In: Horizons in Neuroscience Research. Volume 16 ISBN: 978-1-63463-174-7
Editors: Andres Costa and Eugenio Villalba

Chapter V

Semantics with or without Categorization*

Shin Asakawa†
Tokyo Woman's Christian University, Tokyo, Japan

Abstract

In spite of drastically increasing neuroimaging data, how semantic memory is organized has been a topic of controversy for decades. There still are contradictions shown by children in conceptual development and among symptoms shown by patients with category–specific semantic memory disorders. Several hypotheses have been proposed by cognitive– and neuro–psychologists based on empirical data and their models for semantic memory derived from this evidence. Therefore, it might be worth clarifying disagreements among them. We compared the differences between two major contentions about semantic categorization. One is WITH and the other is WITHOUT categorization hypotheses (Rogers and McClelland, 2011). Two hypotheses, marginalization operation and tree structures expressed in the anterior temporal lobes, were proposed as alternatives. Furthermore, essential problems in modeling semantic memory and categorization were discussed. They were (1) dimensionality of representation, (2) roles of intermediate layers, and (3) computational limitations of error-driven learning algorithms. According to these frameworks, considering that both WITH and WITHOUT categorization hypotheses could be similar might be possible.

Keywords: Semantic memory, Categorization, Neural networks, Concept development, Semantic Dementia

*This chapter was written to honor Dr. Timothy T. Rogers and Dr. James L. McClelland. The author was inspired by their book, Semantic Cognition: A Parallel Distributed Processing Approach (Rogers and McClelland, 2004). The chapter was also named in honor both of their writing "Semantics Without Categorization" (Rogers and McClelland, 2011) and the song of the Irish rock band U2, "With Or Without Your."

†E-mail address: asakawa@ieee.org

1. Introduction

A major contribution of the Parallel distributed processing (PDP) approach might enable the consideration of internal representations for cognitive processes. Until the back–propagation algorithm was developed, in general, constructing any cognitive models with detailed internal representations were impossible. Although perceptrons (Rosenblatt, 1958) could deal with complex tasks, there existed several limitations (Minsky and Papert, 1988). After Rumelhart et al. (1986) developed the generalized delta rule, we could begin to consider meanings of internal representations in which the activation of units in intermediate layers could be observed as magnitudes of activations. From this background, Rumelhart (1990) proposed a model to scrutinize the representation of semantics. He and his colleagues elaborated the model (McClelland and Rogers, 2003; McClelland et al., 2009; Rogers and McClelland, 2008; Rogers and McClelland, 2004, 2011; Rogers et al., 2004) and proved that the model could account simultaneously for both conceptual development and semantic degradation.

Children and normal subjects can acquire "basic" concepts earlier and faster than concepts at other levels (basic concept superiority). Patients with semantic dementia (SD) tend to answer "superordinate" concepts when asked the names of objects presented visually or verbally (superordinate concept preservation). Some literature expresses claims about success in simulating both phenomena. We focused on these problems about internal representations of semantic memory and revealed that the Rumelhart model would be neither the simplest nor the smallest architecture. Moreover, we discussed whether the categorization of semantic memory was required. Furthermore, essential issues concerning cognitive modeling and limitations were discussed.

Neuropsychologists (Warrington, 1981; Warrington and McCarthy, 1983, 1987; Warrington and Shallice, 1984; Warrington and McCarthy, 1994) have continued to explain the double dissociation between living and nonliving things in brain-damaged patients, especially patients with SD. Cognitive psychologists (Collins and Quillian, 1969; Collins and Loftus, 1975), on the other hand, have proposed hierarchical structures of semantic memory[1]. Rogers and McClelland (2004) and McClelland et al. (2010); Rogers and McClelland (2011) have proposed that categorization is not always required to explain semantic memory. Although this WITHOUT categorization hypothesis (henceforth, WITHOUT hypothesis) might be plausible, another WITH categorization hypothesis might also be possible. Here, we propose an alternative hypothesis following a WITH categorization framework. This alternative also implies that our semantic memories might be innately adaptive. If both hypotheses are true, it might also be clarified that semantic memory has duality. In this chapter, we discuss these problems concerning representations of semantic memory and its categorization.

Organization of This Chapter

Here, we address three points as follows:

[1]The spreading activation theories of the 1970s proposed that different category representations were connected in a graph structure that facilitated the "flow of activation" between categories that are "linked" in memory (Collins and Quillian, 1969; Collins and Loftus, 1975).

1. Necessity for hidden and representation layers
2. Dimensionality of data expression
3. Indirectness of simulations

This chapter is mainly intended to shed lights on the structure of semantics. However, we would like to refer to the additional problems as follows:

4. Basic concept superiority and superordinate concept preservation
5. Roles of hidden layers
6. Limitations as psychological models on error-driven learning algorithm

The essence of PDP models is that our knowledge would be represented in a parallel distributed manner. PDP models imply parallel distributed representations of knowledge. Knowledge is embedded as connection weights. As cited above, there are some problems that two-layered perceptrons cannot solve. With recruiting units in an intermediate layer and propagating errors through backward, this three-layered system obtains computationally better performance than that of perceptron. As a consequence, we could consider an internal representation that the model can provide. However, the model should align with biological reality (O'Reilly et al., 2012) and computational requirements (Marr, 1982) as much as possible.

First, in Section 2, we drew a rough sketch of WITHOUT hypothesis (Rogers and McClelland, 2008; Rogers and McClelland, 2004, 2011) and attempted to reveal several problems. After referring to a closely related study (Section 3), we discussed in turn, certain problems (the necessity of hidden and representation layers, the dimensionality of data expression, and the relation to empirical findings of basic concept superiority in development and superordinate concept preservation in SD). Then, we referred to more general and essential problems: roles of intermediate layers and limitations of gradient descent algorithms. These problems might give us a fertile perspective. Our considerations about the structure of semantic memory must make models plausible.

2. WITHOUT Hypothesis

Although Rogers and McClelland (2011) admitted that "Categorization is the core mechanism supporting semantic abilities," they also insisted that "Categorization is not the only efficient mechanism for storing and generalizing knowledge about the world." They enumerated the puzzles of the categorization of semantic memory as follows;

1. Multiple category representation
2. Category coherence
3. Primacy of different category structures in development, maturity, and dissolution
4. Domain-specific patterns of inductive projection

They also explained that "Furthermore, our framework suggests potential mechanisms that account for the variety of phenomena summarized above, and also provides some clues as to how the semantic system may be organized in the brain." However, their approach also invited other puzzles, which we attempted to describe. Some of them might be essential for PDP modeling. For instance, Rogers and McClelland (2011) explained that "The semantic system adopts a 'convergent' architecture in which all different kinds of information, regardless of the semantic domain or the modality of input or output, are processed through the same set of units and weights. This architecture permits the model to exploit high-order patterns of covariation in the sets of visual, tactile, auditory, haptic, functional, and linguistic properties that characterize objects."

However, the Rumelhart model is a feed-forward neural network. In general, there is no relation between the representation of intermediate layers when the input/output information is adopted and the inverse problem when adopted the reverse input/output information is adopted. We can show uncorrelated configurations of multidimensional scaling (Torgerson, 1952, 1965, henceforth, MDS) against their results when we reversely set the attribute data for the input layer and both the relation and attribute data for the output layer (Figure 4).

Figure 1 shows the Rumelhart model that they employed consistently (Rumelhart, 1990; McClelland and Rogers, 2003; McClelland et al., 2010, 2009; Rogers and McClelland, 2008; Rogers and McClelland, 2004, 2011; Rogers et al., 2004).

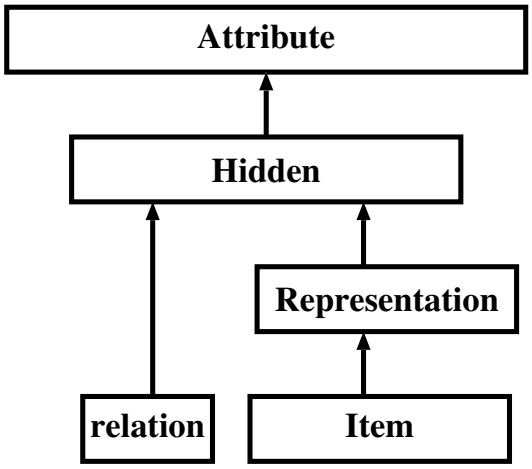

Figure 1. The Rumelhart model with five layers for hypothesis WITHOUT categorization.

Because Farah and McClelland (1991) employed a bidirectional model (Section 3) of Boltzmann machines, their model would be confirmed as the "convergent" property.

Furthermore, the Rumelhart model employs error-driven learning algorithms. Rogers and McClelland (2011) insisted that they modeled slow-learning semantic memory only. However, semantic memory must include declarative knowledge. For instance, "Penguins are in the Antarctic" or "Platypuses are mammals, but oviparous." There is no affinity between error-driven learning algorithms and declarative knowledge. Declarative knowledge has an all-or-nothing nature. We know it well or do not know it at all. Therefore, supposing

semantic memory without declarative knowledge is difficult. Error-driven learning algorithms are adequate for regression and classification. However, we must consider a type of one–shot algorithm to process declarative knowledge.

The data adopted in the Rumelhart model are depicted below for later discussion. Table 1 shows dependency among objects, two superordinate categories (plant and animal), and four subcategories (tree, flower, bird, and fish). Table 2 shows objects' identities. Tables 3,4, and 5 show features of objects, such as "isa", "can", and "has" relations, respectively.

Table 1. A part of data from Rogers and McClelland (2004, page 395, Appendix B.2.). Because all the items are living things, we deleted the two columns corresponding to "living" and "grow," as all their data were 1

Plant	Animal	Tree	Flower	Bird	Fish	
1	0	1	0	0	0	Pine
1	0	1	0	0	0	Oak
1	0	0	1	0	0	Rose
1	0	0	1	0	0	Daisy
0	1	0	0	1	0	Robin
0	1	0	0	1	0	Canary
0	1	0	0	0	1	Sunfish
0	1	0	0	0	1	Salmon

Table 2. Another part of data from Rogers and McClelland (2004, page. 395, Appendix B.2). This matrix is an identity matrix; all the diagonal elements are 1, and the rest are 0

Pine	Oak	Rose	Daisy	Robin	Canary	Sunfish	Salmon
1	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1

Tables 1 and 2 can be regarded as a structure of category. Both the tables make a tree structure. For example, the "pine" on the first row belongs to both plant and tree categories. This information can be written as a nested list similar to the computer language LISP, $((1,0)((1,0)(0,0))((((1,0)(0,0))((0,0)(0,0)))))$. A binary tree structure can be always expressed as a list, and, this list is equivalent to a binary tree structure. The list makes a vector with the 14 dimensions $\left(2^1+2^2+2^3=2+4+8=14\right)$, indicating an object, for

Table 3. "isa..." part of data from Rogers and McClelland (2004, page. 395, Appendix B.2)

Pretty	Big	Green	Red	Yellow	
0	1	1	0	0	pine
0	1	0	0	0	oak
1	0	0	1	0	rose
1	0	0	0	1	daisy
0	0	0	1	0	robin
0	0	0	0	1	canary
0	0	0	0	1	sunfish
0	0	0	1	0	salmon

Table 4. "can..." part of data from Rogers and McClelland (2004, page. 395, Appendix B.2)

Move	Swim	Fly	Sing	Skin	
0	0	0	0	0	pine
0	0	0	0	0	oak
0	0	0	0	0	rose
0	0	0	0	0	daisy
1	0	1	0	1	robin
1	0	1	1	1	canary
1	1	0	0	1	sunfish
1	1	0	0	1	salmon

example, a "pine."

Thus, the data expression adopted in Rogers and McClelland (2004) can be considered to have a hierarchical structure. Although the WITHOUT hypothesis insists on the existence of graphical representations of semantics, another hypothesis can still simultaneously claim a tree-structured representation. Note that both expressions can be translated to each other without the loss of information. Figure 2 shows an equivalent expressions of a three-layered binary tree.

Patterson et al. (2007) proposed the "distributed-plus-hub" hypothesis (Figure 3). According to the distributed-plus-hub view, a shared and amodal hub in ATL connects to various types (including modality-specific representation) of semantic representations, and communicates through them. These researchers considered that at the hub stage, associations between different pairs of attributes would all be processed by a common set of neurons and synapses, regardless of the task. This notion is analogous to the "convergent" architecture of Rogers and McClelland (2008); Rogers and McClelland (2004, 2011). Patterson et al. (2007) might consider that the hub in ATL would play roles which are equiva-

Table 5. "has..." part of data from Rogers and McClelland (2004, page 395, Appendix B.2)

Roots	Leaves	Bark	Branch	Petals	Wings	Feathers	Gills	Scales	
1	0	1	1	0	0	0	0	0	pine
1	1	1	1	0	0	0	0	0	oak
1	1	0	0	1	0	0	0	0	rose
1	1	0	0	1	0	0	0	0	daisy
0	0	0	0	0	1	1	0	0	robin
0	0	0	0	0	1	1	0	0	canary
0	0	0	0	0	0	0	1	1	sunfish
0	0	0	0	0	0	0	1	1	salmon

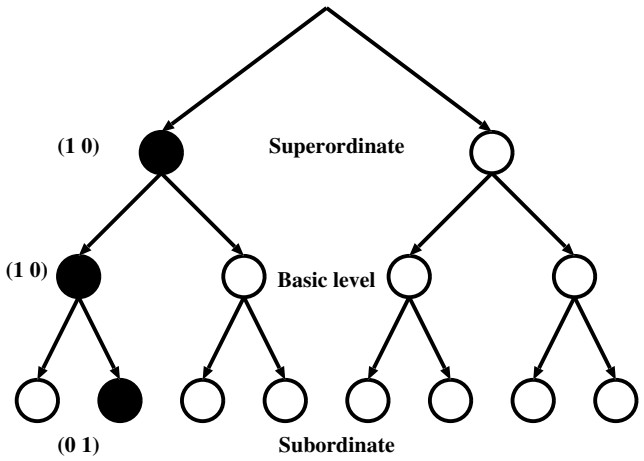

Figure 2. Three-layered binary tree structure equivalent to Tables 1 and 2.

lent to the hidden layer shown in Figure 1 (Patterson et al., 2007, Figure 1b, page 977). In the hidden layer shown in Figure 1, two flows of information converge into the hidden layer; one is from the representation layer and the other is from the relation (to identify a task) layer. Because the attribute (output) layer can be divided into several subgroups, Patterson et al. (2007) might propose such a viewpoint. The attribute layer had six subgroups, according to the information from the relation layer. The relation layer consisted of six units; three of them indicated as "isa" relations, such as the superordinate (general), basic, and subordinate (specific), and the other three units had meanings of relations, such as the ''is" (Table 3), "can" (Table 4), and "has" (Table 5), respectively. The network had to respond by activating one of the units in the subgroup, according to the information indicated by the units in the relation layer indicated.

However, we can propose another interpretation about the role of the hub in ATL. Because a hierarchy of tree structures was included in the data of the Rumelhart model, the

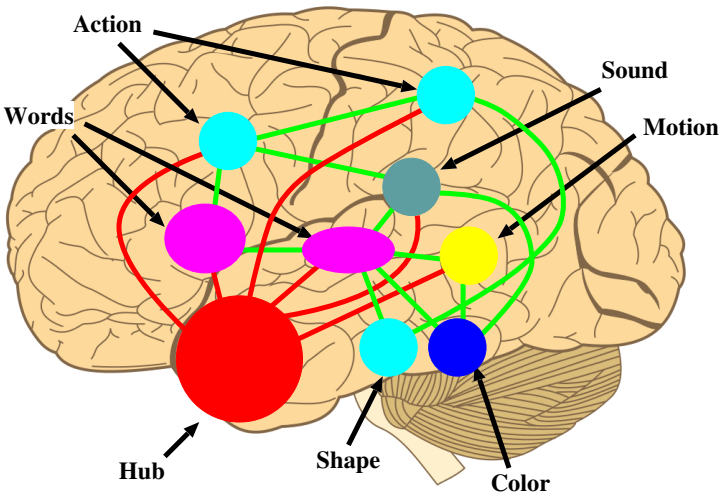

Figure 3. Schematic diagram of the "distributed-plus-hub" hypothesis. Redrawn from Patterson et al. (2007, Figure 1.b, page 977).

hub might encode the dependency of concepts. Guided by the hierarchical information presented in the hub, modality-specific information (sensory, motor, and linguistic) might be integrated into the concept of an object. The Rumelhart model is a feed-forward neural network model. Considering intermediate layers of the Rumelhart model as a convergent zone might not be impossible but might be unreasonable. Note that representations obtained in intermediate layers might be different from those obtained when the input and output data would be reversed. Figure 4 shows the results. The configuration in Figure 4 indicates a configuration of MDS, which was obtained from the original direction of information flow, whereas Figure 5 indicates a configuration of the reverse direction from the attribute to item layers.

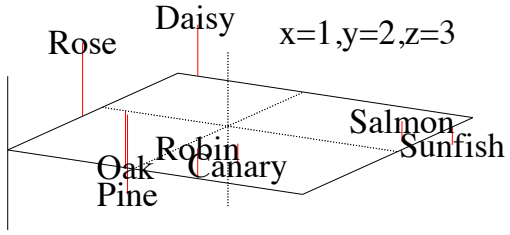

Figure 4. Results of MDS on units activation in a hidden layer. An original feed-forward flow of information from the item to attribute layers.

Each plot was drawn on the configurations of the three dimensions corresponding to the greater three eigenvalues in order, whereas, five eigenvectors which were greater than 1.0 were obtained (see Section 5 for the discussion of the dimensionality of MDS). We

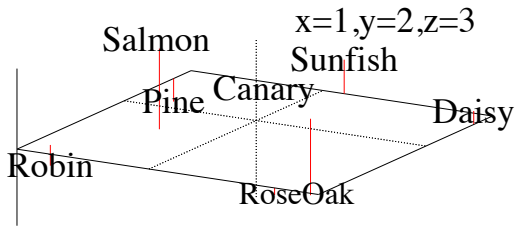

Figure 5. Results of MDS on units activation in a hidden layer. The reverse direction from the attribute to item layers.

must point out that the number of eigenvalues greater than 1.0 should be referred when performing MDS. We could not find any descriptions about these values in the literature (McClelland et al., 2010, 2009; Rogers and McClelland, 2008; Rogers and McClelland, 2004, 2011). It is important to know the number of eigenvalues greater than 1.0, because the number of dimension can be determined based on the number of eigenvalues to be extracted. Two-dimensional configurations might not be validated, or justified when we have more than 3 eigenvalues, whose real parts would be greater than 1.0. Kiani et al. (2007, Fig.5, page 4301) reported that when monkeys were shown natural and artificial object images, a minimum of five dimensions were required to express their neurons in the inferior temporal cortex.

On the contrary, we can ask a question: Can semantic memory be described as a two dimensional configuration, and such a two-dimensional configuration, that is, the best and the only one way to describe semantics? The fact that cortices have two-dimensional maps is only an existing proof. A necessary and sufficient condition must be proved in another way. The fact that the cortex has two-dimensions might have emerged accidentally, by chance, through evolution. Thus, another inference must induce the necessity of two-dimensional representation of semantics.

Furthermore, the relation between forward and inverse functions ($y = f(x) \leftrightarrow x = f^{-1}(y)$) is not always simple. Note that an inverse function, in general, cannot always be interpreted easily. Moreover, there exists a case that the inverse function cannot be determined analytically. Therefore, considering a model that assumes the flow of information in both directions in advance is necessary. We must take into consideration that entities of the inverse function represented are different from those of the original function. Feed forward networks like the Rumelhart model cannot be alternatives to replace bi-directional models like FM91 for the simplicity of computation.

Tables 3, 4, and 5 indicate features of objects. These tables are similar to the data of Plaut and Shallice (1993); Hinton and Shallice (1991); Farah and McClelland (1991). In Plaut and Shallice (1993); Hinton and Shallice (1991), the data were refereed as microfeatures. Rogers and McClelland (2008); Rogers and McClelland (2004, 2011) believed that our semantic memories consist of both the hierarchical tree structure of objects and microfeatures. On the other hand, Plaut and Shallice (1993); Hinton and Shallice (1991); Farah and McClelland (1991) employed only the microfeatures. Figure 6 depicts the correlation matrix among eight objects. The open and closed circles show positive and negative

correlations, respectively. The sizes of circles indicate strengths of correlations. The correlation matrix among microfeatures might be affected by semantic memory or categorization among objects.

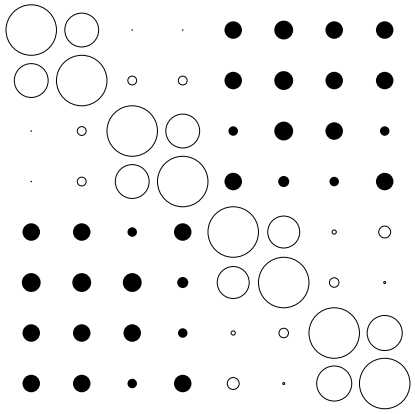

Figure 6. Correlation matrix of microfeatures Rogers and McClelland (2004, p395, Appendix B.2). Open and filled circles indicate positive and negative correlations, respectively. The size of the diameter of each circle represents the magnitude of correlation.

Tyler et al. (2000), for instance, constructed a stimulus set and attempted to explain the double dissociation to use this dataset. Figure 7 shows the correlation matrix they employed by them. The upper left 8×8 small matrix indicates the correlation coefficients among inanimate objects, and the lower right small matrix indicates the correlation coefficients among animate objects. As shown in Figure 7, when compared to animate objects, inanimate objects have relatively lower correlation coefficients within category. This suggests that animate objects share more features than inanimate objects. The correlation matrix among microfeatures might affect the organization of semantic memory. Although there were few studies to control correlation matrix as a dependent variable, an interesting question is how the correlation matrix affects an organization of semantic memory. Proposing a different group of simulations when considering random matrices consisting of correlated random numbers by Cholesky decomposition is possible.

Rogers and McClelland (2004, 2011) drew a configuration as a progressive function of a semantic memory organization, whereas they did not employ such configuration in then case of semantic degradation. If they could explain both phenomena within a framework of MDS, their hypothesis might further be plausible. They succeeded in explaining both basic concept superiority in development and superordinate concept preservation in degradation with one model; however, they did not try to explain both with one method, MDS. From the viewpoint of simplicity of explanation, one model to explain and one method to analyze for both phenomena would be desired.

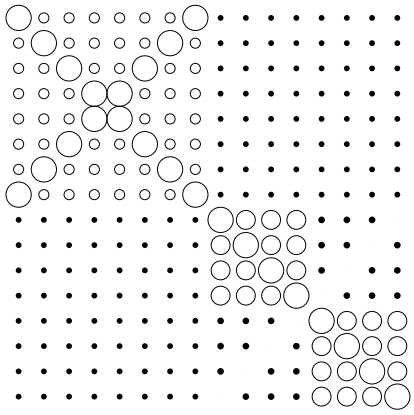

Figure 7. The correlation matrix adopted by Tyler et al. (2000).

3. FM91

The model proposed by Farah and McClelland (1991, henceforth, FM91) could explain picture naming and selection tasks simultaneously. This might be one of the main reasons they employed the Boltzmann machines. In addition, Dilkina et al. (2008) succeeded in explaining task performances of five patients with SD by adjusting parameters in their model.

Einstein said that "Everything must be made as simple as possible, but not simpler." The simpler the model, the better. The smaller the numbers of parameters, the better. If we could explain everything in a simple and single model, the model could be regarded as superior to other models. The Rumelhart model can also simultaneously describe both basic concept superiority in development and superordinate concept preservation in degradation simultaneously. It is worth seeking the integration of all models to explain every cognitive process. Toward a theory as a Grand Unified Theory (GUT) in particle physics or a Theory Of Everything (TOE) unifying gravity, we must explore a principle explaining all of cognitive processes. However, determining a model that very simple to explain everything, or a model that is very complex to describe a simple phenomenon might be difficult. Unfortunately, Einstein never referred to such a criteria about the limitation of simplicity. Because the FM91 model was proposed to simulate the visual/functional hypothesis (see below), and to confirm its validity[2],we first briefly introduce the hypothesis.

In neuropsychology, animate and inanimate concepts are well known to be separately declined by brain damage. There are brain-damaged patients who cannot name, discriminate, identify, categorize, or repeat animate objects, whereas their performance with inanimate objects remains intact. On the contrary, another type of patients has selective impairment with inanimate objects, whereas they show no deficits with animate objects. This

[2]According to the review of Thompson-Schill (2003), there are supportive evidence by neuroimaging studies in which there are correspondences to each visual and functional categories. Further, there also exist response selectivity: in case of selection the frontal lobe was activated, while temporal lobe is activated in case of retrieval.

double dissociation between animate and inanimate objects is suggestive when we think about the nature of human semantic memory organization. Warrington and her colleagues (Warrington, 1981; Warrington and McCarthy, 1983, 1987; Warrington and Shallice, 1984; Warrington and McCarthy, 1994) proposed that the visual/functional hypothesis, i.e., the knowledge about visual and functional features, is separately represented in our brains. Animate objects would share more visual properties than inanimate objects. On the other hand, inanimate objects are often defined by their functional features. Therefore, if visual semantic memory suffered damage, then knowledge about animate objects might selectively emerge. It could also be predicted that damage in functional semantic memory, on the contrary, might cause an inanimate object-specific deficit.

According to the visual/functional hypothesis, Farah and McClelland (1991) proposed a neural network model to explain this double dissociation between living and nonliving things. Figure 8 depicts the FM91 model. The FM91 model succeeded in showing category specificity, when these researchers destroyed the network. In a picture naming task, visual

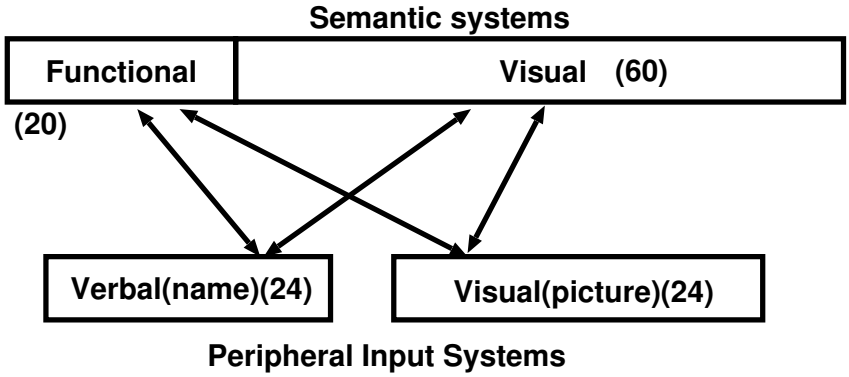

Figure 8. The FM91 model (Farah and McClelland, 1991). Digits in parenthesis indicate the number of units.

(picture) stimuli were set and the system was required to activate corresponding units in the verbal (name) layer. The system was trained to activate correct units in the verbal (name) layer when verbal (name) stimuli were presented. This condition was considered to correspond to the matching to sample task of Warrington and McCarthy (1983, 1987). The verbal (name) and picture layers had 24 units each. These two layers behaved as input and output devices. The peripheral layers were intermediated by semantic systems. The semantic layer could be divided into two components, functional and visual systems. The functional semantic system had 20 units and the visual semantic system had 60 units. Therefore, the visual/functional ratio was 1 : 3. There were no direct connections between the verbal and visual layers. The number of objects that the system had to learn was 10 living and 10 nonliving things. Four simulations were performed. The first two simulations were different in the training epochs. The third simulation was performed without weight decay. In the fourth simulation, functional and visual semantic systems each had 40 units. Farah and McClelland (1991) decided the number of units activated for each object based on the results of behavioral experiments. Living objects had 16.1 for visual and 2.1 functional

units activated in average, whereas nonliving objects were represented by the average of 9.4 visual and 6.7 functional units, respectively. Figure 9 shows the result of the Experiment 2 by Farah and McClelland (1991). This figure can be considered the FM91 model that could confirm the visual/functional hypothesis. We assumed that the curves in Figure9 as

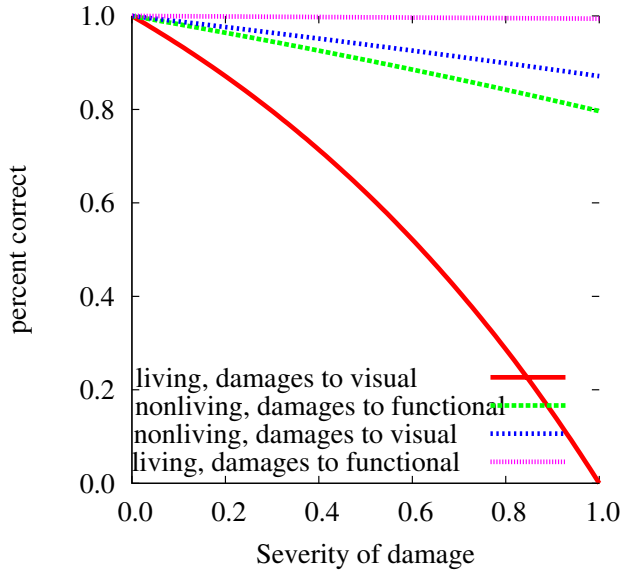

Figure 9. Results redrawn from the data of Farah and McClelland (1991, Figure 2).

functions of the severity of damages, are able to approximate to the function as follows:

$$f(a,x) = 1 - \frac{\exp(ax) - 1}{\exp - 1}, \tag{1}$$

where x indicates the severity of the damage. Then, we performed a least-square regression to estimate parameter values of a for each curve. The value a can be regarded as the indexing severity of the damage, $(0 \leq x \leq 1)$. The estimated values of a are shown in Table 6[3].

[3]We believe that any statistical tests should not be applied to make judgments for differences among conditions obtained in neural network simulations. If a neural network researcher decides a network topology, number of units, training data set, a learning algorithm, values of parameters, a random number generator, and a seed, then the result he or she will obtain is exactly the same deterministically. When he or she performs simulations with different seeds, results may vary. Small differences among seeds make big differences. The results completely depend on the random number generating algorithm they adopted. Statistical inferences, in general, depend on the property of sampling data. Statistical tests have been developed based upon the strong assumption that data reflects population. Average values of sample are assumed to distribute around average values of population. But this assumption does not assure in neural network simulations. There exists no warranty that results of neural network simulations can satisfy with assumptions that statisticians assume. Further, many statistical tests were developed to satisfy with some restrictions of experimental conditions or data resources. But the situation is different in neural network simulations. Neural network researchers can repeat their simulations until they would obtain any results they hope.

Table 6. The estimated values of a calculated from Figure 2 of Farah and McClelland (1991)

concept	damaged to	
living	1.00	0.01
nonliving	0.20	0.30

Several points about the FM91 model should be discussed. If living things have more information of visual semantic memory than nonliving objects, then category-specific disorders of animals might emerge when semantic memory suffers damages. Note that the FM91 model was postulated that visual semantic memory shared a larger area than functional semantic memory. Therefore, neither discussion nor simulation might be required.

The FM91 model has too many parameters to be determined in advance. Farah and McClelland (1991) decided that the ratio of visual to functional features, combining living and nonliving things, is 3 : 1 (60 visual semantic and 20 functional semantic units), based on the results obtained from behavioral experiments[4]. We agree with the hypothesis that visual attributes are more important than functional attributes for defining living things. But the ratio should not be considered strictly. If the FM91 model is adequate for semantic memory, another approach should be worth taking. Farah and McClelland (1991) tried another condition, the same numbers of units of visual and functional semantic systems (40 units each). How could we predict the performance of other ratios? How about changing an average of 16.1 visual and 2.1 functional units for living things to an average of 9.4 visual and 6.7 functional units? It is an interesting "reverse engineering" problem to observe the systems behavior as a function of the average values for visual and functional units for living and nonliving things. Although the average values were determined based on the results of behavioral experiments, the FM91 model might not necessarily work best with the values. Other sets of values might be more adequate for describing the results of behavioral experiments or performances of SD patients. Farah and McClelland (1991) might have not considered the internal correlation matrix among objects in the semantic memory system. When we employ microfeatures as an internal memory representation, we can always calculate the correlation matrix among objects shown in Figures 6 and 7. Correlation matrix can be the basis of similarity judgment, which in turn, can be the basis of category. As described above, different correlation matrices consisting of correlated random numbers by Cholesky decomposition might introduce a new viewpoint concerning semantic memory organization. Considering these possibilities, many parameters must be examined to analyze the FM91 model's performance. Because such a parameter space is too vast to confirm with Brute-Force manner, mathematical consideration must be performed. Necessary and sufficient conditions that the model must possess will be revealed to mimic human performances for both basic concept superiority in categorical development and superordinate concept preservation in SD; or although the numbers of units activated for both living and

[4]The ratio of 1 : 3 might be called a magic number in computer programming. Because Farah and McClelland (1991) had considered this ratio as sacred and inviolable, they did not doubt the meaning of it, except for Experiment 4.

nonliving things are good as a first trial, other supportive evidence might still be required to confirm that these numbers are true, because we do not know the mathematical property of the system. The values Farah and McClelland (1991) employed are just estimators. Therefore, we must be careful in interpreting the results. Otherwise, the values might be considered as a type of black magic. It might be suspected that the conclusion should be postponed until these values are confirmed as adequate. However, no such considerations have thus far been contemplated.

Finally, it should be highlighted that the FM91 model has the same problem as the WITHOUT hypothesis has. The correlation matrix among objects must play an important role in semantic memory. However, Farah and McClelland controlled the average activation ratio of living and nonliving things and the "magic" ratio of the units of visual and functional semantic systems. Farah and McClelland did not seem to pay attention to the correlation among objects. If so, the FM91 model cannot perform similarity judgment tasks. Concepts or categories of living and nonliving things might be affected by similarity among members belonging to the same concept. Therefore, the lack of information about the correlation matrix among objects might be disadvantageous for a model of semantic memory. However, because the FM91 model was derived from Boltzmann machines, the model can retrieve or rebuild complete information from incomplete data. This aspect of the model might be valuable to reemphasize.

4. Necessity for Hidden and Representation Layers

The five-layered model is neither the simplest nor the smallest architecture. We propose here a marginalization hypothesis as an alternative for semantic representation.

In sense of machine learning, model should be selected in accordance with the complexity of the given data. The data employed by the Rumelhart model (Rumelhart, 1990; McClelland and Rogers, 2003; McClelland et al., 2010, 2009; Rogers and McClelland, 2008; Rogers and McClelland, 2004, 2011; Rogers et al., 2004) can be solved with a perceptron, i.e., with no hidden layers. We can write that mapping information from relation and item layers to the expression on the attribute layer was simple and easy. To hypothesize hidden and representation layers, another reason is required. Again we quote that "the simpler the model, the better." The existence of hidden and representation layers is suspected, as they are convenient for computing inner representation and performing MDS. Because Rumelhart and his colleagues wanted to know internal expressions, they wanted to introduce these intermediate layers. We admire that the greatest advantage of the back-propagation algorithm is its ability to calculate all values of all units in all intermediate layers. However, the computational requirement differs from what our brains compute. The reason they want to know inner representations appears to be other from what is going on in real implementation. Rumelhart needed hidden and representation layers, and this might be the reason he introduced intermediate layers into his model. The existence of a representation layer cannot be justified as a computational requirement in the Rumelhart model.

Referring to Farah and McClelland (1991), we suspect that a picture-naming task that can deal with microfeatures as input data and a system must answer the corresponding names of items. Moreover, we can consider another task: picture selection task, i.e., when parts of microfeatures are given, the system must answer whole of them. Both tasks will

be required to be modality independent, and reaction method independent representations. We can settle a question such as "How can we get this type of an attribute-independent representation, whenever whatever we want?"

An operation to marginalize for an input modality, or an attribute of environment, is worth considering, because an operation of marginalization appears to be performable, every time we want. Marginalization, though an attribute, can also be regarded as to identify an abstraction. This operation might be similar to computing posterior probabilities to sum up any variables in Bayesian inference:

$$p(a) = \int_x p(a,x)\ dx, \tag{2}$$

where p does not always mean a probability.

If we sum up the computation of the relation layer, or marginalize for an input modality, we would obtain the same activation or configurations on one single hidden layer. We suspect that two intermediate layers (representation and hidden) were employed to draw a figure of an internal representation, independent from the relation information. The representation layer was supposed to be relation information free, whereas the hidden layer might contain the relation information. However, such a dichotomy might be a convenience for researchers. We can question the nature of internal representations with PDP models. Note that the task employed was so easy that no hidden layers were required. We can consider a perceptron as a model of semantic memory. Those who would like to apply a rather complex model should justify for doing so with plausible reasons. In its simplest form, Occam's Razor states that one should not make more assumptions than what are needed. When multiple explanations are available for a phenomenon, "the simplest version is preferred" (`http://www.spaceandmotion.com/Ockhams-Razor.htm`). The fewest assumptions should be selected among competing hypotheses; otherwise, any positive, possible, and strong reasons for the Rumelhart model having a five-layered structure should be required. Perceptrons cannot possess intermediate layers. It might be inadequate to employ any perceptrons as models of internal representations. However, when we consider that the task to be solved is a simple problem of classification or identification, a perceptron might be able to be one possible candidate. In the area of statistical decision theory, there are proposed criteria, AIC(Akaike, 1974), BIC(Schwarz, 1978), MDL(Rissanen, 1978; Grünwald, 2005), and NIC(Murata et al., 1994) are partially based on the log likelihoods, or Kullback-Leibler divergence $\int \ln \frac{p(x)}{q(x)}\ px$, and penalty terms as functions of the number of free parameters. The statistical decision theory can give us an optimal solution among models; however, it does not mean that the solution might be also be implemented in our brain. Therefore, we must question psychological meanings of intermediate layers of the Rumelhart model whether the Rumelhart model is the only model that can explain semantics. If we allow other models with direct connections between input and output layers, then the plots or configurations calculated from activations of units in the hidden layers might change drastically (Rumelhart and McClelland, 1986, compare Figure 4, page 64 with Figure 2, page 321 in volume 1)).

Can we conceive neural correlates of the representation layer in the Rumelhart model? McClelland et al. (2009) distinguished two kinds of semantic degradation:

Semantic Dementia (ATL damage), loss of central semantic knowledge.

Semantic Aphasia (temporoparietal or prefrontal damage) showing multimodal semantic impairments[5].

McClelland et al. (2009) also wrote that "We have hypothesized that the patient groups reflect the two primary ingredients in semantic cognition: semantic dementia reflects a degradation of the core conceptual knowledge, whereas semantic aphasia arises from a deficit in the regulation of semantic cognition." As McClelland et al. (2009) explained, "ATL might be considered as an action independent representation that is responsible for semantic dementia. However, the "convergent" property might still be required even in this case, although it might be possible to assume an intermediate representation."

One attractive feature of the constructive approach such as neural network modeling is that researchers can consider constructing any possible architectures concerning a phenomenon without any restrictions. Figure 10 shows one of the possible alternatives with direct connections.

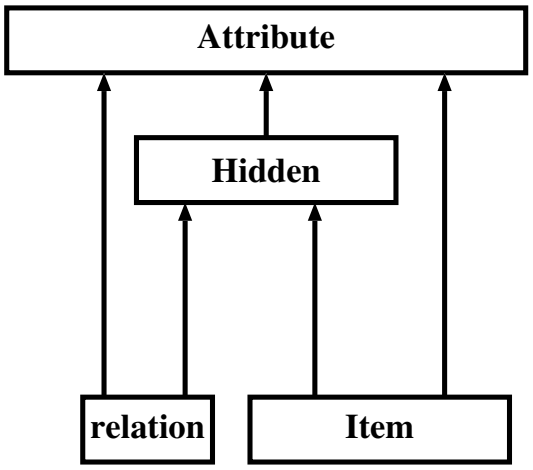

Figure 10. A model with direct pathways.

This alternative is worth considering, because there are no assurances that the Rumelhart model is the one and the only model to explain the phenomena of concern. Moreover, no disadvantages might not occur, even when assuming direct pathways. Rather, it is consistent with the spirit of the PDP approach. As Rumelhart and McClelland (1986) described, the role of intermediate layers changes drastically (see Section 7). We do not have any reasonable evidence, excluding direct connections between the input and output layers. Notably, direct connections across all the layers exist in the brain (Felleman and Essen, 1991, Figure 3).

Once we allow all direct connections similar to that done by Felleman and Essen (1991), we must consider further feedback, and within layer connections. That would make models move toward Boltzmann machines. When we allow all the direct connections like Felleman and Essen (1991), further feedback and within-layer connections should be considered.

[5]This distinction might correspond to the findings of Thompson-Schill (2003).

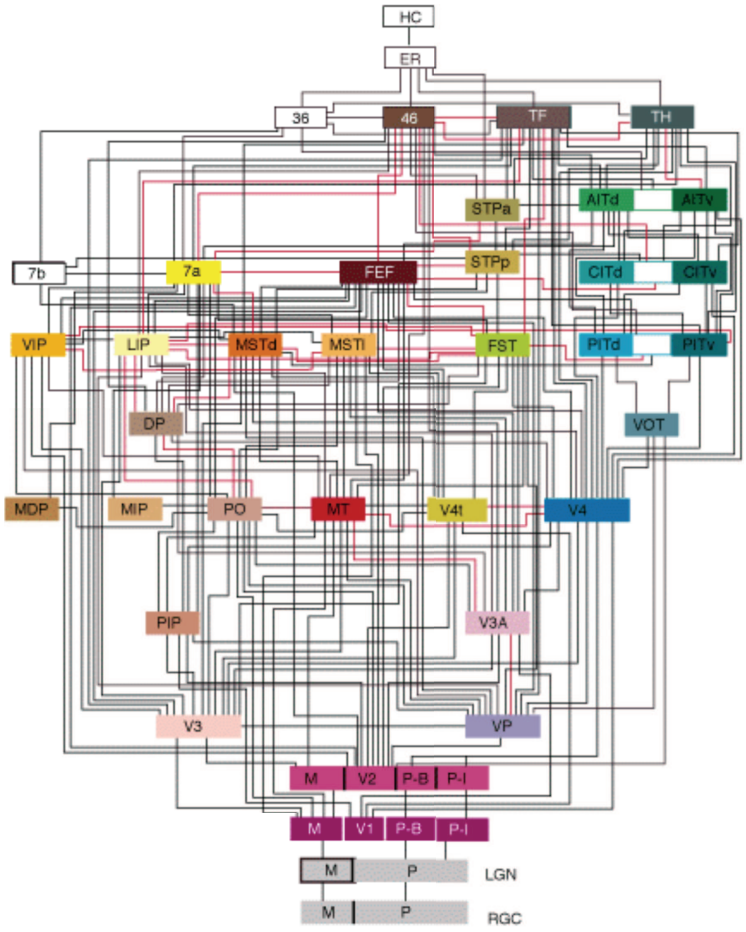

Figure 11. Redrawn from Felleman and Essen (1991) page 30, Figure 4.

Otherwise, this might lead us to a concept of the microcluster as in (Rumelhart and Zipser, 1985, Figure 2, page 85). Rumelhart and Zipser (1985) described that the microclusters having inhibitory connections within clusters and excitatory connections between layers. Direct pathways among layers might make it difficult to direct comparison among models. However, the biological reality insists that there are many connections among layers, as shown in Figure 11. Researchers should not complain about the inconvenience for model comparison. Without direct pathways, we can assume that all the information should be inspected by one layer. This suggests advantage for some system administrations. Because a direct pathway model cannot provide any configuration of MDS (at least MDS might not be applicable to models with direct and/or bypassed pathways among layers), Rogers and McClelland (2004, 2011) would be expected not to adopt any models with direct pathways. God never selects any models for the convenience of researchers.

Also, Dilkina et al. (2008, Figure 3) constructed a model with a direct connection between orthography and phonology in addition to a central hidden layer. However, these researchers could not succeed in introducing the direct pathway. A related statement was

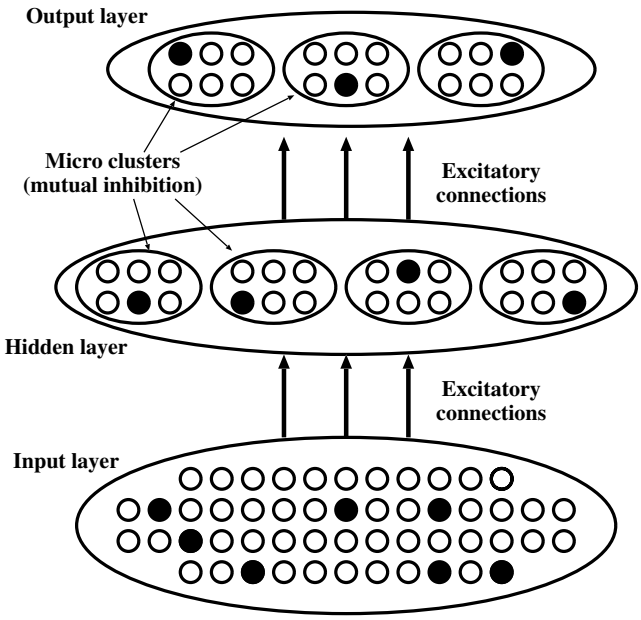

Figure 12. The architecture of the competitive learning mechanism. Redrawn from Rumelhart and Zipser (1985, Figure 2, page 85).

given by McClelland et al. (2009) as follows:

> Although the direct route tends to specialize in capturing typical spelling-to-sound correspondences whereas the pathway through the integrative layer tends to specialize in idiosyncratic word-specific information, the partitioning is not absolute, and neither pathway corresponds to a strict rule system or a strictly lexical system.

However, McClelland et al. (2009) did not describe any plausible reasons at all. They did not prepare any mechanism in advance to divide orthographic space into subspaces for conquering each word adequately. In multilayered perceptrons, in general, all units will participate to process all data. Moreover, all units are responsible for all data in cooperation. Therefore, another algorithm is required to process their tasks completely (see Section 8). Other alternatives are also indicated in Figure 13. Biological reality might be necessary to refute these alternatives (O'Reilly, 1998, see also).

Rogers and McClelland (2004, Figure 9.1, page 356) indicated the original Rumelhart model, Figure 9.1b meant a localist representation, which considered a log-linear model with second-order interactions. Figure 9.1c was the model with four subgroups of an intermediate layer. Each subgroup corresponded to an attribute. This model could be considered a model dealing with each attribute independently therefore, as the one that gathered a set of unrelated networks meaninglessly. On the other hand, Figure 9.1d can be regarded as significant. If we could sum up all the relations, we could get the accountable results accumulated through all the experiences. Nevertheless, such trails with no representation layers

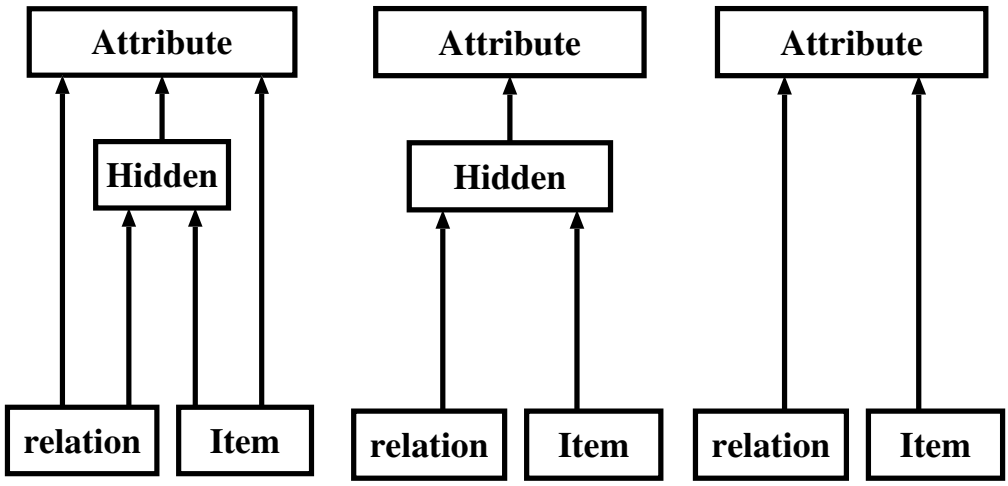

Figure 13. Alternative models of the Rumelhart model.

could not be performed. In addition, McClelland et al. (2009) considered the effect of numbers of units in the intermediate layer. The effect of numbers of layers might be uncertain when we adopt the standard back-propagation procedure (refer to discussions about the restricted Boltzmann machines (Hinton, 2002; Hinton et al., 2006; Salakhutdinov and Hinton, 2006, RBM)), because errors would diffuse to all units in the layers below.

We insist here that there is no necessity for the Rumelhart model to have five layers. The representation layer in the Rumelhart model exists only because researchers wanted to acquire representations for their own purposes. This might be the only reason they introduced the representation layer. We cannot think of any other plausible reasons. Models that satisfy only one's own purpose should be avoided. However, neural network modeling and simulations would still be required to understand human cognitive functions, because we do not know types of microfunctions and/or data representation will be required in detail before we write a well-functioning program in advance. However, even such a program might not assure that our brains would employ the same algorithm. Functional resemblance gives us a necessary but not an efficient condition. Therefore, the meaningfulness of a model should be measured by other criteria. Basic concept superiority and superordinate concept preservation should be considered, along with the existence of the representation layer, because we can get reasonable solutions without any hidden layers for their tasks. Procedures for neural network modelings are very powerful such that almost all problems might be solved without any constraints. This might imply that we should not employ any models only because they can solve any given task. Generalization, plausibility, adequateness, and other possible reasons should be considered in order for modelers to identify their models as cognitive processes. Processes of acquisition and the decay of knowledge should also be considered. We can criticize the Rumelhart model because of the lacks of two types of connections: within layer and feedback connections. Because these connections can emerge time-dependent properties as dynamical systems and different principles must be considered to understand such systems, we do not use these connections here. As O'Reilly

et al. (2012) wrote:

> In this sense, a model of cognitive neuroscience is just like any other 'theory', except that it is explicitly specified and formalized, forcing the modeler to be accountable for their theory if/when the data don't match up. Conversely, models can sometimes show that when an existing theory is faced with challenging data, the theory may hold up after all due to a particular dynamic that may not be considered from verbal theorizing.

5. Dimensionality of Data Expression

The classical MDS (Torgerson, 1952, 1965) results in the eigenvalue problem. Magnitudes of eigenvalues are important to be noted to determine the dimensionality of the given data on MDS. Because Torgersons MDS was developed based on the eigenvalue problem, the number of dimensions with nonnegative real maximum eigenvalues greater than 1.0 must be considered to properly evaluate the dimensionality of configurations of MDS. Otherwise, we might incorrectly interpret the dimensionality of the results acquired. Kiani et al. (2007) performed, for example, MDS to neural activities in IT neurons of monkeys and obtained five dimensional configurations. Because monkeys were given only pictures while recordings and no declarative knowledge, visual information might include a minimum of five-dimensional information. Considering visual information as a type of information about microfeatures, more than five dimensions are required to express all information for human semantic knowledge. However, the matrix of Appendix B3 in Rogers and McClelland (2004, page 396) consisted of two submatrices, "isa", and microfeatures Rogers and McClelland (2004) showed only two-dimensional configurations. Although Rogers and McClelland (2004, Figure 3.6, page 100) discussed eigenvectors, they did not mention that eigenvalues corresponded to eigenvectors. Therefore, their discussions cannot be justified because of the lack of information about the dimensionality of MDSs they performed. As shown in Figure 4, we could obtain a minimum of three-dimensional configuration with the data of Rogers and McClelland (2004). This result might be worth reconsidering in terms of the dimensionality of the representation of semantic knowledge. It would not mean that the configuration of semantics must obtain two-dimensional mappings because our cortex consists of two-dimensional space.

As described earlier, while constructive approaches have been attracting researchers, there is no assurance for the correctness of such models, even if they provide good performance. Good performance does not always indicate that a model is good. In other words, a good model always satisfies necessary and sufficient conditions because the possible presence of a network and its information processing in vivo might not be uniquely determined. There are no reasons to exclude other possibilities or hypotheses. About a quarter-century ago, Crick (1989) described the back propagation as an "alien technology." Crick (1989) criticized that neural network models as alien reverse engineering. Neural network modelers believe that computers can mimic humans cognitive functions. However, people asking emphatically that "What are calcium channels and how does fluorescence imaging of neural activity work?" is a large population in the related fields. However, the further we go with this notion, the sooner we can reconcile the alien technology with the much older notion

the "Turing test" (Turing, 1950). Turing wrote:

> The fact that Babbage's Analytical Engine was to be entirely mechanical will help us rid ourselves of a superstition. Importance is often attached to the fact that modern digital computers are electrical, and the nervous system is also electrical. Since Babbage's machine was not electrical, and since all digital computers are in a sense equivalent, we see that this use of electricity cannot be of theoretical importance. [...] If we wish to find such similarities we should look rather for mathematical analogies of function.

Multiple realizability is another aspect of a contemporary interpretation of the Turing test. Moreover, multiple realizability and biological reality (O'Reilly, 2001; O'Reilly and Munakata, 2000) are sometimes cooperative, but sometimes exclusive each other. The computer implementation of a model inevitably includes some degree of abstraction. Resemblance is a key to imitation and mimicry. Functional resemblance is also a key to simulation. Validity, descriminability, predictability, reproductivity, and biological realizability must be considered as much as possible. Aims are different from psychology, cognitive science, and machine learning. Differences in aims might produce different implementations. Although jet planes were developed on the aerodynamics of birds wings, their operating principles and performances differ. When we analyze crashed airplanes, we do not obtain any suggestions about methods of repairing birds wings. Does the Rumelhart model resemble with the human or the aliens brain? Do diagnoses of network performances with removed units correspond to diagnoses of brain-damaged patients?

The reason why MDS was not applied to superordinate concept preservation in neuropsychology would need to be explained. Similarly, the reason that the "eight times problem" was not adopted to the basic concept superiority in developmental psychology might also remain unclear. For the latter, there may be a tacit agreement that identifies the learning of the back propagation with the developmental psychology of semantic concepts. A final configuration of MDS and its progressive process on the way to obtain such a configuration appear quite different. For example, when we learned the knowledge "Penguins are in the Antarctica," it did not imply that penguins living in the Amazon jungle had moved gradually to Antarctica. Although Amazon.com would deliver almost anything, even it might not be able to deliver penguins to Antarctica on our semantic space. This type of declarative knowledge must be learned by the one–shot algorithm, not a gradient decent algorithm, such as back-propagation. Therefore, we can not compare an average error rate percentage of each graders with an error curve of a single simulation.

6. Basic Concept Superiority and Superordinate Concept Preservation

Basic concept superiority effect can be enumerated as in Rogers and McClelland (2011, chapter 5, page 176);

1. Children first learn to label objects with their basic-level name, instead of with more general or more specific names.

2. In free naming tasks, adults prefer to name at the basic level even when they know more general or specific names.

3. Adults are faster in verifying category membership at the basic level.

4. Adults are quicker to verify properties of objects that are shared by exemplars of a 'basic' category.

Rogers and McClelland (2004) considered progressive expansions of MDS as corresponding changes of concept development; "basic concept superiority." On the other hand, they introduced another method to show degradations of semantic memory, "superordinate concepts preservation." One principle might be better to describe both phenomena than the application of different principles. Changes of MDS configurations might be observed when some units in hidden and/or representation layers after learning completing learning. Comparing normal configurations of MDS with damaged ones must provide interesting insights about whether the categorization of semantics should be required. The explanation of basic concept superiority and superordinate concept preservation with a single discipline would provide a supportive evidence to the WITHOUT hypothesis. We could not find any explanations of reasons that Rogers and McClelland (2004, 2011) did not conduct MDS with unit destruction to simulate patients behaviors with SD.

Furthermore, if Rogers and McClelland (2004, 2011) had hypothesized that an error-driven learning algorithm could simulate concept developments, they, at least, should provide some evidence that it would be superior to other algorithms, such as the Hebbian (Hebb, 1949), or Contrastive Divergence (Salakhutdinov and Hinton, 2006), because the progress of learning and the configuration of MDS obtained as a result are quite different. To identify learning processes of the back-propagation algorithm with the algorithm of conceptual development, some assumptions are required. There exist limitations and conditions of the back-propagation algorithm (see Section 8).

6.1. Basic Concept Superiority in Development

Because graphical representations (i.e., MDS) are configured from correlation matrices among objects, all information is presented in the matrices. The classical multidimensional scalings proposed by Torgerson (1952, 1965) can be regarded as rewritings of eigenvalues and corresponding eigenvectors of the matrices. Therefore, units in the models's intermediate layers might contain similar types of information as the matrices. We previously showed the correlation matrix in Figure 6, which indicates two large clusters, and that each cluster has two sub clusters; the Rumelhart model appears to reflect this.

Patterson et al. (2007) proposed the "distributed-plus-hub" hypothesis (Figure 3). According to this hypothesis, modality-specific semantic memories are distributed in the brain. Pieces of information would each be processed in the brain. The ATL would play an important role to integrate them as an entity. If the "distributed-plus-hub" hypothesis is true, both microfeatures and tree structures would be necessary for normal functioning of semantic memory. The Rumelhart model also contains a tree structure, as a nested list expressions, such as LISP (cited as Table 1 and 2). Note that expressions of semantic categories in the ATL might be interpreted as a hierarchical list. If the "distributed-plus-hub" hypothesis

is true, and the ATL plays a role as a hub, then it appears to be an interesting idea that tree structures might be represented in the ATL and any other properties (visual, verbal, functional, and other modality-specific information) are represented in other areas.

6.2. Superordinate Concept Preservation

Rogers and McClelland (2004, 2011) performed MDS to make configurations among concepts and their progressions. However, they did not perform MDS to demonstrate the degradation of semantics. It appears consistent when the same method for both concept development and degradation was employed. If they succeeded in explaining the "superordinate concept preservation" in SD in MDS configurations when destroying units in intermediate layers, their WITHOUT hypothesis might strengthen further. However, we could not find any evidence that they employed MDS for describing SD, or superordinate concept preservation. We would like to reintroduce Einsteins thought:

If you can't explain it simply, you don't know the subject well enough.

Scientific reasoning must follow such simplicity.

Rogers and McClelland (2004, 2011), rather, introduced the "eight times problems" (see Section 6.3) to explain superordinate concept preservation. The "eight times problem" appear to be indirect for proving their claim. The drawing configuration of MDS with destroying units in intermediate layers might be more direct than the "eight times problems." Deleting units must be performed in accordance with the information of the Hessian matrix (Hassibi et al., 1993; LeCun et al., 1990). The greater the amount of information the unit can transfer, the more important the unit is. When an error gradient descent method was employed, effects of deleting units in intermediate layers should be evaluated in accordance with the Hessian information; at least, items that were likely to be reported as dog had to be investigated. If Rogers and McClelland (2004, 2011) could succeed in showing that some specific or basic levels of concepts were likely to be reported as superordinate concepts, their assumption would be validated. However, a question whether frequency is the only reason that superordinate concepts were reported more frequently than those of other levels of concepts would remain unresolved. If we could assume that the information about the tree structure of concepts was stored in the ATL, another explanation for superordinate concept preservation might be possible.

The Input data of Rogers and McClelland (2004, 2011) consisted of four components: a tree structure equivalent (Table 1), an identity (Table 2), microfeatures (Table 3,4, and 5), and relation matrices. The identity ("item" layer in Figure 1) and relation matrices were for input, and the tree and microfeatures matrices were for output. We can point out that there is regularity among the items in the matrix to represent the tree structure, whereas we can hypothesize that there is no correlation between items in the microfeature matrix. When "plant" is "1", "animal" is always "0", either "tree" or "flower" is "1", neither "bird" nor "fish" is "1". The probability of an item belonging to a "general" or superordinate category is 0.5, to a basic category is 0.25, and to a "specific" or subordinate category is 0.125. Therefore, we can predict that the system to tend to respond the most frequent category among all the possible answers when they suffered damages. That might be one reason we can observe the superordinate concept superiority in neuropsychology. As described above,

the output matrix consisted of the tree structure and microfeatures data. It can be supposed that the tree structure was dominated by the first-order statistics, frequencies. On the other hand, microfeatures were dominated by the second-order statistics, correlations. Like FM91, microfeatures of Plaut and Shallice (1993); Hinton and Shallice (1991) had greater correlations within category than those between categories. The system might respond to the most frequent item when suffering damages. This might be one possible explanation of superordinate concept preservation in degradation of semantic categories. Therefore, we can propose a hypothesis about the role of the ATL: the ATL integrates all modalities and represents hierarchies of semantic concepts.

6.3. Eight Times Problems

Rogers and McClelland (2004, 2011) introduced a frequency effect to explain "superordinate concept preservation" in SD. It might be considered that systems degradation for semantic memory would be described as a function of times that subjects and models were exposed to environmental stimuli. Because their hypothesis implies that the more times they were exposed, the more they would lose contents and details, they appeared to perform simulations with eight times more frequent "general" items, whereas other items remained at the same frequency. We, here, refer their hypothesis to the eight times problem. As the eight times problem, we refer, for example, to McClelland et al. (2009, Figure 72.6) Rogers and McClelland (2004, Figure 5.11, page 214), and Rogers and McClelland (2004, Figure 5.12, page 216). It appears strange to us that they did not employ eight times problem in case where Rogers and McClelland (2004, 2011) demonstrated "basic concept superiority" effects. On the other hand, they employed "eight times problem" in the case of degradation to simulate brain-damaged patients performances with SD. If frequency effect, or the eight times problem, is a key that affect all performances, both development and degradation should show the same effects. Then, MDS should describe trajectories of any performances for both learning and decaying processes. Rogers and McClelland (2004) gave an account for this eight times problem as follows (Rogers and McClelland, 2004, Figure 5.12, page 216)):

> Because the network is trained so frequently with the input dog paired with various relations, it learns to use the name dog even before it has successfully differentiated this item from other animals. Consequently, similarity based generalization leads the network to over-extend the name to other animals early in learning. As the dog representation is differentiated from the other animals (first from the birds and fish, and later from the other mammals), the network can learn to associate different responses with each, and the tendency to over-extend the name is gradually eliminated.

However, their account is for description about overlearning, not for the preservation of a level of concept when suffering damage. Neural networks, in general, can learn a proposition: $P \rightarrow Q$, and they also learn $\neg P \rightarrow \neg Q$, because a value to be minimized is often defined as a sum of squared differences between output and teacher signals. This implies that "if P then Q," and "if not P then not Q," where there is no asymmetry and anisotoropy between P and $\neg P$.

When we consider neural network models as mappings from input $((x_i, y_i), i \in 1, \ldots, m)$ to output (y) spaces, units in an intermediate layer form a basis, but it is neither an orthogonal nor orthonormal basis. Therefore, each unit is not independent. Each unit in an intermediate layer cooperates and contributes to the distribution of all the output signals. However, this does not always imply that all units acquire the same function. In case of an exclusive OR problem, for example, a three–layered perceptron with two hidden units gets a solution that one unit can solve OR, and another unit can solve AND (see 7). The weight between the OR and output units is positive, whereas that between the AND and output units is negative. The AND unit inhibits its activity, when two input units are "1."

Units in an intermediate layer form a basis, which is, however, not orthonormal nor orthogonal. All units in an intermediate layer contribute each datum equally. Items with low frequency might be dealt with exceptions. This is a "credit assignment problem" (Plaut et al., 1996). But we cannot know in advance what type of credit will be assigned. Suppose that a unit in an intermediate layer suffered damage. Other undamaged units will work precisely because these units suffered no damage. The systems behavior will be changed by the role played by the damaged neuron. Intact units will follow a probability to behave corresponding to the frequency of learned items. Figure 14 shows a schematic of such a situation in an exclusive OR problem. In Figure 14, the left 4 units ($\sum_{i=1}^{4} w_{oh_i} = 1$) play a

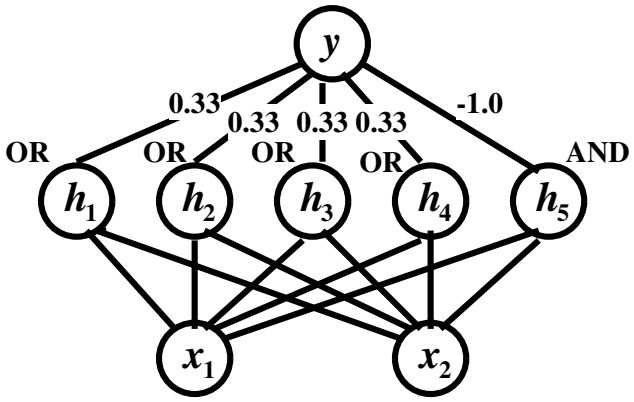

Figure 14. A possible solution for XOR problem with hidden units of more than 3.

role equivalently to an AND circuit. Notably, $\sum_i w_{oh_i} > \alpha$, and $w_{oh_5} \ll \alpha$ always hold. If the AND circuit is intact, then the damaged network will be able to provide right answers to three among four learned items. Therefore, the OR circuit is likely to be acquired as a most frequent rule (75%), whereas the AND circuit is dealt with as an exception (25% of frequency). In other words, an exception rule (AND circuit in this case) must be acquired defeating frequent rules (OR). This analogy can be expressed as "Gaussian ocean with Dirac's islands" (Asakawa, 2014).

Whether a unit in intermediate layers will be activated when a certain input is exposed is determined stochastically. Because all units in intermediate layers contribute to all data, numbers of units for exceptions would appear inevitably and relatively decreased. If such units are damaged, the system will fail to answer such exceptions correctly. However, it

is questionable whether answers to such exceptions would be replaced by more frequent items. Although the system learns a proposition , such as “if P then Q”, it does not learn what should be answered in case of ”if not P.”

As asserted in the Bible: “God is fair,” there is no inequality among units. Every unit has the potential to play any role equally. Again, because “God is fair,” output units cannot distinguish routes of connections: which connection comes from which layer, because the output units have no monitors. Thus, we consider meanings of units in intermediate layers in the next section.

7. Roles of Hidden Layers

The Rumelhart model (Rogers and McClelland, 2004, 2011) and its tasks were so easy that no hidden layers were required for completion. If so, why and how do hidden layers exist? What are the roles or meanings of hidden layers? Here, we consider possible reasons for hidden layers in a general framework. Introducing direct pathways makes a drastic change in internal representations. In an exclusive OR problem, we need at least two hidden units for a solution. However, only one unit is required for a solution when we allow direct connections between input and output units (Rumelhart and McClelland, 1986). We will take this problem into consideration in detail in Section 7.

At least 6 roles of hidden layers can be enumerated:

1. Higher-Order Interactions by Log-Linear Models
2. Controlling Dimensionality (Reducing and Expanding of Dimensions) like Principal Component Analysis (PCA) and RBM (Hinton, 2002; Hinton et al., 2006; Salakhutdinov and Hinton, 2006)
3. Blind Source Separation, Independent Component Analysis (ICA) (Hyvärinen et al., 2001; Hyvärinen and Oja, 2000; Bell and Sejnowski, 1995; Comon, 1994)
4. Planning, Prediction, Control, Retrieval, Restoration, or Processing Priming Stimulus by the Simple Recurrent Networks (SRN) (Elman, 1990, 1991, 1993; Elman et al., 1996)
5. Extraction of Invariant InformationMarr (1982)
6. Topological Mapping, Self-Organization Mapping(Kohonen, 1985)

Feed-forward neural network models can be regarded as log-linear models. Learning in feed-forward neural networks can be identified as finding adequate connection strengths between units. These values (including higher interaction terms) consist of a group of basis functions. We will consider this problem in section 7. In controlling dimensionality, reducing dimensionality may include the principle of parsimony (or Occam’s razor) such as PCA. In deep learning, multilayered architectures have an important role for improving their performances, whereas the standard back-propagation algorithm does not possess such advantages. Salakhutdinov and Hinton (2006) and Hinton et al. (2006) observed that errors would be propagated to all units in all layers below, therefore spreading errors to

all units that might show worse convergence properties. Because the restricted Boltzmann machines (Hinton, 2002; Salakhutdinov and Hinton, 2006) consist of binary units that will be activated stochastically, we cannot introduce the back-propagation rule that is differentiable. And, once an RBM was trained, all the connection weights between units would be fixed. Therefore, in principle, the overspreading of errors might not occur in RBM. RBM has been employed for both reducing and expanding dimensionality. Blind source separation and ICA have been used as denoising techniques. These are regarded as signal separators or rectifiers. ICA plays an important role in reducing dimensionality. Increasing dimensionality or an addition of dimensions would give us drastically improvements, as an exclusive OR problem (Rumelhart and McClelland, 1986). Reducing dimensionality, or the reduction of dimension, and/or the denoising process would also play an important role in information processing.

SRN has abilities to predict, plan, control, retrieve, restore, and so on. Moreover, attractor networks proposed by Hinton and Shallice (1991) and Plaut and Shallice (1993) can also be regarded as recurrent networks. Attractor networks could be employed to simulate performances of patients with deep dyslexia. Thus, memory retrieval and its disorder can be mimicked with these hidden layers.

The extraction of invariant information is essential for perception(Marr, 1982), such as location, color, rotation (Affine transform), or modality free information. Orientations, edges, circles, or gradation detections can be also applied to this operation.

Topological mapping, or self-organization mapping Kohonen (1985), are also another important role of hidden layers. Topological mapping, in other words, the self-organization mapping principle is often observed in primary cortices in visual, auditory, tactile, and somatosensory information processing. This principle appears to be general and universal in neural information processing. As Rumelhart and Zipser (1985) indicated, winner–take–all circuits in Figure 12 (or mutual inhibition) in microclusters might be an important computational principle in our brains. Softmax function $f(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$ might be employed to implement winner–take–all circuits. We can point out a relation between microclusters(Rumelhart and Zipser, 1985) and convolutional deep belief networks for scalable unsupervised learning of hierarchical representation (http://videolectures.net/icml09_lee_cdb/).

These roles of intermediate layers are not independent of each other. But all of them should be considered as candidates for implementing cognitive functions.

Thought Experiments of XOR

Here, we focus on a case of an exclusive OR problem. Figure 15 indicates an example.
To solve the exclusive OR problem in three-layered-feedforward networks, we need two units in a hidden layer. One unit stands for an OR circuit and the other for an AND circuit. In this case, the connection weight from the OR to the output is positive, and the weight from the AND to the output is negative. Therefore, the system behaves like an AND circuit in the case of the input signals being (0,0), (0,1), (1,0). However, the systems output is inhibited when the input signals are (1,1). Although the OR and the AND units in the hidden layer play different roles, the only difference between the two hidden units is the thresholds values. When we suppose the formal neuron proposed by McCulloch and Pitts (1943) (in case of a logistic function, $f(x) = 1/(1+\exp(-\alpha x))$, the logistic function is

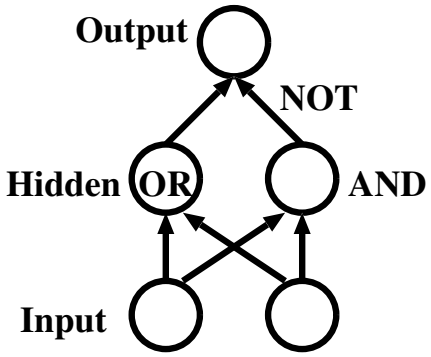

Figure 15. A possible solution for XOR problem without direct pathways.

approximated to a step function in the limitation of $\alpha \rightarrow \infty$), both logic functions, OR and AND, can be expressed as the difference between the threshold values. Thus, a combination between two logic functions makes an exclusive OR circuit. We can here consider the OR unit as dominant and the AND circuit as irregular, because 75 % of answers were correct when the OR unit was destroyed. On the other hand, when the AND unit was destroyed, 50 % would be correct in (0,0) $\rightarrow$ (0), and (1,1) $\rightarrow$ (1).

Another solution is indicated in Figure 16 (Rumelhart and McClelland, 1986, redrawn from, Figure 2, page 321, chapter 8, vol. 1). Direct pathways were introduced between the input and output units. Note that only one unit was required in this case, although the exclusive OR problem is not liner separable in the two-dimensional space composed of input units.

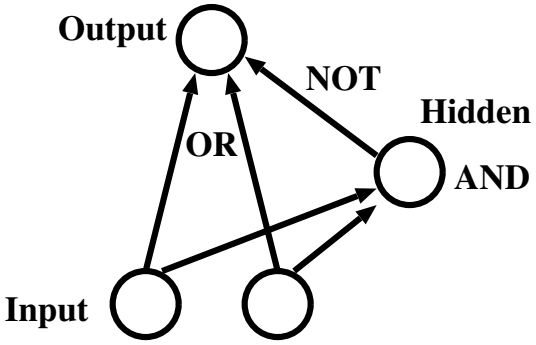

Figure 16. Another solution of XOR problem with direct pathways.

We can summarize the two examples as follows:

1. The system with direct pathways must have an AND circuit, and the weight from the hidden to the output units is negative, which represents NOT.

2. The system without direct pathways requires at least two units in the hidden layer. One unit in the hidden layer represents an AND, and the other represents an OR.

As described above, there is no reason for favoring direct pathways, or indirect pathways. When considering the Rumelhart model, we also have no direct reason to reject to the introduction of direct pathways. There is also no reason to introduce hidden and representation layers. Therefore, there is no reason to accept the Rumelhart model, nor to reject it.

Introduction of Log-Linear Models

Log-linear models can be introduced to formalize roles of intermediate layers as higher interactions terms. A log-linear model without interactions can be described as follows:

$$p(y|x) = \frac{1}{\mathcal{Z}} \exp\left(\theta_0 + \sum_i \theta_i x_i\right), \tag{3}$$

where $\mathcal{Z}$ is a normalization factor to interpret this equation as a probability. x_i indicates each input, and θ_i indicates each parameter to be estimated. When we consider higher interactions, we can obtain the equation as follows:

$$p(y|x) = \frac{1}{\mathcal{Z}} \exp\left(\theta_0 + \sum_i \theta_i x_i + \sum_i \sum_j \theta_{ij} x_i x_j + \sum_i \sum_j \sum_k \theta_{ijk} x_i x_j x_k + \cdots\right). \tag{4}$$

Because the equation (4) has many parameters to be estimated, it makes an ill-posed problem. One advantage of the standard back propagation is to acquiring solutions of equations with many parameters.

In case of XOR with direct connections (see Figure 16), we can obtain the equation as follows:

$$p(y|x) = \frac{1}{\mathcal{Z}} \exp\left(\theta_0 + \sum_i \theta_i x_i + \theta_{11} x_1 x_2\right), \tag{5}$$

Note that each $x \in (0,1)$ is a binary; therefore $x_1 x_2 = 1$ when and only when both x_1 and x_2 are 1. This second-order interaction makes it an AND circuit. Therefore, we can predict $\theta_{11} \neq 0$. In particular, in the case of parameters for both single terms of $x_1 > 0$ and $x_2 > 0$, these two terms make an OR circuit and also make θ_{11} to be negative to inhibit the output of the AND circuit when input signals are $x_1 = 1$ and $x_2 = 1$.

The back-propagation algorithm can calculate parameters of any higher-order interactions if sufficient units are in an intermediate layer, even in the case that there are fewer numbers of the training dataset than the number of parameters to be estimated[6]. However, log-linear models with higher interactions present the difficulty that we cannot know in advance how many interactions should to be considered. The standard back propagation might have a possibility of calculating unknown higher-order interactions without explicit specifications.

Each unit in an intermediate layer can be regarded as a single or an interaction terms in log-linear model. If sufficient units, more than a task requires, can be provided, then the

[6]Plaut et al. (1996) might be able to be formulated by the first and the order interaction terms, because their model had three components, onset, vowel, and coda. These three components made regular and irregular words. Irregular words might be possible to describe as much lager values of parameters of third order interaction terms.

system can perform the task; otherwise, the system will fail to complete the task. Therefore, we can claim that a task can determine the lowest number of units in an intermediate layer. The number of units can be determined in accordance with the task's complexity. This also implies that the number of units required in intermediate layers does not vary performances at all, if the minimal number of units is supplied. Many studies failed to account for the effects of number of the units, (Dilkina et al., 2008, for example). This might be a trap or a pit in which many modelers were prone to fall. The minimal number of units in a model grantees that their models will solve given tasks; however, increasing unit numbers in intermediate layers does not always assure better performances. There might be no meanings if simulations were performed with variance in the number of units in intermediate layers as an independent variable. On the other hand, the study of Ueno et al. (2011) is valuable, because different roles were assigned layer by layer.

8. Limitations of Error-Driven Learning Algorithms

We emphasize again that the Rumelhart model is worth considering, because it could explain the development and degradation of semantic category within one framework. As explained in the previous sections, this model cannot be considered as the simplest, nor the most adequate. Here, we summarize points about the standard back-propagation algorithm. We admire the standard back propagation as one of the most powerful algorithms in the world. It has been applied to not only psychology and cognitive science but also computational intelligence and machine learning. It is useful because it is not a theory or a hypothesis, but an algorithm proposed for models to adapt the external environment. Because It is an architecture independent algorithm, it has unlimited possibilities for applications to many problems.

However, it also has problems. There are several impossibilities without additional assumptions. We summarize here these shortcomings. Note that these shortcomings would be caused inevitably because it improves performance based on the gradient vector of errors. This implies that all algorithms based on gradient descent (in the same meaning 'ascent') methods have the same problems. We show a list of problems that such gradient descent (in other words, error driven) algorithms cannot avoid. This may be a checklist for constructing models.

Either minimizing sums of squared errors or maximizing log likelihoods, gradient descent/ascent algorithms have to move in a small step-by-step manner toward minimal/maximal points. The direction that the gradient vector indicates at any point may not always be the optimal point. The small value to determine size for moving in a direction that a gradient vector at a point indicates is called a learning coefficient. The size of this learning coefficient is preferably, much smaller to avoid overshooting or the system will oscillate or diverse. For instance, in the standard back propagation, weights are updated as follows:

$$\Delta w = \varepsilon \frac{\partial E}{\partial w} = \eta \delta f'(x) = \varepsilon \delta (1-\delta) x, \tag{6}$$

where x indicates input vector, δ is a difference between a teacher and output signals ($\delta = (t-x)$). η indicates a learning coefficient. When the learning has converged, we have, $\delta = 0$, then we have $\Delta w = 0$, where no change would occur. Thus, after learning is

complete, more training might show no changes. For animals including our human beings, rats, monkeys, and others, this is not true. Everyday iterative trainings are useful to maintain a level of performances and to avoiding forgetting. However, machines cannot show such overtraining and erasure resistance effects without additional assumptions.

Below, we show a list of effects that machines trained by gradient descent algorithms will not show without additions of extra assumptions.

1. Abstract Concept Acquisition or Formation
2. Adaptive Algorithm, or Constructive Architecture
3. Age of Acquisition Effect
4. Catastrophic Interference
5. Divide and Conquer Method
6. Increasing Erasure Resistance by Repeated Practices
7. Insight Learning
8. Interactions Among Hidden Units
9. One–Shot Algorithm
10. Reward
11. Sparse Coding

As an example of abstract concept acquisition, suppose a child is learning a concept of addition and its commutative law. Laws of operations of numbers (associative, commutative, anti-commutative, idempotent, and so on) cannot be learned by gradient decent algorithms. McCloskey and Cohen (1989) trained the networks with two input units for numbers to be added and one output unit for the answer. After learning was completed, they presented to their models the dataset in which they replaced a quantity to be added with a quantity to add. They found that their models could not acquire the concept of commutative law. Abstract concepts, such as commutative law, cannot be explain in terms of generalization. In a context of machine learning, generalization is measured as the amount of errors when replacing a training dataset with test dataset. A sufficiently powerful algorithm can obtain correct answers in cases of interpolation, even in cases of extension. However, the acquisition of commutative law must be out of the scope of such an algorithm. Even highly intelligent people might find it difficult to say whether commutative law holds for all numbers (natural numbers, integers, rational numbers, real numbers, complex numbers, metrics, quaternions, and so on). When extending this question to query an algebraic structure, it is suspectable whether an operation will be true a semigroup or an abelian group. Neural networks are useful to some extent in cases of regression and classification, but cannot exaggerate too much.

Adaptive algorithm, or constructive architecture, is supposed to be a promising approach for the survival of the fittest. Organization, in general, possesses such adaptive

ability that it can metamorphose according to the complexity of tasks. Metamorphosis is quite another thing from gradient descent algorithms, but worth considering. Fahlman and Lebiere (1990) proposed an adaptive neural network model called "cascade correlation." Shultz (2003) applied the cascade correlation to Piaget's tasks (for example, a balance beam task) and succeeded in simulating developmental stages. Such adaptive algorithms are considered as candidates for additional assumptions.

The Age of acquisition effect has often been reported in various psychological fields. However, the standard back propagation does not have memory devices to store when an item was learned. Therefore, another additional assumption is required to implement this effect.

Catastrophic interference (McCloskey and Cohen, 1989; Ratcliff, 1990) is a good evidence that gradient descent algorithms sufficiently strong abilities. Any content that systems learned once would be overwritten with a new learning content. Therefore, the system often forgets what happened in the past. To avoid catastrophic interference and to implement the age of acquisition effect, another additional assumption is required.

Divide-and-Conquer algorithms (Jordan and Jacobs, 1994) "attack a complex problem by dividing it into simpler problems whose solutions can be combined to yield a solution to the complex problem. This approach can often lead to simple elegant and efficient algorithms." However, one entire system dealing with a complex problem requires many computational resources to obtain correct answers.

Increasing erasure resistance by repeated practices must also be difficult to implement for gradient descent algorithms without additional assumptions. If a system learns an item, then the error between the output and teacher signals would be almost zero. Therefore any updates might not be needed at all.

Insight learning was discovered by Köhler, a German Gestalt psychologist. This is quite different from trial and error learning. Inspiration appears to occur suddenly, in a case where no gradual approximation can be observed at all. Insight learning is one of the difficult cases to implement with gradient descent algorithms.

One-shot algorithm corresponds to episodic and auto autobiographic memories. Such impressive events for these memories might happen only once in ones life. Therefore, no gradual updates are likely to b repeated. O'Reilly et al. (2012) also discussed the one-shot algorithm and the reason of existence of it(McClelland et al., 1995).

Reward is an opposite concept to punishment, and shaping is based only on reward not on punishment. Because gradient descent algorithms are based on the gradient of error function, there are no spaces for praise or reward to modify behavior. Sutton and Barto (1998) formalized reinforcement learning, in which learning, learning take place to maximize the expected reward. This differs from gradient descent algorithms

The four algorithms, such that sparse coding, competitive learning (Figure 12), winner-take-all, and softmax function, must be implemented independently from gradient descent algorithms, for the same reason as divide-and-conquer algorithms.

9. Discussion

Below is the summary of this study:

1. Demonstration of the negation of intermediate layers to express internal representation
2. Proposal of marginalization as the alternative of internal representations
3. Indication of equivalency among tree structures, list expressions, and graphical (MDS) representations
4. ATL might represent such tree structures.
5. Dimensionality of data expression. Eigenvalues must be indicated whenever configuration is drawn.
6. Possible roles of intermediate layers
7. Computational limitation of gradient descent algorithms

The Rumelhart model is neither the simplest nor smallest architecture. We insisted that there was no necessity to assume the hidden and representation layers to express the internal representation of a semantic category. Rather of assuming intermediate layers, an operation to marginalize attributes was proposed. This operation would be equivalent to the roles of intermediate layers, and the operation would be performed whenever necessary. Spatial (graphical) and tree-structured representations have mutual compatibility without any loss of information. In this, in essence, any differences between tree structures and spatial representations. This is neither dispute nor contradiction. Although one hypothesis insists on the existence of graphical representations of semantics, we claim another to be tree-structured representations. However, these differences might be proposed from opposing viewpoints. Our semantic memories are encoded among connections of neurons in IT gyrus or other regions. In meanings that they can be interpretable, there appears no difference between graphical and tree models. The difference in interpretation between them might be whether distances can be expressed explicitly or implicitly. In a framework of graphical representations, it is difficult to explain for exceptions such as "chickens can't fly". On the other hand, implementation is likely easy for tree structures because they can inherit attributes of higher concepts. However, similarity does not always imply hierarchy. It might be safe to claim that both models are equally worth considering for the depiction our semantic memories. Answers may be changed according to how we consider the axiom of distance in semantic memories. When we assume that it holds an Euclid distance, it would be approximated to graphical models. It would be approximated to tree structure representations when we can assume that it holds a Hamming distance.

Rogers and McClelland (2004, 2011) might intend for their model to capture abstract concepts. However, it appears to be unsuccessful for reasons described here. We should ask again what conditions are required when constructing a model for our semantic memories. We proposed a marginalization hypothesis as an alternative. This hypothesis is different from the others because it can hold without the convergence zone (Damasio, 1989). Both

models could obtain similar MDS configurations. Therefore, we could not distinguish between them based only on the empirical data. Further, we propose that the ATL may represented tree structures. This gave another viewpoint of the "ATL-hub-plus" hypothesis in terms of a function. When we assume list expressions to represent semantic memories, this can be interpreted as both MDS configurations and tree structures. We might also consider related findings of Thompson-Schill (2003) and Kiani et al. (2007).

Finally, a question about the roles of the frequency effect, or the eight times problem, (Section 6.3), can be proposed. Plaut et al. (1996) dealt the frequency effect with a logarithmic transformation of words. However, there is no basis for this transformation. No one might have understood how to include the frequency effect into neural network models. Therefore, it may be possible to propose a simulation about how models behave when the frequency is dealt with as an independent variable. It might be interesting to ask how performances of models are varied when the models suffered damages. We can ask this problem further. That means that researchers cited here have not ever defined a standard error curve and a standard decay curve. If someone can define these standard curves as reference points, we can compare frequency and regularity effects of words with the standard curves. It would bring this research area to a higher level than ever.

Conclusion

McClelland et al. (2009); Rogers and McClelland (2008); Rogers and McClelland (2004, 2011); Rogers et al. (2004) proposed that the Rumelhart model could represent semantic category and theoretical consideration greatly advanced. However, the assumption of hidden and representation layers might not be a requirement for obtaining internal representations. If we can postulate an operation to sum up or marginalize attributes related with their interests, there is no need to hypothesize any intermediate layers at all. It might be possible to preserve or keep the superordinate concept when we consider tree structures in semantic memories. Rogers and McClelland (2004) might consider that their data contained the WITH information, although they insisted on the WITHOUT hypothesis. Thus, we can conclude that the WITH expression still plays an important role for memory representation, because tree structure, a type of declarative knowledge, can contain MDS configurations. Once upon a time, the Irish rock band U2 sang "With or Without You",

> With or without you, I can't live

With or without categorization, the author hopes to know the truth of semantics.

Acknowledgments

The author is indebted to Sachiyo Iwafune for help, a number of suggestions, and encouragements concerning this work. This work was supported by JSPS KAKENHI Grant Number 26919002.

References

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley and Sons, New York, USA, 2001.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

Alan Mathison Turing. Computing machinery and intelligence. *Mind A quartely review of psychology and philosophy*, LIX 236:433–460, 1950. doi: 10.1093/mind/LIX.236.433.

Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–247, 1969. doi: 10.1016/S0022-5371(69)80069-1.

Allan M. Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82:407–428, 1975.

Anthony J. Bell and Terrence J. Sejnowski. An information maximisation approach to blind separation and blind deconvolution. *Neural Computaion*, 7(6):1004–1034, 1995.

Antonio R. Damasio. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1:123–132, 1989.

B. Hassibi, D. G. Stork, and G. Wolff. Optimal brain surgeon. In S. J. Hanson, J. D. Cowan, and C. L. Giles, editors, *Advances in Neural Information Processing Systems (Denver)*, volume 5, pages 164–171, San Mateo, 1993. Morgan Kaufmann.

Daniel J. Felleman and David C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

David E. Rumelhart and David Zipser. Feature discovery by competitive learning. *Cognitive Science*, 9:75–112, 1985.

David E. Rumelhart and James L. McClelland. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*. MIT Press, Cambridge, MA, USA, 1986.

David E. Rumelhart, Geoffery E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. In David E. Rumelhart and James L. McClelland, editors, *Parallel Distributed Porcessing: Explorations in the Microstructures of Cognition*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

David E. Rumelhart. Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Daviis, and C. Lau, editors, *An Introduction to Neural and Electronic Networks*, pages 405–420. Academic Press, San Diego, CA, USA, 1990.

David Marr. *Vision*. W. H. Freeman and Company, San Francisco, USA, 1982.

David Plaut and Tim Shallice. Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, 10(5):377–500, 1993.

David Plaut, James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103:56–115, 1996.

Donald O. Hebb. *Organization of behavior, a Neuropsychological Theory*. Lawrence Erlbaum, New York, USA, 1949.

Elizabeth K. Warrington. Neuropsychological studies of verbal semantic systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 295:411–423, 1981.

Elizabeth K. Warrington and Rosaleen A. McCarthy. Category specific access dysphasia. *Brain*, 106:859–878, 1983.

Elizabeth K. Warrington and Rosaleen A. McCarthy. Categories of knowledge further fracitonations and an attempted integration. *Brain*, 110:1273–1296, 1987.

Elizabeth K. Warrington and Timothy Shallice. Category specific semantic impairment. *Brain*, 107:829–854, 1984.

Elizabeth K. Warrington and Rosaleen A. McCarthy. Multiple meaning systems in the brain: A case for visual semantics. *Neuropsychologica*, 32:1465–1473, 1994.

Francis Crick. The recent excitement about neural network. *Nature*, 337:129–132, 1989.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958. In J.A. Anderson and E. Rosenfeld (Eds.) Neurocomputing (1988), MIT Press.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Geoffrey E. Hinton and Tim Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, 98(1):74–95, 1991.

Geoffrey E. Hinton, Simon Osindero, and Wee-Yeh Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Transaction of Autom. Control*, AC-19:716–723, 1974.

James L. McClelland and Timothy T. Rogers. The parallel distributed processing approach to semantic cognition. *Nauture Neuroscience*, 4:310–313, 2003.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

James L. McClelland, Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14:348–356, 2010.

James L. McClelland, Timothy T. Rogers, Karalyn Patterson, Katia Dilkina, and Matthew A. Lambon Ralph. Semantic cognition: Its nature, its development and its neural basis. In Michael Gazaniga, editor, *The Cognitive Neurosciences IV*, chapter 72, pages 348–356. MIT Press, Boston, MA, USA, 2009.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

Jeffrey L. Elman. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225, 1991.

Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 8:71–99, 1993.

Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Simth, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A connectionist perspective on development*. MIT Press, 1996.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

Karalyn Patterson, Peter J. Nestor, and Timothy T. Rogers. Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8:976–987, 2007.

Katia Dilkina, James L. McClelland, and David C. Plaut. A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive neuropsychology*, 25(2):136–164, 2008.

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

Martha J. Farah and James L. McClelland. A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339–357, 1991.

Marvin Minsky and Seymour Papert. *Perceptrons*. MIT Press, Cambridge, MA, expanded edition edition, 1988.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower, editor, *The Psychology of Learning and Motivation*, volume 24, pages 109–164. Academic Press, New York, 1989.

Noboru Murata, Shuji Yoshizawa, and Shun ichi Amari. Network information criterion - determining the number of hidden units for an artificial neural netwrork model. *IEEE Transactions on Neural Networks*, 5(6):865–872, 1994.

Peter Grünwald. A tutorial introduction to the minimum description length principle. In Peter Grünwald, I.J. Myung, and M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*. MIT Press, USA, 2005.

Pierre Comon. Independent component analysis: a new concept? *Signal Processing*, 36 (3):287–314, 1994.

Randall C. O'Reilly. Generalization interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Computation*, 13(6):1199–1241, 2001.

Randall C. O'Reilly. Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455–462, 1998.

Randall C. O'Reilly and Yuko Munakata. *Computational Explorations in Cognitive Neuroscience: Understanding in mind by simulating the brain*. MIT Press, MA, 2000.

Randall C. O'Reilly, Yuko Munakata, Michael J. Frank, Thomas E. Hazy, and Contributors. *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition, URL: http://ccnbook.colorado.edu, 2012. URL http://ccnbook.colorado.edu.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, MA USA, 1998.

Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, 97:285–308, 1990.

Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophsiology*, 97:4296–4309, 2007.

Ruslan Salakhutdinov and Geoffrey E. Hinton. An efficient learning procedure for deep Boltzmann machines. *Neural Computation*, 24(8):1967–2006, 2006.

Scott E. Fahlman and Christian Lebiere. The cascade-correlation learning architecture. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 524–532. Morgan-Kaufman, 1990.

Sharon L. Thompson-Schill. Neuroimaging studies of semantic memory: inferring "how" from "where". *neuropsychologia*, 41:280–292, 2003.

Shin Asakawa. A model for evaluating ratios of contributions between semantics and orthography in reading written words. In *Proceedings of the 28th International Congress of Applied Psychology*, Prais, France, 2014.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, 1985.

Taiji Ueno, Satoru Saito, Timothy T. Rogers, and Matthew A. Lambon Ralph. Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, 72:385–396, 2011. doi: 10.1016/j.neuron.2011.09.013.

Thomas R. Shultz. *Computational Developmental Psychology*. MIT Press, Cambridge, MA, 2003. ISBN 0-262-19483-X.

Timothy T. Rogers and James L. McClelland. Authors' response: A simple model from a powerful framework that spans levels of analysis. *Behavioral and Brain Sciences*, 31: 729–749, 2008.

Timothy T. Rogers and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. The MIT press, Cambridge, MA, 2004.

Timothy T. Rogers and James L. McClelland. Semantics without categorization. In Emmanuel M. Pothos and Andy J. Wills, editors, *Formal Approaches to Categorization*, chapter 5, pages 88–119. Cambridge University Press, Cambridge, UK, 2011.

Timothy T. Rogers, Matthew A. Lambon Ralph, Peter Garrard, Sasha Bozeat, James L. McClelland, John Hodges, and Karalyn Patterson. Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111(1):205–235, 2004.

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of mathematical biophysics*, 5:115–133, 1943.

Warren S. Torgerson. Multidimensional scaling: I. theory and method. *psychometrika*, 17 (4):401–419, 1952.

Warren S. Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393, 1965.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 589–605, Denver, WS, USA, 1990. Morgan Kaufmann.