Connectionist approaches to language learning*

GERT WESTERMANN, NICOLAS RUH, AND KIM PLUNKETT

Abstract

In the past twenty years the connectionist approach to language development and learning has emerged as an alternative to traditional linguistic theories. This article introduces the connectionist paradigm by describing basic operating principles of neural network models as well as different network architectures. The application of neural network models to explanations for linguistic problems is illustrated by reviewing a number of models for different aspects of language development, from speech sound acquisition to the development of syntax. Two main benefits of the connectionist approach are highlighted: implemented models offer a high degree of specificity for a particular theory, and the explicit integration of a learning process into theory building allows for detailed investigation of the effect of the linguistic environment on a child. Issues regarding learnability or the need to assume innate and domain specific knowledge thus become an empirical question that can be answered by evaluating a model's performance.

1. Introduction

How is language learned? Connectionist models have in the past twenty years begun to provide novel answers to this question. These models are radically different from traditional linguistic theories in most aspects: they do not contain explicit rules or symbol manipulation processes; they learn from exposure to a language environment and are sensitive to the statistical structure of this environment on different levels; and they exist in the form of implemented computational models, enforcing a very high level of specificity in their underlying assumptions, processing mechanisms, and generated outputs. In this paper we present an overview of the connectionist approach and its application to problems in language development.

Linguistics 47–2 (2009), 413–452
DOI 10.1515/LING.2009.015
0024–3949/09/0047–0413

In the first part of the article we highlight a number of properties of connectionist models that make them attractive for the study of language development and that set them apart from traditional linguistic theories. We then explain in detail the principles of how connectionist models work and how they can be analyzed, using the most common architecture, feed-forward networks as an example. This is followed by describing two additional widely-used architectures, simple recurrent networks, and feature maps, as well as "constructivist" models that change their architecture during learning.

In the second part of the article we discuss the application of neural networks to several aspects of language development. A range of models are reviewed including models of speech sound development, speech segmentation, lexical development, inflectional morphology and the development of syntax. The theoretical stance behind the construction of these models is elaborated. We conclude with a discussion of the connectionist approach to language learning which highlights the conceptual differences to traditional symbolic approaches and provides guidelines to aid evaluation of the strengths and weaknesses of a specific model.

2. What are connectionist models?

Artificial neural network models — also called connectionist models especially when used in psychology — are computer models whose functionality is loosely inspired by neurons in the brain. These network models assume that the main function of biological neurons is to receive activation from other neurons and to become activated themselves if the summed incoming activation is high enough. Neurons (also called "units" or "nodes") are interconnected so that activation flows through the entire neural network. An important property of neural networks is that they can learn from data. Learning happens by changing the strengths of the interconnections (corresponding to synapses in biological systems) between neurons as the result of exposure to a stimulus or a set of stimuli. From these basic principles follow several properties of connectionist models that make them a useful tool for investigating language development, and that set them apart from other linguistic theories.

2.1. *Emergent complex behavior*

Each unit in a neural network functions in a very simple way: it merely adds up the activation arriving through its incoming connections and

transforms this input, generally through a non-linear function, into an activation value that is then passed on to other neurons through its outgoing connections. The complex behavior often observed in neural network models emerges from the interactions of a large number of these neurons (typically ten to several hundred). This is different to linguistic theories in which explicit statements about the combination and transformation of symbolic structures are made. Because neural networks do not contain explicit symbols or rules this type of processing is also called subsymbolic (Chalmers 1992).

2.2. *Knowledge in the weights*

Knowledge in a connectionist network is not stored in one specific location but is encoded across the network in the strengths of the connections (''weights'') between the units. Weights vary continuously and are adapted in response to a learning algorithm, such as Hebbian learning or ''backpropagation'' (see below). A change in the weight matrix corresponds to a change in the network's knowledge. In a neural network there is therefore no physical separation between memory and process. The weights that encode knowledge are the same weights through which activation flows when a stimulus is processed by the network.

2.3. *Learning from the environment*

Learning is driven by exposing the network to a training environment that is representative of the problem of interest. In the linguistic domain such a training environment might consist of sequences of words in syntax learning tasks, or verb stems and their corresponding past tense forms in inflection learning tasks (e.g., Elman 1990; Plunkett and Marchman 1993). Initially the weight values in a network are set to random values, which results in unsystematic patterns of activity propagating through the network. However, over successive exposures to training patterns, the learning algorithm configures the weight matrix so that the network responds in a systematic fashion. The network thus learns exclusively from exposure to a simulated environment by adjusting its connection weights, and the nature and frequency of the stimuli will have an effect on the developing weight matrix and the behavior of the model. This stands in contrast to linguistic theories which often postulate a weaker and less precise link between learning and the environment.

2.4. *Generalization to new data*

After a neural network has learned by adapting its connection weights, it can be used to examine generalization to novel stimuli that do not form part of the training environment. The way in which the network generalizes will depend on the relationship between training and test stimuli. For example, learning about one regular verb will help the network to inflect other regular verbs. In other cases, the network will overgeneralize a transformation to inappropriate patterns. For example in past tense learning, if a network has learned that the past tense of *swim* is *swam* it might generate the past tense of *bring* as *brang* (McClelland and Patterson 2002). Investigating the generalization pattern of a model can give valuable insights into the role of previous knowledge on performance.

2.5. *Developmental modeling*

Neural networks learn from environmental data and therefore provide an excellent tool to study the processes of learning and development in children. Many domains of language development are characterized by specific changes in proficiency and error patterns, and the aim of the connectionist modeler is to replicate these changes in the model. In this way, connectionist models can give insight into the mechanisms underlying developmental change and explain how change arises from interactions between the learning organism and a structured environment (Elman 2005).

2.6. *Linking brain and behavior*

Artificial neural networks are inspired by the functioning of the brain (McLeod et al. 1998). Although the model of a neuron in a connectionist model is a gross abstraction of biological neurons leaving out specific processing properties such as temporal spikes, complex processing of incoming activation and modulation of activations through different chemicals, and ignores other properties such as the exclusively excitatory or inhibitory nature of a neuron and the specifics of synaptic adaptation, the modeler assumes that the essential properties of neural processing have been retained. In this way, connectionist models serve to link brain and behavior: they can help answer the question of how a specific type of processing can be achieved in a brain-like architecture (in the broadest sense). As such, connectionist models offer accounts of behavior that

often cut across traditional levels of description in a manner that highlights the importance of the implemented details of computational processes.

2.7. *Specificity*

In common with all formal, implemented models of cognitive processing, connectionist models provide highly specific, testable hypotheses. A connectionist model, given a certain input, will generate a specific output, and this level of specificity allows for a detailed examination of the validity of a model and its underlying theory of processing. The output of a model can, for example, be compared to the production of a child learning language both in terms of specific outputs and the statistical properties of a range of outputs (such as percentage of errors in a set of utterances), and it can be used to generate predictions of how a child would generalize to novel circumstances (for example being asked to inflect novel words such as *wug*). The level of specificity provided by computational models implies a caveat: implemented neural networks will sometimes generate predictions that are wrong, whereas vaguer theories might not even address this level of specificity and therefore be less prone to criticism of producing false predictions. It would be the wrong approach in this case to abandon a fully specified model in favor of a more vague and underspecified theory.

3. How connectionist models work

A specific neural network model is defined by its ARCHITECTURE, the way in which this architecture is adapted in response to the processed stimuli (LEARNING), as well as the type and form of the stimuli processed by the model (INPUT/OUTPUT). In the following sections we will elaborate on these features on the basis of the most common network architecture, the three-layer feed-forward error-backpropagation (''backprop'') network.

3.1. *Feed-forward networks*

The units in most connectionist networks are organized in different layers of units. Figure 1 shows a standard 3-layer feed-forward network. This architecture consists of an input layer that receives inputs from the environment, an output layer that provides output to the environment, and a

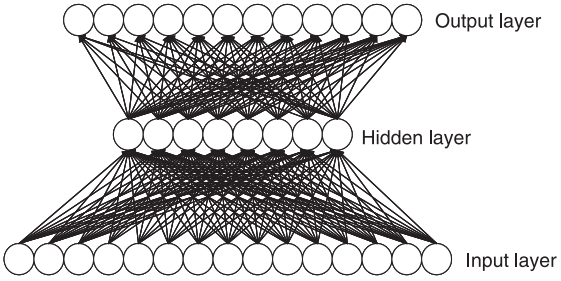

Figure 1. *A three-layer feed-forward network. Activation flows from the input layer through the hidden layer to the output layer.*

hidden layer located between the input and output layers that does not directly interact with the environment. All input units are connected to all hidden units which, in turn, are fully connected to the output units. The flow of activation in such feed-forward networks is unidirectional, resulting in each stimulus being processed in two steps: first, a numerical representation of a stimulus is presented to the input layer and the activation of the input units is propagated forward through the weighted connections into the hidden layer, where the hidden units calculate their own activation state from this incoming activation. The transformation from incoming to outgoing activations in a unit is generally achieved by a non-linear function. In the second step, the activation state of the hidden layer is then sent through the weighted connections to the output units, thus determining the pattern of activation in the output layer which constitutes the network's response to the input stimulus.

It is, of course, possible to conceive of different network architectures. Recurrent networks, for example, have feedback connections from higher layers to lower ones, or they allow for a unit to be connected to itself. This leads to the recirculation of activation within the network and adds a time component to the processing of an individual input exemplar because the production of a stable output pattern might take much more than two steps. Also, units within specific layers might be connected to one another, thus introducing an element of lateral competition. Other networks have been developed that add and delete units and connections during learning. Even the distinction between input, hidden, and output units need not necessarily be strictly upheld, as a specific unit can act as either, depending on the task or even the pattern processed. We will introduce some of these more sophisticated models and their theoretical motivation in later sections. For the purpose of explaining the basic principles, however, we will stick to the classical architecture described above.

3.1.1. *Learning in feed-forward networks.* Prior to learning, the activation of the output units in a neural network will be unsystematic because initially all connection weights are typically initialized to small random values. To pick a specific example, if a network given the task of learning the past tense of verbs is presented with the stem *walk*, activation is sent through connections with random weight values, leading to a random activation of the output units. In order to learn to produce the correct output (a numerical representation of *walked*), the model has to adjust its connection weights appropriately. In the (most) standard case, this learning process works as follows: first, the discrepancy (often called "error") between the output activation and the desired activation pattern (the "target" pattern) is calculated, and then all weights in the network are adjusted by a small amount so that the resulting error would be smaller were the same stimulus to be processed again. Over successive exposures to training patterns, the learning algorithm configures the weight matrix so that the network responds in a systematic fashion and, eventually, produces the desired outputs for each of the stimuli — it has learned the task.

The mathematical equation according to which the changes to the weights are calculated is termed the "learning algorithm". Similar to and often in conjunction with different network architectures, a large variety of learning algorithms have been devised over the years. Feed-forward models tend to use SUPERVISED learning algorithms, the most common of which is called ERROR BACKPROPAGATION (Rumelhart et al. 1986). As described above, these algorithms require a target signal that is used to calculate and ultimately to minimize the observed error through incremental weight adjustments. This fact has led to some controversy about the use of these models in language tasks, because in natural language learning there is often no explicit feedback available to the child. One possible justifications for a teaching signal comes from conceptualizing the output of the network as corresponding to a prediction made by the child (e.g., that the past tense form of *eat* is *eated*), and subsequent exposure to the correct form (in this case, *ate*) would lead to detection of the discrepancy between self-generated and perceived forms.

Learning in a connectionist model is thus a general process that consists exclusively in the gradual adaptation of connection weights in response to exposure to environmental stimuli. The meaning of these stimuli is irrelevant to the model's adaptation; adaptation proceeds in the same manner for speech-like sounds, words, word classes and so on. Importantly however, the statistical distribution of stimuli has an effect on the developing weight matrix and the resulting behavior of the model. This is because a frequently occurring stimulus will lead to more weight

adaptation steps than a less frequent one. The order in which stimuli are experienced also has an effect on learning. This is because subsequent patterns are processed through the connections that have already adapted to previously experienced patterns, and the nature of this previous knowledge will affect the adaptation to the current stimulus. Returning to a past tense example, a model that has already adapted to learn the past tense forms of *sing*, *ring* and *swim* will need to adapt more to learn the correct past tense form of *bring* than a model that has not seen any of these forms. The sensitivity of neural network models to the order in which stimuli are learned has been used to explain age of acquisition effects in adult language processing (Ellis and Lambon-Ralph 2000).

An important factor in a connectionist model's ability to learn a given task is controlled by its "learning rate", that is, the small amount by which the weights are adjusted in response to a given training stimulus. Because each of these adjustments is the equivalent of the network's attempt to become better at processing one specific exemplar, too large a learning rate can be problematic: the same connections are used to process all exemplars in a learning task, and adapting to one specific stimulus often undermines the network's ability to deal with other stimuli. Too small a learning rate, on the other hand, will reduce the speed of convergence to a correct solution and might even prevent the network from finding an optimal weight setting at all, due to the learning process getting stuck in a "local minimum": a network algorithm aims to reduce error at every step, but sometimes accepting a slightly higher error temporarily in order to then achieve a greater error reduction would be necessary. A larger learning rate can help avoid these local minima. Many learning algorithms include other mechanisms that reduce the risk of getting trapped in local minima, for example a so-called "momentum term" in standard backpropagation (Rumelhart et al. 1986).

The process of learning, then, can be seen as an effort to find a single configuration of weights that supports the mapping for all exemplars. Unless there are very few exemplars, a network will not be able to learn this task by rote learning of all required mappings but instead will have to extract regularities that are implicit in the given mapping. This extraction of an abstract regularity in the mapping from input to output also means that the model becomes able to generate meaningful outputs for novel exemplars, i.e., stimuli that the model has not seen during training. This emergent generalization ability has proven useful in studying different aspects of language learning.

3.1.2. *Input and output coding.* One of the first steps in either constructing or evaluating a specific connectionist model consists in deter-

mining the type and form of the input and output patterns with which it is trained. Similar to any other model or theory, the type of data will depend on the process under investigation and will be mainly driven by theory. For example, words might be suitable as the basic coding unit for a model of sentence comprehension whereas phonemes, possibly combined with stress patterns or explicit morphological markers, seem adequate for a model of inflectional morphology. As computational models, however, connectionist networks require the chosen content to be converted into some kind of numerical representation, and this can be done in several ways.

Consider a model whose input should consist of a phonemic representation of words. The simplest idea would be to employ a ''localist'' coding scheme in which there is one dedicated input unit per phoneme. This choice results in the need to have as many input units as there are possible phonemes in the language under investigation. More importantly, however, it means that similarity relations between different phonemes are not reflected: for example, if the phoneme /p/ activates one specific input unit, and the phoneme /b/ a different input unit, and the phoneme /a/ yet a different unit, there is no overlap between the activation patterns for all three phonemes, and to the model, /b/ will be as (dis)similar to /a/ as it is to /p/. It seems clear that such a coding scheme is unsuitable when modeling processes that are expected to depend on phonological similarity, such as past tense formation. In this case it is desirable to employ a scheme that preserves some of the similarities between different inputs. A distributed, feature based coding scheme, for example, could encode individual phonemes based on a combination of (binary) phonological features such as ''aspirated'', ''voiced'', ''labial'', etc. In this case, /b/ and /p/ would only differ in their value of the feature unit for ''voiced'' but would both activate input units e.g., for ''plosive'' and ''bilabial''. This makes it evident that the choice of coding scheme is a very important step in developing a connectionist model. The specific set of features will impose a similarity structure on the model, thus biasing a system that is highly sensitive to such similarities in a very specific way. In situations where the ''true'' similarity structure is uncertain it might be preferable to use a localist coding scheme, because a feature based coding might introduce an incorrect bias and thus present a possible confounding factor.

At this point a note on terminology might be warranted. Note that a distributed encoding, e.g., of phonemes through phonetic features, is in fact localist on the level of phonetic features where each feature is represented by one distinct unit. It is therefore important to consider at which level it is desirable for a representation to be distributed and reflect

similarity relationships between stimuli, and at which level a localist encoding is sufficient.

3.1.3. *Hidden layer representations.* When activation flows through a feed-forward network to generate an output pattern, it can be very instructive to observe the activation profile of the units in the hidden layer. This is because we can understand the mapping from an input to an output as a representation of this input, and the main role of the hidden layer is to allow for complex such representations. As the same connections adjust to solve the input-output mapping task for all patterns, observing the hidden layer representations for all input patterns allows us to investigate how the model solves this mapping problem. In most cases each input to the model will lead to activation of all hidden units to some degree and it is unusual to find that a specific hidden unit comes to represent any meaningful concept. Instead each unit acts together with all the other hidden unit activations to enable the network to produce the correct responses. These distributed patterns of activation do, however, possess an amount of internal structure because the representations evoked by different stimuli overlap systematically with one another. In the model that produces the corresponding past tense forms when presented with a verb, for example, rhyming verb families such as *drink*, *sink*, *stink*, will form overlapping representations because of their phonological similarity in both stem (input) and past tense (output) form. An exception (regular) word like *blink*, however, will show less overlap with its phonological family, but might share more resources with other regularly inflected words that also require *-ed* suffixation. Hidden layer representations thus reflect the tension between similarity of inputs and similarity or dissimilarity of their corresponding outputs.

A common way of visualizing the distributed representations that emerge in a neural network during the process of learning is to conceptualize the hidden activation pattern for each stimulus as a point in a multidimensional space. By probing the network with different stimuli it is thus possible to record several of these points and to analyze their relationship with each other. Applying mathematical dimension reduction or clustering techniques (e.g., principle component analysis, multidimensional scaling or cluster plots) gives a snapshot of the inner workings of the model which allows the modeler to draw conclusions about the kind of regularities the network has extracted at a given point in the learning process. This is especially interesting with regard to generalization, i.e., the network's treatment of previously unseen exemplars. If, for example, the above network has formed a tight cluster of representations for words like *drink*, *sink*, *stink*, etc., how will it treat a rhyming nonword such as

nink? The answer to this question is far from obvious. On the one hand, phonological neighbors such as the above, but also more distant relatives like *ring* or *sing* make it likely that the connection weights are configured such that the novel stimulus *nink* will be close to the existing cluster in representational space and thus produce the output *nank*. However, the network could equally treat *nink* as similar to *blink* or *think*, especially if these competitors are highly frequent and have had a larger impact on the weight configuration as compared to less frequent stimuli. The network's response to this novel exemplar is the product of many different forces such as the number of the exemplar's phonological friends and enemies, their respective degrees of similarity, their frequencies, and even how recently they have left their trace in the weight configuration. How the model generalizes thus becomes an empirical question which is decided by the output that it produces. Various tests with many models have shown, however, that connectionist models often perform comparably to people in generalization tasks (Daugherty and Seidenberg 1992; Westermann 1998).

3.2. *Other network architectures*

Having discussed the main principles of connectionist models on the basis of feed forward networks, we turn now to other connectionist architectures that have been developed and used in models of language development. Indeed we see a trend away from simple 3-layer error backpropagation models and towards using more complex architectures and approaches that are more constrained by neurobiological considerations (Westermann et al. 2006). Two of these alternative architectures are simple recurrent networks and feature maps.

3.2.1. *Simple Recurrent Networks.* Feed-forward models cannot directly represent time and are therefore unable to process temporal sequences of stimuli. This limitation can be circumvented by converting a temporal sequence into a spatial representation. For example, in a feed-forward model of speech perception, the speech sounds of a word can be presented to the model all at the same time. However, a relatively simple modification to the network's architecture can equip the model with an ability to deal directly with sequential input. True sequence processing is possible by extending a feed-forward model with recurrent connections, the approach taken in simple recurrent networks (SRN; Elman 1990).

The architecture of an SRN is similar to a regular 3-layer feed-forward network but for one important extension: the hidden units are connected

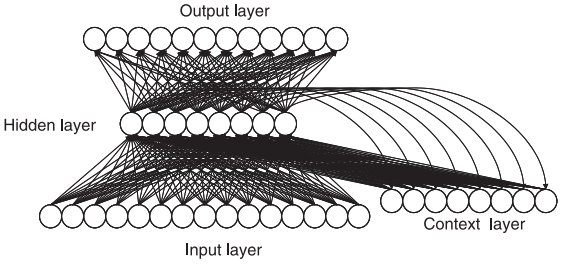

Figure 2. *A simple recurrent network. The hidden layer activation is copied to the context layer and fed back into the hidden layer at the next time step. Connections from the hidden units to the context units are one-to-one copy connections. All other connections are adjustable.*

to a context layer, and this context layer feeds back into the hidden layer (Fig. 2). In effect, an SRN retains a copy of the hidden activations from a specific time step in its context layer, and adds this activation profile to the input for the following time step. This provision equips the network with a memory capacity so that previous activation states of the network can influence subsequent activation states. The content of the context layer at a certain point, however, consists of a similar superposition of the previous input and the previous context layer activation, which explains why SRNs can also show sensitivity to information from even earlier processing steps (this is decisive in mastering "long distance" dependencies (e.g., number agreement) between words in the syntactic string, see Elman 1993). The connections from the context to the hidden layer are adjusted in the same way as other connections in the network. Because learning is driven by the pressures of the task, the context layer will come to represent and maintain only those pieces of information from past processing steps that are useful with regard to a future output.

SRNs are most commonly used to carry out PREDICTION TASKS: Given a sequence of events, the network is trained to predict the next event in the sequence. Training is achieved by presenting an event to the input units. Activity propagates through the network to produce a pattern of activation across the output units which constitutes the network's prediction of the next event in the sequence. Insofar as this prediction is accurate, the connections in the network remain unchanged. When the network's prediction does not correspond to the following event in the sequence, weight adaptation is usually done with the same error backpropagation algorithm that is also used for many feed-forward networks. As a consequence, predictions become increasingly accurate with repeated presentations of the training stimuli. Note that the most accurate prediction

achievable by such a network corresponds to the conditional probability of an event to occur next. If, for example, event A is always followed by either event B1 or event B2, the successfully trained network will activate both representations to an extent that reflects the relative frequency of the respective transitions.

SRNs are useful for capturing the statistical distribution of events in a structured sequence. For example, given a sequence of phonemes in a single utterance, the SRN will learn to predict the order of the phonemes in the utterance. Given a range of utterances, the SRN will learn to predict which phoneme sequences are most likely to occur in running speech. In particular, it will learn which phoneme sequences go together to make up words, and which phoneme transitions are probable in a language, and by implication, those which never occur or are highly improbable. As the network is presented with more and more information about a word, its prediction concerning the next phoneme becomes increasingly accurate and constrained, demonstrating a "cohort effect" (Marslen-Wilson and Welsh 1978). However, when the network reaches the end of the word it will be very poor at predicting the next sound since there are many possibilities for the following words. In essence, the network is learning to segment speech into words by discovering the phonotactic regularities of the language.

Network architectures like SRNs will automatically exploit the co-occurrence relations in a sequence of events in order to achieve the goal of accurate prediction. Co-occurrence relations may simply be the sequence of phonemes that combine to make a word, or the general likelihood, within a whole corpus of utterances, of a phoneme following another phoneme. However, SRNs can also calculate the correlations between regularities ACROSS different levels of structure. For example, both prosodic information and phonotactics can yield parallel and converging cues to word boundaries. An SRN can exploit these converging constraints to assist in the prediction task (Christiansen et al. 2005).

3.2.2. *Feature maps.* Feature maps are very different from feed-forward architectures. They are inspired by the topographic mappings found in many areas of the cortex such as visual, auditory, sensory and motor cortices. These maps are characterized by their topographical organization where neighboring neurons respond to similar stimuli, e.g., spatially close visual inputs or tones of similar pitch. Feature maps do not have a teaching signal but instead self-organize and cluster their inputs on a two-dimensional grid of neurons in a topographic manner. This property makes them useful for projecting high-dimensional data onto two dimensions while preserving similarity relations (albeit not the

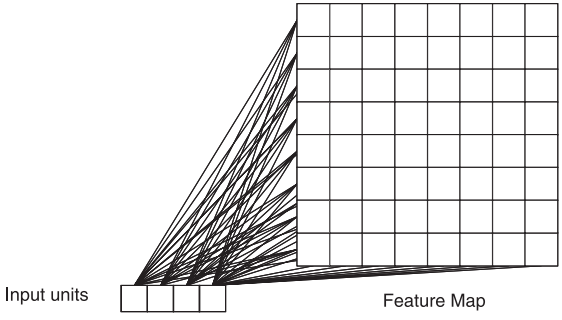

Figure 3. *A self-organizing feature map. Four-dimensional inputs are projected on the two-dimensional feature map.*

distances between different inputs). While several feature map models were developed on the basis of neural information processing in the brain (e.g., Willshaw and von der Malsburg 1976), the best known feature map model is the Kohonen Self Organizing Feature Map (SOFM, Kohonen 1982). This map consists of an input layer which sends activation through weighted connections to a map layer typically consisting of a square grid of neurons (Fig. 3). A map unit is maximally activated when the input vector corresponds to its incoming weight vector. Training a SOFM consists in presenting an input, determining the maximally active (winning) map unit, and adjusting the weights to this unit so that they become more similar to the input vector. This will lead to the winning unit becoming even more active when the same input is presented subsequently. A topographic mapping is achieved by not only adjusting the weights to the winning unit but also to all units in a surrounding neighborhood, albeit by a smaller amount. Therefore these neighboring units will come to respond to similar inputs as the winning unit. During training the radius in which neighboring units are adjusted is gradually shrunk to zero, leading to a progressive fine-tuning of the map organization.

Feature maps are useful for studying how a neural system organizes and represents sensory inputs, and how the statistical structure of these inputs affects map organization. In the domain of language, phonological, orthographic or semantic maps can be modeled. On these maps, words with similar orthography, phonology and meaning, respectively, will cluster closely together. Such models have been used to account for category-specific deficits in dyslexia and acquired aphasia (Miikkulainen 1997), development of a semantic space (Ritter and Kohonen 1989) as well as lexical development (Li et al. 2007).

3.2.3. *Constructivist models.* In most connectionist models, the architecture of the network — the number of units and connection patterns — are predetermined and static. However, while in feed-forward networks the number of input and output units is determined by the encoding of the stimuli, it is often difficult to choose the "correct" number of hidden units, although the size of the hidden layer can make a dramatic difference to the way in which a network learns. With too few hidden units it is possible that a network will be unable to learn the required task, but too many units might encourage rote-learning with a loss of generalization ability. Likewise, in feature maps, the size of the map is important to achieve an expected grainedness of clusters, and this might change across development. Finally, research in developmental cognitive neuroscience has shown that many areas of the cortex adapt to the environment in an experience-dependent manner by rewiring themselves in response to environmental input (Johnson 2005).

Constructivist models (Shultz 2003; Westermann et al. 2006) address these issues by allowing the network architecture to develop during learning by adding and removing units and connections. In feed-forward models the learning process often works as follows: the model begins with a minimal architecture, with no or few hidden units, and it attempts to learn a task with this architecture. If this is not possible and the error reduces no further, a new unit is inserted into the network, and learning proceeds in this new architecture until the error stagnates again. Then another unit is inserted, and so on until the task has been learned. This method of developing the network architecture ensures that only as many hidden units as necessary are used, but it also allows for further analysis of the learning process. For example, a question of interest is which stimuli can be learned in a smaller architecture, and for which other stimuli more units have to be inserted. Learning in constructivist networks can thus lead to different learning trajectories from those in static models and may provide insights into learning in a highly plastic brain. Similar constructivist learning exists in feature maps where units are added in various ways to take account of a need for more resources (Fritzke 1994; Li et al. 2004).

4. Connectionist models in linguistic research

Connectionist models have become a useful tool for exploring a number of aspects of linguistic theories. As models of statistical learners they have contributed to our understanding of how much information can be usefully extracted from the environment (e.g., Saffran et al. 1996), and in

doing so they have challenged the poverty of the stimulus argument in language learning (e.g., Pullum and Scholz 2002; MacWhinney 2004). They have also raised the question of how much domain-specific knowledge has to be assumed to be innate. Connectionist models learn aspects of language on the basis of domain-general associative learning mechanisms and therefore minimize the role of innate domain-specific knowledge. A second important contribution of connectionist models to linguistic research is to raise the question whether apparent rule-governed behavior is based on mental symbolic rules or whether this behavior can be explained on the basis of complex associative processes. In many instances this perspective on linguistic processing has motivated experiments that examine in great detail to what degree human performance shows influences of frequency and similarity in apparently symbolic processes (e.g., Seidenberg and Bruck 1990; Marchman 1997; Ullman 1999; Abbot-Smith et al. 2004; Joanisse and Seidenberg 2005). Thirdly, connectionist models can provide an integrated account of language learning, adult processing and deficits in acquired or progressive disorders. This is because a connectionist model moves through a learning process to reach an adult-like state. Then, a neural network model can be artificially lesioned by removing some of the processing units or connections or by adding "noise" to the input data, and the patterns of breakdown can be compared with those of brain-damaged patients (e.g., Plaut et al. 1996; Joanisse and Seidenberg 1999; Penke and Westermann 2006).

A large number of connectionist models have been developed to account for a range of phenomena in language development. We now describe a number of such models spanning linguistic levels from speech sounds to syntax that have made a contribution to explaining how different aspects of language can be learned in associative models.

4.1. *Speech sound development*

A number of connectionist models (Yoshikawa et al. 2003; Westermann and Miranda 2004; Guenther et al. 2006) have shown how a repertoire of speech sounds specific to the infant's native language can emerge from links between sensory and motor brain areas that emerge through babbling. In broad terms, these models have in common a neural map on which unit activation patterns encode articulatory commands (speech gestures) and an auditory map representing speech sounds. These maps are linked by adjustable weights. The models "babble" by articulating random speech sounds which lead to activation patterns on both the articulatory and auditory maps. Connections between these maps are then

tuned to reinforce the mapping from articulation to sound. Weight adaptation proceeds in variants of Hebbian learning (Hebb 1949), a biologically plausible method of weight adaptation in which connections between units on the different maps that are co-active are strengthened. The mechanisms for adaptation to the native language differ between these models: the model by Yoshikawa et al. (2003) is based on the assumption that caregivers tend to imitate the sounds produced by infants, thereby providing a native-language target to the babbling infant. The advantage of this approach is that it solves the problem of speaker normalization when the same speech sounds are uttered by speakers with different articulatory systems. The disadvantage is that learning relies heavily on the assumption that caregivers reliably imitate infants. The model by Westermann and Miranda (2004) assumes that babbling first creates a broad mapping between articulation and perception and that the links for native speech sounds are then selectively reinforced through exposure to the ambient language. This mechanism bears close resemblance to the "articulatory filter hypothesis" (Vihman 1993, 2002) which suggests that after the onset of canonical babbling an articulatory filter begins to highlight those speech sounds in the environment that correspond to vocal patterns produced by the infant herself and facilitates motoric recall of these patterns. As a consequence these patterns become particularly salient to the infant and can serve as building blocks for first words.

Although these models involve links between perception and production they are different from the motor theory of speech perception (Liberman et al. 1957; Liberman and Mattingly 1985). Whereas the motor theory assumes an innate link between perception and production that allows the direct perception of articulatory gestures, the connectionist models show that these mappings can be learned and need not be prespecified. Given the principles of learning in connectionist systems the models also predict that the precise nature of mappings is sensitive to the statistical structure of the language environment.

In a different model that only involved the auditory domain, Guenther and Gjaja (1996) used a variant of the self-organizing feature map to explain the emergence of a perceptual magnet effect (PME) in vowel perception. The PME (Kuhl 1991) describes the organization of the perceptual vowel space where the regions around prototypical vowels are compressed so that two stimuli close to a prototype cannot be well discriminated, but speech sounds falling in two different classes can. Based on their model, Guenther and Gjaja showed that a PME can occur without explicit storage of category prototypes (Kuhl 1995), but solely on the basis of the statistical structure of the sounds experienced by the infant. According to this model the PME arises from a warping of the perceptual

space which is a consequence of adapting the firing preferences of auditory neurons. This explanation sees the perceptual magnet effect as an emergent consequence of the formation of cortical maps in the auditory system. In this way, the PME was explained as a simple perceptual phenomenon instead of a high-level linguistic phenomenon in which the infant would make linguistic category decisions for each heard sound (Lacerda 1995) or construct explicit prototypes for each phonetic category (Kuhl 1995).

An evolutionary perspective on the categorical perception of speech sounds was pursued in a model by Nakisa and Plunkett (1998). Through a process akin to natural selection they applied the evolutionary approach to modelling a fundamental ability in infant speech perception — the ability to discriminate categorically between speech stimuli that vary along a continuum, such as the syllables /ba/ and /pa/.

Nakisa and Plunkett (1998) used real speech input taken from a database consisting of a phonetically tagged corpus of speech (Garofolo et al. 1990). A large population of networks was generated so that a wide range of learning algorithms could combine with a wide range of architectures in a random fashion. Each network in the population was presented with the speech samples. Whilst the speech was presented, the learning algorithms in a network responded to the activity patterns by updating their connection weights. There was no error signal to indicate whether the activity patterns were accurate. At the end of the training period, the speech corpus was presented again, but this time without any learning and the networks' internal representations of each phoneme in the corpus of speech were recorded. A network was deemed to have a good phonetic code if its representations of different tokens of the same phoneme were similar to each other and tokens of different phonemes were dissimilar to each other. The networks with the best representations were allowed to REPRODUCE and the worst networks were removed from the population. It is important to note that parent networks did not transfer any of their "lifetime" experiences to their offspring, i.e., no information about the changes in the parents' connections were inherited. The only information that was inherited concerned the ARCHITECTURE of the parent networks (with a little bit of mutation thrown in).

The process of selection and reproduction continued for 10,000 generations. At the end of this evolutionary period, the best networks were exposed to just two minutes of speech, and then tested on various speech continua (such as the /ba/–/pa/ continuum). The networks exhibited categorical perception of these continua (though only for consonants and not for vowels), mimicking the pattern observed in humans (Lisker and Abramson 1971). The networks also confused phonemes when presented

against a background of white noise in a manner which resembled human performance (Miller and Nicely 1955). Furthermore, it was found that training the evolved networks on a limited sample of speech produced very similar results regardless of the language of training. Hence, a training sample of two minutes of speech produced the same outcome irrespective of whether the language was Cantonese, Swahili or Hungarian. However, training the network on white noise or low-pass filtered speech failed to reproduce the categorical perception and confusability results.

These modeling results indicate that the evolutionary process had selected a network architecture that was well-adapted to the categorical perception of speech. Furthermore, the architecture did not seem to be language-dependent. Any speech was equally good at generating the right kind of internal connectivity in the network. However, the acoustic stimulation needed to be speech — at least, white noise and low-pass filtered speech did not work. This finding shows that the proper configuration of the network was reliant on the structure inherent in the speech signal itself. The Nakisa and Plunkett (1998) model is a good example of what Elman et al. (1996: 27) refer to as ARCHITECTURAL INNATENESS. Inspection of the network architectures revealed that the best networks all exploited some version of Hebbian learning and used recurrent connections. It is known that this type of learning algorithm and architecture is involved in many other human brain processes. It is perhaps, then, not so surprising that the heavily simplified evolutionary process described here came up with the same results. What is less clear is whether the details of the architectures evolved in the model are specific to the processing of speech. For example, the network architectures may have been well-suited to the processing of non-linguistic stimuli or even tactile or visual information. These questions, however, were not pursued here and are subject to future research.

4.2. *Lexical segmentation and word learning*

Many linguistic theories operate at the word level. They are concerned with questions regarding the storage, manipulation or ordering of words. From a learning perspective, however, another question needs to be addressed first: How do children acquire the ability to break down the continuous auditory stream into words (Jusczyk 1999)? Several connectionist researchers claim that a child's sensitivity to distributional features of the speech input can explain not only how lexical segmentation is performed, but also how it is acquired.

The sequential nature of the domain requires models that are capable of processing sequences of inputs as temporal information. The SRN is

useful for capturing the statistical distribution of events in a structured sequence and it is this feature that is exploited in several existing models of speech segmentation (Elman 1990; Cairns et al. 1997; Christiansen and Allen 1997; Christiansen et al. 1998). Importantly, this approach does not rely on the assumption of inbuilt linguistic knowledge. Rather, it is an SRN's exploitation of co-occurrence relations in a sequence for the purpose of performing the immediate task of generating accurate predictions that gives rise to an emergent concept of word boundaries in terms of the locations in the speech signal where the extracted regularities do not hold. In the simplest case, an SRN performing a prediction task on a continuous stream of phonemes will come to extract a notion of the phonotactic regularities of a language (Elman 1990; Cairns et al. 1997). Phonotactics, however, are clearly not the only type of information that can be exploited for speech segmentation. There is a host of additional sublexical cues as to what constitutes a word, such as prosodic patterns (stress, pauses), utterance boundary information, co-occurrence with referential objects, etc. None of these cues is in itself a reliable predictor of word boundaries, but in combination they lead to improved speech segmentation performance of an SRN model when included in the input signal. Christiansen et al. (1998) provided a predictive SRN with additional cues concerning utterance boundaries and metrical stress and found that their model was able to detect 74% of word boundaries (see also the contribution of Hockema and Smith, this volume).

From a mechanistic point of view, we can understand the advantage of using multiple probabilistic cues in terms of their constraining effect on the weights. Any kind of additional information, even if unreliable, constrains the possible solutions in terms of viable weight configurations, thus often facilitating learning and increasing the likelihood that the solution generalizes well (Christiansen et al. 2005). This insight into the utility of multiple probabilistic cues is likely to apply to other domains of language learning as well (Morgan and Demuth 1996), and implemented models of this principle are redefining our understanding of what is learnable.

Models such as the above imply that a system with an inbuilt sensitivity to the distribution of multiple segmentation cues in a continuous speech signal can give rise to an emergent notion of word boundaries. In the models described so far, however, this notion of word boundaries is captured indirectly in terms of high prediction error. Furthermore, these models experience problems in specific situations such as segmenting words that also constitute onsets of longer words — for example the word *cap* that can be embedded in the word *captain*. These situations pose a serious challenge for a system without top-down influences from

a lexical level, and empirical evidence suggests that such top-down influences do play a role in adult speech segmentation (Gow and Gordon 1995). Davies (2003) took this as motivation to augment a predictive SRN model with an additional output layer that attempts to produce a static representation of the words in a sentence from the ongoing stream of phonemes. Observing which words are recognized (this is defined as their output activation surpassing a specific threshold) gives a notion of the model's growing receptive vocabulary during the learning process and includes an element of vocabulary acquisition in the model that matches the developmental profile in children, e.g., it exhibits a vocabulary spurt (Fenson and Pethick 1994). Furthermore, it is possible to analyze how many phonemes the network needs to have seen prior to recognizing a specific word. Davies (2003) showed that his model provided a good fit to eye-movement data in that (a) words were recognized increasingly earlier with increased training on phoneme sequences despite the growing vocabulary and (b) recognition points varied according to a word's lexical environment, i.e., the amount of other words with similar onsets. These results concerning the lexical identification of words from a stream of phonemes were found to be independent of whether the network simultaneously tried to predict the next phoneme or not. Including the predictive part, however, not only helped the model by providing indirect segmentation cues (prediction error) for words that were not yet assimilated in the receptive vocabulary, but it also increased performance in lexical identification. Again, additional cues — this time in the form of an extra output layer and task — imposed further constraints on the weight configuration, leading to faster vocabulary acquisition, earlier recognition points and increased discriminability of phonemes.

The connectionist approach to speech segmentation demonstrates that SRNs can integrate multiple and partially unreliable cues across different levels and timescales, making the most of the information available. This, of course, casts doubt on the concept of distinct linguistic levels that can be investigated independently from one another.

4.3. *Lexical development*

Several models have addressed the question of how the mental lexicon develops (Li 2003; Li et al. 2004; Li et al. 2007). The lexicon is here generally conceptualized as a link between phonological and semantic word representations, and the questions of interest are how word space and semantic space as well as the links between these spaces develop. One recent model (Li et al., 2004) used two self-organizing maps together with links

between these maps to account for several phenomena in the development of lexical categories. The aim of this model was to provide a unified view of developing cortical maps and the dynamics of the developing vocabulary. The most recent version of this model (Li et al. 2007) was augmented with a third self-organizing map with sequential characteristics to allow for production of phoneme sequences. The model was trained on 591 words extracted from the CDI (MacArthur-Bates Communicative Development Inventories, Dale and Fenson 1996), including nouns, verbs, adjectives and closed-class words. The model's input phonology map developed abstract topological representations of heard words on the basis of distributed phonetic features. The semantic map developed in response to distributed semantic representations of these words that were obtained by calculating the co-occurrence statistics for the chosen words from a large corpus of child-directed speech (CHILDES, MacWhinney 2000). The sequence output map dealt with a sequential version of the phonological input where one phoneme was presented at a time and the map, in addition to mapping the phoneme onto its topological structure, attempted to establish an activation gradient that encoded the order in which the phonemes were activated.

In this model all three maps self-organize in an attempt to optimally accommodate the structures inherent in their respective input data. At the same time, however, the links between the maps are updated via simple Hebbian learning. Over time, these links come to perform a mapping from one similarity space to another so that, for example, a perceived word will activate its corresponding semantic representation (comprehension), and this activation of the semantic map, in turn, can drive the activation of an ordered sequence of phonemes (production). The overall system thus organizes as an outcome of multiple interacting constraints.

The dynamic interaction of the increasingly structured self-organizing maps and the simultaneously developing links give rise to a number of quantitative phenomena that have been observed in lexical development. When analyzing average receptive and productive vocabulary size, for example, the model showed a clear vocabulary spurt for comprehension, which was mirrored by a subsequent similar spurt in production. This nonlinear change in the rate at which new words are acquired is related to the fact that consistent associations between maps can only be formed when the self-organization process within the individual maps has reached a somewhat stable state (cf. the ''critical mass'' hypothesis, Marchman and Bates 1994). At this point the activation conveyed by the links becomes meaningful and can be used as an additional cue by the receiving map, thus leading to more efficient self-organization and rapidly increasing performance. While all networks showed non-linear changes

in vocabulary growth, details such as the onset or slope of the spurt varied considerably as a function of the random weight initialization and the density of connections between maps. This variability of the networks matched well with empirically observed patterns of individual differences in lexical development (Thal et al. 1997). Other more detailed phenomena captured by this model include effects of word length and frequency on age of acquisition, the impact of phonological short-term memory on word production, and patterns of recovery after early brain injury. The latter point is of specific interest because it provides a good example of how connectionist models can address typical and a-typical development within one integrated account.

More specifically, with regard to lexical development it has been observed that lesions acquired very early in life have less severe consequences on the final outcome than lesions at a later stage. The model showed a similar pattern when lesioned at different stages during the learning process, where lesions were simulated by resetting a proportion of the weights to random values. The explanation for this phenomenon derives from yet another general principle of connectionist models, sometimes captured by the term "entrenchment" of weights (Altmann 2002; Elman 2005). This term is used to describe the fact that the amount of plasticity exhibited by neural networks tends to decrease with increasing experience. Initial learning requires coarse changes to the networks weights matrix, eventually resulting in a relatively stable configuration that captures the general structure of the task. Further training will usually lead to more fine grained adjustments that may continue to improve performance on a more detailed level. However, at this stage it becomes increasingly difficult to radically reorganize the model's weight configuration. Events that require such a drastic reorganization thus are easier to cope with during the initial stage of basic organization, before a stable state has been reached. Note that this kind of explanation not only speaks to the differential effect of otherwise comparable lesions, but it might also offer a perspective on differences between first and second language acquisition (for an overview see Thomas and van Heuven 2003). Importantly, the changes in network plasticity are not driven by some maturational schedule but rather result from the weights becoming increasingly entrenched as a function of experience.

4.4. *Morphological learning*

Learning inflectional morphology has been a core aspect of connectionist modeling in language acquisition since publication of the first such model

on the acquisition of the English past tense (Rumelhart and McClelland 1986). Their conclusion "that a reasonable account of the acquisition of past tense can be provided without recourse to the notion of a "rule" as anything more than a *description* of the language" (p. 267) has provoked an extended and still ongoing debate about the necessity to stipulate the psychological reality of explicit rules as underlying human linguistic abilities (Marcus et al. 1995; Pinker 1997; Marcus 1999; Seidenberg and Elman 1999a, 1999b; Pinker 2001; McClelland and Patterson 2002; Pinker and Ullman 2002; Seidenberg and Joanisse 2003). The English past tense has been a particular focus, although other paradigms (noun plural, progressive) and other languages (German, Arabic) have also been addressed (Marcus 1995; Plunkett and Nakisa 1997; Clahsen 1999).

Traditional accounts have postulated that the verbs of a language fall into a number of inflectional classes with distinct rules governing the generation of the past tense form. Following this tradition, dual-mechanism (or "words-and-rules") theories (Marslen-Wilson and Tyler 1998; Clahsen 1999; Pinker and Ullman 2002) postulate the existence of two qualitatively different mechanisms that govern the production of regular and irregular forms and explain observed differences through the two mechanisms. More specifically, regular verbs are held to be inflected trough the application of a simple explicit rule, such as (in English) "attach *-ed*". Irregular verbs, conversely, are stored individually along with their corresponding past tense forms. A verb is treated as regular (default) unless a past tense form is found as an entry in the lexicon, in which case application of the default rule is blocked and the retrieved form is used instead.

One of the main challenges for any model of past tense learning is to account for children's overregularization errors and U-shaped learning in which irregular forms that are initially produced correctly are overregularized before being produced correctly again. Dual mechanism approaches posit that the period in which children are susceptible to overregularization errors coincides with their discovery of the default rule which is applied too widely until the appropriate entries for exception words have been acquired (Marcus et al. 1992).

Connectionist models of morphological learning, conversely, are based on the assumption that a single process underlies the production of all verb forms, be they regular or irregular. Differences in the processing of these verbs arise from distributional features such as frequency, neighborhood of similar sounding verbs with the same or different past tense forms, phonological complexity, age of acquisition, etc., all of which will impact on the development of the weight matrix within the network and act as soft constraints on the regularities emerging in the model. Many connectionist models of past tense inflection have been simple three-layer

feed-forward models (MacWhinney and Leinbach 1991; Plunkett and Marchman 1991, 1993; Plunkett and Juola 1999) with a phonological representation of the verb stem as input, and a phonological representation of the past tense form as output, and the task of the model to map between stem and past tense. Sometimes outputs are encoded more abstractly by inflection classes, so that all regular verbs activate one output unit, all verbs sharing the same past tense such as *sing*, *cling*, *ring* another, and so on (Westermann 1998; Hahn and Nakisa 2000).

U-shaped learning according to the connectionist approach arises from the competition between regular and irregular inflection resulting from the sharing of computational resources in the form of units and connections within the network. On this view, the initial phase of correct performance derives from the fact that processing resources are ample in relation to the small vocabulary at this stage, and the model is thus able to solve the mapping by rote learning. As the active vocabulary increases, however, the model comes under increasing pressure to adjust its weight matrix such that it captures general regularities, rather than treating each exemplar on its own merits. Because of their lower (type) frequency, irregular verbs are more likely to be the victims of the competition for representational resources, and the network develops a transient tendency to treat them like regular verbs. With increasing practice, however, the weight matrix will eventually settle into a configuration that does justice to all training patterns, resulting in good performance for both regular and irregular verbs.

This account of U-shaped learning has been criticized because it crucially requires the vocabulary size to be increased during the acquisition process. While the original simulation by Rumelhart and McClelland (1986) included a sudden step from the 10 most frequent words to the full set of 420 monosyllabic verbs, subsequent models have employed an incremental expansion of the training corpus (Plunkett and Marchman 1993; Plunkett and Juola 1999). Similar to Elman (1993), another possibility entails capturing the gradual vocabulary expansion not in terms of a change in the child's language environment, but rather to conceptualize the child itself as undergoing changes which impact on the way in which the static environment is processed. This idea is exemplified by a constructivist neural network model (Westermann 1998): the model starts out with predominantly direct connections from the input to the output layer, and additional hidden layer units are inserted during the acquisition process. Exception words, i.e., those verbs that are disfavored by the distributional factors mentioned above, are more likely to rely on the additional processing power provided by the (growing) hidden layer. For this reason these "hard" verbs are also more affected by the

reorganization process in response to the expansion of the architecture and therefore susceptible to being temporarily overgeneralized, even though the training set is kept static. Through time, the model shows an emergent functional dissociation in that harder verbs, many of which are irregular, come to rely more on the route through the hidden layer, whereas processing of easier verbs, most of which are regular, depends mainly on the direct connections. This seems to fit in with brain imaging studies that appear to reveal differences in the localization of processes relating to regular and irregular inflections (e.g., Jaeger et al. 1996; Beretta et al. 2003) but that on closer inspection differentiate between easier and harder verbs (Seidenberg and Arnoldussen 2003; Joanisse and Seidenberg 2005).

Connectionist approaches to morphological learning thus deny the existence of an explicit default rule and, by the same token, deem the a priori distinction into inflectional categories on grammatical grounds unnecessary. Instead they attempt to demonstrate the emergence of implicit categories within a neural network, where category membership is on a continuous scale between poles that should rather be labeled "easy" and "hard", as opposed to regular/irregular, because they derive exclusively from distributional factors (phonological similarity, frequency, etc.). The networks employ a single associative mechanism for both discovering the underlying regularities and performing the mappings onto the past tense form, thus conceptualizing acquisition and performance (including generalization) as intricately connected. Models following these general principles have been used to investigate different inflectional paradigms (e.g., noun plural, Plunkett and Nakisa 1997; Plunkett and Juola 1999), behavioral breakdown due to neurological impairment (Joanisse and Seidenberg 1998; Joanisse 2004; Penke and Westermann 2006; Plunkett and Bandelow 2006), and historical change in morphology (Hare and Elman 1995). A strong contribution of the connectionist approach to inflectional morphology has also been to lead to a re-examination of empirical data and to motivate experiments that strive to distinguish between single and dual mechanism accounts.

4.5. *Learning grammatical categories*

Some connectionist approaches to grammar learning take pre-parsed sentences as input, usually with the aim of investigating whether a specific grammar can be learned from exposure to a set of syntactically annotated example sentences. These models presuppose that the syntactic structure of a sentence (i.e., the syntactic roles of its constituents) is already given

and teach the model to apply a specific transformation, e.g., from passive to active constructions (Chalmers 1990).

A more ambitious approach, exemplified by the influential work of Elman (1990, 1991, 1993) explores the capabilities of a connectionist network to extract syntactic structure from exposure to sequences of words, without supplying any kind of additional information. Elman trained predictive SRNs on sequentially presented strings of words that made up well-formed sentences. Training sentences were generated from a small vocabulary using a simple context-free grammar, resulting, in the simplest case, in a set of very basic SV(O) sentences that showed variations in verb argument structure (transitive, intransitive and optionally transitive verbs) and number agreement between subject and verb.

Similar to the models of speech segmentation described earlier, the models extracted a notion of the permissible order of events from the sequential input signal. This was demonstrated by observing the network's predictions from a specific point in a sentence. For example, after the network had encountered first a noun and then a verb, different scenarios occurred: when the verb was transitive, the network predicted a noun as the next word by activating all nouns on the output layer. When the verb was intransitive, the network predicted the end of a sentence. For an optionally transitive verb a mixture of noun and end-of-sentence activations was observed. Similarly, the network expected a sentence initial singular noun to be followed by a singular verb, but not by a plural verb, noun or the end of a sentence.

Analyzing the distributed representations that developed in the fully trained network provided another way of looking at the regularities extracted by the model. At a coarse level, this revealed two large clusters in representational space, one for verbs and one for nouns. The network had learned to distinguish between these two grammatical categories solely on the basis of their co-occurrence relations. Furthermore, transitive and intransitive verbs formed distinct groups within the verb cluster, with optionally transitive verbs falling between these groups, indicating that the relative position of their representations also carried information about argument structure.

It might be argued that these results do not show more than the network's ability to deal with the very limited corpus it had been trained on in terms of transition probabilities between individual words. However, Elman provides two further arguments to support his claim that the network had acquired an emergent representation of hierarchically structured grammatical categories, both based on the model's generalization abilities. In the first of these generalization tests (Elman 1990), the fully trained model was tested on a number of sentences in which one of the

nouns was replaced by a novel word that the network had not seen before. A subsequent cluster analysis showed that the novel word's internal representation was very proximate to the word it had replaced, indicating that the network's representations were based to a large extent on where in a sentence a word appeared, rather than on the identity of the word itself. Secondly (Elman 1998), a network was trained on a corpus in which a specific noun was excluded from ever occurring as the direct object of a sentence, although it did occur in subject position. Following training, the model nevertheless treated the excluded word as a possible successor for transitive verbs, based on its experience that other subject nouns could also appear in object position.

Subsequent research has carried this approach to syntax learning further by introducing more sophisticated syntactic constructions in the training set. The main thrust of several models was to challenge Chomsky's (1957) claim that the existence of recursion in natural language implies the reality of explicit recursive rules as part of human linguistic competence, thus ruling out associative models of language processing. Christiansen and Chater (1999) trained predictive SRNs on artificially generated language samples that included recursive constructions such as counting recursion, centre-embeddings, cross-dependencies, and right-branching recursions. The models learned to perform correctly in all cases, with the limitation that performance broke down gradually with increasing depth of embedding. This, however, can be seen as supporting the model's validity because human participants show similar deteriorations when asked to process deep embeddings (Marks 1968). The SRNs employed in this and other studies furthermore captured the differences in relative difficulty between types of recursive constructions and, also in parallel to human data (Blaubergs and Braine 1974), could be shown to profit from semantic bias (Weckerly and Elman 1992). This body of research shows that connectionist networks are able to acquire limited recursion to an extent that closely mirrors human performance. In the light of these results, the necessity of postulating unbounded recursion as underlying linguistic performance might be questioned, as could, in a more radical step, the necessity to make a distinction between linguistic competence and performance at all (Christiansen and Chater 1999).

5. Related approaches

Here we briefly compare and contrast connectionist approaches to development with two related approaches: dynamical systems (e.g., Thelen and Smith 1994, see also the papers by Hockema and Smith and Hohenberger

and Pelzer-Karpf in this issue), and Bayesian inference (e.g., Oaksford and Chater 2007).

Connectionism and dynamical systems share many aspects of their approach to explaining development (Elman 2003; Thelen and Bates 2003). Both view development as an emergent process shaped by biological and environmental constraints instead of the maturational triggering of innate knowledge. In both approaches, behavior is seen as emerging from interactions across multiple domains, explaining functional modularity as an outcome of development rather than an innate structure. Both approaches differ radically from symbol system theories in which abstract symbolic representations stand for entities in the world and are manipulated irrespective of their content (e.g., Fodor and Pylyshyn 1988). However, although the connectionist and the dynamical systems approach can be partly mapped onto each other (e.g., Smolensky et al. 1996; Rodriguez et al. 1999), there are also important differences between them (Elman 2003; Thelen and Bates 2003). Perhaps the most significant difference between these approaches concerns the role of representations. Whereas dynamical system theory has traditionally de-emphasized the role of internal, unobservable representations (Thelen and Bates 2003), in connectionist approaches representations take a central role and their analysis aids the characterization of developmental processes (Elman et al. 1996; Mareschal et al. 2007a). Second, dynamical systems tend to stress the role of the body as a source of developmental and cognitive constraints, but connectionist models have used their brain-inspired functionality to highlight the role of the brain. More recently, however, connectionist models have begun to consider the body as a constraining factor in cognitive development more seriously (Mareschal et al. 2007a; Mareschal et al. 2007b). Thirdly, where connectionist approaches aim to provide precise mechanistic accounts of developmental change, dynamical systems so far lack a formalized account of how experiences with an environment change the processing system, that is, how the system learns. Whether these differences between connectionism and dynamical systems are of a principled nature or merely reflect the research focus and modeling decisions made by researchers in these fields is subject to debate (Smith and Samuelson 2003; Thelen and Bates 2003).

Another formal approach to cognitive development that has recently begun to receive wider attention is Bayesian inference (Gopnik and Tenenbaum 2007). This approach uses Bayes' equation of conditional probabilities to explain the role of prior knowledge in learning and reasoning. Like connectionist models the Bayesian approach is based on precise mathematical formulations of cognitive processing, but unlike connectionism these formalisms are expressed on a higher level of abstraction,

without making reference to how they might emerge from brain processing. An obstacle that has to be overcome for Bayesian models to account for transitions in development is a mechanistic explanation of the origins of prior knowledge and its change as an outcome of experience (Shultz 2007; Mareschal and Westermann, in press). Current Bayesian models do not specify where initial prior knowledge comes from (Gopnik et al. 2004; Xu and Tenenbaum 2007), and there is no mechanism by which this knowledge can change with experience. Without such mechanisms, Bayesian models offer a snapshot of children's behavior at certain points in development, but they cannot account for developmental change per se.

6. Discussion

Over the past twenty years, connectionism has presented the field with a wide range of working models of language learning, some of which have been discussed here. These models have offered precise implementations of specific language processes which in many cases have mimicked behavioral characteristics of human language processing. The precision with which these models implement predictions of behavioral patterns makes them falsifiable. This fact represents a great strength of the modeling approach, as ultimately every valid scientific theory needs to be falsifiable[1] (Popper 1959). One advantage of implemented models therefore is to enforce precision and concreteness in the theory to be implemented, thus enabling the theory to be evaluated against the criterion of whether it matches the available data. By the same token, a theory that is implemented as a functional model may generate distinct predictions, which then can be tested empirically. In some cases, connectionist models are used exclusively in this sense: as a research tool with the purpose of implementing and evaluating an existing theory and generating predictions from it.

Many connectionists, however, would attribute additional import to their use of this specific type of model, casting neural network models as a move towards uncovering the general principles of information processing that underpin not only language learning, but also acquisition and performance in other cognitive domains. Properties such as associative learning and self-organization in response to exposure to environmental stimuli, and the inbuilt ability to generalize from these known stimuli to novel exemplars, are held to reflect brain-style computation. For proponents of this perspective, the ASSOCIATIVE stance behind connectionist models forms an integral part of any theory that is instantiated within a

specific neural network model. Connectionism thus is often advocated as an alternative to the "symbols and rules" approach of traditional linguistics and cognitive psychology.

The differences between these two approaches are especially apparent with the concepts of learning and development. Within the symbolic framework, development is usually construed as maturation, where new rules come online at certain stages of development. Learning, if addressed at all, takes the form of explicit hypothesis testing, a process that can refine existing rules or even generate novel ones. Both of these concepts are forced to rely on considerable amounts of innate knowledge, be it in the form of innate rules, the schedule according to which the rules are activated or, more indirectly, the rules that govern the hypothesis testing process.

The connectionist approach, conversely, casts both learning and development as dynamic change within an associative system, driven by exposure to stimuli in the environment. No domain-specific innate knowledge is assumed. Instead, it is the interaction between the general associative learning principles and domain specific stimuli that leads to the extraction of domain specific knowledge in the form of regularities, stored in the configuration of the model's weights. The fact that neural networks MUST acquire proficiency in a task by such a learning process makes them particularly well suited to investigate processes of learning and development.

From a connectionist perspective, actually, LEARNING and DEVELOPMENT are almost indistinguishable. A neural network learns by adjusting its weights (and sometimes — in constructivist networks — even its architecture) in response to the stimuli it processes. The learning algorithm that determines the exact nature of the adjustment operates on the basis of individual stimuli; essentially it attempts to make the model better at processing the currently available item. The incremental adjustments in response to the processing of many exemplars, however, will eventually lead to the development of a configuration that implements a compromise between the different, often conflicting pressures induced by the individual stimuli. Through these incremental adjustments the resulting network thus will have developed from an initial, unstructured state into a state/configuration that is optimally adapted to the task at hand, including an inbuilt ability to generalize to novel stimuli. Because both learning and development in a neural network model are a direct result of the interplay between the environment (stimuli), the network's architecture (including the current weights configuration which, in turn, is a consequence of previous adjustments) and the learning algorithm, they are closely intertwined — and most connectionist researchers would claim this also to be true with respect to cognition.

Building a connectionist model involves many steps of abstraction, and each of these abstractions may of course be critically evaluated. In the following we point out some important aspects to consider when attempting to judge the strengths and weaknesses of a particular model.

Are the assumptions behind the composition of the training set justifiable? This question concerns issues such as whether the level at which the input and output data are encoded is appropriate, whether the specific coding scheme used introduces an unwarranted bias, or whether it is reasonable to assume the availability of the target signal. Implicit pre- or post-processing steps at either end of the model might also be of concern. Ultimately, of course, the aim should be for the training set to faithfully reflect the structure of the human learner's environment. In general, there is a clear trend for neural network models to operate on increasingly larger and more realistic data sets, thus giving an additional motivation to collect large scale corpora. This is not to say that smaller simulations, often trained on small subsets of actual linguistic data or artificial grammars are without scientific value. On the contrary, due to their relative simplicity they often enable a better understanding of the underlying principles at work in the model. However, the question whether such simple models can be made to scale up to more naturalistic input scenarios is one of the major challenges for the connectionist approach.

Is the way in which the weights are adjusted biologically plausible? In general, this can be confirmed in the case of Hebbian learning which is ubiquitous in the brain (Kandel et al. 2000). Note, however, that some architectures introduce additional mechanisms that are less easily justified, for example the shrinking neighborhood radius in classic SOMs. Learning algorithms that are based on error correction (e.g., the widely used backpropagation algorithm), on the other hand, often lack detailed biological plausibility because it is unclear how the error information calculated in subsequent layers could reach neurons in earlier layers. This problem can be circumvented by adding backwards connectivity and employing alternative, Hebbian style algorithms (O'Reilly 1996) that are functionally equivalent to backpropagation, though conceptually and mathematically less transparent. The use of simple backpropagation could thus be justified by the existence of such functionally equivalent but biologically more plausible solutions.

The above considerations might give reasons to reject individual implementations, even if they capture the available data. Rather than invalidating the general approach such criticisms are often an incentive to develop enhanced implementations with a wider data base and better contact to neurobiological constraints. Note, however, that with regard to modeling, more is not always better. If we had, for example, a full

scale model of the brain, incorporating all the biological facts and dealing with entirely naturalistic data, not much would have been gained, because such a model is bound to be as complex as the original brain, thus not furthering our understanding. A certain trade-off between detail and parsimony is inherent in any modeling endeavor and it is because of this that the evaluation of the "right" abstractions is extremely important, but also very difficult.

In this paper we have discussed the connectionist approach to language learning and development. Our aim has been to explain the essential aspects of the functioning of connectionist models and the design considerations involved, and to describe and evaluate a range of specific models addressing different aspects of language learning. We have argued that connectionist models with their ability to learn from data, their sensitivity to the statistical structure of the environment, and the link they make between brain and cognitive processing, imply that the connectionist approach with its ASSOCIATIVE stance provides a viable alternative to the SYMBOL AND RULES approach of traditional linguistics and cognitive psychology.

We see three important developments in the future of connectionist approaches to language development. First, connectionist models are ideally suited to integrate our understanding of development with adult processing and impaired processing after brain damage. While in empirical work these three aspects of language processing are not normally connected, a connectionist model can give an account of how development as an outcome of multiple interacting constraints leads to an adult processing state. Damaging the same model can then lead to insights into the deficits arising from brain damage in human patients. A model that can, in the same system, account for all three aspects of language processing would present powerful evidence for the implemented hypothesis. However, as yet connectionist modelers have generally not taken this integrated approach but have focused on simulating only one or two of these aspects.

Second, we see a trend of moving from simple, three-layer backpropagation models towards a) more complex architectures such as multicomponent models in which the unfolding interactions between components provide additional information about a learning process, b) constructivist systems in which experience-dependent brain development is entered as an important aspect of cognitive development, and c) more biologically realistic architectures such as feature maps which can make more direct predictions about language processing in the brain (Westermann et al. 2006; Mareschal et al. 2007a).

Third, the role of the body for different aspects of language evolution, development and processing has recently become a focus of research (e.g.,

Davis and MacNeilage 2000; Zuidema and Westermann 2003; Gibbs 2006). Unlike dynamical systems, connectionist approaches so far have often not taken embodiment into account (though see Mareschal et al. 2007 and Mareschal et al. 2007b for examples of models that do). Insofar as embodied views as well as situated models offer additional insights into language development connectionist models will have to take these additional constraints into account.

Connectionist modeling of language development has been an active field of research for just over 20 years, leading to novel explanations for many aspects of language and cognitive development and to a plethora of new experimental data. With an increasing understanding of the brain mechanisms underlying language learning and processing and ever more data to constrain explanatory hypotheses we see an active and fruitful future for this approach.

Received 25 August 2005
Revised version received
13 November 2007

Oxford Brookes University
Oxford University

Notes

* The writing of this paper was supported by ESRC grant RES-061-23-0129 to Gert Westermann. Correspondence address: Gert Westermann, Department of Psychology, Oxford Brookes University, Gipsy Lane, Oxford OX3 0BP, E-mail: gwestermann@brookes.ac.uk.

1. According to widely accepted criteria for what counts as a "good theory" in philosophy of science, the virtue of a theory to be falsifiable is ranked higher than the virtue of a theory to be correct.

References

Abbott-Smith, Kirsten, Elena Lieven & Michael Tomasello. 2004. Training 2;6-year-olds to produce the transitive construction: the role of frequency, semantic similarity and shared syntactic distribution. *Developmental Science* 7. 48–55.

Altmann, Gerry T. M. 2002. Learning and development in neural networks — the importance of prior experience. *Cognition* 85. B43–B50.

Beretta, Alan, Carrie Campbell, Thomas H. Carr, Jie Huang, Lothar M. Schmitt, Kiel Christianson & Yue Cao. 2003. An ER-fMRI investigation of morphological inflection in German reveals that the brain makes a distinction between regular and irregular forms. *Brain and Language* 85. 67–92.

Blaubergs, Maija S. & Martin D. S. Braine. 1974. Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology* 102. 745–48.

Cairns, Paul, Richard Shillcock, Nick Chater & Joe Levy. 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology* 33. 111–153.

Chalmers, David J. 1990. Syntactic transformations on distributed representations. *Connection Science* 2. 53–62.

Chalmers, David J. 1992. Subsymbolic computation and the Chinese room. In John Dinsmore (eds.), *The Symbolic and Connectionist Paradigms: Closing the Gap*, 25–48. Hillsdale, NJ: Lawrence Erlbaum Associates.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.

Christiansen, Morten H. & Joseph Allen. 1997. Coping with variation in speech segmentation. In Antonella Sorace, Caroline Heycock & Richard Shillcock (eds.), *Language acquisition: Knowledge representation and processing*, 327–332. Edinburgh: University of Edinburgh Press.

Christiansen, Morten H., Joseph Allen & Mark S. Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes* 13. 221–268.

Christiansen, Morten H. & Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23. 157–205.

Christiansen, Morten H., Christopher M. Conway & Suzanne Curtin. 2005. Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. In James W. Minett & William S.-Y. Wang (eds.), *Language acquisition, change and emergence: Essay in evolutionary linguistics*, 205–249. Hong Kong: City University of Hong Kong Press.

Clahsen, Harald. 1999. Lexical Entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22. 991–1013.

Dale, Philip S. & Larry Fenson. 1996. Lexical development norms for young children. *Behavior Research Methods, Instruments & Computers* 28. 125–27.

Daugherty, Kim & Mark S. Seidenberg. 1992. Rules or connections? The past tense revisited. In David P. Corina (ed.), *Proceedings of the 14th annual conference of the cognitive science society*, 259–264. Hillsdale, NJ: Lawrence Erlbaum.

Davies, Matt H. 2003. Connectionist modelling of lexical segmentation and vocabulary acquisition. In Philip Quinlan (eds.), *Connectionist models of development: Developmental processes in real and artificial neural networks*, 151–187. Hove: Psychology Press.

Davis, Barbara L. & Peter F. MacNeilage. 2000. An embodiment perspective on the acquisition of speech perception. *Phonetica* 57. 229–241.

Ellis, Andrew W. & Matthew A. Lambon-Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology-Learning Memory and Cognition* 26. 1103–1123.

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

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

Elman, Jeffrey L. 1993. Learning and development in neural networks: the importance of starting small. *Cognition* 48. 71–99.

Elman, Jeff L. 1998. Generalization, simple recurrent networks, and the emergence of structure. In Morton Ann Gernsbacher & Sharon J. Derry (eds.), *Twentieth Annual Conference of the Cognitive Science Society*, 6–12. Mahwah, NJ: Lawrence Erlbaum.

Elman, Jeff L. 2003. Development: It's about time. *Developmental Science* 6. 430–433.

Elman, Jeffrey L. 2005. Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences* 9. 111–117.

Elman, Jeffrey L., Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi & Kim Plunkett. 1996. *Rethinking innateness. A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fenson, Larry & Stephen J. Pethick. 1994. *Variability in early communicative development* (Monographs of the Society for Research in Child Development 59). Oxford: Blackwell.

Fodor, Jerry F. & Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28. 73–193.

Fritzke, Bernd. 1994. Growing cell structures — A self-organizing network for unsupervised and supervised learning. *Neural Networks* 7. 1441–1460.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, David S. Pallett & Nancy L. Dahlgren. 1990. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. Tech. Rep. No. NISTIR 4930. Gaithersburg, MD: National Institute of Standards and Technology.

Gibbs, Raymond W. 2006. *Embodiment and cognitive science*. Cambridge: Cambridge University Press.

Gopnik, Alison, Clark Glymour, David M. Sobel, Laura E. Schulz, Tamar Kushnir & David Danks. 2004. A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111. 3–32.

Gopnik, Alison & Joshua B. Tenenbaum. 2007. Bayesian networks, Bayesian learning and cognitive development. *Developmental Science* 10. 281–287.

Gow, David W. & Peter C. Gordon. 1995. Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance* 21. 344–359.

Guenther, Frank H. & Marin N. Gjaja. 1996. The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustic Society of America* 100. 1111–1121.

Guenther, Frank H., Satrajit S. Ghosh & Jason A. Tourville. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96. 280–301.

Hahn, Ulrike & Ramin C. Nakisa. 2000. German inflection: Single route or dual route? *Cognitive Psychology* 41. 313–60.

Hare, Mary & Jeffrey L. Elman. 1995. Learning and morphological change. *Cognition* 56. 61–98.

Hebb, Donald O. 1949. *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Jaeger, Jeri J., Alan H. Lockwood, David L. Kemmerer, Robert D. Van Valin, Brian W. Murphy & Hanif G. Khalak. 1996. Positron emission tomographic study of regular and irregular verb morphology in English. *Language* 72. 451–497.

Joanisse, Marc F. & Mark S. Seidenberg. 1998. Specific language impairment: a deficit in grammar or processing? *Trends in Cognitive Sciences* 2. 240–47.

Joanisse, Marc F. & Mark S. Seidenberg. 1999. Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America* 96. 7592–7597.

Joanisse, Marc F. 2004. Specific language impairments in children — Phonology, semantics, and the English past tense. *Current Directions in Psychological Science* 13. 156–160.

Joanisse, Marc F. & Mark S. Seidenberg. 2005. Imaging the past: Neural activation in frontal and temporal regions during regular and irregular past-tense processing. *Cognitive Affective & Behavioral Neuroscience* 5. 282–296.

Johnson, Mark H. 2005. *Developmental cognitive neuroscience*. Oxford: Blackwell.

Jusczyk, Peter W. 1999. How infants begin to extract words from speech. *Trends in Cognitive Science* 3. 323–328.
Kandel, Eric R., James H. Schwartz & Thomas M. Jessel. 2000. *Principles of Neural Science*. McGraw-Hill New York: Elsevier.
Kohonen, Teuvo. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* 43. 59–69.
Kuhl, Patricia K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50. 93–107.
Kuhl, Patricia K. 1995. Mechanisms of developmental change in speech and language. In Kjell Elenius & Peter Branderud (eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 131–139. Stockholm: KTH and Stockholm University.
Lacerda, Francisco. 1995. The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In Kjell Elenius & Peter Branderud (eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 140–147. Stockholm: KTH and Stockholm University.
Li, Ping. 2003. Language acquisition in a self-organizing neural network model. In Philip Quinlan (ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks*, 115–149. New York: Psychology Press.
Li, Ping, Igor Farkas & Brian MacWhinney. 2004. Early lexical development in a self-organizing neural network. *Neural Networks* 17. 1345–1362.
Li, Ping, Xiaowei Zhao & Brian MacWhinney. 2007. Dynamic self-organization and early lexical development in children. *Cognitive Science* 31. 581–612.
Liberman, Alvin M., Katherine S. Harris, Howard S. Hoffman & Belver C. Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54. 358–368.
Liberman, Alvin M. & Ignatius G. Mattingly. 1985. The motor theory of speech-perception revised. *Cognition* 21. 1–36.
Lisker, Leigh & Arthur S. Abramson. 1971. Distinctive features and laryngeal control. *Language* 47. 767–785.
MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.
MacWhinney, Brian. 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* 31. 883–914.
MacWhinney, Brian & Jared Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40. 121–157.
Marchman, Virginia A. 1997. Children's productivity in the English past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science: A Multidisciplinary Journal* 21. 283–304.
Marchman, Virginia A. & Elizabeth Bates. 1994. Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language* 21. 339–366.
Marcus, Gary F. 1995. The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition* 56. 271–279.
Marcus, Gary F. 1999. Connectionism: With or without rules? *Trends in Cognitive Sciences* 3. 168–170.
Marcus, Gary, Ursula Brinkmann, Harald Clahsen, Richard Wiese & Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive Psychology* 29. 189–256.
Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen & Fei Xu. 1992. *Overregularization in language acquisition* (Monographs of the Society for Research in Child Development, Serial No. 228, Vol. 57 no. 4). Oxford: Blackwell.

Mareschal, Denis, Mark H. Johnson, Sylvain Sirois, Michael W. Spratling, Michael Thomas & Gert Westermann. 2007a. *Neuroconstructivism: How the brain constructs cognition*. Oxford: Oxford University Press.

Mareschal, Denis, Sylvain Sirois, Gert Westermann & Mark Johnson (eds.). 2007b. *Neuroconstructivism, vol II: Perspectives and prospects*. Oxford: Oxford University Press.

Mareschal, Denis & Gert Westermann. In press. Mixing the old with the new and the new with the old: Combining prior and current knowledge in conceptual change. In Scott P. Johnson (ed.), *Neoconstructivism: The new science of cognitive development*. New York: Oxford University Press.

Marks, Lawrence E. 1968. Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior* 7. 965–967.

Marslen-Wilson, William & Lorraine K. Tyler. 1998. Rules, representations, and the English past tense. *Trends in Cognitive Sciences* 2. 428–435.

Marslen-Wilson, William & Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* 10. 29–63.

McClelland, James L. & Karalyn Patterson. 2002. 'Words or rules' cannot exploit the regularities in exceptions. *Trends in Cognitive Sciences* 6. 464–65.

McLeod, Peter, Kim Plunkett & Edmund T. Rolls. 1998. *Introduction to connectionist modelling of cognitive processes*. Oxford: Oxford University Press.

Miikkulainen, Risto. 1997. Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language* 59. 334–366.

Miller, George A. & Patricia E. Nicely. 1955. An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America* 27. 338–352.

Morgan, James L. & Katherine Demuth (eds.). 1996. *From Signal to Syntax*. Mahwah, NJ: Lawrence Erlbaum.

Nakisa, Ramin C. & Kim Plunkett. 1998. Evolution of a rapidly learned representation for speech. *Language and Cognitive Processes* 13. 105–127.

Oaksford, Mike & Nick Chater. 2007. *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

O'Reilly, Randall C. 1996. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation* 8. 895–938.

Penke, Martina & Gert Westermann. 2006. Broca's area and inflectional morphology: Evidence from Broca's aphasia and computer modeling. *Cortex* 42. 563–576.

Pinker, Steven. 1997. Words and rules in the human brain. *Nature* 387. 547–548.

Pinker, Steven. 2001. Four decades of rules and associations, or whatever happened to the past tense debate? In Emmanuel Dupoux (ed.), *Language, brain, and cognitive development: Essays in honor of Jacques Mehler*, 157–179. Cambridge, MA: MIT Press.

Pinker, Steven & Michael T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences* 6. 456–463.

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

Plunkett, Kim & Stephan Bandelow. 2006. Stochastic approaches to understanding dissociations in inflectional morphology. *Brain and Language* 98. 194–209.

Plunkett, Kim & Patrick Juola. 1999. A connectionist model of English past tense and plural morphology. *Cognitive Science* 23. 463–490.

Plunkett, Kim & Virginia A. Marchman. 1991. U-Shaped learning and frequency-effects in a multilayered perceptron — Implications for child language-acquisition. *Cognition* 38. 43–102.

Plunkett, Kim & Virginia A. Marchman. 1993. From rote learning to system building — acquiring verb morphology in children and connectionist nets. *Cognition* 48. 21–69.
Plunkett, Kim & Ramin C. Nakisa. 1997. A connectionist model of the Arabic plural system. *Language and Cognitive Processes* 12. 807–836.
Popper, Karl. 1959. *The logic of scientific discovery*. London: Hutchinson.
Pullum, Geoffrey K. & Barbara C. Scholz. 2002. Empirical assessment of poverty of stimulus arguments. *The Linguistic Review* 19. 9–50.
Ritter, Helge J. & Teuvo Kohonen. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61. 241–54.
Rodriguez, Paul, Janet Wiles & Jeffrey L. Elman. 1999. A recurrent neural network that learns to count. *Connection Science* 11. 5–40.
Rumelhart, David E., Geoffrey E. Hinton & Ronald J. Williams. 1986. Learning internal representations by error propagation. In David E. Rumelhart & James L. McClelland (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, volume 1: Foundations*, 318–362. Cambridge, MA: MIT Press.
Rumelhart, David E. & James L. McClelland. 1986. On learning the past tense of English verbs: implicit rules or parallel distributed processing? In James L. McClelland, Dave Rumelhart & the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 195–248. Cambridge, MA: MIT Press.
Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274. 1926–1928.
Seidenberg, Mark. S. & Aimee Arnoldussen. 2003. The brain makes a distinction between hard and easy stimuli: Comments on Beretta et al. *Brain and Language* 85. 527–530.
Seidenberg, Mark & M. Bruck. 1990. Consistency effects in the generation of past tense morphology. Paper presented at the 31st meeting of the Psychonomic Society, New Orleans, LA.
Seidenberg, Mark S. & Jeffrey L. Elman. 1999a. Do Infants Learn Grammar with Algebra or Statistics? *Science* 284. 433.
Seidenberg, Mark S. & Jeffrey L. Elman. 1999b. Networks are not 'hidden rules'. *Trends in Cognitive Sciences* 3. 288–289.
Seidenberg, Mark S. & Marc F. Joanisse. 2003. Show us the model. *Trends in Cognitive Sciences* 7. 106–107.
Shultz, Thomas R. 2003. *Computational developmental psychology*. Cambridge, MA: MIT Press.
Shultz, Thomas R. 2007. The Bayesian revolution approaches psychological development. *Developmental Science* 10. 357–364.
Smith, Linda B. & Larissa K. Samuelson. 2003. Different is good: connectionism and dynamic systems theory are complementary emergentist approaches to development. *Developmental Science* 6. 434–439.
Smolensky, Paul, Michael Mozer & David E. Rumelhart (eds.). 1996. *Mathematical perspectives on neural networks*. Mahwah, NJ: Lawrence Erlbaum.
Thal, Donna, Elizabeth E. Bates, Judith Goodman & Jennifer Jahn-Samilo. 1997. Continuity of language abilities in late- and early-talking toddlers. *Developmental Neuropsychology* 13. 239–273.
Thelen, Esther & Elizabeth Bates. 2003. Connectionism and dynamic systems: are they really different? *Developmental Science* 6. 378–391.
Thelen, Esther & Linda B. Smith. 1994. *A dynamic systems approach to the development of cognition and action* (Bradford Books). Cambridge, MA: MIT Press.
Thomas, Michael S. C. & Walter J. B. van Heuven. 2003. Computational models of bilingual comprehension. In Judith F. Kroll & Annette M. B. De Groot (eds.), *Handbook of bilingualism: Psycholinguistic approaches*, 202–225. Oxford: University Press.

Ullman, Michael T. 1999. Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes* 14. 47–67.

Vihman, Marilyn May. 1993. Variable paths to early word production. *Journal of Phonetics* 21. 61–82.

Vihman, Marilyn May. 2002. The role of mirror neurons in the ontogeny of speech. In Maxim Stamenov & Victor Gallese (eds.), *Mirror Neurons and the Evolution of Brain and Language*, 305–314. Amsterdam: John Benjamins.

Weckerly, Jill & Jeffrey L. Elman. 1992. A PDP approach to processing center-embedded sentences. In John Kruschke (ed.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 414–419. Hillsdale, NJ: Lawrence Erlbaum.

Westermann, Gert. 1998. Emergent modularity and U-shaped learning in a constructivist neural network learning the English past tense. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Morton Ann Gernsbacher & Sharon. J. Derry (eds.), 1130–1135. Hillsdale, NJ: Lawrence Erlbaum.

Westermann, Gert & Eduardo Reck Miranda. 2004. A new model of sensorimotor coupling in the development of speech. *Brain and Language* 89. 393–400.

Westermann, Gert, Sylvain Sirois, Thomas R. Shultz & Denis Mareschal. 2006. Modeling developmental cognitive neuroscience. *Trends in Cognitive Sciences* 10. 227–233.

Willshaw, David J. & Christoph von der Malsburg. 1976. How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London B* 194. 431–445.

Xu, Fei & Joshua B. Tenenbaum. 2007. Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10. 288–297.

Yoshikawa, Yuichiro, Minoru Asada, Koh Hosoda & Junpei Koga. 2003. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science* 15. 245–258.

Zuidema, Willem & Gert Westermann. 2003. Evolution of an optimal lexicon under constraints from embodiment. *Artificial Life* 9. 387–402.