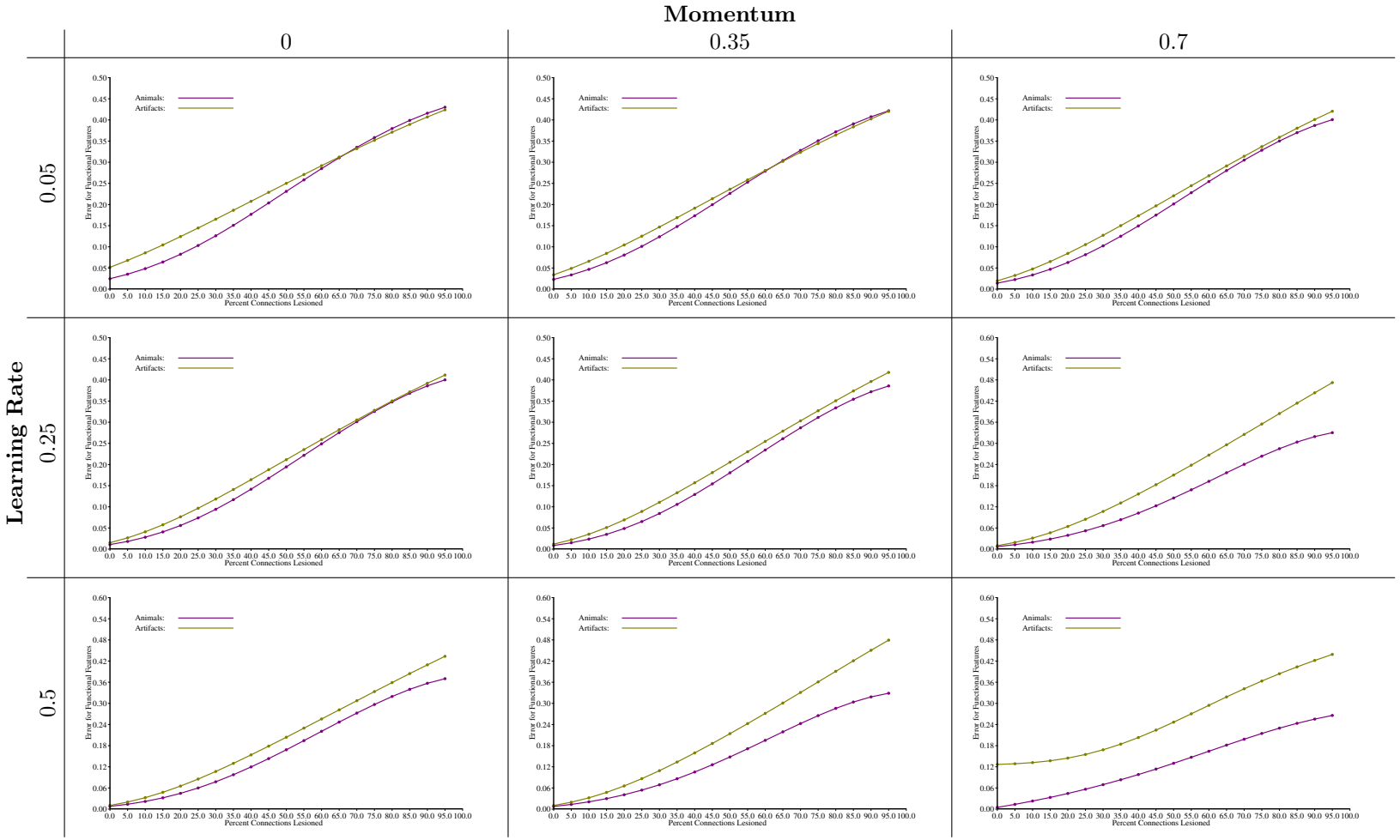


Table C.42: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 6.)

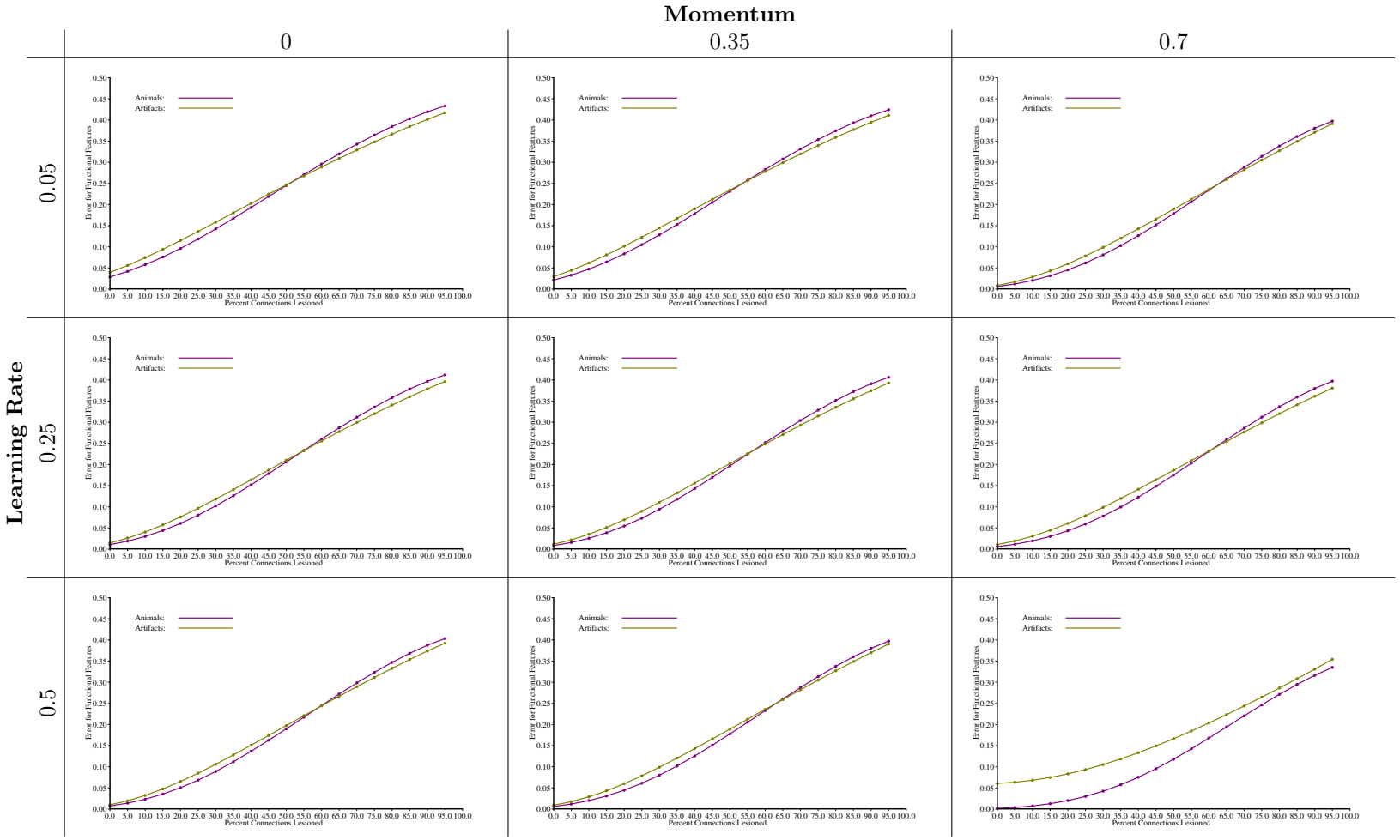


Table C.43: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 6.)

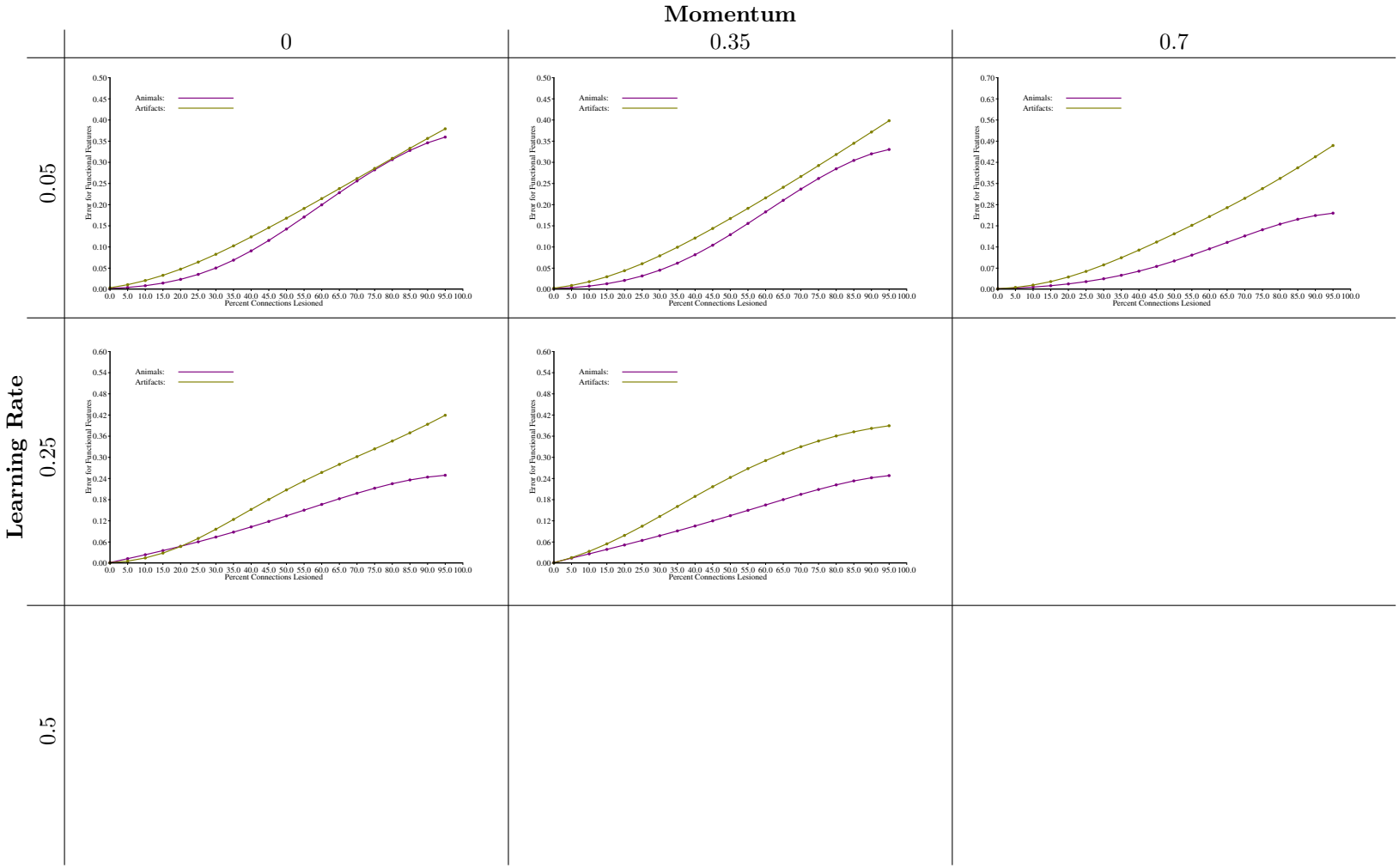


Table C.44: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 6.)

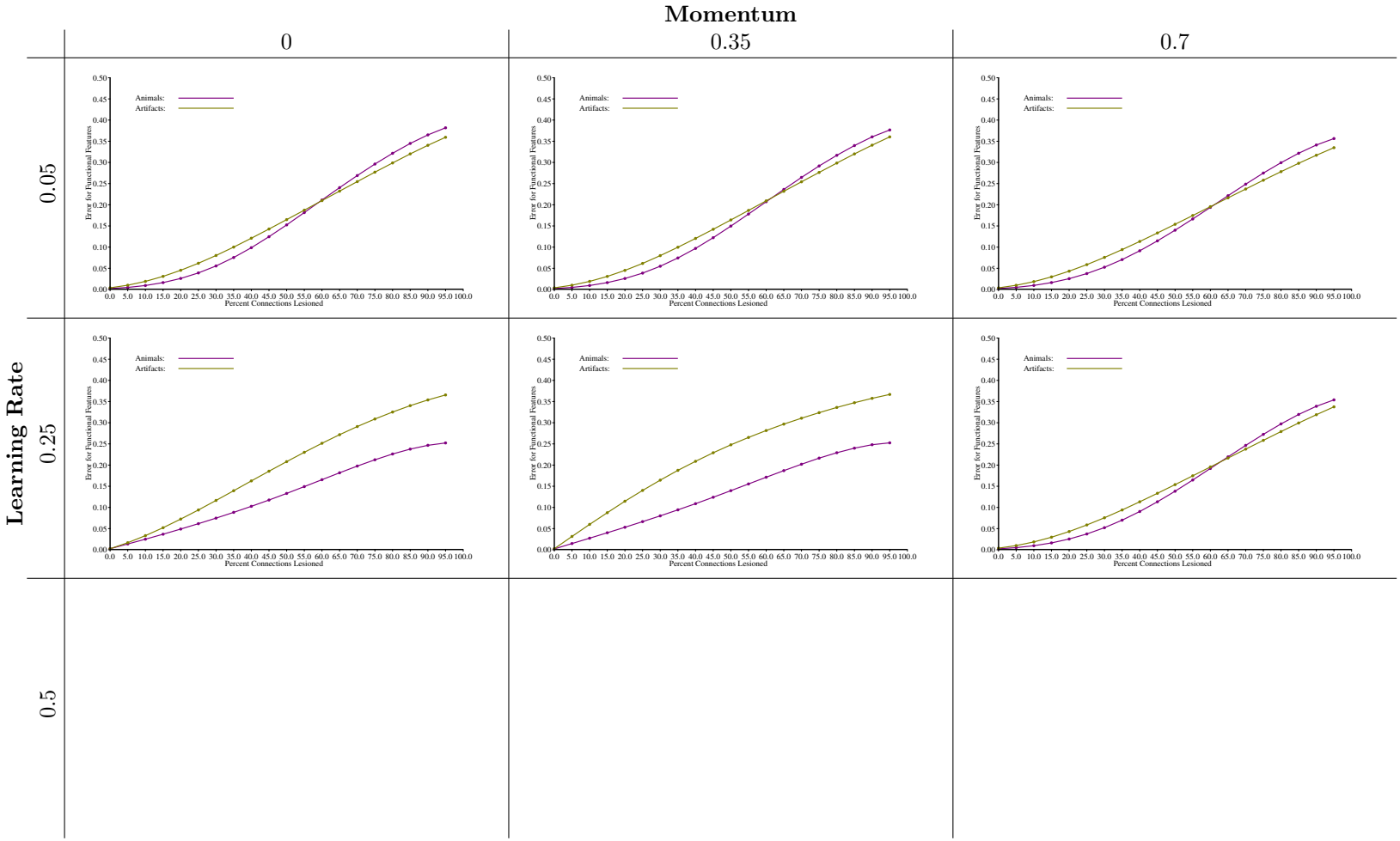


Table C.45: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 6.)

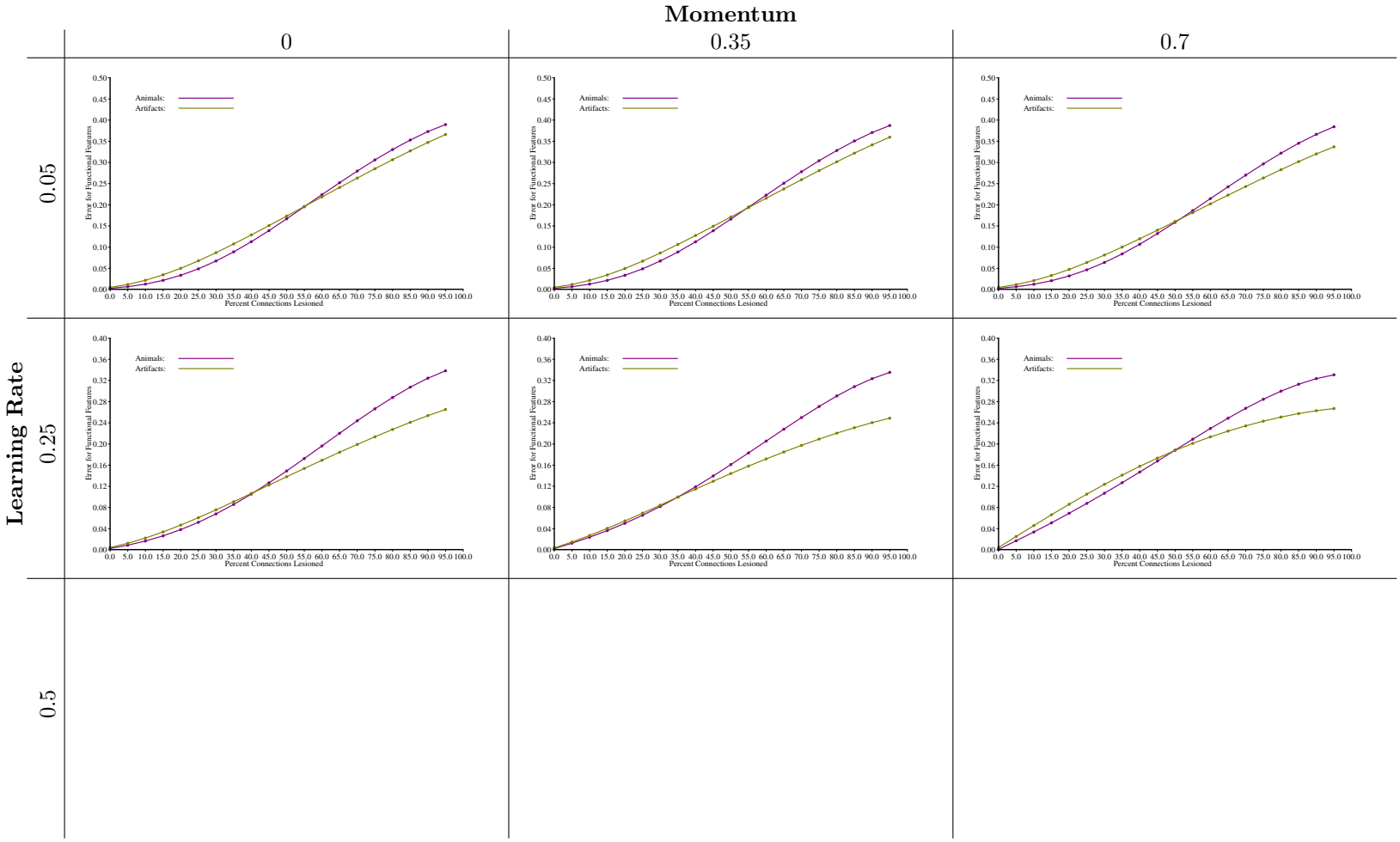


Table C.46: Mean absolute error for functional features for artefacts and living things for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 6.)

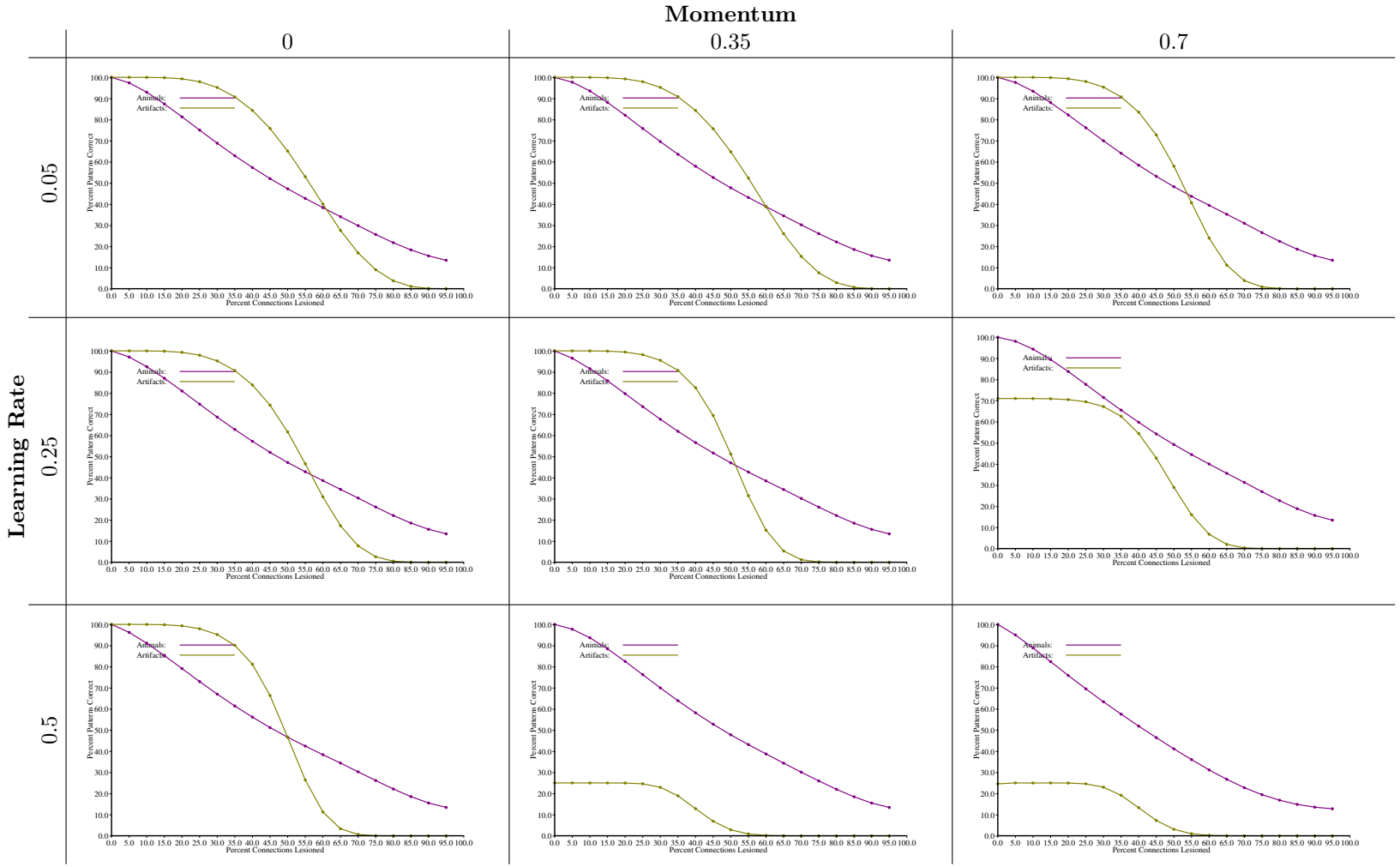


Table C.47: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 7.)

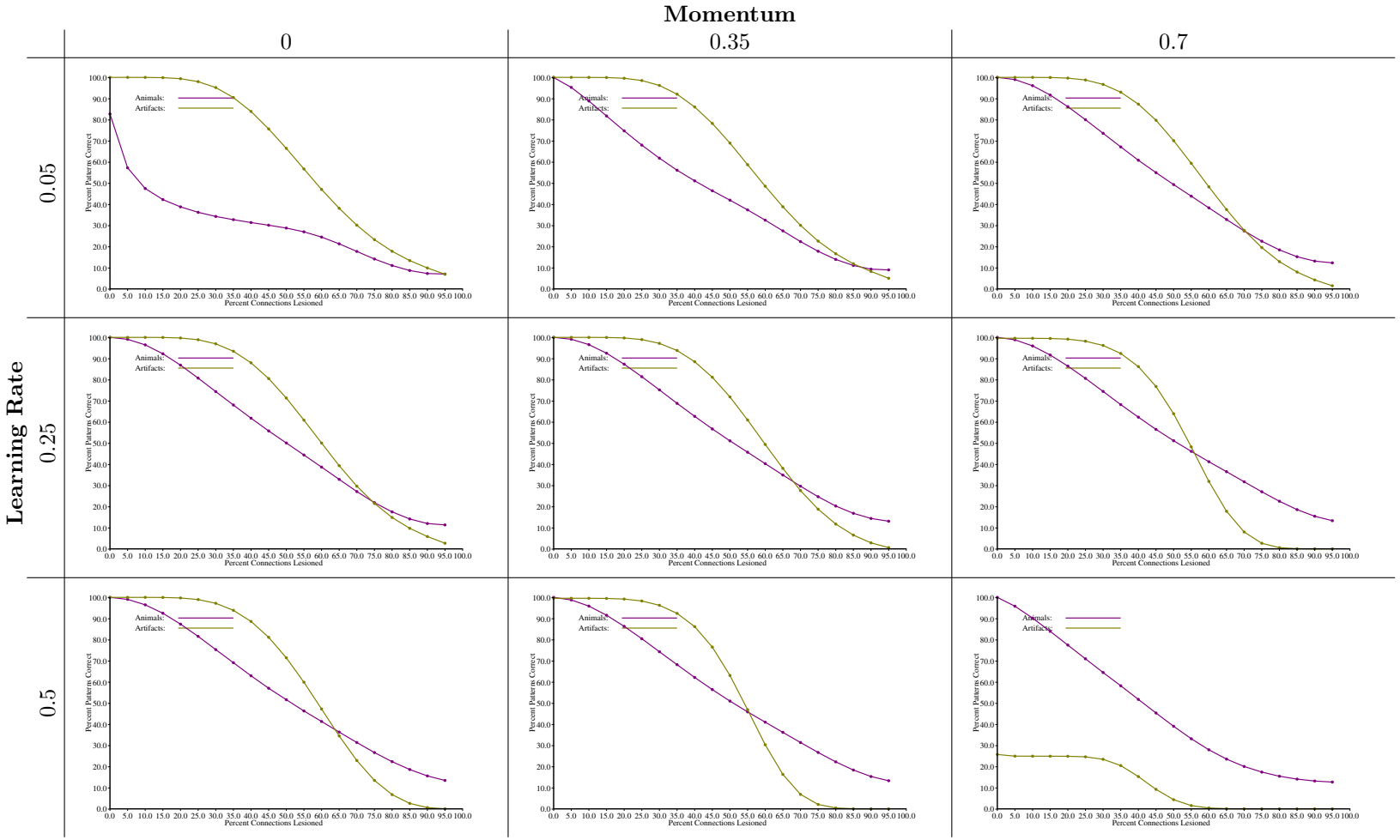


Table C.48: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 7.)

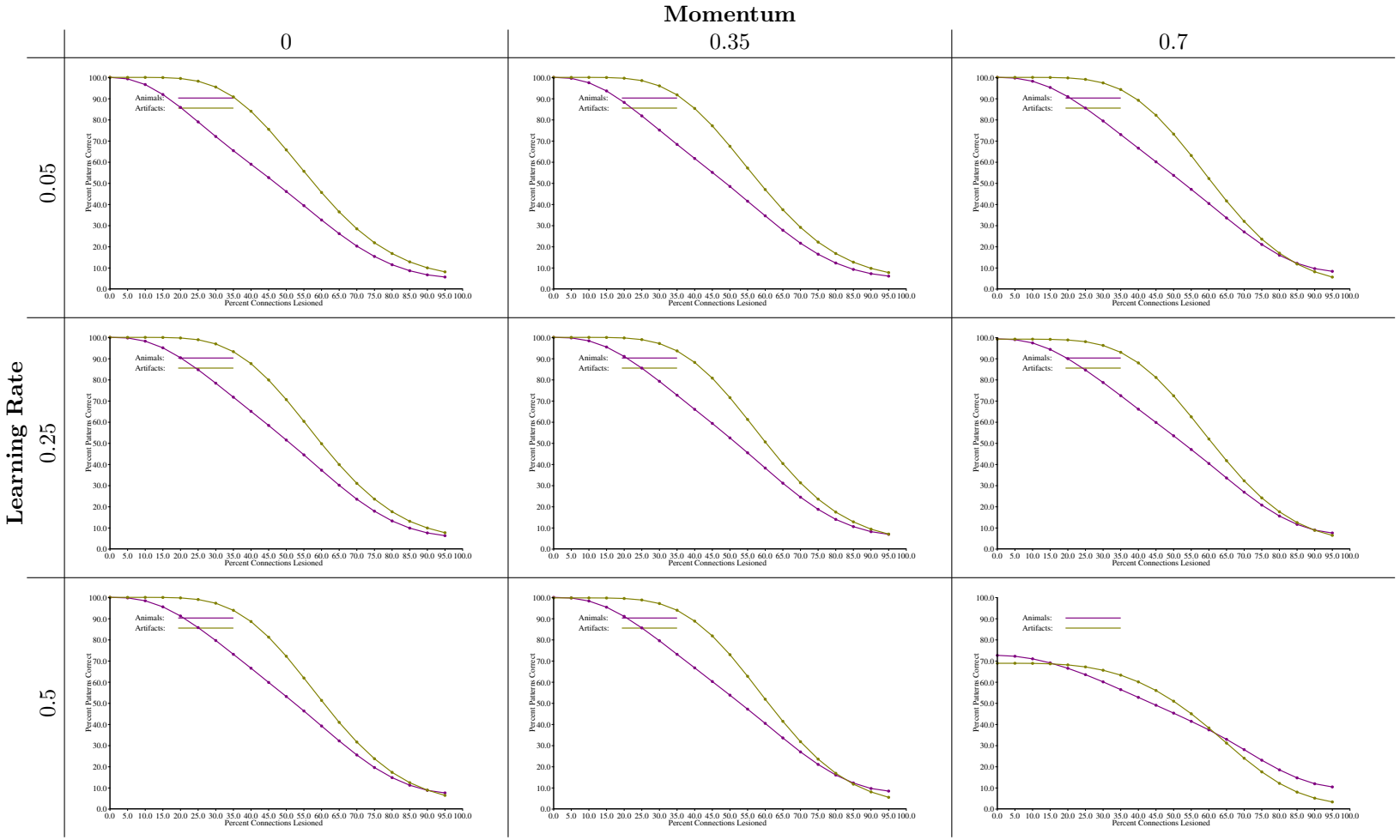


Table C.49: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 7.)

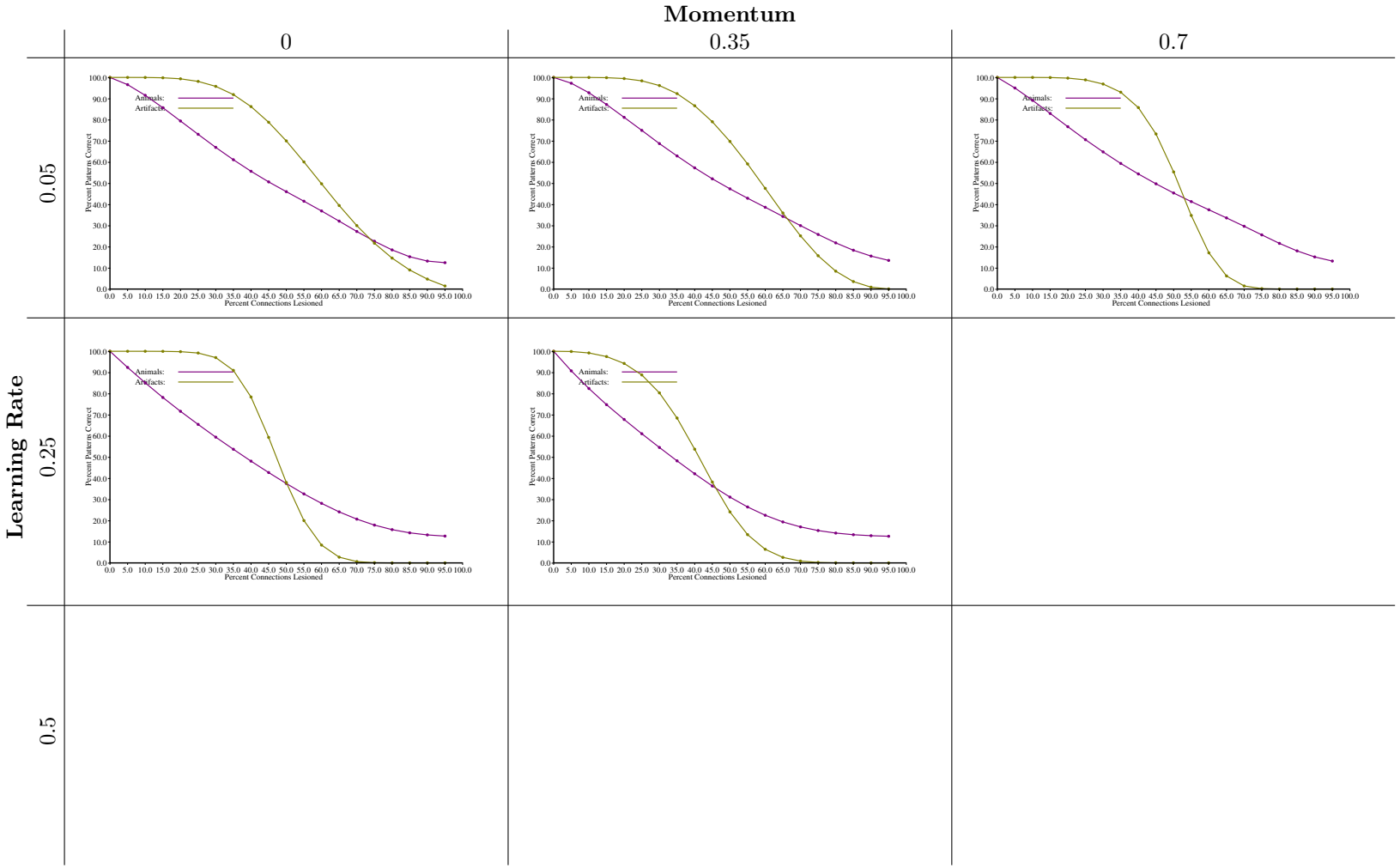


Table C.50: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 7.)

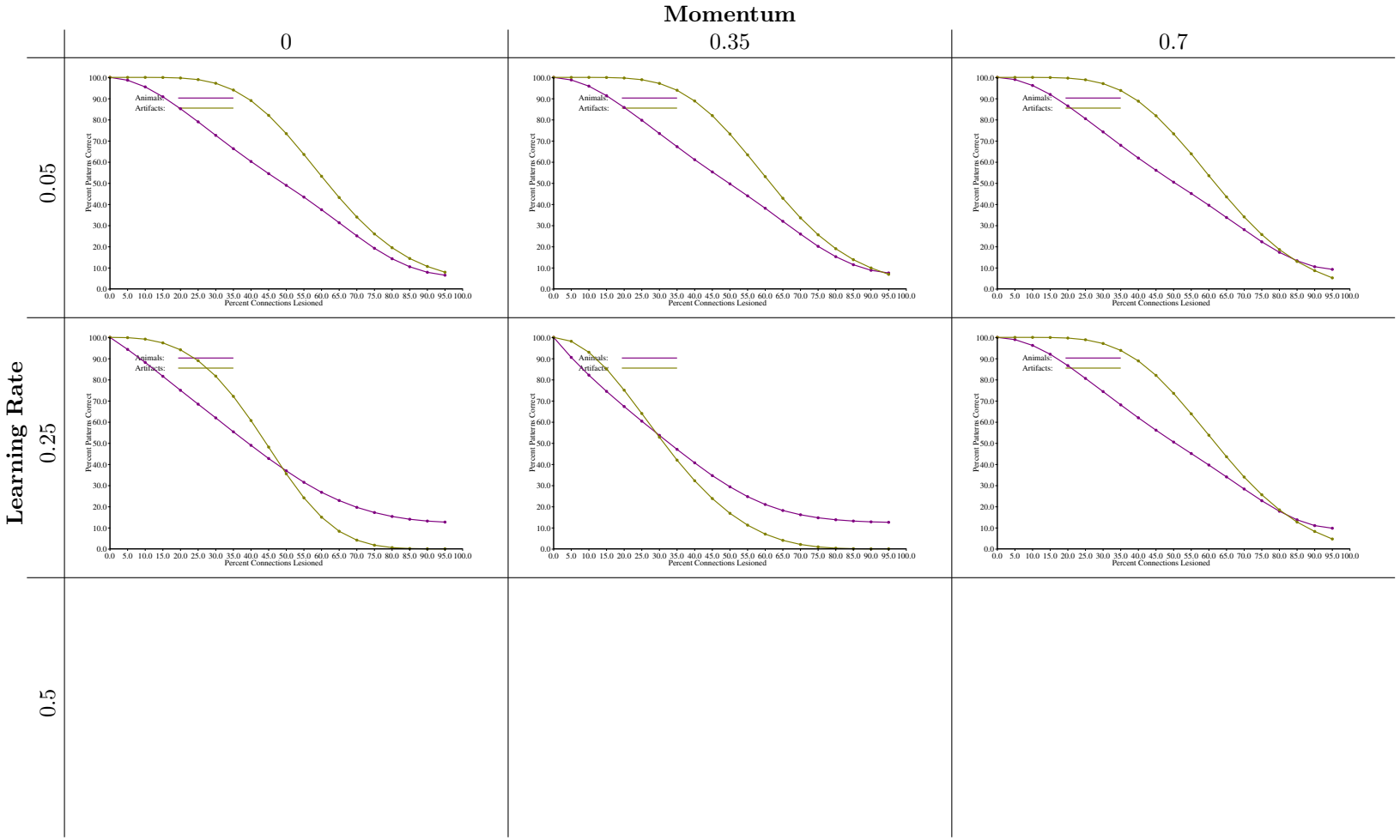


Table C.51: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 7.)

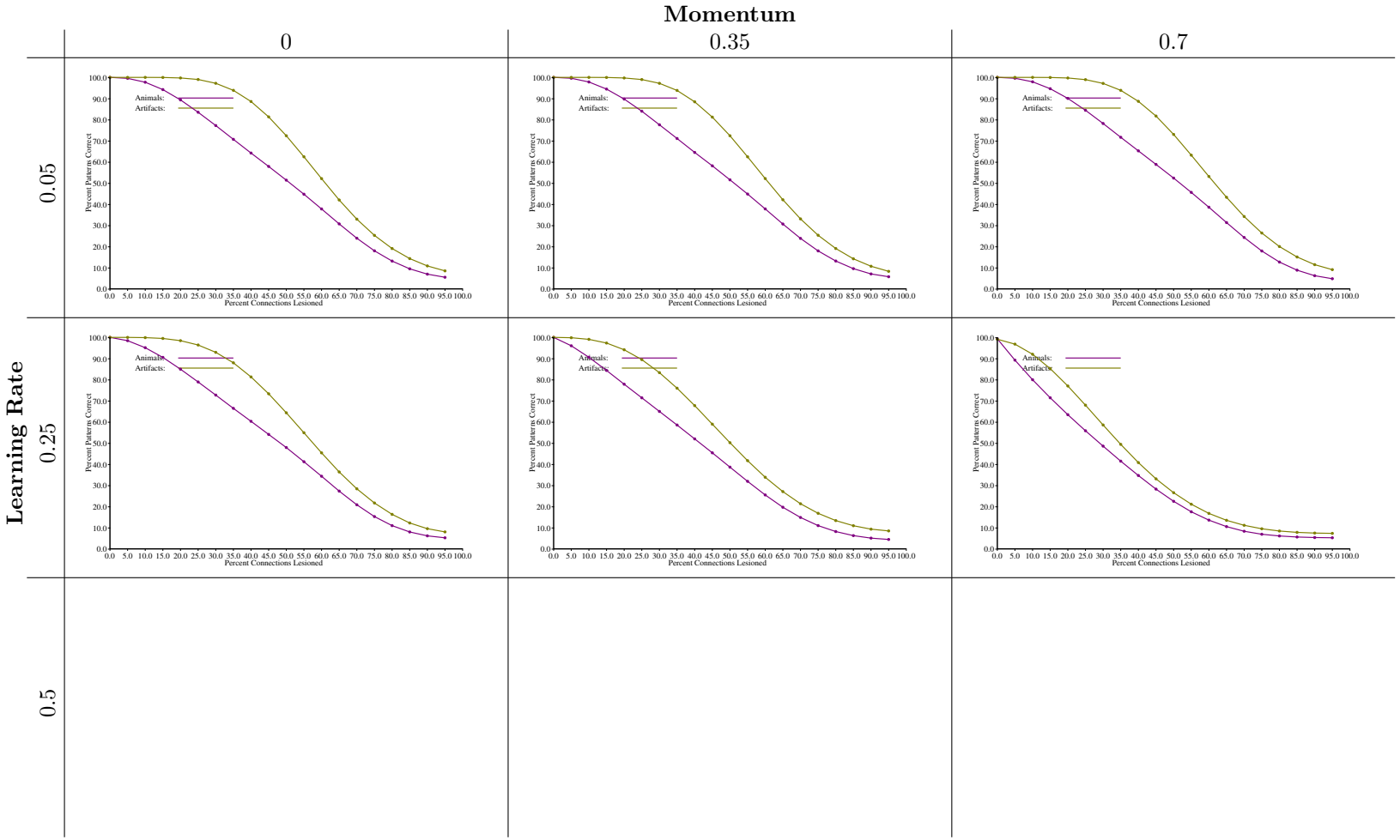


Table C.52: Percentage of patterns correctly mapped for the two domains for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 7.)

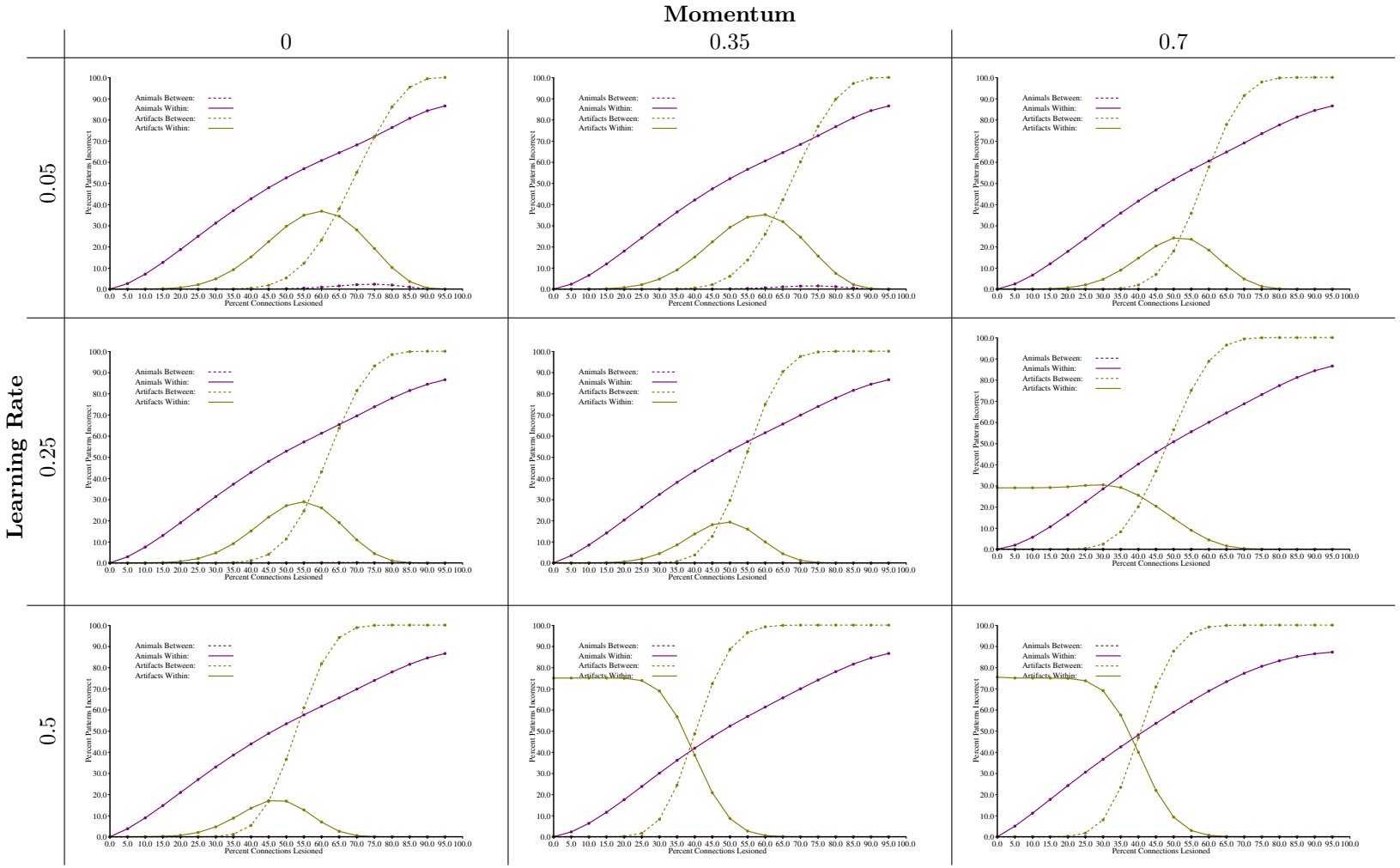


Table C.53: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 8.)

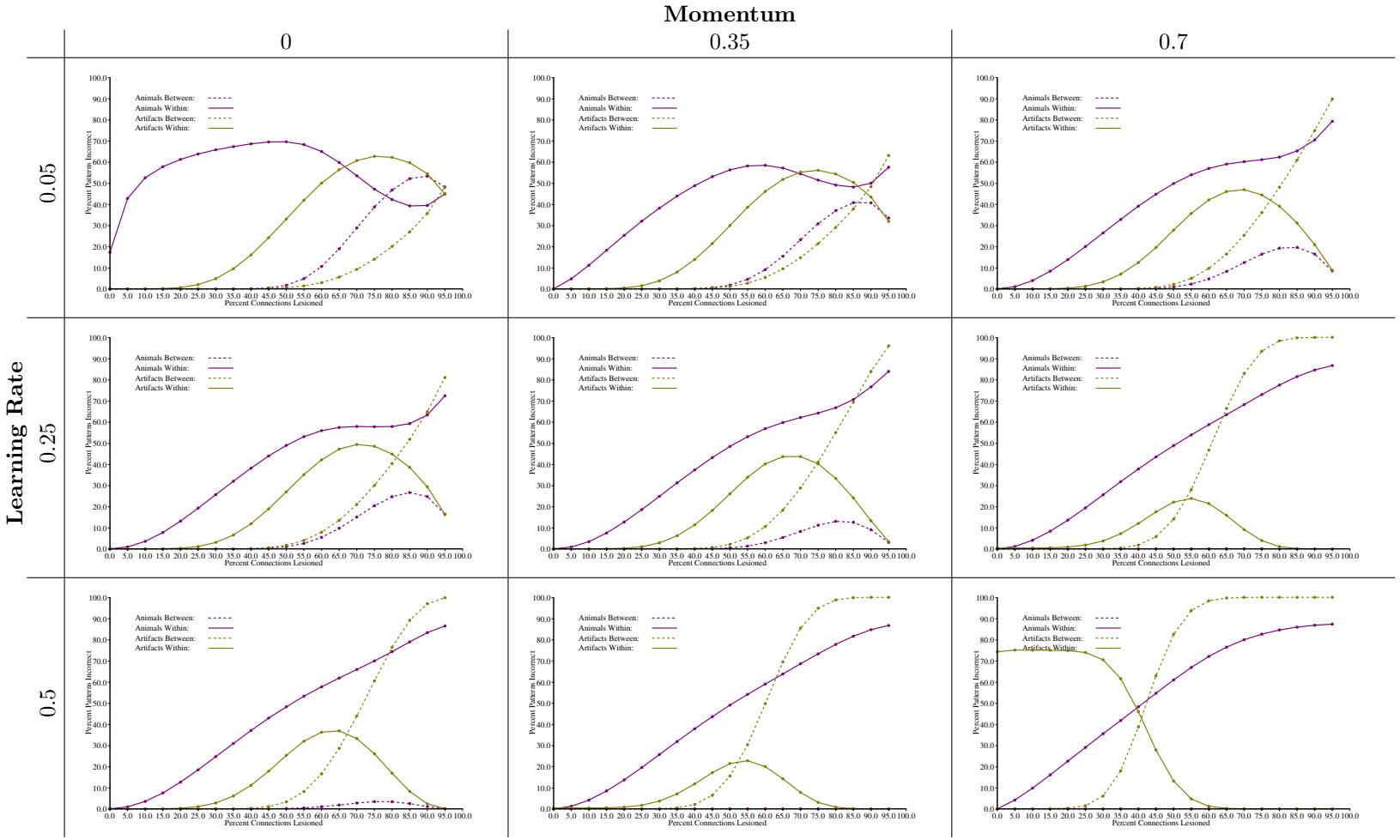


Table C.54: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 8.)

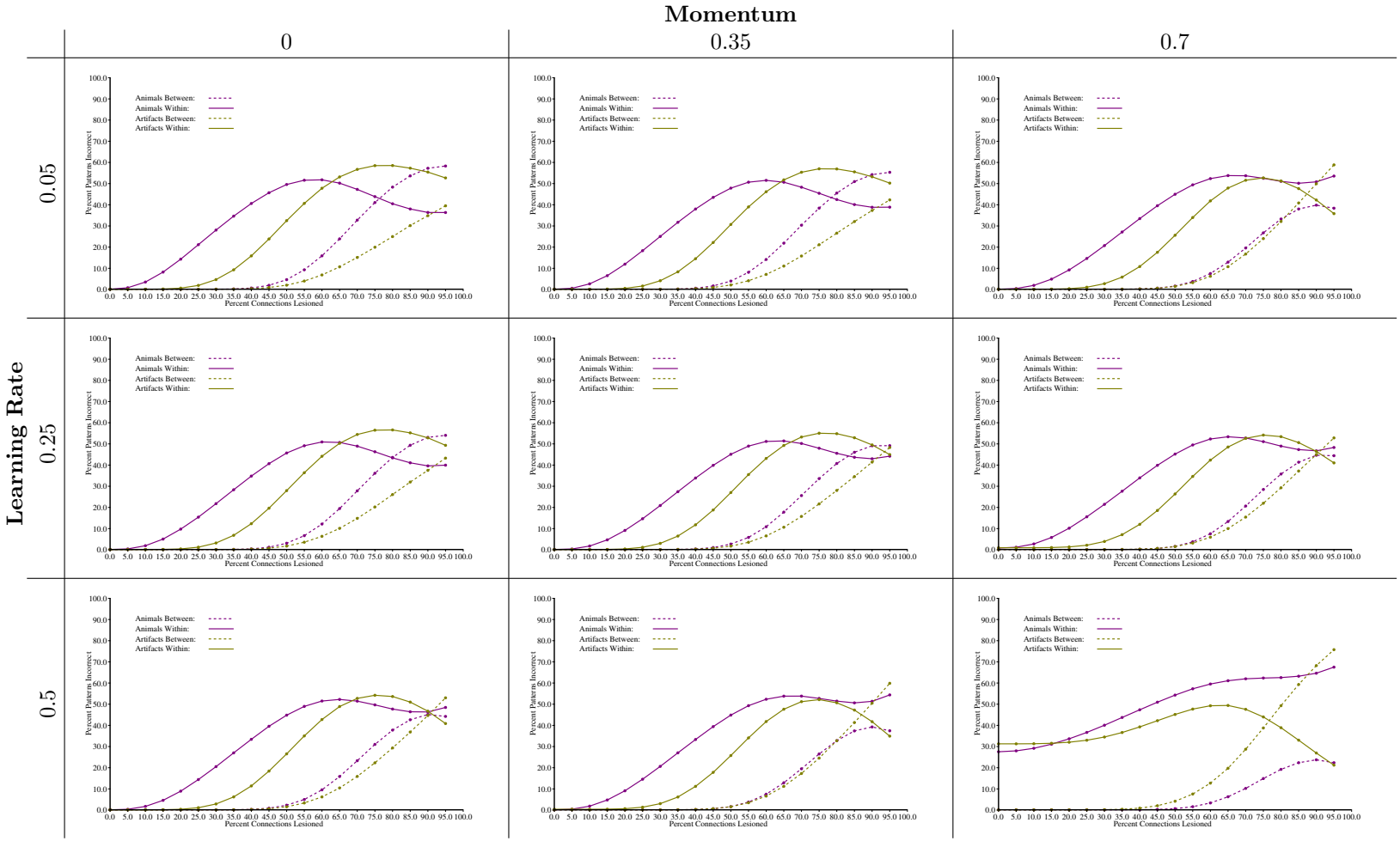


Table C.55: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using SSE (compare with Tyler et al., 2000, fig. 8.)

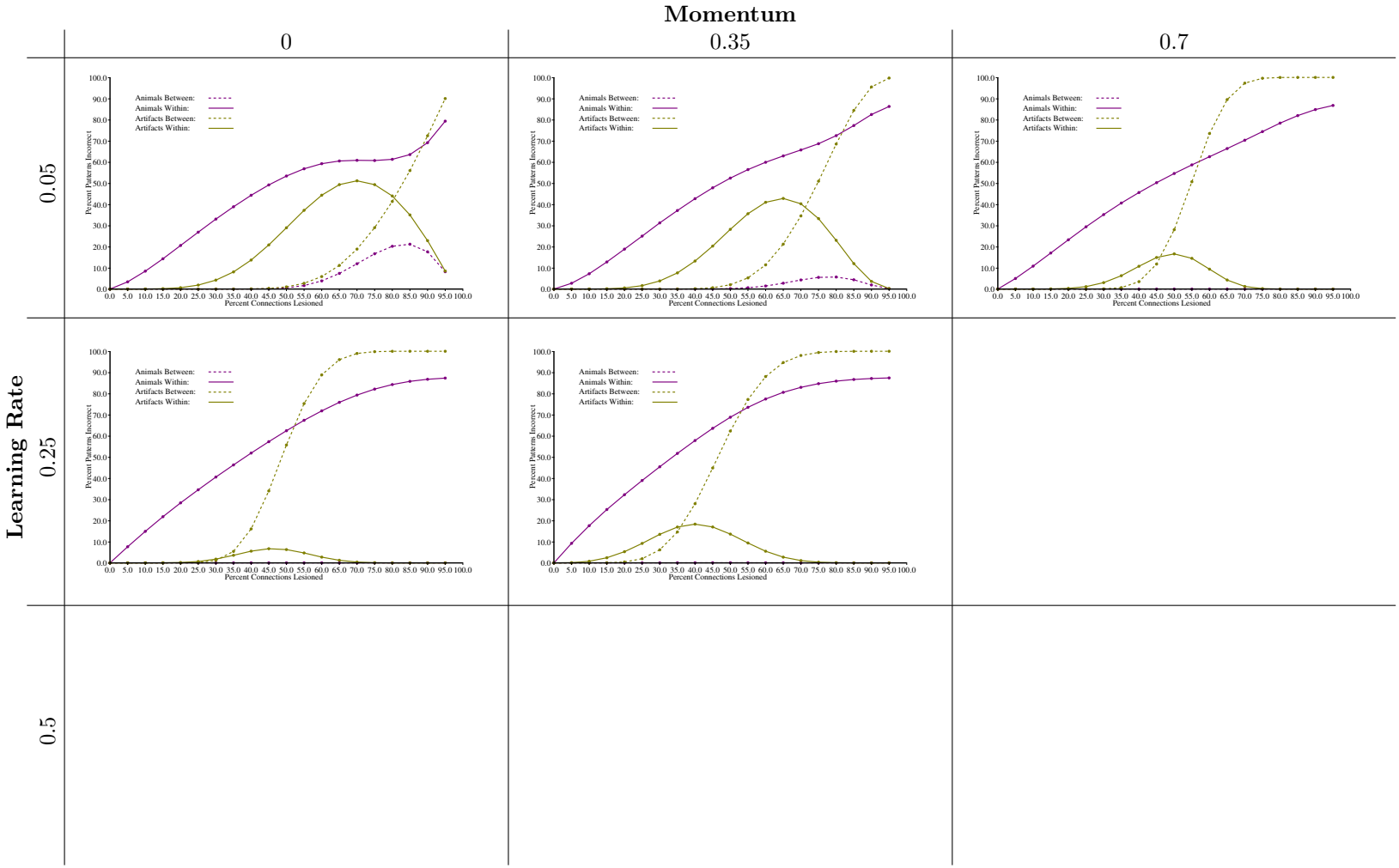


Table C.56: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.001, 0.001] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 8.)

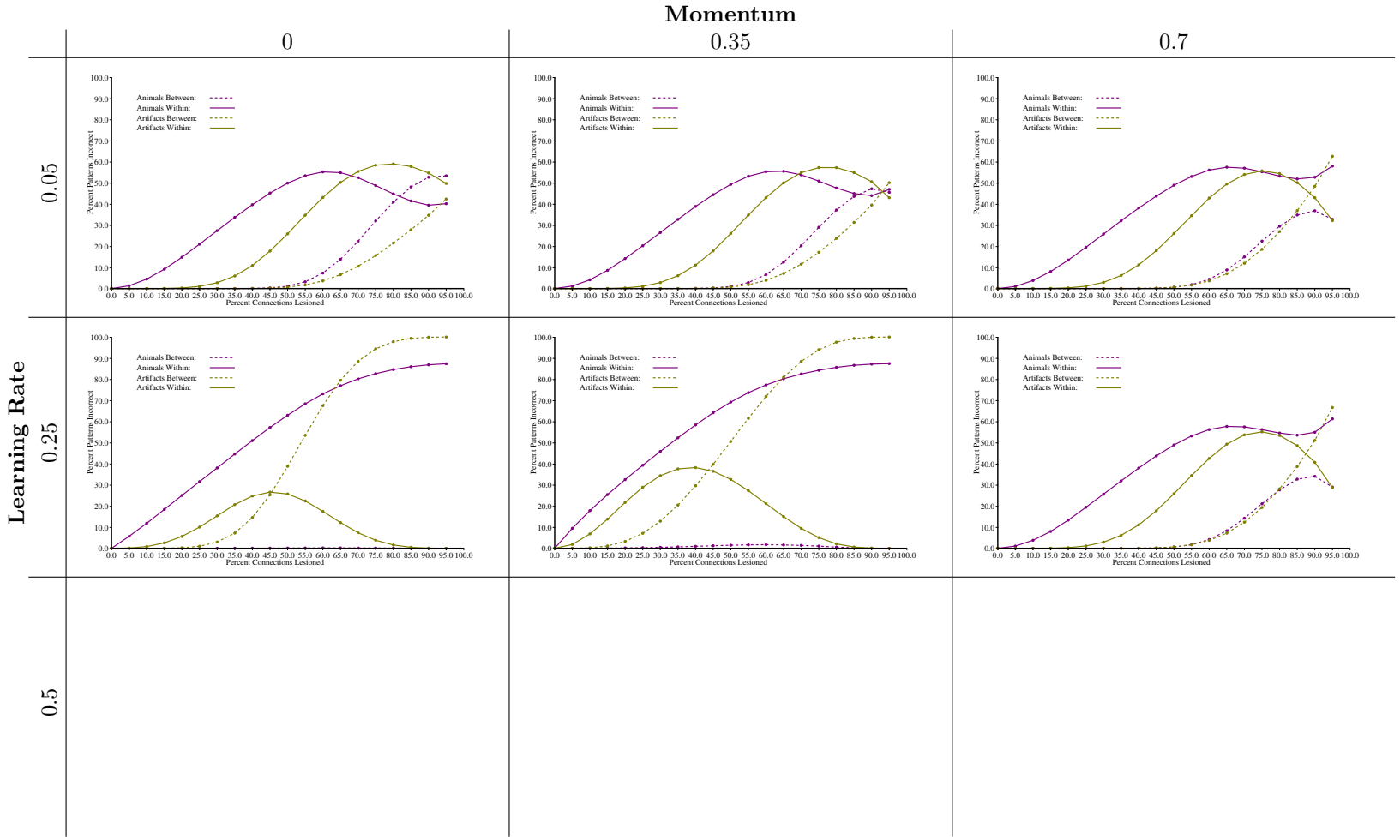


Table C.57: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.05, 0.05] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 8.)

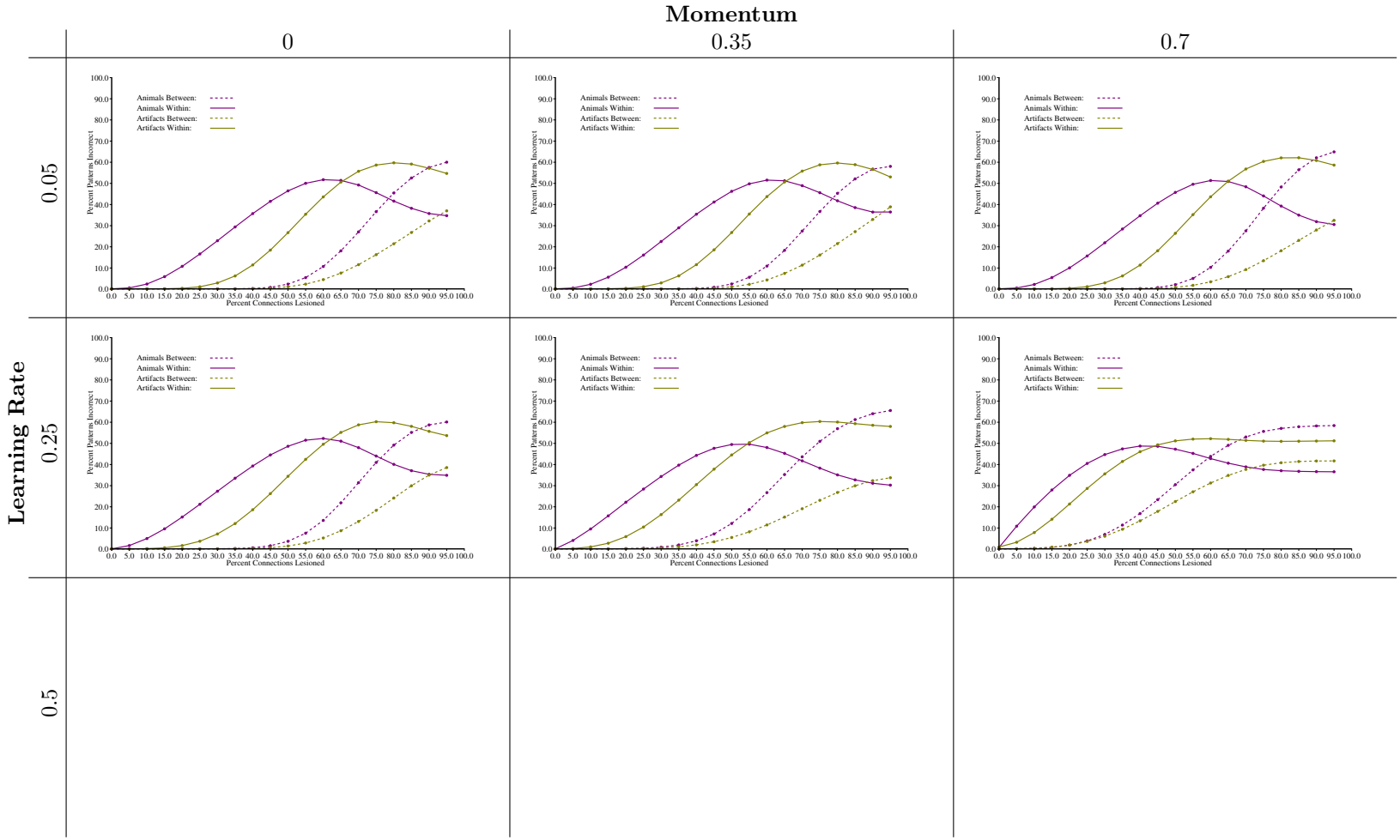


Table C.58: Percentage of between- and within-category errors per domain for networks with weights initialised to a range of [-0.5, 0.5] and error signals generated using CEE (compare with Tyler et al., 2000, fig. 8.)

# References


Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science*, *9*, 147–169.

Adlam, A.-L., Patterson, K., Rogers, T., Nestor, P., Salmond, C., Acosta-Cabronero, J., & Hodges, J. (2006). Semantic dementia and fluent primary progressive aphasia: two sides of the same coin? *Brain*, *129*(11), 3066–3080.

Allport, D. A. (1985). Current perspectives in dysphasia. In S. K. Newman & R. Epstein (Eds.), (p. 32-60). Churchill Livingstone.

Anonymous. (1994). Consensus statement. clinical and neuropathological criteria for frontotemporal dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *4*, 416-418.

Antonucci, S., & Alt, M. (2011). A lifespan perspective on semantic processing of concrete concepts: does a sensory/motor model have the potential to bridge the gap? *Cognitive, Affective, & Behavioral Neuroscience*, *11*, 551–572.

Arbib, M. A. (2008). From grasp to language: embodied concepts and the challenge of abstraction. *Journal of Physiology-Paris*, *102*(1), 4–20.

Arévalo, A. L., Baldo, J. V., & Dronkers, N. F. (2012). What do brain lesions tell us about theories of embodied semantics and the human mirror neuron system? *Cortex*, *48*(2), 242–254.

Aurenhammer, F. (1991). Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, *23*(3), 345–405.

Aziz-Zadeh, L., & Damasio, A. (2008). Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology-Paris*, *102*(1), 35–39.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, *22*(04), 637–660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.

Barsalou, L. W. (2010). Grounded cognition: past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84 - 91.

Basso, A., Capitani, E., & Laiacona, M. (1988). Progressive language impairment without dementia: a case with isolated category specific semantic defect. *Journal of Neurology, Neurosurgery & Psychiatry*, *51*(9), 1201.

Berger, J. R., & Houff, S. (2008). Neurological complications of herpes simplex virus type 2 infection. *Archives of Neurology*, *65*(5), 596-600.

Bi, Y., Han, Z., Shu, H., & Caramazza, A. (2005). Are verbs like inanimate objects? *Brain and language*, *95*(1), 28–29.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *The Journal of Neuroscience*, *2*(1), 32–48.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527 - 536.

Biran, I., Chatterjee, A., & Glosser, G. (2002). Are verbs like inanimate objects? *Brain and Language*, *83*, 217-220.

Bonner, M., Ash, S., & Grossman, M. (2010). The new classification of primary progressive aphasia into semantic, logopenic, or nonfluent/agrammatic variants. *Current Neurology and Neuroscience Reports*, *10*(6), 484-490.

Bonner, M. F., Peelle, J. E., Cook, P. A., & Grossman, M. (2013). Heteromodal conceptual processing in the angular gyrus. *Neuroimage*, *71*, 175–186.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., ... Hodges, J. R. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive Neuropsychology*, *20*(1), 27–47.

Bozeat, S., Ralph, M. A. L., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207 - 1215.

Breedin, S., Saffran, E., & Coslett, H. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology*, *11*(6), 617–660.

Bright, P., Moss, H., & Tyler, L. (2004). Unitary vs multiple semantics: Pet studies of word and picture processing. *Brain and language*, *89*(3), 417–432.

Bunn, E. M., Tyler, L. K., & Moss, H. E. (1998). Category-specific semantic deficits: The role of familiarity and property type reexamined. *Neuropsychology*, *12*(3), 367.

Cairns, N. J., Bigio, E. H., Mackenzie, I. R., Neumann, M., Lee, V. M.-Y., Hatanpaa, K. J., ... others (2007). Neuropathologic diagnostic and nosologic criteria for frontotemporal lobar degeneration: consensus of the consortium for frontotemporal lobar degeneration. *Acta neuropathologica*, *114*(1), 5–22.

Campo, P., Poch, C., Toledano, R., Igoa, J. M., Belinchón, M., García-Morales, I., & Gil-Nagel, A. (2013). Anterobasal temporal lobe lesions alter recurrent functional connectivity within the ventral pathway during naming. *The Journal of Neuroscience*, *33*(31), 12679-12688.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? a critical review of the clinical evidence. *Cognitive Neuropsychology*, *20*(3-6), 213–261.

Caramazza, A. (1991). Some aspects of language processing revealed through the analysis of acquired aphasia: The lexical system. In *Issues in reading, writing and speaking* (pp. 15–44). Springer.

Caramazza, A. (1998). The interpretation of semantic category-specific deficits: What do they reveal about the organization of conceptual knowledge in the brain? *Neurocase*, *4*(4-5), 265–272.

Caramazza, A. (1999). The new cognitive neurociences. In M. S. Gazzaniga (Ed.), (pp. 1037–1046). MIT Press.

Caramazza, A., Hillis, A., Rapp, B., & Romani, C. (1990). The multiple semantics hypothesis: Multiple confusions? *Cognitive Neuropsychology*, *7*(3), 161–189.

Caramazza, A., & Mahon, B. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7*(8), 354–361.

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

Carbonnel, S., Charnallet, A., David, D., & Pellat, J. (1997). One or several semantic system (s)? maybe none: evidence from a case study of modality and category-specific "semantic" impairment. *Cortex*, *33*(3), 391–417.

Cardebat, D., Demonet, J. F., Celsis, P., & Puel, M. (1996). Living/nonliving dissociation in a case of semantic dementia: A spect activation study. *Neuropsychologia*, *34*, 1175-1179.

Castillo, M., & Rumboldt, Z. (2012). Brain imaging with mri and ct: An image pattern approach. In Z. Rumboldt, M. Castillo, B. Huang, & A. Rossi (Eds.), (p. 41-42). Cambridge.

Chertkow, H., Bub, D., & Seidenberg, M. (1989). Priming and semantic memory loss in Alzheimer's disease. *Brain and Language*, *36*(3), 420–446.

Chiou, R., & Rich, A. N. (2014). The role of conceptual knowledge in understanding synaesthesia: Evaluating contemporary findings from a 'hub-and-spoke'perspective. *Name: Frontiers in Psychology*, *5*, 105.

Choe, Y., Sirosh, J., & Miikkulainen, R. (1996). Laterally interconnected self-organizing maps in hand-written digit recognition. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems 8: Proceedings of the 1995 conference* (chap. Laterally Interconnected Self-Organizing Maps in Hand-Written Digit Recognition). MIT Press.

Cipolotti, L., & Warrington, E. (1995). Towards a unitary account of access dysphasia: a single case study. *Memory*, *3*(3), 309–332.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, A. M., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, *8*(2), 240–247.

Collins, A. M., & Quillian, M. (1972). Experiments on semantic memory and language comprehension. *Cognition in learning and memory*, *1969*, 263.

Compston, A. (2011). From the archives. *Brain*, *134*(9), 2444-2446.

Cooper, R. P., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. *Artificial Intelligence*, *85*(1), 3–44.

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42-49.

Cummings, J. (1991). Clinical neurology. In M. Swash & J. Oxbury (Eds.), (p. 131-139). Churchill Livingstone.

Cummings, J., & Benson, D. (1983). *Dementia: a clinical approach* (J. Cummings & D. Benson, Eds.). Butterworth.

Cummings, J., & Duchen, L. (1981). Kluver-bucy syndrome in pick's disease: clinical and pathological correlations. *Neurology*, *31*, 1415-1422.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62.

Damasio, A. R., & Damasio, H. (1994). Large-scale neuronal theories of the brain. In C. Koch & J. Davis (Eds.), (pp. 61–74). MIT Press.

Daum, I., Riesch, G., Sartori, G., & Birbaumer, N. (1996). Semantic memory impairment in alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *18*(5), 648–665.

Davies, R. R., Graham, K. S., Xuereb, J. H., Williams, G., & Hodges, J. R. (2004). The human perirhinal cortex and semantic memory. *European Journal of Neuroscience*, *20*, 2441-2446.

Davies, R. R., Hodges, J. R., Kril, J. J., Patterson, K., Halliday, G. M., & Xuereb, J. H. (2005). The pathological basis of semantic dementia. *Brain*, *128*(9), 1984–1995.

Davis, L. E., & Johnson, R. T. (1979). An explanation for the localization of herpes simplex encephalitis? *Ann Neurol*, *5*, 2-5.

Dejerine, J., & Sérieux, P. (1897). Un cas de surdité verbale pure terminée par aphasie sensorielle, suivi d'autopsie. *C R Soc Biol*, *49*, 1074-1077.

Dennis, M. (1976). Dissociated naming and locating of body parts after left anterior temporal lobe resection: An experimental case study. *Brain and Language*, *3*(2), 147–163.

De Renzi, E., Liotti, M., & Nichelli, P. (1987). Semantic amnesia with preservation of autobiographic memory. a case report. *Cortex*, *23*, 575-597.

De Renzi, E., & Lucchelli, F. (1994). Are semantic systems separately represented in the brain? the case of living category impairment. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior*, *30*(1), 3–25.

Dewar, B.-K., & Gracey, F. (2007). "am not was": Cognitive-behavioural therapy for adjustment and identity change following herpes simplex encephalitis. *Neuropsychological Rehabilitation*, *17*(4-5), 602-620. (PMID: 17676537)

Dickerson, F., Stallings, C., Sullens, A., Origoni, A., Leister, F., Krivogorsky, B., & Yolken, R. (2008). Association between cognitive functioning, exposure to herpes simplex virus type 1, and the {COMT} val158met genetic polymorphism in adults without a psychiatric disorder. *Brain, Behavior, and Immunity*, *22*(7), 1103 - 1107.

Dinn, J. J. (1980). Transolfactory spread of virus in herpes simplex encephalitis. *British medical journal*, *281*(6252), 1392.

Duffau, H. (2001). Acute functional reorganisation of the human motor cortex during resection of central lesions: a study using intraoperative brain mapping. *Journal of Neurology, Neurosurgery & Psychiatry*, *70*(4), 506–513.

Elbert, T., Flor, H., Birbaumer, N., Knecht, S., Hampson, S., Larbig, W., & Taub, E. (1994). Extensive reorganization of the somatosensory cortex in adult humans after nervous system injury. *Neuroreport*, *5*, 2593-2597.

Esiri, M. M. (1982). Herpes simplex encephalitis: an immunohistological study of the distribution of viral antigens within the brain. *Journal of Neurological Science*, *54*, 209-226.

Farah, M. J., Hammond, K. M., Mehta, Z., & Ratcliff, G. (1989). Category-specificity and modality-specificity in semantic memory. *Neuropsychologia*, *27*(2), 193–200.

Farah, M. J., & McClelland, J. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339.

Farah, M. J., & Wallace, M. A. (1992). Semantically-bounded anomia: Implications for the neural implementation of naming. *Neuropsychologia*, *30*(7), 609–621.

Farkas, I., & Miikkulainen, R. (1999). Modeling the self-organization of directional selectivity in the primary visual cortex. In *9th international conference on artificial neural networks: Icann '99* (p. 251-256(5)).

Fenker, D., Waldmann, M., & Holyoak, K. (2005). Accessing causal relations in semantic memory. *Memory & cognition*, *33*(6), 1036–1046.

Ferrari, S., Toniolo, A., Monaco, S., Luciani, F., Cainelli, F., Baj, A., ... Vento, S. (2009). Viral encephalitis: Etiology, clinical features, diagnosis and management. *Open Infectious Diseases Journal*, *3*, 1-12.

Fletcher, P. D., & Warren, J. D. (2011). Semantic dementia: a specific network-opathy. *Journal of Molecular Neuroscience*, *45*(3), 629-636.

Fodor, J. A. (1975). *The language of thought* (J. J. Katz, D. T. Langendoen, & G. A. Miller, Eds.). Thomas Y. Crowell Company, Inc.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189 - 198.

Forde, E. M. E., D. Francis, M. J. R., Rumiati, R. I., & Humphreys, G. W. (1997). On the links between visual knowledge and naming: A single case study of a patient with a category-specific impairment for living things. *Cognitive Neuropsychology*, *14*(3), 403-458.

French, R. M., & Mareschal, D. (1998). Could category-specific semantic deficits reflect differences in the distributions of features within a unified semantic memory. In *Proceedings of the twentieth annual cognitive science society conference* (pp. 374–379).

Friston, K. J., Frith, C. D., & Frackowiak, R. S. J. (1993). Principal component analysis learning algorithms: A neurobiological analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *254*(1339), 47-54.

Funnell, E., & De Mornay Davies, P. (1996). A reassessment of concept familiarity and a category-specific disorder for living things. *Neurocase*, *2*(6), 461–474.

Funnell, E., & Sheridan, J. (1992). Categories of knowledge? unfamiliar aspects of living and nonliving things. *Cognitive Neuropsychology*, *9*(2), 135–153.

Gaffan, D., & Heywood, C. A. (1993). A spurious category-specific visual agnosia for living things in normal human and nonhuman primates. *Journal of Cognitive Neuroscience*, *5*(1), 118–128.

Gainotti, G. (2005). The influence of gender and lesion location on naming disorders for animals, plants and artefacts. *Neuropsychologia*, *43*(11), 1633–1644.

Gainotti, G., & Silveri, M. C. (1996). Cognitive and anatomical locus of lesion in a patient with a category-specific semantic impairment for living beings. *Cognitive Neuropsychology*, *13*, 357-389.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3-4), 455–479.

Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M., Williams, G., Antoun, N., ... Hodges, J. (2001). Differing patterns of temporal atrophy in alzheimer's disease and semantic dementia. *Neurology*, *57*(2), 216-225.

Garrard, P., Lambon Ralph, M. A., & Hodges, J. R. (2002). Semantic dementia: A category-specific paradox. In *Category specificity in brain and mind.* Psychology Press.

Garrard, P., Patterson, K., Watson, P. C., & Hodges, J. R. (1998). Category specific semantic loss in dementia of alzheimer's type. functional-anatomical correlations from cross-sectional analyses. *Brain*, *121*(4), 633-646.

Girling, D. M., & Berrios, G. E. (1994). On the relationship between senile cerebral atrophy and aphasia. *History of Psychiatry*, *5*, 542-547.

Girling, D. M., & Berrios, G. E. (1997). On the symptomatology of left-sided temporal lobe atrophy. *History of Psychiatry*, *8*, 149-159.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, *43*(3), 379–401.

Goedert, M., Ghetti, B., & Spillantini, M. G. (2012). Frontotemporal dementia: Implications for understanding alzheimer disease. *Cold Spring Harbor Perspectives in Medicine*, *2*(2), 1-21.

Gonnerman, L., Andersen, E., Devlin, J., Kempler, D., & Seidenberg, M. (1997). Double dissociation of semantic categories in alzheimer's disease. *Brain and Language*, *57*(2), 254-279.

Gorno-Tempini, M. L., Dronkers, N. F., Rankin, K. P., Ogar, J. M., Phengrasamy, L., Rosen, H. J., ... Miller, B. L. (2004). Cognition and anatomy in three variants of primary progressive aphasia. *Ann Neurol*, *55*, 335-346.

Gorno-Tempini, M. L., Hillis, A. E., Weintraub, S., Kertesz, A., Mendez, M., Cappa, S. F., ... others (2011). Classification of primary progressive aphasia and its variants. *Neurology*, *76*(11), 1006-1014.

Gotts, S. J., & Plaut, D. C. (2002). The impact of synaptic depression following brain damage: A connectionist account of "access/refractory" and "degraded-store" semantic impairments. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(3), 187–213.

Graff-Radford, N., Damasio, A. R., & Hyman, B. e. a. (1990). Progressive aphasia in a patient with pick's disease. *Neurology*, *40*, 423-429.

Graham, N., Patterson, K., & Hodges, J. (2000). The impact of semantic memory impairment on spelling: evidence from semantic dementia. *Neuropsychologia*, *38*(2), 143 - 163.

Grasemann, U., Sandberg, C., Kiran, S., & Miikkulainen, R. (2011). Impairment and rehabilitation in bilingual aphasia: A som-based model. In J. Laaksonen & T. Honkela (Eds.), *Advances in self-organizing maps* (Vol. 6731). Springer Berlin Heidelberg.

Greer, M. J., Van Casteren, M., McLellan, S. A., Moss, H. E., Rodd, J., Rogers, T., & Tyler, L. (2001). The emergence of semantic categories from distributed featural representations. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 358–363).

Grossman, M. (2010). Primary progressive aphasia: clinicopathological correlations. *Nat Rev Neurol*, *6*, 88-97.

Grossman, M., Mickanin, J., Onishi, K., Hughes, E., D'Esposito, M., Ding, X.-S., ... Reivich, M. (1996). Progressive nonfluent aphasia: language, cognitive, and pet measures contrasted with probable alzheimer's disease. *Journal of Cognitive Neuroscience*, *8*(2), 135–154.

Grydeland, H., Walhovd, K. B., Westlye, L. T., Due-Tønnessen, P., Ormaasen, V., Sundseth, Ø., & Fjell, A. M. (2010). Amnesia following herpes simplex encephalitis: Diffusion-tensor imaging uncovers reduced integrity of normal-appearing white matter. *Radiology*, *257*(3), 774–781.

Guest, O., & Cooper, R. P. (2012). Semantic cognition: A re-examination of the recurrent network "hub" model. In *In proceedings of the ninth international conference on cognitive modelling.*

Guest, O., Cooper, R. P., & Davelaar, E. J. (2014). Computational models of cognitive processes. In J. Mayor & P. Gomez (Eds.), (p. 155-169). World Scientific.

Gustafson, L. (1987). Frontal lobe degeneration of non-alzheimer type. ii. clinical picture and differential diagnosis. *Archives of Gerontology and Geriatrics*, *6*, 209-223.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, *13*, 49-52.

Hagberg, B. (1987). Behaviour correlates of frontal lobe dysfunction. *Arch. Gerontol. Geriatr.*, *6*, 311-321.

Hagoort, P. (1993). Impairments of lexical-semantic processing in aphasia: Evidence from the processing of lexical ambiguities. *Brain and Language*, *45*(2), 189–232.

Harciarek, M., & Kertesz, A. (2009). Longitudinal study of single-word comprehension in semantic dementia: A comparison with primary progressive aphasia and alzheimer's disease. *Aphasiology*, *23*(5), 606-626.

Hart, J. J., Berndt, R. S., & Caramazza, A. (1985). Category-specific naming deficit following cerebral infarction. *Nature*, *316*(6027), 439–440.

Hart, J. J., & Gordon, B. (1992). Neural subsystems for object knowledge. *Nature*, *359*(6390), 60–64.

Hauk, O., & Tschentscher, N. (2013). The body of evidence: what can neuroscience tell us about embodied semantics? *Frontiers in psychology*, *4*, 1–14.

Hillis, A., & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain*, *114*(5), 2081.

Hinton, G., & Sejnowski, T. (1986). Learning and relearning in boltzmann machines. *MIT Press, Cambridge, Mass.*, *1*, 282–317.

Hodges, J. R. (1994). Pick's disease. In A. Burns & R. Levy (Eds.), *Dementia* (p. 739-752). Springer US.

Hodges, J. R., Bozeat, S., Ralph, M. A. L., Patterson, K., & Spatt, J. (2000). The role of conceptual knowledge in object use evidence from semantic dementia. *Brain*, *123*(9), 1913-1925.

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory*, *3*(3-4), 463–495.

Hodges, J. R., & Miller, B. (2001). The classification, genetics and neuropathology of frontotemporal dementia. introduction to the special topic papers: Part i. *Neurocase*, *7*(1), 31–35.

Hodges, J. R., Miller, B., et al. (2001). The neuropsychology of frontal variant frontotemporal dementia and semantic dementia. introduction to the special topic papers: Part ii. *Neurocase*, *7*(2), 113–121.

Hodges, J. R., Mitchell, J., Dawson, K., Spillantini, M. G., Xuereb, J. H., McMonagle, P., ... Patterson, K. (2010). Semantic dementia: demography, familial factors and survival in a consecutive series of 100 cases. *Brain*, *133*(1), 300–306.

Hodges, J. R., & Patterson, K. (1996). Non-fluent progressive aphasia and semantic dementia: a comparative neuropsychological study. *Journal of the International Neuropsychological Society*, *2*, 511-524.

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *Lancet Neurology*, *6*, 1004-1014.

Hodges, J. R., Patterson, K., Oxbury, S., & Funnell, E. (1992). Semantic dementia: Progressive fluent aphasia with temporal lobe atrophy. *Brain*, *115*, 1783–1806.

Hoffman, P., Jones, R. W., & Ralph, M. A. L. (2012). The degraded concept representation system in semantic dementia: damage to pan-modal hub, then visual spoke. *Brain*, *135*(12), 3770-3780.

Hoffman, P., & Lambon Ralph, M. A. (2011). Reverse concreteness effects are not a typical feature of semantic dementia: Evidence for the hub-and-spoke model of conceptual representation. *Cerebral Cortex*, *21*(9), 2103-2112.

Hokkanen, L., & Launes, J. (2007). Neuropsychological sequelae of acute-onset sporadic viral encephalitis. *Neuropsychological rehabilitation*, *17*(4-5), 450–477.

Hokkanen, L., Poutiainen, E., Valanne, L., Salonen, O., Iivanainen, M., & Launes, J. (1996). Cognitive impairment after acute encephalitis: comparison of herpes simplex and other aetiologies. *Journal of Neurology, Neurosurgery & Psychiatry*, *61*(5), 478–484.

Holland, A., McBurney, D., Moossy, J., & Reinmuth, O. (1985). The dissolution of language in pick's disease with neurofibrillary tangles: a case study. *Brain Lang.*, *24*, 36-58.

Howard, D., Best, W., Bruce, C., & Gatehouse, C. (1995). Operativity and animacy effects in aphasic naming. *International Journal of Language & Communication Disorders*, *30*(3), 286–302.

Howard, D., & Patterson, K. (1992). *The pyramids and palm trees test: A test of semantic access from words and pictures.* Bury St Edmunds, UK: Thames Valley Test Company.

Hudgins, L., & Kaspersen, J. (1999). Wavelets and detection of coherent structures in fluid turbulence. In *Wavelets in physics* (Vol. 1, p. 201).

Humphreys, G. W., Forde, E., et al. (2001). Hierarchies, similarity, and interactivity in object recognition:" category-specific" neuropsychological deficits. *Behavioral and Brain Sciences*, *24*(3), 453–475.

Humphreys, G. W., & Riddoch, J. M. (2003). A case-series analysis of category-specific deficits of living things: the hit account. *Cognit Neuropsychol*, *20*, 263-306.

Hwang, K., Hallquist, M. N., & Luna, B. (2013). The development of hub architecture in the human functional brain network. *Cerebral Cortex*, *23*(10), 2380–2393.

Ishai, A., Ungerleider, L., Martin, A., Schouten, J., & Haxby, J. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379.

Jefferies, E. (2013). The neural basis of semantic cognition: converging evidence from neuropsychology, neuroimaging and tms. *Cortex*, *49*(3), 611–625.

Jefferies, E., Patterson, K., Jones, R. W., Bateman, D., & Lambon Ralph, M. A. (2004). A category-specific advantage for numbers in verbal short-term memory: Evidence from semantic dementia. *Neuropsychologia*, *42*, 639-660.

Jefferies, E., Rogers, T. T., & Ralph, M. A. L. (2011). Premorbid expertise produces category-specific impairment in a domain-general semantic disorder. *Neuropsychologia*, *49*(12), 3213–3223.

Johnson, R. T., Olson, L. C., & Buescher, E. L. (1968). Herpes simplex virus infections of the nervous system: problems in laboratory diagnosis. *Archives of neurology*, *18*(3), 260.

Kapur, N., Barker, S., Burrows, E., Ellison, D., Brice, J., Illis, L., ... Loates, M. (1994). Herpes simplex encephalitis: long term magnetic resonance imaging and neuropsychological profile. *Journal of Neurology, Neurosurgery & Psychiatry*, *57*(11), 1334–1342.

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (som) papers: 1981-1997. *Neural Computing Surveys*, *1*, 102-350.

Katzman, R. (1986). Differential diagnosis of dementing illness. *Neurol. Clin. North Am.*, *4*, 329-340.

Kennedy, P. G. E., & Chaudhuri, A. (2002). Herpes simplex encephalitis. *Journal of Neurology, Neurosurgery & Psychiatry*, *73*(3), 237-238.

Khanna, K. M., Lepisto, A. J., Decman, V., & Hendricks, R. L. (2004). Immune control of herpes simplex virus during latency. *Current opinion in immunology*, *16*(4), 463–469.

Kirshner, H., Tanridag, O., Thurman, L., & Whetsell, W. (1987). Progressive aphasia without dementia: two cases with focal spongiform. *Ann. Neurol.*, *22*, 527-532.

Knibb, J. A., Xuereb, J. H., Patterson, K., & Hodges, J. R. (2006). Clinical and pathological characterization of progressive aphasia. *Ann Neurol*, *59*, 156-165.

Knopman, D., Christensen, K., & Schut, L. e. a. (1989). The spectrum of imaging and neuropsychological findings in pick's disease. *Neurology*, *39*, 362-368.

Koenig, P., Smith, E., Glosser, G., DeVita, C., Moore, P., McMillan, C., ... Grossman, M. (2005). The neural basis for novel semantic categorization. *Neuroimage*, *24*(2), 369–383.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kohonen, T. (1993). Things you haven't heard about the self-organizing map. In *Ieee international conference on neural networks* (pp. 1147–1156).

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.; T. S. Huang, T. Kohonen, & M. R. Schroeder, Eds.). Springer-Verlang.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1996). *SOM_PAK: The Self-Organizing Map Program Package* (Tech. Rep.). Espoo, Finland: Helsinki University of Technology, Laboratory of Computer and Information Science.

Kohonen, T., Mäkisara, K., & Saramäki, T. (1984). Phonotopic maps insightful representation of phonological features for speech recognition. In *7icpr, int. conf. on pattern recognition.*

Laforce, R. J. (2013). Behavioral and language variants of frontotemporal dementia: A review of key symptoms. *Clinical Neurology and Neurosurgery*, *115*(12), 2405 - 2410.

Laiacona, M., Barbarotto, R., & Capitani, E. (1993). Perceptual and associative knowledge in category specific impairment of semantic memory: A study of two cases. *Cortex*, *29*(4), 727-740.

Laiacona, M., Capitani, E., & Barbarotto, R. (1997). Semantic category dissociations: A longitudinal study of two cases. *Cortex*, *33*(3), 441–461.

Lambon Ralph, M. A. (2014). Neurocognitive insights on conceptual knowledge and its breakdown. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 1–11.

Lambon Ralph, M. A., Graham, K. S., Patterson, K., & Hodges, J. R. (1999). Is a picture worth a thousand words? evidence from concept definitions by patients with semantic dementia. *Brain and Language*, *70*, 309-335.

Lambon Ralph, M. A., & Howard, D. (2000). Gogi aphasia or semantic dementia? simulating and assessing poor verbal comprehension in a case of progressive fluent aphasia. *Cognitive Neuropsychology*, *17*(5), 437–465.

Lambon Ralph, M. A., Howard, D., Nightingale, G., & Ellis, A. (1998). Are living and non-living category-specific deficits causally linked to impaired perceptual or associative knowledge? evidence from a category-specific double dissociation. *Neurocase*, *4*(4-5), 311–338.

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

Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2003). Semantic dementia with category specificity: Acomparative case-series study. *Cognitive Neuropsychology*, *20*(3-6), 307–326.

Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (1997). The relationship between naming and semantic knowledge for different categories in dementia of alzheimer's type. *Neuropsychologia*, *35*, 1251-1260.

Lambon Ralph, M. A., Pobric, G., & Jefferies, E. (2009). Conceptual knowledge is underpinned by the temporal pole bilaterally: Convergent evidence from rtms. *Cerebral Cortex*, *19*(4), 832-838.

Lambon Ralph, M. A., Sage, K., Jones, R. W., & Mayberry, E. J. (2010). Coherent concepts are computed in the anterior temporal lobes. *Proceedings of the National Academy of Sciences*, *107*(6), 2717–2722.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation* (J. Aravind, Ed.). MIT.

Liepert, J., Bauder, H., Miltner, W. H., Taub, E., & Weiller, C. (2000). Treatment-induced cortical reorganization after stroke in humans. *Stroke*, *31*(6), 1210–1216.

Lilly, R., Cummings, J. L., Benson, D. F., & Frankel, M. (1983). The human klüver[U+2010]bucy syndrome. *Neurology*, *33*(9), 1141.

Mahon, B. Z., & Caramazza, A. (2003). There are facts... and then there are facts. *TRENDS in Cognitive Sciences*, *7*, 481–482.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology-Paris*, *102*(1), 59–70.

Mahon, B. Z., & Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, *15*, 97–103.

Mandler, J. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development*, *1*(1), 3–36.

Martin, A. (2007). The representation of object concepts in the brain. *Annu. Rev. Psychol.*, *58*, 25–45.

Martin, A., & Chao, L. (2001). Semantic memory and the brain: structure and processes. *Current opinion in neurobiology*, *11*(2), 194–201.

Mathewson Commission. (1929). *Epidemic encephalitis: etiology, epidemiology, treatment.*

Mauri, A., Daum, I., Sartori, G., Riesch, G., & Birbaumer, N. (1994). Category-specific semantic impairment in alzheimer's disease and temporal lobe dysfunction: A comparative study. *Journal of Clinical and Experimental Neuropsychology*, *16*(5), 689–701.

McCarthy, R. A., & Warrington, E. K. (1986). Visual associative agnosia: A clinico-anatomical study of a single case. *Journal of Neurology, Neurosurgery and Psychiatry*, *49*, 1233-1240.

McCarthy, R. A., & Warrington, E. K. (1988). Evidence for modality-specific meaning systems in the brain. *Nature*, *334*(6181), 428–430.

McClelland, J. (2011). Explorations in parallel distributed processing: A handbook of models, programs, and exercises [Computer software manual].

McClelland, J., & Rogers, T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322.

McGrath, N., Anderson, N., Croxson, M., & Powell, K. (1997). Herpes simplex encephalitis treated with acyclovir: diagnosis and long term outcome. *Journal of Neurology, Neurosurgery & Psychiatry*, *63*(3), 321-326.

McRae, K., & Cree, G. S. (2002). Category-specificity in brain and mind. In E. M. E. Forde & G. W. Humphreys (Eds.), (pp. 211–249). Psychology Press.

McRae, K., Seidenberg, M. S., & de Sa, V. R. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*, 99-130.

Mesulam, M. M. (1982). Slowly progressive aphasia without generalized dementia. *Ann Neurol*, *11*, 592-598.

Mesulam, M. M., Rogalski, E., Wieneke, C., Cobia, D., Rademaker, A., Thompson, C., & Weintraub, S. (2009). Neurology of anomia in the semantic variant of primary progressive aphasia. *Brain*, *132*, 2553-65.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788–804.

Miikkulainen, R. (1990). A distributed feature map model of the lexicon. In *12th annual conference of the cognitive science society.* Lawrence Erlbaum.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory.* Cambridge, MA: MIT Press.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, *59*(2), 334 - 366.

Miikkulainen, R., & Kiran, S. (2009). Modeling the bilingual lexicon of an individual subject. In J. Príncipe & R. Miikkulainen (Eds.), *Advances in self-organizing maps* (Vol. 5629, p. 191-199). Springer Berlin Heidelberg.

Milberg, W., Blumstein, S., & Dworetzky, B. (1987). Processing of lexical ambiguities in aphasia. *Brain and Language*, *31*(1), 138–150.

Milberg, W., & Blumstein, S. E. (1981). Lexical decision and aphasia: Evidence for semantic processing. *Brain and language*, *14*(2), 371–385.

Miller, B., Cummings, J., & Villanueva-Meyer, J. e. a. (1991). Frontal lobe degeneration: clinical, neuropsychological, and spect characteristics. *Neurology*, *41*, 1374-1382.

Miller, K. D., & MacKay, D. J. (1994). The role of constraints in hebbian learning. *Neural Computation*, *6*(1), 100–126.

Mingazzini, G. (1913–1914). On aphasia due to atrophy of the cerebral convolutions. *Brain*, *36*, 493-524.

Montanes, P., Goldblum, M. C., Boller, F., et al. (1995). The naming impairment of living and nonliving items in alzheimer's disease. *Journal of the International Neuropsychological Society*, *1*(1), 39–48.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of acquisition norms for a large set of object names and their relation to adult estimates and other variables. *Quarterly Journal of Experimental Psychology*, *50A*, 528-559.

Moss, H. E., Rodd, J., Stamatakis, E., Bright, P., & Tyler, L. (2005). Anteromedial temporal cortex supports fine-grained differentiation among objects. *Cerebral Cortex*, *15*(5), 616-627.

Moss, H. E., & Tyler, L. (2000). A progressive category-specific semantic deficit for non-living things. *Neuropsychologia*, *38*(1), 60–82.

Moss, H. E., Tyler, L., & Jennings, F. (1997). When leopards lose their spots: Knowledge of visual properties in category-specific deficits for living things. *Cognitive Neuropsychology*, *14*(6), 901–950.

Moss, H. E., & Tyler, L. K. (2003). Weighing up the facts of category-specific semantic deficits. *TRENDS in Cognitive Sciences*, *7*, 480–481.

Moss, H. E., Tyler, L. K., Durrant-Peatfield, M., & Bunn, E. M. (1998). 'two eyes of a see-through': Impaired and intact semantic knowledge in a case of selective deficit for living things. *Neurocase*, *4*(4-5), 291–310.

Mummery, C., Patterson, K., Hodges, J. R., & Wise, R. J. S. (1996). Generating 'tiger' as an animal name or a word beginning with t: differences in brain activation. In *Proceedings: Biological sciences.* The Royal Society.

Munoz-Garcia, D., & Ludwin, S. (1984). Classic and generalized variants of pick's disease: a clinicopathological, ultrastructural and immunocytochemical comparative study. *Ann Neural*, *16*, 467-480.

Nahmias, A., & Dowdle, W. (1968). Antigenic and biologic differences in herpesvirus hominis. *Prog Med Virol*, *10*, 110-159.

Nakamura, A., Yamada, T., Goto, A., Kato, T., Ito, K., Abe, Y., ... Kakigi, R. (1998). Somatosensory homunculus as drawn by {MEG}. *NeuroImage*, *7*(4), 377 - 386.

Neary, D., Snowden, J. S., Gustafson, L., Passant, U., Stuss, D., Black, S. a., ... others (1998). Frontotemporal lobar degeneration a consensus on clinical diagnostic criteria. *Neurology*, *51*(6), 1546–1554.

Neary, D., Snowden, J. S., Northen, B., & Goulding, P. (1988). Dementia of frontal lobe type. *Journal of Neurology, Neurosurgery & Psychiatry*, *51*(3), 353-361.

Newell, A. (1990). *Unified theories of cognition* (A. Newell, Ed.). Harvard University Press Cambridge, MA, USA.

Nikkilä, J., Törönen, P., Kaski, S., Venna, J., Castrén, E., & Wong, G. (2002). Analysis and visualization of gene expression data using self-organizing maps. *Neural networks*, *15*(8), 953–966.

Noppeney, U., Patterson, K., Tyler, L., Moss, H., Stamatakis, E., Bright, P., ... Price, C. (2007). Temporal lobe lesions and semantic impairment: a comparison of herpes simplex virus encephalitis and semantic dementia. *Brain*, *130*(4), 1138.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, *15*(3), 267–273.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, *01*(01), 61-68.

Oja, E., & Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of mathematical analysis and applications*, *106*(1), 69–84.

Oja, M., Kaski, S., & Kohonen, T. (2003). Bibliography of self-organizing map (som) papers: 1998-2001 addendum. *Neural Computing Surveys*, *3*, 1-156.

Ojeda, V., Archer, M., Robertson, T., & Bucens, M. (1983). Necropsy study of the olfactory portal of entry in herpes simplex encephalitis. *The Medical journal of Australia*, *1*(2), 79.

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, *130*(7), 1718-1731.

Orrell, M., & Sahakian, B. (1991). Dementia of frontal lobe type. *Psychol. Med.*, *21*, 553-556.

Panksepp, J., Moskal, J., Panksepp, J. B., & Kroes, R. (2002). Comparative approaches in evolutionary psychology: Molecular neuroscience meets the mind. *Neuroendocrinology Letters*, *23*(Suppl 4), 105–115.

Panksepp, J., & Panksepp, J. B. (2000). The seven sins of evolutionary psychology. *Evolution and cognition*, *6*(2), 108–131.

Panksepp, J., & Panksepp, J. B. (2001). A continuing critique of evolutionary psychology: Seven sins for seven sinners, plus or minus two. *Evolution and Cognition*, *7*, 56–80.

Parkin, A. J. (1993). Progressive aphasia without dementia: A clinical and cognitive neuropsychological analysis. *Brain and Language*, *44*, 201-220.

Pascual, B., Masdeu, J. C., Hollenbeck, M., Makris, N., Insausti, R., Ding, S.-L., & Dickerson, B. C. (2013). Large-scale brain networks of the human left temporal pole: A functional connectivity mri study. *Cerebral Cortex*, *epub ahead of print*, 1-23.

Patterson, K. (2007). The reign of typicality in semantic memory. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 813-821.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Patterson, K., Ralph, M. A. L., Jefferies, E., Woollams, A., Jones, R., Hodges, J. R., & Rogers, T. T. (2006). "presemantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience*, *18*(2), 169-183.

Pick, A. (1892). Uber die beziehungen der senilen hirnatrophie zur aphasieick. *Prager Medizinische Wochenschrift*, *17*, 165-167.

Pick, A. (1901). Senile hirnatrophie als grundlage von herderscheinungen. *Wiener klin Wschr*, *14*, 403-404.

Pick, A. (1904). Zur symptomatologie der linksseitigen schläfenlappenatrophie. *Mschr Psychiat Neurol*, *16*, 378-488.

Pick, A. (1906). Uber einen weiterer symptomenkomplex im rahmen der dementia senilis, bedingt durch umschriebene starkere hirnatrophie (gemische apraxie). *Monatschrift fur Psychiatrie und Neurologie*, *19*, 97-108.

Pijnenburg, Y. A. (2011). New diagnostic criteria for the behavioural variant of frontotemporal dementia. *European Neurological Review*, *6*, 234-237.

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

Plunkett, K., & Elman, J. L. (1997). *Exercises in rethinking innateness: A handbook for connectionist simulations*. MIT Press.

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology*, *20*(10), 964–968.

Poeck, K., & Luzzatti, C. (1988). Slowly progressive aphasia in three patients: The problem of accompanying neuropsychological deficit. *Brain*, *111*, 151-168.

Pöllä, M., Honkela, T., & Kohonen, T. (2009). *Bibliography of self-organizing map (som) papers: 2002-2005 addendum* (Tech. Rep.). Helsinki University of Technology.

Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, *17*(9), 458–470.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science* (Z. W. Pylyshyn, Ed.). MIT Press.

Quillian, M. R. (1966). *Semantic memory* (Unpublished doctoral dissertation). Carnegie Institute of Technology.

Rabinovici, G. D., & Miller, B. L. (2010). Frontotemporal lobar degeneration epidemiology, pathophysiology, diagnosis and management. *CNS Drugs*, *24*, 375-398.

Randall, B., Moss, H. E., Rodd, J. M., Greer, M., & Tyler, L. K. (2004). Distinctiveness and correlation in conceptual structure: behavioral and computational studies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 393.

Raschilas, F., Wolff, M., Delatour, F., Chaffaut, C., De Broucker, T., Chevret, S., ... Rozenberg, F. (2002). Outcome of and prognostic factors for herpes simplex encephalitis in adult patients: Results of a multicenter study. *Clinical Infectious Diseases*, *35*(3), 254-260.

Riddoch, M. J., Humphreys, G. W., Coltheart, M., & Funnell, E. (1988). Semantic systems or system? neuropsychological evidence re-examined. *Cognitive Neuropsychology*, *5*(1), 3–25.

Robbins, P., Aydede, M., Clancey, W. J., Gallagher, S., Wilson, R. A., Clark, A., ... MacIver, M. A. (2008). *The cambridge handbook of situated cognition* (P. Robbins & M. Aydede, Eds.). Cambridge University Press.

Robinson, G., & Cipolotti, L. (2001). The selective preservation of colour naming in semantic dementia. *Neurocase*, *7*(1), 65–75.

Rogers, T. T., Garrard, P., McClelland, J., M.A. Lambon Ralph, Bozeat, S., Hodges, J., & Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235.

Rogers, T. T., Hocking, J., Noppeney, U., Mechelli, A., Gorno-Tempini, M., Patterson, K., & Price, C. (2006). Anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience*, *6*(3), 201-213.

Rogers, T. T., Lambon-Ralph, M., Patterson, K., McClelland, J. L., & Hodges, J. (1999). A recurrent connectionist model of semantic dementia. In *Cognitive neuroscience society*

*annual meeting program.*

Rogers, T. T., Patterson, K., & Graham, K. (2007). Colour knowledge in semantic dementia: It is not all black and white. *Neuropsychologia*, *45*(14), 3285–3298.

Rogers, T. T., & Plaut, D. C. (2002). Connectionist perspectives on category-specific deficits. In *Category specificity in brain and mind.* Psychology Press.

Rosazza, C., Imbornone, E., Zorzi, M., Farina, E., Chiavari, L., & Cappa, S. F. (2003). The heterogeneity of category-specific semantic disorders: Evidence from a new case. *Neurocase*, *9*(3), 189–202.

Rosen, H., Allison, S., Ogar, J., Amici, S., Rose, K., Dronkers, N., ... Gorno-Tempini, M. (2006). Behavioral features in semantic dementia vs other forms of progressive aphasias. *Neurology*, *67*(10), 1752–1756.

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automation* (Tech. Rep.). Project PARA, Cornell Aeronautical Laboratory.

Rosenfeld, M. (1909). Die partielle gorsshirnatrophie. *Journal fur Psychologie und Neurologie*, *14*, 115-130.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. *Parallel distributed processing: explorations in the microstructure of cognition*, *1*, 318–362.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing, Vol. 1: Foundations.* The MIT Press. Paperback.

Sacchett, C., & Humphreys, G. (1992). Calling a squirrel a squirrel but a canoe a wigwam: A category-specific deficit for artefactual objects and body parts. *Cognitive Neuropsychology*, *9*(1), 73–86.

Samson, D., Pillon, A., & De Wilde, V. (1998). Impaired knowledge of visual and non-visual attributes in a patient with a semantic impairment for living entities: A case of a true category-specific deficit. *Neurocase*, *4*(4-5), 273–290.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, *2*(6), 459–473.

Sartori, G., Gnoato, F., Mariani, I., Prioni, S., & Lombardi, L. (2007). Semantic relevance, domain specificity and the sensory/functional theory of category-specificity. *Neuropsychologia*, *45*(5), 966–976.

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

Sartori, G., & Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, *16*(3), 439–452.

Sartori, G., Miozzo, M., & Job, R. (1993). Category-specific naming impairments? yes. *The Quarterly Journal of Experimental Psychology*, *46*(3), 489–504.

Schlitt, M., Lakeman, A. D., Wilson, E. R., To, A., Acoff, R. W., Harsh, G. R., & Whitley, R. J. (1986). A rabbit model of focal herpes simplex encephalitis. *Journal of Infectious Diseases*, *153*(4), 732–735.

Schroeter, M. L., Raczka, K., Neumann, J., & von Cramon, D. Y. (2007). Towards a nosology for frontotemporal lobar degenerations—a meta-analysis involving 267 subjects. *NeuroImage*, *36*(3), 497 - 510.

Schwartz, M. F., Marin, O. S., & Saffran, E. M. (1979). Dissociations of language function in dementia: A case study. *Brain and Language*, *7*(3), 277 - 306.

Schwartz, M. F., Saffran, E. M., & Marin, O. S. M. (1980). Fractionating the reading process in dementia: Evidence for word-specific print-to-sound associations. In *Deep dyslexia.* Routledge.

Seeley, W. W., Bauer, A. M., Miller, B. L., Gorno-Tempini, M. L., Kramer, J. H., Weiner, M., & Rosen, H. J. (2005). The natural history of temporal variant frontotemporal dementia. *Neurology*, *64*(8), 1384–1390.

Sérieux, P. (1893). Sur un cas de surdité verbale pure. *Revue de Médicine*, *13*, 733-750.

Shallice, T. (1987). The cognitive neuropsychology of language. In M. Coltheart, G. Sartori, & R. Job (Eds.), (pp. 111–127). Lawrence Erlbaum Associates, Inc.

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

Shallice, T. (1993). Multiple semantics: Whose confusions? *Cognitive Neuropsychology*, *10*(3), 251–261.

Shallice, T., & Cooper, R. (2011). *The organisation of mind*. Oxford, UK: Oxford University Press.

Sheridan, J., & Humphreys, G. W. (1993). A verbal-semantic category-specific recognition impairment. *Cognitive Neuropsychology*, *10*(2), 143–184.

Silveri, M. C., Daniele, A., Giustolisi, L., & Gainotti, G. (1991). Dissociation between knowledge of living and nonliving things in dementia of the alzheimer type. *Neurology*, *41*(4), 545–545.

Silveri, M. C., & Gainotti, G. (1988). Interaction between vision and language in category-specific semantic impairment. *Cognit Neurospsychol*, *5*, 677-709.

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

Sirosh, J., & Miikkulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation*, *9*(3), 577–594.

Skipper, L. M., Ross, L. A., & Olson, I. R. (2011). Sensory and semantic category subdivisions within the anterior temporal lobes. *Neuropsychologia*, *49*(12), 3419–3429.

Sköldenberg, B. (1996). Herpes simplex encephalitis. *Scandinavian journal of infectious diseases. Supplementum*, *100*, 8.

Sköldenberg, B., Alestig, K., Burman, L., Forkman, A., Lövgren, K., Norrby, R., ... others (1984). Acyclovir versus vidarabine in herpes simplex encephalitis: randomised multicentre study in consecutive swedish patients. *The Lancet*, *324*(8405), 707–711.

Small, S., Hart, J., Nguyen, T., & Gordon, B. (1995). Distributed representations of semantic knowledge in the brain. *Brain*, *118*(2), 441.

Smith, M., Lennette, E., & Reames, H. (1941). Isolation of the virus of herpes simplex and the demonstration of intranuclear inclusions in a case of acute encephalitis. *American Journal of Pathology*, *17*, 55-68.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardised set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 174-215.

Snowden, J. S., Bathgate, D., Varma, A., Blackshaw, A., Gibbons, Z. C., & Neary, D. (2001). Distinct behavioural profiles in frontotemporal dementia and semantic dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *70*, 323-332.

Snowden, J. S., Goulding, P. J., & Neary, D. (1989). Semantic dementia: A form of circumscribed cerebral atrophy. *Behavioural Neurology*, *2*, 167-182.

Snowden, J. S., Neary, D., & Mann, D. M. A. (2002). Frontotemporal dementia. *The British Journal of Psychiatry*, *180*, 140-143.

Solomon, K., & Barsalou, L. (2001). Representing properties locally. *Cognitive psychology*, *43*(2), 129–169.

Srinivas, K., Breedin, S. D., Coslett, H. B., & Saffran, E. M. (1997). Intact perceptual priming in a patient with damage to the anterior inferior temporal lobes. *Journal of Cognitive Neuroscience*, *9*, 490-511.

Stewart, F., Parkin, A. J., & Hunkin, N. M. (1992). Naming impairments following recovery from herpes simplex encephalitis: Category-specific? *The Quarterly Journal of Experimental Psychology*, *44*(2), 261–284.

Stroop, W. G., & Schaefer, D. C. (1986). Production of encephalitis restricted to the temporal lobes by experimental reactivation of herpes simplex virus. *Journal of Infectious Diseases*, *153*(4), 721–731.

Suzuki, K., Yamadori, A., & Fuji, T. (1997). Category-specific comprehension deficit restricted to body parts. *Neurocase*, *3*(3), 193–200.

Taylor, K. I., Moss, H. E., & Tyler, L. K. (2007). Neural basis of semantic memory. In J. Hart & M. Kraut (Eds.), (pp. 265–301). Cambridge University Press Cambridge,, UK.

Thomas, M. S. C., & de Wet, N. M. (1999). Stochastic double dissociations in distributed models of semantic memory. In *Connectionist models in cognitive neuroscience* (pp. 170–183). Springer.

Tien, R., Felsberg, G., & Osumi, A. (1993). Herpesvirus infections of the cns: Mr findings. *American journal of roentgenology*, *161*(1), 167-176.

Tippett, L. J., Grossman, M., & Farah, M. J. (1996). The semantic memory impairment of alzheimer's disease: Category-specific? *Cortex*, *32*(1), 143–153.

Tranel, D. (2009). The left temporal pole is important for retrieving words for unique concrete entities. *Aphasiology*, *23*(7-8), 867–884.

Tsapkini, K., Frangakis, C. E., & Hillis, A. E. (2011). The function of the left anterior temporal pole: evidence from acute stroke and infarct volume. *Brain*, *134*(10), 3094–3105.

Tulving, E. (1972). Organization of memory. In E. Tulving & W. Donaldson (Eds.), (p. 381-403). Academic Press.

Tulving, E. (1987). Multiple memory systems and consciousness. *Human neurobiology*, *6*(2), 67–80.

Twomey, J., Barker, C., Robinson, G., & Howell, D. (1979). Olfactory mucosa in herpes simplex encephalitis. *Journal of Neurology, Neurosurgery & Psychiatry*, *42*(11), 983–987.

Tyler, L. K., Bright, P., Dick, E., Tavares, P., Pilgrim, L., Fletcher, P., ... Moss, H. (2003). Do semantic categories activate distinct cortical regions? evidence for a distributed neural semantic system. *Cognitive Neuropsychology*, *20*(3), 541–559.

Tyler, L. K., & Moss, H. (2001). Towards a distributed account of conceptual knowledge. *Trends in cognitive sciences*, *5*(6), 244–252.

Tyler, L. K., Moss, H., & Jennings, F. (1995). Abstract word deficits in aphasia: Evidence from semantic priming. *Neuropsychology*, *9*(3), 354.

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*(2), 195–231.

Tyler, L. K., Stamatakis, E., Bright, P., Acres, K., Abdallah, S., Rodd, J., & Moss, H. (2004). Processing objects at different levels of specificity. *Journal of cognitive neuroscience*, *16*(3), 351–362.

Ultsch, A. (2003a). Maps for the visualization of high-dimensional data spaces. In *Proc. workshop on self organizing maps* (pp. 225–230).

Ultsch, A. (2003b). *U*-matrix: a tool to visualize clusters in high dimensional data.* Fachbereich Mathematik und Informatik.

Ultsch, A., & Siemon, H. (1990). Kohonen's self organizing feature maps for exploratory data analysis. In *Innc'90, int. neural network conf.* (p. 305-308). Dordrecht, Netherlands: Kluwer.

Utley, T. F., Ogden, J. A., Gibb, A., McGrath, N., & Anderson, N. E. (1997). The long-term neuropsychological outcome of herpes simplex encephalitis in a series of unselected survivors. *Cognitive and Behavioral Neurology*, *10*(3), 180–189.

Varela, J., Song, S., Turrigiano, G., & Nelson, S. (1999). Differential depression at excitatory and inhibitory synapses in visual cortex. *Journal of Neuroscience*, *19*(11), 4293–4304.

Verjans, G. M. G. M., Hintzen, R. Q., van Dun, J. M., Poot, A., Milikan, J. C., Laman, J. D., ... Osterhaus, A. D. M. E. (2007). Selective retention of herpes simplex virus-specific t cells in latently infected human trigeminal ganglia. *Proceedings of the National Academy of Sciences*, *104*(9), 3496-3501.

Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. (2000). *SOM Toolbox for MATLAB 5* (Tech. Rep.). Helsinki University of Technology.

Warrington, E. (1975). The selective impairment of semantic memory. *Quartlerly Journal of Experimental Psychology*, *27*(4), 635–657.

Warrington, E. (1981). Concrete word dyslexia. *British Journal of Psychology*, *72*(2), 175–196.

Warrington, E., & Cipolotti, L. (1996). Word comprehension: The distinction between refractory and storage impairments. *Brain*, *119*, 611–625.

Warrington, E., & McCarthy, R. (1987). Categories of knowledge. *Brain*, *110*(5), 1273.

Warrington, E., & McCarthy, R. A. (1983). Category specific access dysphasia. *Brain*, *106*(4), 859.

Warrington, E., & McCarthy, R. A. (1994). Multiple meaning systems in the brain: A case for visual semantics. *Neuropsychologia*, *32*(12), 1465–1473.

Warrington, E., & Shallice, T. (1979). Semantic access dyslexia. *Brain*, *102*(1), 43.

Warrington, E., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829.

Wechsler, A. (1977). Presenile dementia presenting as aphasia. *J. Neurol. Neurosurg. Psychiatry*, *40*, 303-305.

Weintraub, S., Rubin, N., & Mesulam, M. (1990). Primary progressive aphasia. longitudinal course, neuropsychological profile and language features. *Arch Neurol*, *47*, 1329-1335.

Westbury, C., & Bub, D. (1997). Primary progressive aphasia: A review of 112 cases. *Brain and Language*, *60*(3), 381 - 406.

Whitley, R. J. (2006). Herpes simplex encephalitis: Adolescents and adults. *Antiviral Research*, *71*(2–3), 141 - 148.

Whitley, R. J., Alford, C. A., Hirsch, M. S., Schooley, R. T., Luby, J. P., Aoki, F. Y., ... Soong, S.-J. (1986). Vidarabine versus acyclovir therapy in herpes simplex encephalitis. *New England Journal of Medicine*, *314*(3), 144–149.

Whitley, R. J., & Kimberlin, D. W. (2005). Herpes simplex: Encephalitis children and adolescents. *Seminars in Pediatric Infectious Diseases*, *16*(1), 17 - 23.

Whitley, R. J., & Lakeman, F. (1995). Herpes simplex virus infections of the central nervous system: Therapeutic and diagnostic considerations. *Clinical Infectious Diseases*, *20*(2), 414-420.

Whitley, R. J., & Roizman, B. (2001). Herpes simplex virus infections. *The Lancet*, *357*(9267), 1513–1518.

Williams, G. B., Nestor, P. J., & Hodges, J. R. (2005). The neural correlates of semantic and behavioural deficits in frontotemporal dementia. *NeuroImage*, *24*, 1042-1051.

Williams, R., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, *1*(2), 270–280.

Williams, R., & Zipser, D. (1995). *Gradient-based learning algorithms for recurrent networks and their computational complexity* (Y. Chauvin & D. E. Rumelhart, Eds.). Lawrence Erlbaum.

Wilson, B. A. (1997). Semantic memory impairments following non progressive brain injury: a study of four cases. *Brain injury*, *11*(4), 259–270.

Wilson, S. M., Brambati, S. M., Henry, R. G., Handwerker, D. A., Agosta, F., Miller, B. L., ... Gorno-Tempini, M. L. (2009). The neural basis of surface dyslexia in semantic dementia. *Brain*, *132*(1), 71–86.

Wisniewski, H., Coblentz, J., & Terry, R. (1972). Pick's disease. a clinical and ultrastructural study. *Arch Neurol*, *26*, 97-108.

Woollams, A. M., Ralph, M. A. L., Plaut, D. C., & Patterson, K. (2007). Sd-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*(2), 316–339.

Yamada, S., Kameyama, T., Nagaya, S., Hashizume, Y., & Yoshida, M. (2003). Relapsing herpes simplex encephalitis: pathological confirmation of viral reactivation. *Journal of Neurology, Neurosurgery & Psychiatry*, *74*(2), 262-264.

Yi, H., Moore, P., & Grossman, M. (2007). Reversal of the concreteness effect for verbs in patients with semantic dementia. *Neuropsychology*, *21*(1), 9.

Yochim, B. P., Kane, K. D., Horning, S., & Pepin, R. (2010). Malingering or expected deficits? a case of herpes simplex encephalitis. *Neurocase*, *16*(5), 451-460.

Zarafonetis, C., Smodel, M., Adams, J., & Haymaker, V. (1944). Fatal herpes simplex encephalitis in man. *American Journal of Pathology*, *20*, 429-445.

Zeki, S., Watson, J., Lueck, C., Friston, K., Kennard, C., & Frackowiak, R. (1991). A direct demonstration of functional specialization in human visual cortex. *The Journal of Neuroscience*, *3*, 641-9.

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of learning and motivation*, *44*, 35–62.