HOW THE BRAIN LEARNS LANGUAGE

LACUS 1997, BYU

Brian MacWhinney
Carnegie Mellon University
Pittsburgh, PA

Language processing often seems remarkably effortless. We carry on conversations while we drive a car or while we sew on a button without interfering with either activity. Language learning in children also appears effortless and natural. But this apparent effortlessness presents us with a great puzzle. When linguists look at language learning and processing, they find complexity rules, categories, and symbols. Yet, the users of language seem blissfully unaware of the complexities of the rules that they are obeying. How can we reconcile this apparent complexity with this corresponding perceived effortlessness? Recently, language researchers from many different backgrounds have begun to explore ways of addressing this paradox. Their answers call into question the extent to which language learning and processing actually function in obedience to an explicit set of formal rules. According to this new view of language learning and processing, the behaviors that we tend to characterize in terms of rules and symbols are in fact emergent patterns that arise from the interactions of other less complex or more stable underlying systems. We will refer to this new viewpoint on language learning and processing as "emergentism".

Emergentist accounts have been formulated for a wide variety of linguistic phenomena, ranging from segmental inventories, stress patterns, phonotactic constraints, morphophonological alternations, lexical structures, pidginization, second language learning, historical change, online phrase attachment, and rhetorical structures. Formalisms that have been used to analyze the emergent nature of these forms include connectionist networks, dynamic systems theory, neuronal competition models, classifier systems, production-system architectures, Bayesian models, Optimality Theory, corpora studies, and even hermeneutic analysis.

The basic notion underlying emergentism is simple enough. If you spend time watching the checkout lines at a supermarket, you quickly find that the number of people queued up in each line is roughly the same. At peak times, you may find five or six people in a line waiting to check out. At slower times, lines have only two or three waiting. There is no socially imposed rule governing this pattern. Instead, it emerges from other basic facts about the goals and behavior of shoppers and supermarket managers. Or, to take another example, consider the shape of the cells in a honeycomb. There is nothing in the genetic makeup of the honey bee that determines that each cell in the honey comb should be hexagonally shaped. Rather this shape emerges through the interactions of hundreds of cells in terms of principles of packing functions. Nature is replete with examples of emergence. The form of beaches and mountain ridges, the geometry of snowflakes and crystals, the appearance of fata morgana, and the movement of the Jet Stream in the air and the Gulf Stream in the sea -- all of these patterns arise from interactions of physical principles with constraints imposed by physical bodies. Even in the biological world, much of our somatic form is emergent, whether it be the patterns of stripes on the tiger, the formation of teeth into a

uniform bite, the structuring of enzymes to catalyze organic reactions, or our patterns of fingerprints and hair formations.

Basic Assumptions

In this paper, we want to explore three specific types of emergence that operate during the course of language learning. The first involves the acquisition of basic lexical structures in small areas of cortex that we will call "local maps". The second involves the processing of information across longer neural distances in what we will call "functional neural circuits". Within local maps, processing is highly automatic and largely capacity free. However, the strict topological structure of local maps makes certain types of new learning difficult. Across areas connected by functional neural circuits, processing is non-automatic and capacity-restricted. Although this type of processing leads to some instability, it also allows for a certain flexibility in learning. Between the level of the local map and the functional neural circuit, we can distinguish a third level of emergent structure which we will refer to as "emergent lexical properties". This third level uses the information implicit in local lexical maps to support alternative grammatical cryptotypes and argument structure frames. In this paper, we will outline ways in which these three levels of neural processing help to support the complex human activity which we call language.

Principles of neural networks

Connectionist models are implemented in terms of artificial neural networks. Neural networks that are able to learn from input are known as "adaptive neural networks". The architecture of an adaptive neural network can be specified in terms of eight design features:

1. Units. The basic components of the network are a number of simple elements called variously neurons, units, cells, or nodes. In Figure 1, the units are labeled with letters such as "x_1".

2. Connections. Neurons or pools of neurons are connected by a set of pathways which are variously called connections, links, pathways, or arcs. In most models, these connections are unidirectional, going from a "sending" unit to a "receiving" unit. This unidirectionality reflects the fact that neural connections also operate in only one direction. The only information conveyed across connections is activation information. No signals or codes are passed. In Figure 1, the connection between units x_1 and y_1 is marked with a thick line.

3. Patterns of connectivity. Neurons are typically grouped into pools or layers. Connections can operate within or between layers. In some models, there are no within-layer connections; in others all units in a given layer are interconnected. Units or layers can be further divided into three classes:

   a. Input units which represent signals from earlier networks. These are marked as "x" units in Figure 1.

   b. Output units which represent the choices or decisions made by the network. These are marked as "z" units in Figure 1.

   c. Hidden units which represent additional units juxtaposed between input and output for the purposes of computing more complex, nonlinear relations. These are marked as "y" units in Figure 1.

4. Weights. Each connection has a numerical weight that is designed to represent the degree to which it can convey activation from the sending unit to the receiving unit. Learning is achieved by changing the weights on connections. For example, the weight on the connection between x_1 and y_1 is given as .54 in Figure 1.

5. Net inputs. The total amount of input from a sending neuron to a receiving neuron is determined by multiplying the weights on each connection to the receiving unit times the activation of the sending neuron. This "net input" to the receiving unit is the sum of all such inputs from sending neurons. In Figure 1, the net input to y_1 is .76, if we assume that the activation of x_1 and x_2 are both at "1" and the x_1y_1 weight is .54 and the x_2y_1 weight is .22.

6. Activation functions. Each unit has a level of activation. These activation levels can vary continuously between "0" and "1". In order to determine a new activation level, activation functions are applied to the net input. Functions that "squash" high values can be used to make sure that all new activations stay in the range of "0" to "1".

7. Thresholds and biases. Although activations can take on any value between "0" and "1", often thresholds and bias functions are used to force units to be either fully "on" or fully "off".

8. A learning rule. The basic goal of training is to bring the neural net into a state where it can take a given input and produce the correct output. To do this, a learning rule is used to change the weights on the connections. Supervised learning rules need to rely on the presence of a target output as the model for this changing of weights. Unsupervised learning rules do not rely on targets and correction, but use the structure of the input as their guide to learning.

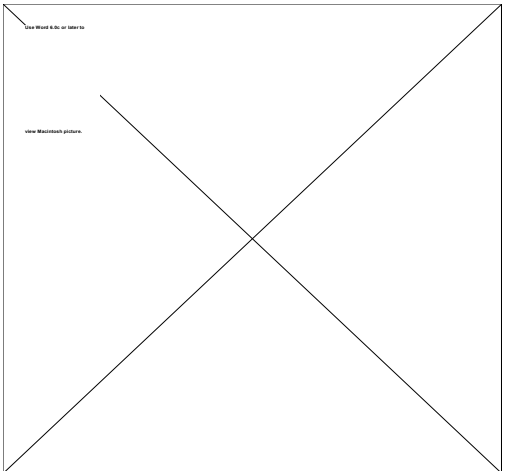


Figure 1: A sample adaptive neural network

All connectionist networks share this common language of units, connections, weights, and learning rules. However, architectures differ markedly both in their detailed patterns of connectivity and in the specific rules used for activation and learning. For excellent, readable introductions to the theory and practice of neural network modeling, the reader may wish to consult Bechtel and Abrahamsen (1991) or Fausett (1994). For a mathematically more advanced treatment, see Hertz, Krogh, and Palmer (1991).

Local Lexical Maps

Nothing is more basic in language than the learning of new words. The child's first word often appears toward the beginning of the second year of life. But word learning is not a sudden process. Rather, it depends on a whole range of experiences and activities in which the child participates during the first year of life. Some of this experience involves producing non-conventional sounds through babbling. Another type of experience involves listening to the cadences and phonetic forms of the words used by the adult community. Still another type of experience involves the slow development of thinking about the various categories of objects and events in the natural world. All of these activities and experiences are prerequisites to the learning of the first words. About two or three months before the first productive words are produced, we find some evidence that the child has begun to acquire a passive comprehension of a few of the most common words of the language. For example, the 14-month-old who has not yet produced the first word, may show an understanding of the word "dog" by turning to a picture of a dog, rather than a picture of a cat, when the word "dog" is uttered. It is difficult to measure the exact size of this

comprehension vocabulary in the weeks preceding the first productive word, but it is perhaps no more than 20 words in size.

During this early period of auditory learning, the child starts to form associations between certain auditory patterns and particular meaningful interpretations. In older models of lexical learning, the process of associating a sound with a meaning involved the trivial formation of a single link. For example, in Morton's (1970) Logogen Model, the learning of a new word requires nothing more than the linking up of one already available pattern or cluster to another. The idea that auditory and semantic patterns form coherent clusters does seem to reflect real facts about the infant's cognition. On the semantic level, one could argue (Mervis, 1984) that the child's previous experience with dogs has served to promote the consolidation of the concept of a "dog". On the phonological level, it also appears that repeated exposure to the consistent pattern of "dog" also leads to the emergence of a consolidated phonological pattern.

The self-organizing feature map (SOFM) framework of Kohonen (1982) and Miikkulainen (1990; 1991) provides us with a way of characterizing these early processes of semantic and phonological consolidation. In the framework of SOFM models, word learning can be viewed as involving the development of maps in which individual patterns can be stored and retrieved reliably. For word learning, three types of local maps are important: auditory maps, meaning maps, and articulatory maps. Each of these three maps has a similar structure. Figure 2 illustrates the activation of a particular node in an auditory map.

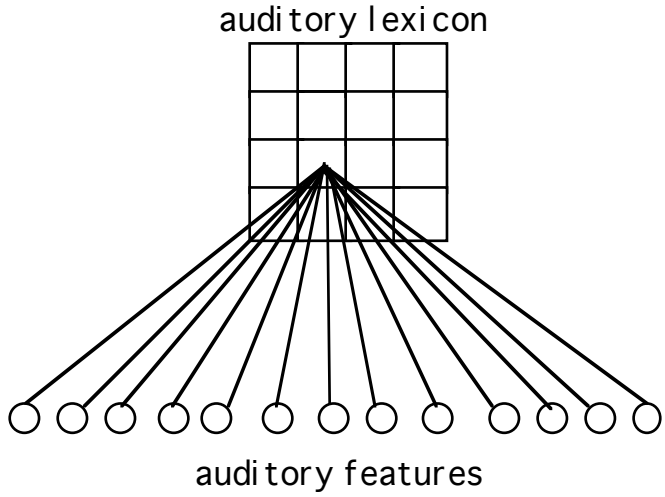


Figure 2: A self-organizing feature map for storing auditory patterns

The input to this feature map involves a large number of auditory phonological features taken from separate domains such as sibilance, formant transition direction, formant duration, formant frequency, stop click timing, and others. These are schematically represented as "auditory features" at the bottom of Figure 2. For the purposes of computational modelling, this rich

multidimensional space can be compressed onto a 2-D topological space. The two dimensions of the graphic representation do not have any direct relation to features in the input data set. However, the learning algorithm works to preserve the topological relations inherent in the high-dimensional space. By itself, a map of the type given in Figure 2 does not model the full learning of a word; it only encodes the possible auditory forms to which the child has been exposed.

What makes this mapping process self-organizing is the fact that there is no pre-established pattern for these mappings and no preordained relation between particular nodes and particular feature patterns. The SOFM algorithm decides which node on the map should be the "winner" for a particular input pattern. At first, the weights on the map are set to small random values. When the first input comes in, the random setting of these weights make it so that, by chance, some particular node is the one that is maximally responsive to the current input pattern. That node then decrements the activation levels on the other nodes. However, the shape of this decrementation takes on the form of a "Mexican hat" or sombrero. Right around the winner, related nodes are not decremented as much as are more distant nodes. Because of the architecture of the relation between the input and the grid, nodes that are nearby in the map come to respond to similar input patterns. For example, words that begin with similar initial segments will tend to be assigned to neighboring units in the map. The Mexican hat shape obeyed by the competitive interactions in the SOFM conforms closely to known facts about lateral inhibition and the redistribution of syntactic resources (Kohonen, 1982) in cortical tissue. The actual computational implementation of this framework uses a computationally efficient algorithm that is faithful to these biological principles (Miikkulainen, 1990).

This system works well to encode large numbers of patterns. In one example simulation, we found that a 100 x 100 network with 10,000 nodes can learn up to 6000 phonological patterns with an error rate of less than 1%. In this implementation, we used eight floating-point numbers to generate the input. At the beginning of learning, the first input vector of eight numbers led by chance to somewhat stronger activation on one of the 10,000 cells. This one slightly more active cell then inhibits the activation of its competitors, according to the Mexican hat function. As a result of this pattern of activation and inhibition, inputs that are close in feature space end up activating cells in similar regions of the map. Once a cell has won a particular competition, its activation is negatively damped to prevent it from winning for all of the inputs. Then, on the next trial, another cell has a chance to win in the competition for the next sound-meaning input pattern. This process repeats until all 6000 sound-meaning patterns have developed some "specialist" cell in the feature map. During this process, the dynamics of self-organization make it so that items with shared features end up in similar regions of the feature map.

We tracked the development of the feature map by computing the average radius of the individual items. After learning the first 700 words, the average radius of each word was 70 cells; after 3000 words, the radius was 8; after 5000 words the radius was 3; and after 6000 words the radius was only 1.5 cells.

Clearly, there is not much room for new lexical items in a feature map with 10,000 cells that has already learned 6000 items. However, there is good reason to think that the enormous number of cells in the human brain makes it so that the size of the initial feature map is not an important limiting constraint on the learning of the lexicon by real children. We have found that there is no clear upper limit on the ability of the SOFM to acquire more items, when it is given a larger dimensionality.

Using maps for retrieval

In order to permit full use of the lexicon in processing, the basic SOFM architecture must be supplemented by additional connections. Miikkulainen (1990) did this by training reciprocal connections on two maps using Hebbian learning. Figure 3 illustrates the relations of these two maps. In this figure, a particular auditory form is associated with a particular semantic form or meaning.

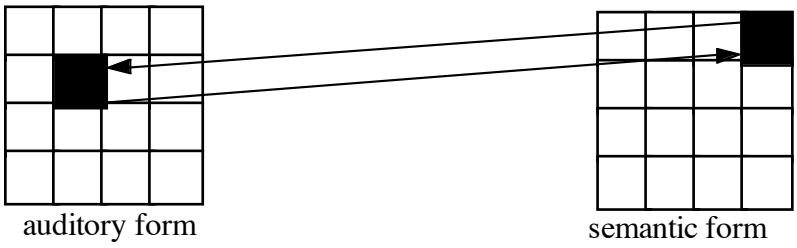


Figure 3: A bidirectional sound-meaning association in a feature map

Neuronal connections can only fire in a single direction. However, we want to have bidirectional connections between sounds and meanings. In order to establish a bidirectional association between an auditory pattern and a particular semantic pattern, training has to be conducted separately in each direction. In our simulations, learning begins with the consolidation of both the auditory and semantic maps according to the SOFM competitive learning algorithm. Once patterns are established on the two basic maps, Hebbian learning strengthens connections between units that are coactive on the sound map and the meaning map. This training is intended to represent the actual process of word learning, during which the child hears a word at the same time some meaningful aspect of the environment is being focused upon.

This proposed model is oversimplified in terms of both structure and process. In structural terms, additional maps are needed to represent additional aspects of lexical knowledge. In addition to the two maps given in Figure 3, there must be a map that encodes output phonological form, since the child must not only associate an auditory form to a semantic form, but must also associate the auditory form to an articulatory form and an articulatory form to the semantic form. Later, when the child learns to read and spell, there will also be maps for orthographic and visual forms. In processing terms, the SOFM given in Figure 3

fails to express important aspects of the serial structure of auditory and articulatory patterns. Later, we will discuss a lexical learning model developed by Gupta and MacWhinney (1996) that deals in a more explicit way with issues of serial ordering.

Articulatory scaffolding

The relation between a pattern in the auditory map and a pattern in the semantic map is essentially arbitrary. There is nothing about the phonological shape of /kaet/ that corresponds in some patterned way to the meaning "cat". However, the relation between auditory and articulatory forms is far more systematic. Once an adult has been exposed to a new auditory form, the corresponding articulatory form is extremely easy to produce. When we hear someone say that their last name is "Tomingo" we can quickly reproduce that name, even after only one trial.

For the child, the mapping from a new auditory form to an articulatory form is a bit more difficult, but it is still the case that audition serves to "scaffold" articulation. What this means is that the auditory form remains an active target as we attempt to match the form in articulation. By then listening to our articulation, we can verify the match of our output to the target auditory form. This allows us to correct errors and to set up an excitatory feedback loop between the two forms that stabilizes the new articulatory shape. Gupta and MacWhinney (1996) show how the development of this correspondence is based primarily on the mapping of correspondences between auditory fragments and articulatory fragments. In the simplest case, these fragments are syllables. For example, once the child has learned how to produce the syllable /go/ of "go", this auditory-articulatory correspondence is available for use in any new word. Even individual segments can be extracted through analysis. Some of this learning occurs during late babbling, but it is consolidated with the first words. Over time, the links between auditory and articulatory forms become more extensive.

Prosody and Time

Both the auditory and the articulatory maps must be structured to deal effectively with multisyllabic patterns. A major limitation of some earlier connectionist models (MacWhinney, Leinbach, Taraban & McDonald, 1989) was their reliance on a grid-based representation for words that required preallocation of a fixed number of syllables. For example, the word "banana" would require three fully specified sets of nodes for each of the three syllables. Within each syllable, the complete mechanism for specifying the features of the onset, nucleus, and coda would have to be repeated. In order to extend the system to deal with a four-syllable word such as "rhinoceros" a whole new set of identical units would then have to be grafted onto the system.

Systems like NetTalk (Bullinaria, 1996; Sejnowski & Rosenberg, 1988) or recurrent networks (Dell, Juliano & Govindjee, 1993; Elman, 1990) deal with the processing of temporal information by imposing an attentional focus that moves across the incoming word left-to-right in time. These systems try to interpret the sound of each letter or phoneme with only a little attention to surrounding context. One major strength of this type of system is the fact that it avoids repetitive encoding of phonological information for different positions in the word. However, like the slot-based approach mentioned above, these systems need to preallocate a certain number of positions in the input space. Words that exceed this length cannot be processed.

An alternative to these systems uses a SOFM to encode individual incoming syllables. This system deals with the problem of having to duplicate phonological knowledge by repeatedly using the same basic syllable processing map. This map stores a large number of identifiable syllabic forms such as /ba/, /kib/, and /Uv/, as well as subsyllabic forms such as /s/ or /n/. The input to this SOFM arrives in a sequential way, but each syllable is processed as a separate temporal chunk. This is easy to do on the level of the syllable, because there and many cues that tell whether a segment is in the position of the onset, the nucleus, or the coda. Because most coarticulation effects occur within the syllable, this is an effective way of dealing with low-level context effects. The syllabic processor operates repeatedly through the word to encode a series of activations of syllables.

The functioning of this syllabic map is supplemented by a process that associates particular syllabic vectors with additional prosodic information. This processor attends not to the segmental forms in the speech wave, but to the overall prosodic structure. Prosodic information works in terms of the system of metrical feet to encode the status of a given syllable as being in a iambic or trochaic foot and being either a strong or weak syllable. It is the union of these prosodic features with the basic segmental syllabic features which then serve as input to the auditory lexical SOFM. In a word like “banana” the syllabic processor operates repeatedly to encode three syllables. However, without the additional metrical information, these three encodings could be perceived as the patterns “nabana” or “nanaba”, as well as “banana”. In order to uniquely encode “banana”, the first syllable /ba/ must be coded as a first foot weak beat, the second syllable /na/ must be coded as the strong beat and the final /na/ must be coded as the second foot weak beat. Thus, the complete input to the lexical map includes both segmental and prosodic information. It is this complete merged pattern which is then associated to the semantic pattern to specify emergent lexical items.

Emergent Lexical Properties

Lexical items provide a substrate that can lead to the emergence of a variety of complex linguistic structures. The force that drives this structural emergence is the linking of words into syntactic combinations. We will call the properties that

emerge from the combinations of words "emergent lexical properties". When words are combined into phrases and sentences, there are a variety of complex interactions that arise in terms of phonology, grammatical relations, and meaning. In this section we will review some of these interactions.

Acquiring inflectional morphology

The local lexical map can be used to acquire both stems such as "dog" or "jump" and affixes such as the plural suffix or the past tense suffix. Stems can be learned directly. However, in order to model the learning of affixes, we need to examine an additional process called "masking" (Burgess, 1995; Burgess & Hitch, 1992; Carpenter, Grossberg & Reynolds, 1991; Cohen & Grossberg, 1987; Grossberg, 1987). Let us use the learning of the English past tense suffix to illustrate how masking works.

1. The net learns a set of present tense verbs, along with the corresponding past tense forms. We can refer to this initial phase of learning as "rote" learning. These rote-learned forms include regular pairs such as "jump - jumped" and "want - wanted", as well as irregulars such as "run - ran" and "take - took".
2. The network then learns a new present tense such as "push" for which the corresponding past tense form has not yet been learned.
3. Then the child hears the word "pushed" with the auditory form /pUS/ and the semantic pattern "push + past". On the auditory map, the node corresponding to /pUS/ is the closest match. On the semantic map, the node corresponding to "push + present" is the closest match.
4. A pattern of bidirectional activation is established between the two maps. It is this bidirectional activation that supports the process of "masking". Masking works to drain activation from nodes and features that are coactive in the two maps. In the current example, the features of the stem on both maps are all masked out, leaving the feature "past tense" as unmasked on the semantic map and the features corresponding to the final /id/ as unmasked on the auditory map.
5. The unmasked phonology is then associated with the unmasked semantics through the same type of Hebbian learning that is used to produce the basic rote-learning of new lexical forms.

This implements in a neuronally plausible way the process of morphological extraction by analysis. In the terms of MacWhinney (1978), affix analysis involves associating the "unexpressed" with the "uncomprehended". This approach to the problem of learning the English past tense solves a number of problems faced by earlier nonlexical models. First, the model succeeds in capturing both rote lexicalization and combinatorial lexicalization within a single connectionist model. Rote forms are picked up directly on the feature map. Combinatorial forms are created by the isolation of the suffix through masking and the use of masking in production.

Having learned to comprehend the past tense in a productive way, the child can then learn the association between the auditory pattern and an articulatory representation. This occurs when the child tries to produce the new form. The activation of a semantic pattern leads to the activation of an auditory pattern which then sets up a temporary excitatory feedback loop to the articulatory map. During the process of scaffolding, the auditory form remains active as we attempt to match the form in articulation. By then listening to our articulation, we can verify the match, correct errors, and set up an excitatory feedback loop between the two forms that stabilizes the new articulatory shape. As we noted earlier, the process of developing a match between the auditory and articulatory forms proceeds syllable by syllable (1996) by relying on prosody to encode the temporal properties of successive syllables.

Using inflectional morphology

Having acquired productive use of inflectional morphology, the child can begin to learn which forms of each inflection can be used with particular stems in particular cases. The emergentist approach to language acquisition claims that these complex patterns emerge naturally from the interaction of already learned phonological structures in the lexical map. To illustrate how this can work, let us take as an example the model of German gender learning developed by MacWhinney, Leinbach, Taraban, and McDonald (1989). This model was designed to explain how German children learn how to select one of the six different forms of the German definite article. In English we have a single word "the" to express definiteness. In German, the same idea can be expressed by "der", "die", "das", "des", "dem", or "den". Which of the six forms of the article should be used to modify a given noun in German depends on three additional features of the noun: its gender (masculine, feminine, or neuter), its number (singular or plural), and its role within the sentence (subject, possessor, direct object, prepositional object, or indirect object). To make matters worse, assignment of nouns to gender categories is often quite nonintuitive. For example, the word for "fork" is feminine, the word for "spoon" is masculine, and the word for "knife" is neuter.

Although these relations are indeed complex, MacWhinney et al. show that it is possible to construct a connectionist network that learns the German system from the available cues. The MacWhinney et al. model, like most current connectionist models, involves a level of input units, a level of hidden units, and a level of output units (Figure 4). Each of these levels or layers contains a number of discrete units or nodes. For example, in the MacWhinney et al. model, the 35 units within the input level represent features of the noun that is to be modified by the article. Each of the two hidden unit levels includes multiple units that represent combinations of these input-level features. The six output units represent the six forms of the German definite article.

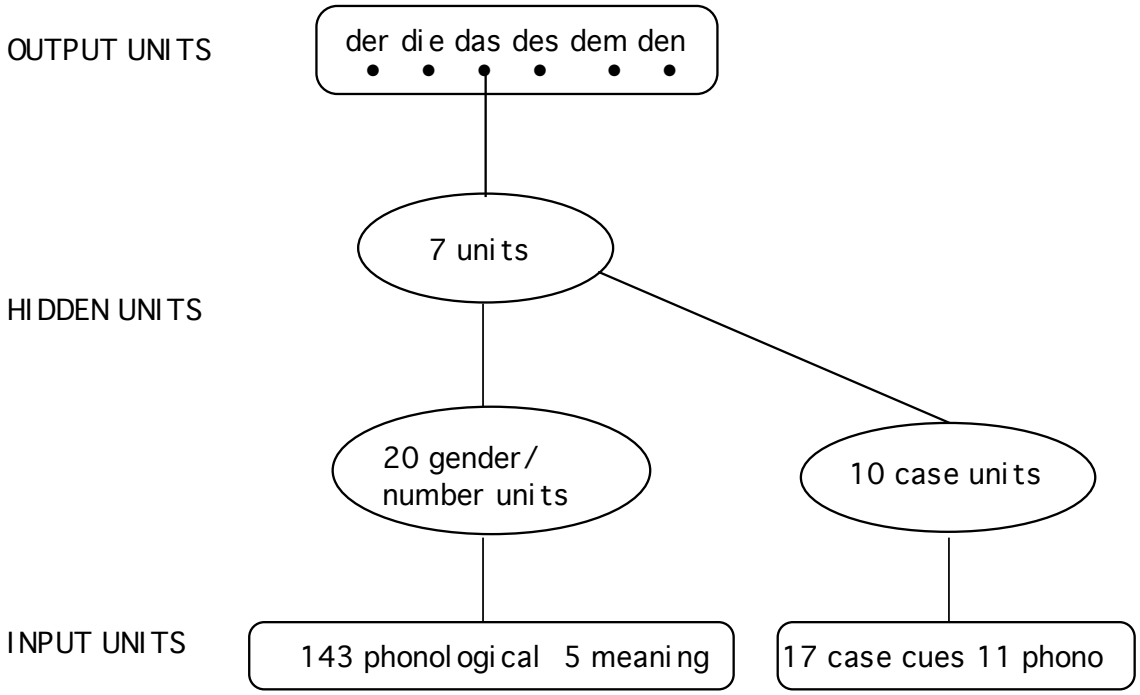


Figure 4: A back propagation model for German declension

As noted above, a central feature of such connectionist models is the very large number of connections among processing units. As shown in Figure 4, each input-level unit is connected to first-level hidden units; each first-level hidden unit is connected to second-level hidden units; and each second-level hidden unit is connected to each of the six output units. None of these hundreds of individual node-to-node connections are illustrated in Figure 4, since graphing each individual connection would lead to a blurred pattern of connecting lines. Instead a single line is used to stand in place of a fully interconnected pattern between levels. Learning is achieved by repetitive cycling through three steps. First, the system is presented with an input pattern that turns on some, but not all of the input units. In this case, the pattern is a set of sound features for the noun being used. Second, the activations of these units send activations through the hidden units and on to the output units. Third, the state of the output units is compared to the correct target and, if it does not match the target, the weights in the network are adjusted so that connections that suggested the correct answer are strengthened and connections that suggested the wrong answer are weakened.

MacWhinney et al. tested this system's ability to master the German article system by repeatedly presenting 102 common German nouns. Frequency of presentation of each noun was proportional to the frequency with which the nouns are used in German. The job of the network was to choose which article to use with each noun in each particular context. After it did this, the correct answer was

presented, and the simulation adjusted connection strengths so as to optimize its accuracy in the future.

After training was finished, the network was able to chose the correct article for 98 percent of the nouns in the original set. Of course, the ability to learn the input set is not a demonstration of true learning, since the network may have simply memorized each presented form by rote. However, when the simulation was presented with a previously encountered noun in a novel context, it chose the correct article on 92 percent of trials, despite the noun's often taking a different article in the new context than it had in the previously encountered ones. This type of cross-paradigm generalization is clear evidence that the network went far beyond rote memorization during the training phase. In addition, the simulation was able to generalize its internalized knowledge to entirely novel nouns. The 48 most frequent nouns in German that had not been included in the original input set were presented in a variety of sentence contexts. On this completely novel set, the simulation chose the correct article from the six possibilities on 61 percent of trials, versus 17 percent expected by chance. Thus, the system's learning mechanism, together with its representation of the noun's phonological and semantic properties and the context, produced a good guess about what article would accompany a given noun, even when the noun was entirely unfamiliar.

The network's learning paralleled children's learning in a number of ways. Like real German-speaking children, the network tended to overuse the articles that accompany feminine nouns. The reason for this is that the feminine forms of the article have a high frequency because they are used both for feminines and for plurals of all genders. The simulation also showed the same type of overgeneralization patterns that are often interpreted as reflecting rule use when they occur in children's language. For example, although the noun Kleid (which means clothing) is neuter, the simulation used the initial "kl" sound of the noun to conclude that it was masculine. Because of this, it invariably chose the article that would accompany the noun if it were masculine. Further, the same article-noun combinations that are the most difficult for children proved to be the most difficult for the simulation to learn and to generalize to on the basis of previously learned examples.

How was the simulation able to produce such generalization and rule-like behavior without any specific rules? The basic mechanism involved adjusting connection strengths between input, hidden, and output units to reflect the frequency with which combinations of features of nouns were associated with each article. Although no single feature can predict which article would be used, various complex combinations of phonological, semantic, and contextual cues allow quite accurate prediction of which articles should be chosen. This ability to extract complex, interacting patterns of cues is a particular characteristic of the back-propagation algorithm.

Despite its successes, this model and a similar model for English (MacWhinney & Leinbach, 1991) faced certain basic problems. These problems all arose from the fact that the model assigned no privileged role to words as lexical items. Instead, all learning was based simply on phonological patterns.

A clear example of this type of problem arises in the case of the sound /rIN/ which is used for three different verb meanings (to ring a bell, the wring out the clothes, and to ring the city). The past tense forms of these verbs will be "rang", "wrung", and "ringed" depending on the meaning of the stem. By itself, the back propagation net cannot solve this type of problem with homophony. However, if the back propagation network is joined to a SOFM, homophony is no longer a problem, because the various homophonic meanings of "ring" are now representationally distinct. When a phonological pattern is activated, the corresponding semantic form is also activated and both are available as input to the back propagation pattern extractor. Gupta and MacWhinney (1992) showed how the addition of lexical information to the back propagation network for German could lead to improved performance

Because this model combines two different architectures, inflectional formations can be produced in several different ways. First, the SOFM can generate both regular and irregular forms by rote. Second, because the SOFM includes affixes along with stems, regular affixation can be produced through combination. Third, the pattern generalization processes found in the back propagation network can help produce irregularizations. For example, the past tense forms "wrung" and "rang" could be produced either directly by rote or by generalization using the back propagation network.

The logical problem - benign cases

The most important aspect of the network we have been discussing is that fact a single lexical feature map can produce both a rote form like "went" and a productive form like "*goed". The fact that both can be produced in the same lexical feature map allows us to begin work on a general solution to the "logical problem of language acquisition" (Baker & McCarthy, 1981; Gleitman, 1990; Gleitman, Newport & Gleitman, 1984; Morgan & Travis, 1989; Pinker, 1984; Pinker, 1989; Wexler & Culicover, 1980). In the case of the competition between "went" and "*goed", we expect "went" to become solidified over time because of its repeated occurrence in the input. The form "*goed", on the other hand, is supported only by the presence of the combinational -ed form. Figure 5 illustrates this competition:

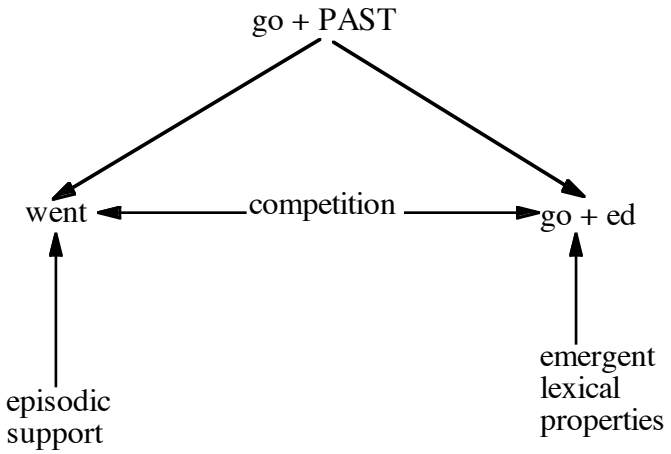


Figure 5: Competition between episodic and combinatorial knowledge

This particular competition is an example of what Baker (1981) calls a "benign exception to the logical problem". The exception is considered benign because the child can learn to block overgeneralization by assuming that there is basically only one way of saying "went". This Uniqueness Constraint is thought to distinguish benign and non-benign exceptions to the logical problem. However, from the viewpoint of the Competition Model account we are constructing here, all exceptions are benign.

The basic idea here is that, when a child overgeneralizes and produces "*goed", the system itself contains a mechanism that will eventually force recovery. In cases of overgeneralization, alternative expressions compete for the same meaning. One of these forms receives episodic support from the actual linguistic input. This episodic support grows slowly over time. The other form arises productively from the operation of analogistic pressures. When episodic support does not agree with these analogistic pressures, the episodic support eventually comes to dominate and the child recovers from the overgeneralization. This is done without negative evidence, solely on the basic of positive support for the form receiving episodic confirmation.

Argument frame induction

Earlier versions of the Competition Model (MacWhinney, 1988) emphasized the role of lexical argument ("valency" or "dependency") relations as the basic controllers of syntactic structure. This analysis was in tune with the theories of Lexical Functional Grammar (LFG) (Bresnan, 1982) and Head-driven Phrase Structure Grammar (HPSG) (Pollard & Sag, 1994) that developed during the 1980s. The role of lexical predicates in determining syntactic structure is now widely accepted. However, there is still no agreement regarding the ways in which children learn to attach argument frames to lexical items or groups of

lexical items. Symbolic proposals regarding this learning can be found in MacWhinney (1988), Pinker (1984), and Brent (1994). Within a connectionist framework, the major attempts to deal with syntactic processing include McClelland and Kawamoto (1986), St. John and McClelland (1988), Elman (1990), and Miikkulainen (1993). However, none of these accounts come to grips with the relation between argument frames and specific lexical items.

We know that the induction of argument relations must occur in parallel with the process of learning new words. To illustrate this process, consider an example in which the child already knows the words "Mommy" and "Daddy", but does not know the word "like". Given this state of lexical knowledge, the sentence "Daddy likes Mommy" would be represented in this way:

d a d i l aI k s m a m i
Daddy | unknown | Mommy

For the first and last phonological stretches, there are lexical items that match. These strings and the semantics they represent are masked. The unknown stretch is not masked and therefore stimulates lexical learning of the new word "likes". The core of the learning for "likes" is the association of the sound /laIk/ to the meaning "like". In addition to this association, the central association feature map now constructs links not only to sound and meaning, but also to argument relations. The initial argument frame for the word "likes" is:

arg1: preposed, "Daddy", experiencer
arg2: postposed, "Mommy", experience

Further exposures to sentences such as "Daddy likes pancakes" or "Billy likes turtles" will soon generalize the dependency frame for "likes" to:

arg1: preposed, experiencer
arg2: postposed, experience

No theoretical weight is placed on the notion of "experiencer" or "experience" and different learners may conceptualize this role in different ways.

Adjectives typically have only one argument. Prepositions have two -- one for the object of the preposition and a second for the head of the prepositional phrase. Verbs can have as many as three arguments. For each lexical item, we can refer to these arguments as arg1, arg2, and arg3. When a group of words share a common set of semantic relations with a particular argument, they form an argument frame group. For example, words like "send" or "promise" share the syntactic property of permitting a double object construction as in "Tim promised Mary the book". Pinker (1989) has argued that there are a variety of semantic cues which work together to decide which verbs allow this type of double object construction.

Argument frames and the logical problem

Children often produce double object argument frame overgeneralizations such as "I recommended him the book ". Bowerman (1988) and Pinker (1989) argue that, because children do not receive or process negative evidence correcting these errors, the process of recovery constitutes a difficult case of the

logical problem of language acquisition. However, in the Competition Model framework, this learning is just as "benign" as recovery from errors such as "*goed". Figure 6 illustrates the situation:

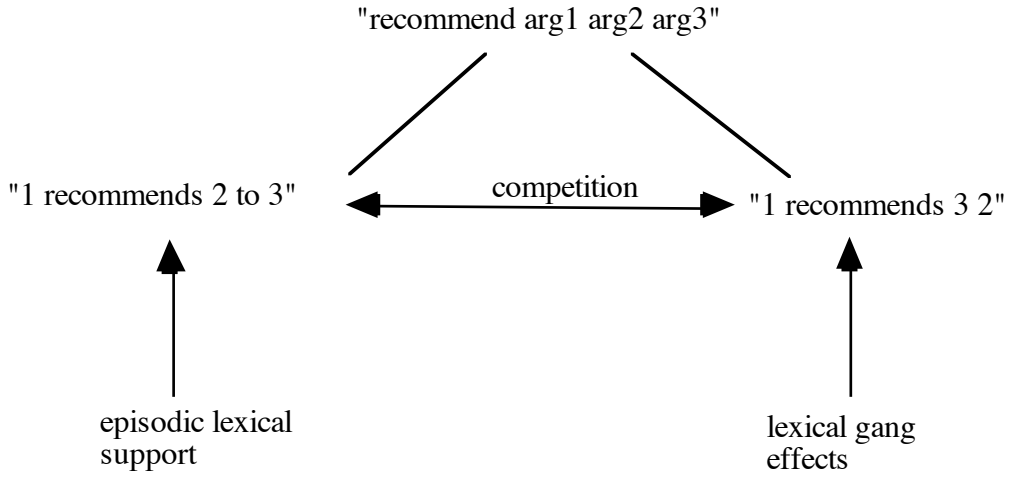
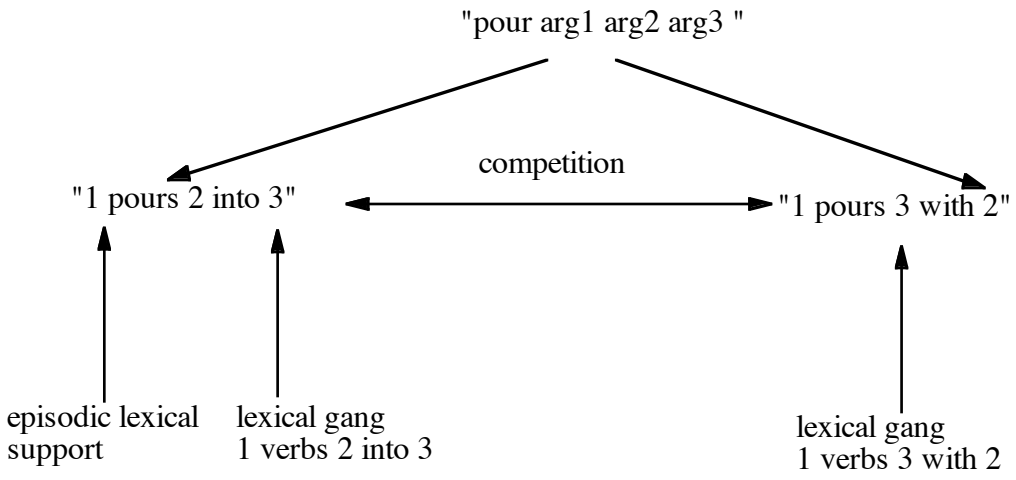


Figure 6: Competition between lexical argument frames and lexical gang effects

In this case, the child receives episodic support for the construction "X recommends Y to Z". However, there is also analogistic pressure from the argument frame of words such as "send" or "promise" for the double object argument frame "X recommends Z Y". Because the verb "recommend" shares many semantic features with transfer verbs such as "give" and "offer", it becomes attracted by that lexical gang and is subject to lexical gang effects. In the case of "recommend", the double object frame is incorrect and receives no episodic support. Over time, the continuing growth of positive episodic support for the prepositional dative form will lead to a decrease in overgeneralizations of the double object form, without any need to invoke negative evidence. Thus, a competitive system of this type learns on the basis of positive evidence.

Perhaps the most complicated form of argument frame competition arises when there are two competing argument patterns. For example, verbs like "pour" or "dump" have a frame in which arg2 is the thing transferred and arg3 is the goal. Thus, we say "Tim poured water into the tub." Another group of verbs like "fill" or "cover" have a frame in which arg2 is the goal and arg3 is the thing transferred. Thus, we say "Tim covered the lawn with gypsum."



Argument frames and lexical items

The process of learning argument frames follows the logic developed by Gupta and MacWhinney (1992) for the acquisition of German declensional marking. That model used a SOFM for the extraction of cooccurrence patterns between articles and nouns. Using these patterns, the full shape of the German declensional pattern emerged inside the SOFM. Nodes in the map took on the role of associating particular constellations of case and number marking on the article with one of the three grammatical genders of German. This system was then linked to a back-propagation system that responded to additional phonological cues to gender. The general shape of this type of model is given in Figure 7.

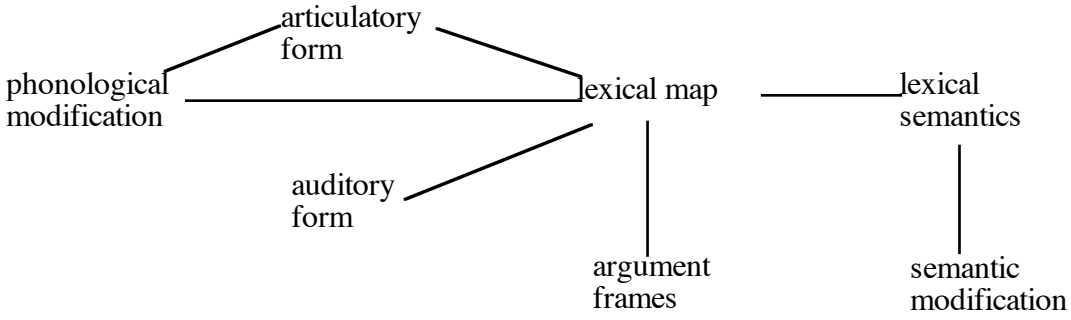


Figure 7: A connectionist account of the lexical knowledge

The lexical map produces two argument frame effects. One is the activation of the correct argument frame for a specific lexical item. The other is the activation of argument frames for semantically related groups of words or lexical "gangs". Words that have similar meanings will tend to activate similar argument structures.

Sentence Interpretation

In addition to the basic maps for lexical associations and argument frames, Figure 7 includes systems for phonological and semantic modification. Semantic modification works to adapt meanings when words are linked together. Because connectionist systems are constraint satisfaction systems, rather than rule systems, they can deal with partial violations in the combinations of words. Consider a combination like "another sand." Typically, the word "another" requires a count noun and "sand" is a mass noun. However, when the listener is confronted with this particular combination, it is still possible to retrieve an interpretation by treating "sand" as a count noun. This can be done by thinking of bags of sand, types of sand, alternative meanings of the word "sand", or even the act of applying sandpaper to something. MacWhinney (1989) talks about these semantic extension effects in terms of a process of "pushy polysemy".

Functional Neural Circuits

Pushy polysemy is one of several processes that helps us understand how sentence interpretation arises during online sentence processing. As each predicate is linked to its several arguments, the listener shifts focus away from the individual lexical items onto the emerging sentence interpretation (Gernsbacher, 1990; MacWhinney, 1977). In effect, every word that is linked to the growing interpretation is "masked" from the short-lexical memory. However, some syntactic structures place a heavy demand on working memory. For example, in a sentence such as "The dog the cow the pig chased kicked barked", the listener cannot construct interpretations by linking each word to its neighbor. Instead the string of three nouns and three verbs have to be stored in virtually unassociated form in working memory while the listener attempts to find meaningful clusters. Sentences of this type, while technically grammatical, are notoriously difficult to process. Accumulations of unattached nouns in relative clauses are a well-known problem for speakers of SOV languages such as Hungarian (MacWhinney & Pléh, 1988) and Japanese (Hakuta, 1981).

Language processing also places demands on general systems for memory storage when we need to remember lists of words (Gathercole & Baddeley, 1993; Gupta & MacWhinney, 1994). There is now a fair amount of evidence from functional magnetic resonance imagery (fMRI) and positron emission tomography (PET) indicating that word rehearsal places demands on both frontal cortex and superior temporal cortex. Gupta and MacWhinney (1996) argue that the link between frontal and temporal cortex constitutes a functional neural circuit that subserves some of the more difficult, non-lexical aspects of sentence processing.

Recently, we have been studying sentence processing in children with early focal lesions. Despite the fact that large areas of cortex are missing in these

children, their basic language abilities are quite normal. Only when we attempt to measure their abilities more closely through more demanding tasks or precise reaction time methodology do we discover important gaps in online processing. This suggests that the basic local encoding of lexical knowledge is maintained in all of these children, and that it is functional neural circuits which are most prone to damage and disorder. Let us look at four example cases.

The 10-year-old subject imaged in Figure 8 has experienced a massive left hemisphere lesion that approximates the loss found in a full hemispherectomy, although major areas of the left frontal area are spared. Despite this massive lesion, his language processing is within the low normal range. However, higher level abilities acquired only later in life, such as reading, are markedly impaired.

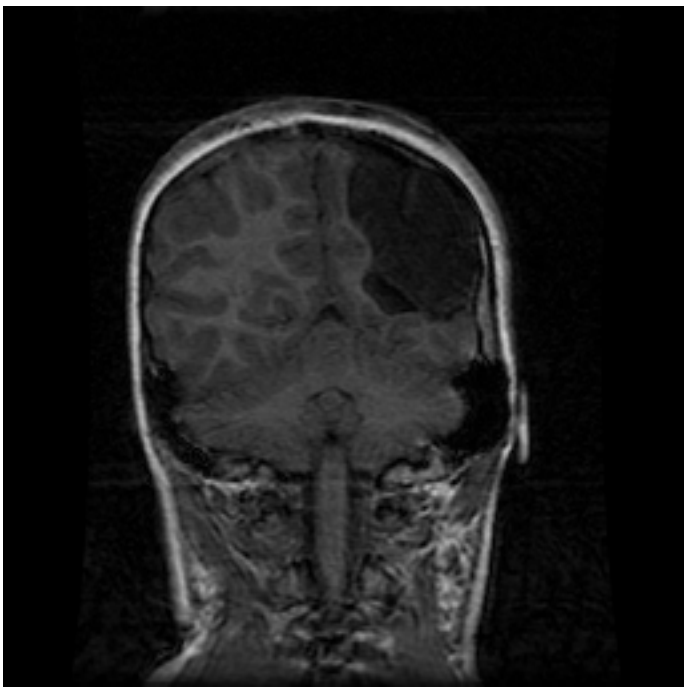


Figure 8: Magnetic resonance imaging scan coronal section from Subject #1. Section is taken from center of parietal area. Lesion is in left parietal and temporal cortex. The left hemisphere appears on the right side.

The 10-year-old subject whose scan is given in Figure 9 has a clearly discernible left unilateral focal lesion including the supramarginal gyrus of the temporal lobe and the inferior parietal. Her performance lies within the normal to high-normal range on all of our on-line measures with the exception of the two that require repetition of an auditory signal. On the other hand, she is able to name pictures and remember strings of digits at normal levels. It appears that the lesion to the pathways that connect the posterior auditory area and anterior output phonological processing area in this subject has markedly impaired her ability to repeat words. This is because the repetition task requires the subject to quickly transfer information between these primary phonological processing areas

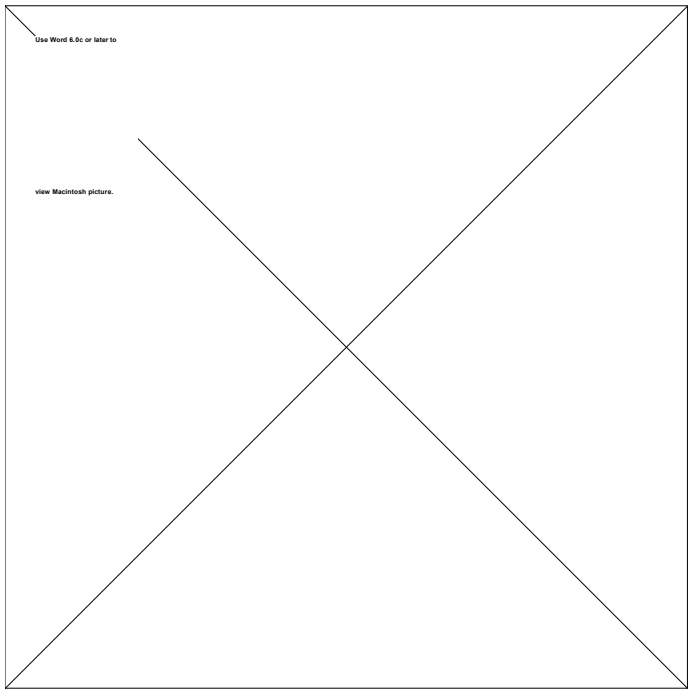


Figure 9: Magnetic resonance imaging scan coronal section from Subject #2. Section is taken from center of parietal area. Lesion is in left superior temporal cortex. The left hemisphere appears on the right side.

The subject whose scan is given in Figure 10 has an extensive periventricular lesion affecting the anterior parietal. This subject has achieved low normal levels in all of our basic processing tasks. Moreover, he scored quite high on the PPVT. On the other hand, he scored particularly low on the grammatical subtests of the CELF and performed at a low age level in our two on-line grammatical interpretation tasks. Thus, he illustrates a pattern of some below-normal language skills and others that are well above normal.

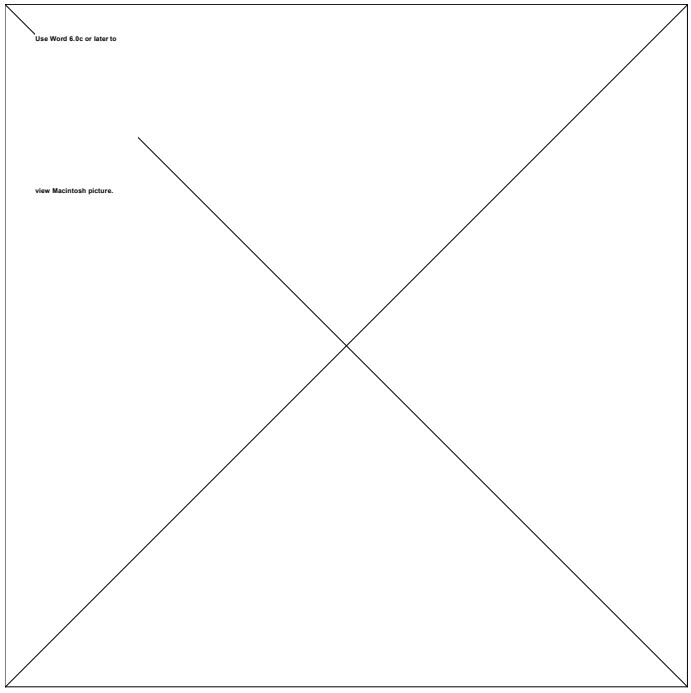


Figure 10: Magnetic resonance imaging scan coronal section from Subject #3. Section is taken from center of parietal area. Lesion is in the area around the left ventricle. The left hemisphere appears on the right side.

The subject whose scan is given in Figure 11 has a clear left focal lesion in the area of the frontal lobe just anterior to Broca's area. His basic PPVT and Leiter scores are normal, but his auditory choice reaction times and his naming reaction times are slow, suggesting an expressive slowdown. He also had low scores on the subtest of the CELF that requires subjects to formulate and complete sentences. This would seem to echo the interpretation of this brain area as related to expressive disorders.

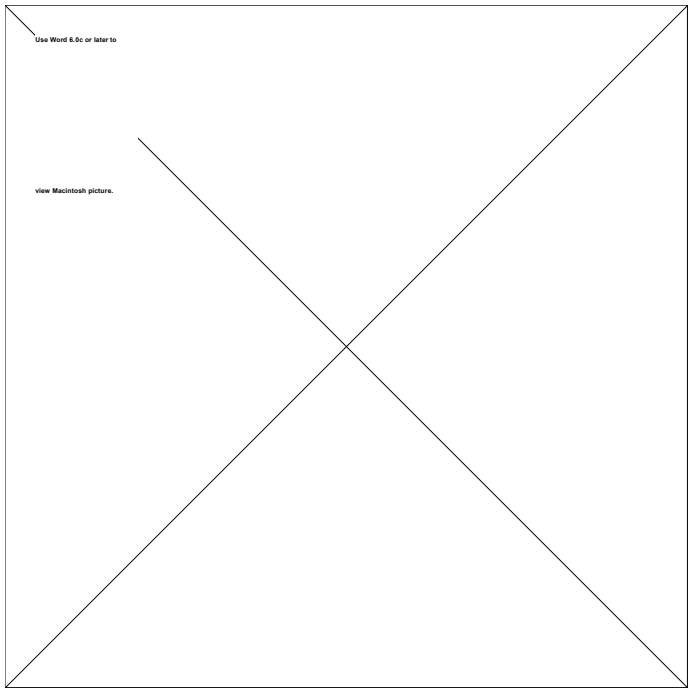


Figure 11: Magnetic resonance imaging scan coronal section from Subject #4. Section is taken from prefrontal area. Lesion is in the area anterior to Broca's area. The left hemisphere appears on the right side. The blurring in the scan is from movement artifact.

Conservatism, Functional Circuits, and the Logical Problem

Although we have emphasized the extent to which lexical competition can solve the logical problem of language acquisition, there are certain complex syntactic structures for which the lexical solution is more questionable. For example, O'Grady (1987) notes that children learn these positive contexts for wh-movement in this order:

What did the little girl hit __ with the block today?
What did the boy play with __ behind his mother?
What did the boy read a story about __ this morning?

Although one might be able to formulate a lexical basis for the processing of these wh-movement patterns, it is more likely that they involve a form of sentence memory that relies rather more on functional neural circuits and less on lexically-organized information. What is interesting is the fact that, precisely in these non-lexical contexts, children's tendency toward conservatism seems to be maximized. Children are never presented with contexts such as this:

*What did the boy with ___ read a story this morning?

Because children approach the learning of these contexts conservatively, they seldom or never make overgeneralizations of this type and never attempt wh-movement in this particular context. Because of this conservativism, attribution of these acquisitional patterns to innate knowledge of a condition blocking subjacency violations seem to be unmotivated.

Summary

At this point, it may be useful to summarize the core assumptions being made in this emergentist account of how the brain learns language.

1. The models assumes an auditory processing mechanism that can extract information regarding the onset, nucleus, and coda elements of individual syllables.
2. The information from the syllabic processor is supplemented by information from the prosodic processor which marks the position of each syllable in terms of feet and beats.
3. Auditory and semantic information about words is encoded in a self-organizing feature map.
4. Associations between sound and meaning are formed through Hebbian learning.
5. Auditory information can be used to scaffold the construction of an articulatory representation. This is done in terms of syllables and prosodic structures.
6. Masking in lexical recognition provides the support for the extraction of new affixes.
7. Changes in stems and affixes can be controlled through a system of modifications using the back propagation algorithm.
8. Sentence interpretation requires the linking of words in terms of argument structures. These structures are learned through frame generalization and are pegged to items in the lexical map.
9. The processing of complex syntactic structures and lists of words requires the involvement of functional neural circuits including frontal attentional processing and temporal lobe verbal memory and rehearsal.

In this model of language development, the first commitment that the brain makes is to the encoding of auditory, articulatory, and lexical information in localized maps. After this information is consolidated, back propagation systems develop

to fine tune the interactions of lexical items and functional neural circuits control capacity-intensive aspects of sentence processing.

Although the developments we have discussed lead to a great complexity of patterns and constructions, the underlying elements of feature maps, masking, argument frames, and rehearsal loops from which these patterns emerge are themselves fairly simply structures grounded in basic properties of neural structure and functioning. Some aspects of these structures are probably basic to all of mammalian cognition. However, the great elaboration of lexical structures that we find in human language point to the extensive elaboration of earlier structures during the million years of human evolution. Most recently, the overlay of functional neural circuitry between areas such as the frontal attentional areas and the temporal auditory areas has led to further species-specific advances in the capacity for learning and using language. Hopefully, the continuing rapid advances in our understanding of brain function and structure will allow us to move forward rapidly in elaborating this emergentist account of how language is learned by the brain.

References

Baker, C. L., & McCarthy, J. J. (Eds.). (1981). The logical problem of language acquisition. Cambridge: MIT Press.

Bechtel, W., & Abrahamsen, A. (1991). Connectionism and the mind: An introduction to parallel processing in networks. Cambridge, MA: Basil Blackwell.

Bowerman, M. (1988). The "no negative evidence" problem. In J. Hawkins (Ed.), Explaining language universals. London: Blackwell.

Brent, M. (1994). Surface cues and robust inference as a basis for the early acquisition of subcategorization frames. Lingua, 92, 433-470.

Bresnan, J. (Ed.). (1982). The mental representation of grammatical relations. Cambridge, MA: The MIT Press.

Bullinaria, J. A. (1996). Modelling reading, spelling and past tense learning with artificial neural networks. Unpublished mss.

Burgess, N. (1995). A solvable connectionist model of immediate recall of ordered lists. In G. Tesauro, D. Touretzky, & J. Alspector (Eds.), Neural Information Processing Systems 7, . San Mateo, CA: Morgan Kaufmann.

Burgess, N., & Hitch, G. (1992). Toward a network model of the articulatory loop. Journal of Memory and Language, 31, 429-460.

Carpenter, G., Grossberg, S., & Reynolds, J. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Networks, 4, 565-588.

Cohen, M., & Grossberg, S. (1987). Masking fields: A massively parallel neural architecture for learning, recognizing, and predicting multiple groupings of patterned data. Applied Optics, 26, 1866-1891.

Dell, G., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. Cognitive Science, 17, 149-195.

Elman, J. (1990). Finding structure in time. Cognitive Science, 14, 179-212.

Fausett, L. (1994). Fundamentals of neural networks. Englewood Cliffs, NJ: Prentice Hall.

Gathercole, V., & Baddeley, A. (1993). Working memory and language. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gernsbacher, M. A. (1990). Language comprehension as structure building. Hillsdale, NJ: Lawrence Erlbaum.

Gleitman, L. (1990). The structural sources of verb meanings. Language Acquisition, I(1), 3-55.

Gleitman, L. R., Newport, E. L., & Gleitman, H. (1984). The current status of the motherese hypothesis. Journal of Child Language, 11, 43-79.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, 11, 23-63.

Gupta, P., & MacWhinney, B. (1992). Integrating category acquisition with inflectional marking: A model of the German nominal system, Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, . Hillsdale, NJ: Lawrence Erlbaum Associates.

Gupta, P., & MacWhinney, B. (1994). Is the articulatory loop articulatory or auditory? Re-examining the effects of concurrent articulation on immediate serial recall. Journal of Memory and Language, 33, 63-88.

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. Brain and Language, xx, xx-xx.

Hakuta, K. (1981). Grammatical description versus configurational arrangement in language acquisition: The case of relative clauses in Japanese. Cognition, 9, 197-236.

Hertz, J., Krogh, A., & Palmer, R. (1991). Introduction to the theory of neural computation. New York: Addison-Wesley.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.

MacWhinney, B. (1977). Starting points. Language, 53, 152-168.

MacWhinney, B. (1978). The acquisition of morphophonology. Monographs of the Society for Research in Child Development, 43, Whole no. 1, pp. 1-123.

MacWhinney, B. (1988). Competition and teachability. In R. Schiefelbusch & M. Rice (Eds.), The teachability of language, (pp. 63-104). New York: Cambridge University Press.

MacWhinney, B. (1989). Competition and lexical categorization. In R. Corrigan, F. Eckman, & M. Noonan (Eds.), Linguistic categorization, (pp. 195-242). New York: Benjamins.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. Cognition, 29, 121-157.

MacWhinney, B., & Pléh, C. (1988). The processing of restrictive relative clauses in Hungarian. Cognition, 29, 95-141.

MacWhinney, B. J., Leinbach, J., Taraban, R., & McDonald, J. L. (1989). Language learning: Cues or rules? Journal of Memory and Language, 28, 255-277.

McClelland, J. L., & Kawamoto, A. (1986). Mechanisms of sentence processing: Assigning roles to constituents. In J. L. McClelland & D. E. Rumelhart (Eds.), Parallel Distributed Processing, . Cambridge, MA: MIT Press.

Mervis, C. (1984). Early lexical development: The contributions of mother and child. In C. Sophian (Ed.), Origins of cognitive skills, . Hillsdale, N.J.: Lawrence Erlbaum.

Miikkulainen, R. (1990). A distributed feature map model of the lexicon, Proceedings of the 12th Annual Conference of the Cognitive Science Society, . Hillsdale, NJ: Lawrence Erlbaum Associates.

Miikkulainen, R. (1993). Subsymbolic natural language processing. Cambridge, MA: MIT Press.

Miikkulainen, R., & Dyer, M. (1991). Natural language processing with modular neural networks and distributed lexicon. Cognitive Science, 15, 343-399.

Morgan, J., & Travis, L. (1989). Limits on negative information in language input. Journal of Child Language, 16, 531-552.

Morton, J. (1970). A functional model for memory. In D. A. Norman (Ed.), Models of Human Memory, . New York: Academic Press.

O'Grady, W. (1987). Principles of grammar and learning. Chicago: Chicago University Press.

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

Pinker, S. (1989). Learnability and cognition: the acquisition of argument structure. Cambridge: MIT Press.

Pollard, C., & Sag, I. (1994). Head-driven phrase structure grammar. Chicago: Chicago University Press.

Sejnowski, T. J., & Rosenberg, C. R. (1988). NETtalk: A parallel network that learns to read aloud. In J. A. Anderson & E. Rosenfeld (Eds.), Neurocomputing: Foundations of research. Cambridge, MA: MIT Press.

St. John, M. F., & McClelland, J. L. (1988,). Learning and applying contextual constraints in sentence comprehension. Paper presented at the Artificial Intelligence.

Wexler, K., & Culicover, P. (1980). Formal principles of language acquisition. Cambridge, Mass.: MIT Press.