Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 17 (2004) 1345–1362
Neural Networks
www.elsevier.com/locate/neunet

2004 Special issue

Early lexical development in a self-organizing neural network

Ping Li[a,*], Igor Farkas[b], Brian MacWhinney[c]

[a]University of Richmond, Richmond, VA 23173, USA
[b]Comenius University, Bratislava, Slovak Republic
[c]Carnegie Mellon University, Pittsburgh, PA 15213, USA

Received 12 December 2003; accepted 14 July 2004

Abstract

In this paper we present a self-organizing neural network model of early lexical development called DevLex. The network consists of two self-organizing maps (a growing semantic map and a growing phonological map) that are connected via associative links trained by Hebbian learning. The model captures a number of important phenomena that occur in early lexical acquisition by children, as it allows for the representation of a dynamically changing linguistic environment in language learning. In our simulations, DevLex develops topographically organized representations for linguistic categories over time, models lexical confusion as a function of word density and semantic similarity, and shows age-of-acquisition effects in the course of learning a growing lexicon. These results match up with patterns from empirical research on lexical development, and have significant implications for models of language acquisition based on self-organizing neural networks.

Keywords: Language acquisition; Self-organizing neural network; Lexical development

1. Introduction

Connectionist modeling of language acquisition has made significant progress since Rumelhart and McClelland's pioneering model of the acquisition of the English past tense (Rumelhart & McClelland, 1986). However, three major limitations need to be considered for the further development of neural network models of language acquisition.

First, some language acquisition models use artificially generated input representations that are isolated from realistic language uses. Other models use representations that are limited to small sets of vocabulary with handcrafted phonological and semantic representations. Such artificially generated or limited sets of vocabulary prohibit us from understanding the structure of realistic lexicons in language learning.

Second, most previous models have used supervised learning through back-propagation as the basis for network training (see models reviewed in Elman et al., 1996; Quinlan, 2003). Although these types of networks have been able to model many aspects of children's language learning, their biological and psychological implausibility limits their explanatory adequacy. In real language learning, children do not receive constant feedback about what is incorrect in their speech, or the kind of error corrections provided to the network by supervising learning algorithms (see Li, 2003; MacWhinney, 2001a; Shultz, 2003 for discussion).

Third, neural network models of lexical learning (Li, 2003; Li & MacWhinney, 1996; MacWhinney, 2001a; Plunkett, Sinha, Møller, & Strandsby, 1992) have not yet devised a method for modeling the incremental nature of lexical growth. Children's vocabularies expand gradually by adding a few words each day. Currently, there are no neural network models that are capable of modeling this gradual expansion of vocabulary. The core problem here is the phenomenon of catastrophic interference (see French, 1999 for a review). If we train a network to acquire a vocabulary of 100 words, for example, and then train it on another 100 words, the addition of the second set will disrupt (or interfere catastrophically with) the learning of the first 100 words. Although we know that learning new

* Corresponding author. Tel.: +1 804 289 8125; fax: +1 804 287 1905.
E-mail address: pli@richmond.edu (P. Li).

0893-6080/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2004.07.004

words can lead to confusion with old words for children (Gerskoff-Stowe & Smith, 1997; see discussion later), these effects only involve local competitions between pairs of words and are never catastrophic.

To address these three problems, we developed DevLex, a self-organizing neural network model of the development of the lexicon. DevLex is designed to combine the dynamic learning properties of connectionist networks with the scalability of representation models (such as HAL, Burgess & Lund, 1997). It is able to acquire a continually expanding vocabulary whose representation enriches in a realistic way with linguistic contexts over time. As such, the model itself also evolves dynamically with learning. DevLex relies on corpus-based speech data to establish the sequence as well as the structure of the input, using phonological and semantic representations that correspond to actual language use.

Previous work by Li (2003), MacWhinney (2001a), Miikkulainen (1993, 1997), and Ritter and Kohonen (1989) has shown that self-organizing neural networks, especially SOMs, are particularly suitable as models of the human lexicon. In our earlier work we used SOM to simulate language acquisition in various tasks: Li (1999, 2000) simulated the acquisition of lexical categories along with morphological acquisition (prefixes *un-* and *dis-* and suffixes *-ing* and *-ed*; see Li, 2003 for a summary); Farkas and Li (2001) modeled lexical category representation in an artificial corpus and a natural speech corpus based on parental input from the CHILDES database (MacWhinney, 2000); Farkas and Li (2002a) used growing nodes in SOM on the basis of the increasing vocabulary sizes during learning; Farkas and Li (2002b) modeled word confusion in production as a function of word frequency, word density, and rate of vocabulary increase; and Li and Farkas (2002) modeled lexical development in bilingual children. In all cases, the simulated patterns captured the development of basic linguistic categories from the statistical characteristics of the input data. Building on the results of these studies, the current model attempts to account for three important phenomena in language acquisition: (a) the emergence and organization of linguistic categories in early lexical representation, (b) early lexical confusion in children's productive speech during naming, and (c) age-of-acquisition effects in early lexical development. Before providing details of the model, we will briefly discuss the psycholinguistic phenomena that have motivated our computational model.

Early lexical development has intrigued cognitive and developmental psychologists for a number of important reasons. First, a central issue in the cognitive neuroscience of language has been how the brain represents linguistic categories. Neuropsychological and neuroimaging studies have demonstrated distinct areas of cortical activation in response to nouns, verbs, and other linguistic categories (e.g. Caramazza & Hillis, 1991; Damasio Grabowski, Tranel, Hichwa, & Damasio, 1996; Pulvermüller, 1999). Nativists (Chomsky, 1975; Fodor, 1983; Pinker, 1994) take these findings as evidence that humans have a species-specific genetic code that determines modular lexical organization in the brain. However, an alternative account for these same findings holds that the neural modularization of linguistic categories is an emergent process (Elman et al., 1996; MacWhinney, 1998). In this account, it is the distinct syntactic and semantic functions of groups of words during mental processing that account for their dynamic organization during development into different areas of the brain. For example, the processing characteristics of nouns and verbs in Chinese differ from those in English, and thus we would expect different patterns of neural activities for the two languages (see Li, Jin, & Tan, 2004). One of the goals of our model is to use SOM's topography preserving properties to study the emergence and organization of linguistic categories across stages of lexical learning.

A second issue that has intrigued developmentalists is the changing developmental landscape of early vocabulary (see Elman et al., 1996 for discussion). For example, it has been observed that between 18 and 20 months, children's productive vocabulary increases rapidly, a phenomenon often described as the 'vocabulary spurt', a sudden and rapid acceleration in the amount of words produced in a short period (Bates & Carnevale, 1993; Dromi,1987). Associated with such a spurt is a brief period of confusion regarding the uses of some words, during which time the child calls some objects by the wrong name. This confusion has also been labeled the 'naming deficit' (Gerskoff-Stowe & Smith, 1997). One prominent account of this phenomenon is that children's lexical confusions reflect a retrieval difficulty. The idea is that children may have encoded the correct lexical semantic representations, but because these representations are being densely packed in memory due to vocabulary spurt (and are therefore in strong competition with each other), successful and efficient retrieval of the stored items is temporarily disrupted (Gerskoff-Stowe & Smith, 1997). Although this explanation appears reasonable, other factors may also be important in the process of lexical confusion. Bowerman (1978, 1982) has suggested that children undergo periods of organization and reorganization, in which semantically related word pairs are more likely to be substituted for one another when children recognize their shared meaning components (e.g. *put* substituted for *give*, or *take* substituted for *bring*). In other words, at this stage word pairs that are not initially recognized as semantically related now move closer together in semantic space, but their fine-grained distinctions have not been worked out. Thus, semantic relatedness in the representation may play an important role in triggering early lexical confusion. Although Bowerman's proposal was not used specifically to explain the early 'naming deficit', according to this perspective, confused words, whether they occur early or late during lexical acquisition, should tend to be those that are neighbors in

the same densely populated semantic neighborhood, rather than random pairs of words. Another goal of the current study is to examine these different accounts of early lexical confusions in the context of an explicit computational model.

Third, an emerging focus in recent psycholinguistic research is the age at which a word is learned, or its age of acquisition (AoA; Ellis & Morrison, 1998; Morrison & Ellis, 1995). This research suggests that AoA is often a better predictor of a word's processing latency than is frequency: people are faster at reading and naming words that are acquired early as compared to words that are acquired late. Several connectionist models based on feed-forward networks have attempted to simulate the AoA effects. These models used an auto-association task in which randomly generated patterns (representing words) are considered learned if the patterns can be accurately reconstructed at the output. For example, Ellis and Lambon-Ralph (2000) used a two-staged learning process, in which the training items were split into the early versus late stages. They showed that when the words were presented to the network in this staged fashion, that is, with one set of words trained first and a second set added to the training set, the network would display strong AoA effects, as shown in the lower reconstruction errors for early learned words. They suggested that the AoA effects were due to the lost plasticity in their network as a result of learning. Smith, Cottrell and Anderson (2001) showed that, using the same model as that of Ellis and Lambon-Ralph but without staged learning, the earlier a word was learned, the lower its final reconstruction errors would be. AoA effects are important in our view because they impose significant constraints on connectionist models: a model has to possess both plasticity in learning and stability in representation in order to display AoA. Minimally, the model has to be able to overcome catastrophic interference, with early-learned structures undisrupted by new learning. Thus, a final goal of the current study is to develop a computational model that can deal with the plasticity-stability dilemma and lend itself naturally to AoA effects.

The DevLex model was designed to address each of these three central issues: cortical topography, vocabulary dynamics, and AoA effects. At the same time, it is designed to correct the three modeling problems we noted earlier by using realistic input, self-organization, and a dynamically expanding lexicon. With these goals in mind, we are now ready to turn to the computational details of the DevLex model.

2. The DevLex Model

2.1. A Sketch of the Model

Fig. 1 presents a diagrammatic sketch of DevLex. The model has been inspired by Miikkulainen's (1993, 1997)

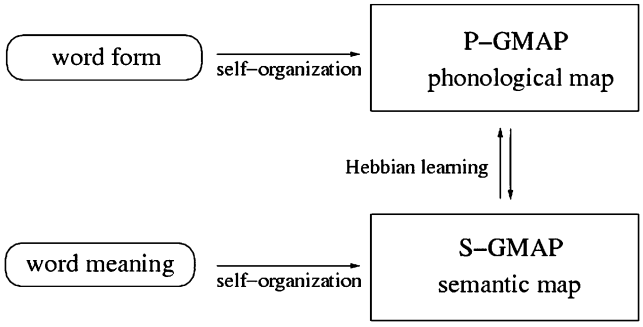

Fig. 1. The DevLex model of lexical development. Two growing self-organizing maps (GMAPs) are connected via associative links updated by Hebbian learning; the P-GMAP self-organizes on word form (phonological information) and the S-GMAP self-organizes on word meaning. Both form and meaning are presented to the network in the input.

DISLEX, in which multiple SOMs are connected by Hebbian learning. DevLex consists of two identical growing maps (GMAPs) - a phonological map that processes phonological information of words (P-GMAP), and a semantic map that processes lexical-semantic information (S-GMAP).

Formally, a GMAP is defined as a graph $G=(A,C)$, where A is a set of nodes, and $C \subset A \times A$ is a set of connections between the nodes. Each node k in a GMAP has an input weight vector $\mathbf{m}_k$ associated with it. Given a stimulus $\mathbf{x}$ (distributed word representation), the localized output response of a node k is computed as

$$a_k = \begin{cases} 1 - \dfrac{||\mathbf{x} - \mathbf{m}_k|| - d_{\min}}{d_{\max} - d_{\min}} & \text{if } k \in N_c \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where N_c is the set of neighbors of winner c (whereas $a_c = \max_k\{a_k\}$), $d_{\min}$ and $d_{\max}$ are the smallest and the largest Euclidean distances of $\mathbf{x}$ to node's weight vectors within N_c. Upon training of the network, a phonological representation of the word ('word form') is presented to the network, and simultaneously, the semantic representation of the same word ('word meaning') is also presented to the network. Through self-organization, the network forms a pattern of activation on the P-GMAP in response to the word form, and a pattern of activation on the S-GMAP in response to the word meaning. Weights of nodes around the winner are updated (self-organized) as

$$m_{kj}(t+1) = m_{kj}(t) + \alpha(t)[x_j - m_{kj}(t)] \quad \text{for all } k \in N_c \quad (2)$$

The two GMAPs are bidirectionally linked with associative connections. Simultaneously with input weights, the associative weights between active nodes in both GMAPs are updated using Hebbian learning (Hebb, 1949)

$$\Delta w_{kl} = \alpha(t) a_k^S a_l^D \quad (3)$$

where w_{kl} is the unidirectional associative weight leading from node k in the source map to node l in the destination map, and a_k^S and a_l^D are the associated node activations.

The associative weight vectors are then normalized:

$$w_{kl}(t+1) = \frac{w_{kl}(t) + \Delta w_{kl}}{\{\sum_l [w_{kl}(t) + \Delta w_{kl}]^2\}^{1/2}} \quad (4)$$

The combination of Hebbian learning with self-organization can account for the process of how the learner establishes relationships between word forms and lexical meaning representations on the basis of how often they co-occur and how strongly they are co-activated in the representation. Links between the maps are then used to propagate the activation from one map to the other to model production (from semantics to word form), and comprehension (from word form to semantics). Formally, the response in the destination map can be evoked by propagation of activity from the source map:

$$a_l^D = g(y_l) = g\left(\sum_k w_{kl} a_k^S\right) \quad (5)$$

where the activation function $g(y) = y/y_{\max}$ scales down the activations in the destination map linearly into the interval of 0–1.

In production, the S-GMAP first creates its output activation pattern, as a response to its input stimulus (word meaning representation). This activation pattern then propagates to the P-GMAP via the semantic-to-phonological associative links and creates an activation pattern concentrated around the winner. The winner's weight vector is then compared with all word forms from the current lexicon, and the one closest to it (in Euclidean distance) becomes the retrieved word form. The comprehension process is analogous and runs in the opposite direction.

While the overall architecture of DevLex is similar to that of DISLEX, the major difference between the two is that DISLEX uses standard SOM, whereas DevLex uses GMAPs. The use of GMAPs is based on the consideration of the growing nature of the lexical learning task: both the size of the lexicon and the input space (the number of non-zero components in the semantic vectors) grow as learning progresses. In addition, the GMAPs in DevLex combine the advantages of SOM and ART2 in learning (see discussion in Section 2.3). DevLex also allows for the dynamic creation of lexical semantic representations as inputs to the S-GMAP, using the WCD (word co-occurrence detector) module described below.

DevLex operation involves three processes: (1) formation of distributed word representations (both phonological and semantic), (2) GMAP organization, and (3) formation of associative links between form and meaning. The second and third processes occur simultaneously. The first process occurs relatively independently of the other two and can be thought of as the process in which the child extracts phonological and semantic information from lexical contexts (sentences) during listening. Our working hypothesis here is that the formation of word meanings is a gradual process and that it takes much longer than the extraction of word forms. Before providing detailed descriptions of DevLex's mechanics, let us first describe how input data—both phonological word forms and lexical meanings—are created and used.

2.2. Phonological and semantic representations of the input lexicon

We used the PatPho system to construct the phonological patterns for word forms. PatPho is a generic phonological pattern generator for neural networks that fits every word (up to trisyllables) onto a template according to its vowel–consonant structure (Li & MacWhinney, 2002). It uses the concept of syllabic template: a word's representation is made up by combinations of syllables in a metrical grid, and the slots in each grid are made up by bundles of features that correspond to phonemes, consonants (C) and vowels (V). For example, a full trisyllabic template would be CCCVVCCCVVCCCVVCCC, with each CCCVV representing one syllable and the last CCC representing final consonant clusters. This template has 18 C and V units. PatPho uses articulatory features of phonemes (Ladefoged, 1982) to represent each C and V, and a phoneme-to-feature conversion process produces real valued or binary feature vectors for any word up to three syllables in length. We decided to use the binary option, because binary coding provided better discrimination of phonemes. To save computational time, the feature vectors were dimension-reduced to 54 dimensions using PCA (accounting for 99% of variance). PatPho shows advantage over traditional phonemic representations in its ability to accurately capture phonological similarities of multisyllabic words.

We assume that children's auditory representations of word forms are close to those of the target language (Menn & Stoel-Gammon, 1995). In contrast, children's semantic representations are built up gradually during the development of the lexicon. We constructed two sets of semantic word representations in qualitatively different ways: the first set was constructed from word co-occurrence probabilities using the WCD network (word co-occurrence detector; see also Farkas & Li, 2001, 2002a), and the second set was derived from the WordNet database using special feature-extracting routines (Harm, 2002). Our best results were achieved when these two sets of representations were combined and normalized for vector uniformity, as discussed below.

2.2.1. WCD-based meanings

WCD is a special recurrent neural network that learns the lexical co-occurrence constraints of words (see Appendix A for details). WCD reads through a stream of input sentences (one word at a time) and learns the transitional probabilities between words which it represents as a matrix of weights. Given a total lexicon sized N, all word co-occurrences can be represented by an $N \times N$ contingency table, where the representation for the ith word is formed by concatenation

of *i*th column vector and *i*th row vector in the table. This procedure is similar to the method used within the HAL model of Burgess and Lund (1997). Hence, the two vectors correspond to the left and the right context, respectively. The WCD method allows us to build semantic representations on the fly, incorporating more and more different words in a context, until the size of the lexicon (n) reaches the target N.

This incremental scenario entails that the number of non-zero components in the semantic vectors will grow as the child learns new words (and is always $2n$). For example, given a total target lexicon, N, of the size 500 and the current lexicon, n, of the size 50, only the first 100 components of the 1000-dimensional word-representation vectors can be non-zero. The rest of the components are zero, since the remaining words are ignored because they are not yet available on the S-GMAP, despite their appearances in the corpus. This scheme can be considered as corresponding to the situation in which the young child does not understand some context words and treats them as noise. Metaphorically one can think of this scenario as filling the holes in a Swiss cheese: initially there may be more holes than cheese but the holes get filled up gradually, as more words are incorporated and the co-occurrence context expands.

To make the WCD-based representations uniform across different vocabulary sizes, all representation vectors were submitted to random mapping (Ritter & Kohonen, 1989) to achieve vector normalization in terms of number of (non-zero) components. The mathematical details of this random mapping process are given in Appendix B.

2.2.2. WordNet-based meanings

WordNet-based features were derived by a feature generation system (Harm, 2002) that can produce a set of binary features for each of the 500 words. Harm's software incorporates semantic features mainly from WordNet, a computational thesaurus that provides semantic classification of the English lexicon in terms of hyponyms, synonyms, and antonyms, as well as searchable word entries with semantic definitions (Fellbaum, 1998; Miller, 1990). Harm extracted relevant semantic features from WordNet for nouns and verbs, but for adjectives he hand-coded the semantic features according to a taxonomy of features given by Frawley (1992). About two-dozen adjectives in our simulation vocabulary were not found in Harm's coding. For these, we hand-coded a set of features, also according to Frawley. The above method yielded a list of 459 binary features in total, with the number of features for any given word ranging from 1 to 12. For computational consistency, WordNet-based representation vectors were also submitted to random mapping for dimension reduction ($D=100$ dimensions). Note that, however, unlike WCD-based representations, WordNet-based representations are static and thus do not evolve along with the growing vocabulary.

One could consider the addition of WordNet-based features as providing the learning process with semantic information grounded on concepts and percepts in the real world. One would expect that both sets of information should be important for lexical acquisition, and that simulations using both WCD-based and WordNet-based representations should end up with an increased accuracy in capturing linguistic categories. We will discuss the effects of such a combination in Section 3.

2.3. Growing map (GMAP)

DevLex was designed to allow us to explore the three core theoretical issues we discussed earlier. Specifically, it should (1) allow for an incremental lexicon to self-organize topographically into categorical representations in a growing map architecture, (2) model word confusion rates in production at early stages of learning, and (3) capture the age of acquisition of the words in the map in which learning needs to be stable despite subsequent lexical growth.

Our pilot simulations indicated that SOM accounts well for (1) and (2), but has difficulty in accounting for (3). This is because SOM uses a neighborhood concept that allows for the formation of a topographic map, which makes it difficult to set SOM learning parameters appropriately for an incremental lexicon. As new words enter the lexicon for learning (i.e. the pool of words used for training at current time), they tend to 'fluctuate' in the map (i.e. changing their winners due to neighborhood influence) and merge with other words in the map during growth. Such fluctuations and mergers result in catastrophic interference, a problem we mentioned in Section 1.

To solve this problem, our model implements GMAPs by a novel combination of SOM with the Adaptive Resonance Theory for real-valued patterns (model ART2; Carpenter & Grossberg, 1987). Although SOM and ART2 have been designed for different purposes, both models are classical self-organizing neural networks and features of both can be exploited in a model of lexical development such as DevLex. The building block, GMAP, consists of nodes that form a planar graph (rather than a grid), whose connectivity structure determines the neighborhood relations, as illustrated in Fig. 2. All GMAP nodes are

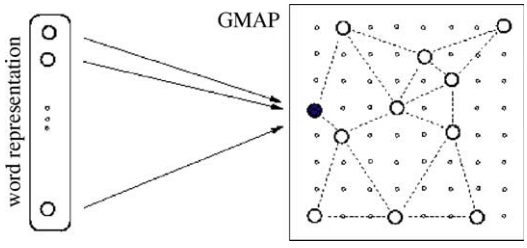

Fig. 2. A diagrammatic sketch of GMAP (growing map). All recruited nodes (shown as large empty circles) in GMAP fit an underlying grid, and new nodes can only be recruited in yet unoccupied positions (small circles). Nearest nodes are connected with each other to define the GMAP topology. Each GMAP node has connections from all inputs (as in SOM). Only the links to one node (filled circle) are shown here.

defined on an underlying grid, so that each node can be identified by its two integer coordinates in that grid.

Although there have been several growing neural network models in the literature (Blackmore & Miikkulainen, 1995; Fritzke, 1994; Marsland, Shapiro, & Nehmzow, 2002), no previous model has combined SOM and ART within a growing architecture.[1] We see such a combination as combining the virtues of both SOM and ART2 for effectively modeling lexical acquisition. Thus, we designed DevLex to function in two modes: first as a SOM for map organization, and then in ART mode during vocabulary growth.[2] Since ART does not have the topography preserving features as does SOM, we enhanced it with a novel mechanism of topographic placement of newly recruited nodes in GMAP (see Appendix C). In addition, the transition between the two learning modes is not sharp but gradual (see Section 2.5), allowing us to model the gradual nature of developmental phenomena. Algorithmically, the two modes (SOM and ART) of DevLex differ from one another in two respects: node neighborhood and node recruitment. In SOM mode, neighborhood structure exists but new nodes cannot be recruited, whereas in ART mode, there is no neighborhood function but new nodes can be recruited. Other properties such as weight update and learning rate are kept the same for both modes. In short, such a dynamic instead of a fixed map structure solves the combined problems of (1)–(3), and also facilitates training (e.g. overall typographic order can be more easily established) because the model need not to update all possible weights right from the beginning.

We hypothesize that the transition from SOM to ART corresponds to maturational processes (both functional and anatomical) known to occur early in mammalian brains which, in language-related areas may be associated with early vocabulary growth (see Quartz & Sejnowski, 1997 and references therein).[3] In particular, acceleration in vocabulary growth in early child language may be associated with the increase in the number of synaptic connections within and across cortical regions in the brain (Elman et al., 1996). As the child learns more and more words, lexical confusions tend to occur. To resolve these confusions, a mechanism is needed to accommodate further word learning. DevLex does this by recruiting new resources (i.e. new nodes in the ART mode), while keeping the overall structure of already learned words stable. During SOM learning, the network can mark out the overall shape of the basic lexical space (i.e. basic structure and topography of the lexicon), whereas during ART learning, the network can add in more words and fine-tune this initial structure. This consideration is based on the hypothesis that children can quickly acquire a growing vocabulary if the basic lexicon space is established (Li & Zhao, 2004; Elman, 2004).

2.3.1. GMAP development in SOM mode

When DevLex is in SOM mode, GMAP formation is expected to exploit the topography preserving properties of SOM to develop lexical organization. Self-organization in SOM occurs in a layer of processing units, arranged in a two-dimensional regular sheet (a planar graph in our case), where each processing unit in the network is a location on the map that can uniquely represent one or several input patterns (Kohonen, 1982, 2001). In SOM, two major parameters are modulated to help achieve convergence and order: the learning rate for weight update, and the neighborhood radius that determines the area of organization. DevLex uses these ordering properties of SOM to establish a coarse map of early words, allowing the model to simulate the early emergence of coarsely organized linguistic categories.

2.3.2. GMAP development in ART mode

In ART mode, node neighborhood becomes zero, and node recruitment is turned on. The elimination of node neighborhood reduces the chance of node shifting for winners, as each node now has a better chance to preserve its status as winner for the same word. Node recruitment is configured so that new nodes can only be recruited in free grid positions. At the beginning of learning, the GMAP starts with a subset of nodes fitting an underlying grid and new nodes are recruited to free grid positions in its limit. This process continues until the GMAP becomes a SOM in terms of connectivity, that is, a regular grid in which all positions are taken (see Fig. 2). We do not view recruiting as involving the physical addition of new nodes, but rather incorporating new nodes by sprouting their lateral connections to neighboring nodes and allowing all inputs to sprout connections to them. This weight sprouting models the process of synaptogenesis that may be associated with early vocabulary growth (see Elman et al., 1996; Quartz & Sejnowski, 1997).

ART2, as one of several ART family models, has been designed as a biologically plausible model of clustering (categorization) of real-valued input vectors. The model reads input vectors one at a time and whenever the input is not sufficiently close to any of the existing nodes

[1] Marsland et al.'s (2002) model used a similar mechanism as ours for recruiting new nodes (e.g. allowing recruitment at every single iteration), but their model did not combine SOM and ART and differed from DevLex in other aspects.

[2] Strictly speaking, we can only speak of the ART mode instead of the ART network for DevLex. The model borrows from ART only the capacity of node recruitment, while its GMAP architecture remains unchanged.

[3] While contentious debates exist on the facts and the functional role of synaptogenesis and neurogenesis in the brain, recent evidence suggests that both types of neural genesis occur across the life span and have important developmental and cognitive consequences for the learning species (see Gould, Tanapat, Hastings, & Shors, 1999; Quinlan, 2003; Shultz, 2003 for discussion).

(more precisely, to any of their weight vectors as in SOM), a new node becomes recruited. Otherwise, only the winner's weight vector is slightly changed toward the current input without any consideration of neighborhood structure as represented in the GMAP. In the original ART2, the closeness between an input and nearest weight vector is based on the cosine of the two vectors, which is then compared to a pre-specified vigilance parameter. Low vigilance leads to coarse clustering of inputs, in which a set of similar inputs will tend to form a cluster, and be represented by a single node whose weight vector will converge to the centroid of these inputs. In contrast, high vigilance leads to very fine-grained clustering, where each input (word) can be assigned its own representing unit. In DevLex, we use Euclidean distance as a similarity measure, which is compared to a node insertion threshold ($q_{\text{pho}}(t)$ for P-GMAP and $q_{\text{sem}}(t)$ for S-GMAP) being inversely related to the vigilance parameter. A node may be inserted if $\min_k\{||\mathbf{x}-\mathbf{m}_k||\} > q(t)$). In order to facilitate smooth transition between the two modes (and hence, allow for modeling development), linearly decreasing insertion thresholds are used for both GMAPs.

Unlike SOM, ART2 does not represent topographic relations between nodes. However, topographic representations are important for our study, as discussed earlier. Since we wish to examine the topographical organization of lexical representations, we need to find the position of the new node in the existing GMAP that would best fit the nearest neighbors' preservation from input to output. We do this by using the Distance Ratio Preservation (DRP) procedure. Appendix C provides the details of the DRP procedure, and Appendix D the pseudocode of the whole DevLex algorithm.

2.4. Structuring the input lexicon

To model early lexical acquisition by children, we used as our basis the vocabulary from CDI, the MacArthur–Bates Communicative Development Inventories (Dale & Fenson, 1996; Fenson et al., 1994). From the Toddler's List, we extracted 500 words and sorted them according to order of acquisition (the original Toddler's List contains 680 words; we excluded the homographs, word phrases, and onomatopoeias from our analysis). The order of acquisition of the 500 words was determined by the CDI lexical norms at the 30th month (Dale & Fenson, 1996), as 500 words may correspond to the size of the vocabulary of an average 30-month-old child (Bloom, 2000). To model an incremental lexicon and to make the modeling more tractable, we divided the 500 words into 10 major growth stages, at 50, 100, 150, 200, 250, 300, 350, 400, 450, and 500 words. These major stages were necessary for computing WCD-based meanings (see below). Within each stage, the new 50 words were further divided into five sub-stages, each having 10 words. As a result, the vocabulary grew in a more fine-grained fashion with 10 new words at each increment.[4]

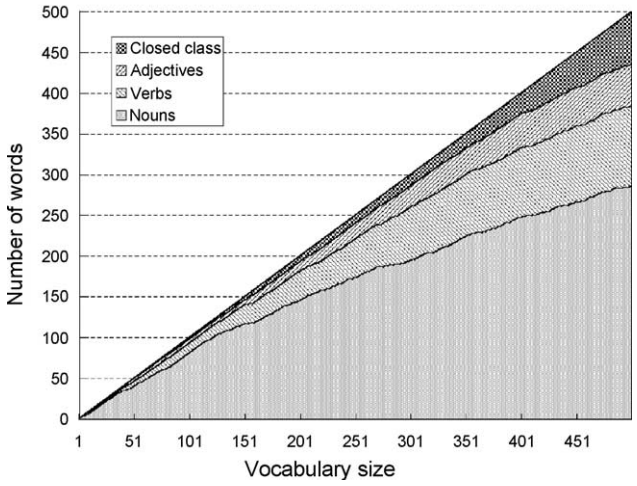

Fig. 3. Vocabulary growth profile according to CDI order of acquisition for the four major categories.

Unlike previous simulations our study used a growing vocabulary whose composition is consistent with known observations regarding variation in growth across grammatical categories (Bates et al., 1994). Fig. 3 presents the vocabulary composition regarding the four major grammatical categories used in our simulations. In the original CDI, words are divided into 22 smaller categories. Four CDI categories—games and routines, sound effects, time words, and places to go were excluded from our analyses (see Bates et al., 1994 for why these words should be excluded from a normal analysis of vocabulary development). The major categories in Fig. 3 thus collapsed the remaining 18 CDI categories, as follows: (1) nouns, including 10 nouns subcategories: animals, body parts, clothing, food, household, outside, people, rooms, toys, and vehicles, (2) verbs, (3) adjectives, and (4) closed-class words, including six subcategories: auxiliary verbs, connecting words, prepositions, pronouns, quantifiers, and question words. The figure shows that the number of nouns linearly increases toward the end, and the number of verbs also increases but with a much slower rate. Given the constraints of these earliest words, there is a clear 'noun-bias' in this vocabulary (286 words), as compared to the number of verbs (98), adjectives (51), and closed-class words (65). Adjectives and closed-class words rarely come into the vocabulary until the 100-word mark. Thus, the accelerating profile of vocabulary growth at early stages of learning might be associated with this noun-bias (see GoldField & Reznick, 1990).

[4] We used this stage-like growth in vocabulary size only as a way of streamlining the computation and organizing the presentation of the results. It would have also been possible to break the training set down into maximum, 500 stages with one-word increments at a time. This would not have qualitatively altered the general results, but would have prolonged the training time (due to the number of evaluations to be run at the end of each stage).

Table 1
DevLex parameters that modulate vocabulary development during training

Vocabulary size	50	100	150	200	250	300	350	400	450	500
GMAP mode coef.	1	1	1	0.8	0.6	0.4	0.2	0	0	0
Nbhd radius	3	2	1	0	0	0	0	0	0	0
q_sem	1	0.85	0.7	0.55	0.4	0.25	0.25	0.25	0.25	0.25
q_pho	3.5	3	2.5	2	1.5	1	0.9	0.9	0.9	0.9

To represent the 500 words as input to our network, we first generated the phonological forms for the words using the PatPho generator, in a left-justified template with binary encoding (Li & MacWhinney, 2002; see also Section 2.2). Second, the WCD-based word meaning representations for the 500 words were computed by using the parental input from the CHILDES corpus (Li, Burgess, & Lund, 2000). The parental CHILDES corpus contains the speech transcripts from child-directed adult speech in the CHILDES database (MacWhinney, 2000). WCD was performed on words at each of the 10 major growth stages, resulting in 10 different data sets with increasing number of entries (i.e. from 50 to 500). Each set was used during the corresponding five sub-stages of 10-word increments. As discussed earlier, WCD learns the transitional probabilities between co-occurring words and transforms these probabilities into normalized vector representations for word meanings. Because WCD relies on the use of lexical contexts, the same word will have different representation vectors at different stages as word contexts fill out with the growing lexicon, resulting in representations that converge with the growing vocabulary.

2.5. Network initialization and simulation parameters

Both the S-GMAP (word meaning) and the P-GMAP (word phonology) in DevLex were initialized with 300 nodes, a number chosen heuristically to be large enough to span the whole grid and support the finding of a 2D manifold early on, but small enough to allow for node recruitment given the lexicon size. The nodes were randomly scattered on an underlying 50×50 grid for each map. Both input and associative links were set to small random numbers. During each lexical growth stage (i.e. 50, 100, and so on to 500), words entered the lexicon one by one: at every iteration a word was picked from the pool—the current vocabulary at the stage—according to its word (token) frequency in the parental CHILDES corpus. Since the word frequency distribution follows Zipf's law (Zipf, 1932), we calculated the logarithms of these frequencies to force a more even distribution of words in the input.

A set of six simulations was run with identical modeling parameters. Some of the parameters are shown in Table 1. GMAP mode coefficient modulates the transition from SOM to ART mode. Its relative value (between 0 and 1) denotes the switching moment from SOM to ART mode within each sub-stage. Hence, DevLex spends first three stages exclusively in SOM mode, followed by gradual transition to ART mode during stages 4 (200 words) to 7 (350 words), after which DevLex stays completely in ART mode. The second parameter, neighborhood radius, sets the node interaction range between the nodes during learning, and its non-zero values at early stages were used to allow for GMAP reorganization as new words come in. It becomes zero during the transition to the ART mode and remains zero thereafter. The next two parameters, *q_sem* and *q_pho*, are the threshold distances that modulate the rate of node recruitment in the semantic and phonological GMAPs. The linear decrease during the first half was chosen to allow for an initially lower node recruitment rate for both GMAPs. Node recruitment starts at around 200 words, with low rates until 250 words, followed by a faster increase until the end. On average, each GMAP ended up having roughly twice as many nodes (600) as when it began (300).[5] Finally, each stage was trained with the same number of 50 epochs, and with a constant learning rate of 0.05 for input weights and 0.1 for associative links.

The parameters in Table 1 were chosen heuristically to allow for the modeling of vocabulary development in terms of (a) coarse map organization and reorganization during SOM mode, (b) vocabulary spurt associated with initial word confusions and later de-confusions during the transition to ART mode, and (c) stabilized vocabulary growth and convergence during ART mode.

3. Simulation results

All results reported below were based on the averages of the six simulations with identical parameters as described in Section 2.5.

Before presenting the simulation results, we briefly discuss our choice of semantic input representations for the network. In Section 2.2, we discussed two types of semantic representations, the WCD-based co-occurrence results and the WordNet-based semantic features, and hypothesized that they could be combined to yield more accurate representations of word meanings than the ones generated by either the WCD or the WordNet method alone. To verify this hypothesis, we used a simple, *k*-nearest neighbor

[5] It turned out that some nodes were always left unused, so both GMAPs typically ended up with a larger number of nodes than the number of acquired words.

(*k*-NN) classifier (Duda, Hart, & Stork, 2000) to determine the existence of compact category clusters for the training materials (the 500 toddler words from the CDI database). A 5-NN classifier was built for each word based on all the remaining words in the considered lexicon. The label of the test word was predicted according to the most frequent label among the *k* nearest neighbors, that is, the words being closest to the test word in input space. Ties in prediction were broken randomly. The classification rate for words in each of the 18 categories was evaluated against all 500 words.

Results from this *k*-NN analysis indicate that a combined representation that incorporates both the WCD-based and the WordNet-based features provides higher accuracy in classification for most of the original 18 categories in the CDI database. Because the combined representation includes both dynamic context features and static semantic features, it is no surprise that it yields a better data structure than either type of features used alone. In realistic language learning, this could be thought of as having both the statistical information of the sequential linguistic input and the perceptual information of words as cues to word meanings. In what follows, therefore, we report simulations based only on the combined representations as the semantic input.

3.1. Category emergence and reorganization

As discussed earlier, DevLex presents an emergentist alternative to the nativist assumption that lexical categories are hardwired in the brain. Our simulations show that the representation of linguistic categories can emerge in the topology of the network as a natural outcome of the self-organizing process in lexical learning. Using the *k*-NN classifier as discussed above, we were able to determine if major grammatical categories emerged in the network at various points during training.

Fig. 4 depicts the representation of the four major categories (nouns, verbs, adjectives, and closed-class words) in terms of their compactness in the S-GMAP space, computed by a 5-NN classifier. It can be seen that, by the end of the training, the S-GMAP formed good representations for all four categories. Except for nouns, all the categories have shown some degree of development, that is, moving from a less compact to a more compact category (hence higher classification rates). The high compactness of the noun category, starting at the beginning, is possibly an artifact of the *k*-NN measure, due to the overwhelming number of nouns at early stages of learning, relative to other word classes (see Fig. 3 for the structure of the input data and the 'noun-bias' discussed there). In contrast to the nouns, closed-class words are poorly classified early on because of their relatively late entry into the lexicon (again see Fig. 3), and their classification rate increased only toward the end. Moreover, nouns, verbs, and adjectives formed compact clusters during the SOM mode and were fine-tuned later when new words were added to their existing coarse structure; by contrast, closed-class words formed compact clusters only in the ART mode of

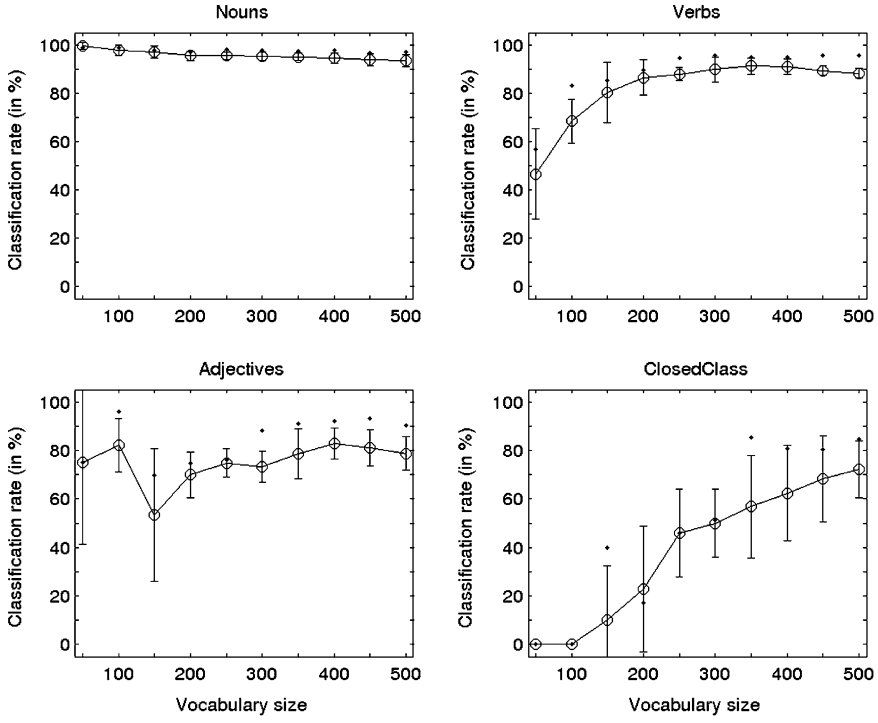

Fig. 4. Classification rates for four grammatical categories using a 5-NN classifier, evaluated at regular stages of 50-word increments in the S-GMAP. Dots in the graphs correspond to the results of a 5-NN classification performed for the input space, which can serve as a reference (the baseline) for the amount of distortions in data structure due to network's 2D mapping, and as shown, the amount of distortions is small.

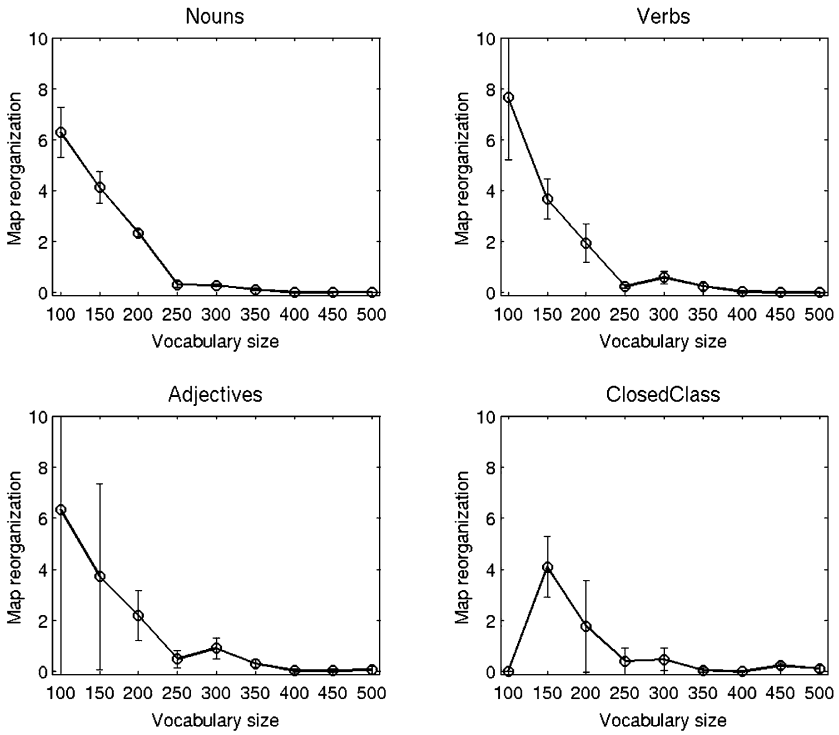

Fig. 5. S-GMAP reorganization as a function of vocabulary size, computed as the amount of word shifts in the map's underlying grid. Each map was compared with its predecessor at the previous stage (the index of x-axis refers to the successor), comparing the positions of the words common to both maps. As shown, all categories underwent significant reorganization at initial stages of development when GMAPs were in the SOM mode. The zero reorganization of closed-class words between 50- and 100-word stages was due to the absence of this category in the early vocabulary. High standard deviations at early stages indicate that category reorganization was more variable as well as more flexible early on.

the model when the vocabulary size reached 300–400. Finally, the temporary decrease in accuracy for adjectives during 100–150 words was due to the poor data structure in the input (which can be seen in the dotted lines), and not due to the inability of the S-GMAP learning—it is unclear, however, why the meanings of adjectives are particularly difficult to represent for this stage (but see Gasser & Smith, 1998, for an in-depth analysis of children's late acquisition of adjectives, as compared with nouns).

Category organization as well as category emergence is the focus of our model, and we used a map reorganization measure to monitor it explicitly. For any pair of two adjacent stages (e.g. 100-word and 150-word stages), we evaluated map reorganization for each major category as the Euclidean distance of the same word in the two maps, averaged over all words currently present in the category. For example, in the two maps for 100 and 150 words, we compared only the 100 words that were common to both maps. Results of this procedure are shown in Fig. 5. They show that all categories underwent significant reorganization at initial stages of development when GMAPs were in the SOM mode.[6] The SOM-mode learning with non-zero neighborhood encouraged the network to reorganize the map structure. This higher plasticity for reorganization, on the other hand, leads to lower stability initially in the map. The generally higher standard deviations for early stages shown in Fig. 5 also indicate that category reorganization was more variable as well as more flexible early on. At later stages, however, map reorganization gradually disappeared, and as new words entered the lexicon they were simply added to the existing structure of the network. Decreased word shifting at later stages was further reinforced by the converging WCD-based word representations towards the end of growth, which implies that once the winner for a certain word was chosen, there was no need to change it later. There is yet not much empirical information about how children organize and reorganize their word categories in the mental lexicon in the fashion as we have described here, but this early-plasticity-and-late-stability scenario would be consistent with the standard view about language acquisition in many other domains (Bates, 1999; Elman, 1993; see also Section 4).

3.2. Lexical confusion and naming deficit

The k-NN measure provides us with only a coarse picture of how the lexical categories are distributed on the S-GMAP. High classification rates for a given category suggest that this category forms a compact cluster in

[6] As a reference, a completely random remapping of words could be calculated as an average distance between two randomly chosen grid positions in the map, yielding a considerably higher number that depends on the grid size (50×50 in our simulations).

the map. However, this measure does not give us more fine-grained information at the individual word level such as how words are differentiated within each category. For example, although Fig. 4 indicates that nouns formed compact categories from almost the beginning, an inspection of the individual nouns showed that many of these items were confused with other items within the noun category (that is, represented by the same nodes in the map). Examination of this type of lexical confusion is important, as empirical literature suggests that young children experience temporary periods of confusion of word uses, often resulting in naming errors (Bowerman, 1978; Gerskoff-Stowe & Smith, 1997; see discussion in the introduction). Word confusions at early stages have often been linked to the so-called 'vocabulary spurt', a sudden and rapid increase in the rate at which new words are produced in children's spontaneous speech. Although vocabulary increase in our model is based on the CDI data and has no such abrupt vocabulary spurt, it would still be informative to see what types of confusion the network makes with various words.

To monitor word confusion rates across categories in DevLex, we used a set of four measures to quantify the network's representation accuracy, according to Miikkulainen (1997). Two of these are accuracies of map representations (semantic and phonological) that quantify the proportion of words being uniquely represented in either GMAP. For example, if *car* and *truck* are mapped onto the same node in S-GMAP, this is interpreted as indicating that the network cannot differentiate between these two concepts. The other two measures are associations between the two GMAPs, allowing for the modeling of comprehension (via phonology-to-semantics links) and production (via semantics-to-phonology links). For example, if the node with the highest activation in the S-GMAP is consistent with the activated node in the P-GMAP (e.g. *truck* in S-GMAP and /tr@k/in P-GMAP), this is interpreted as indicating that the network has correctly named the concept in production.

Fig. 6 presents the results of the four measures. The lower two curves were related to the individual maps and the upper two curves related to the associations between the maps and can be interpreted as production and comprehension rates, respectively. Our analyses here focus on the production errors although the model displayed very similar comprehension profiles. As can be seen, the SOM mode of the model produces coarse category formation and reorganization, as discussed in Section 3.1, and it was this mode that is associated with high confusion rates. Word confusions started to decrease when GMAPs gradually switched to the ART mode. The decrease of confusion errors lagged behind for the associations between the two maps as compared with individual maps, which shows that the production and comprehension associations required more time as they were trained by Hebbian learning. Fig. 6 also shows that confusions occur for both the map representations and the map associations, which suggests that both lexical representations and form-meaning associations are responsible for lexical confusions in the model.

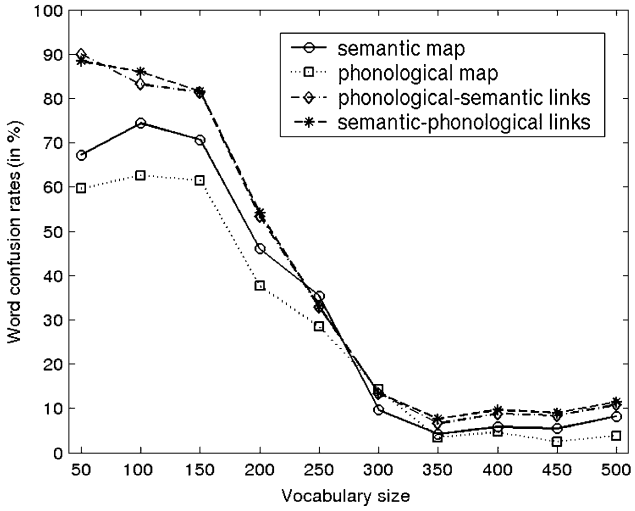

Fig. 6. Word confusion rates during early vocabulary growth. Frequent confusions occur during initial map formation and map reorganization, but then decrease as GMAPs gradually switch to ART mode. By 350 words in the vocabulary, word confusion rate reached a minimum. Standard deviations are not shown here as they were relatively small.

In addition to displaying lexical confusion, our model lends itself nicely to explaining the effect of word density on word confusion: most of the confusions in our model come from densely populated areas in the map (with many words in the neighborhood), which is consistent with predictions based on studies of similarity neighborhood effects that lexical access and recognition are more effortful and error-prone when they are associated with higher lexical neighborhood density (Charles-Luce & Luce, 1990).[7] In the context of S-GMAP, a densely populated area can be understood in terms of either a set of neighboring nodes, or, more locally, within single nodes that serve as best matching nodes for more than one word. In both cases, the higher word density represented by these nodes resulted in higher word confusion rates, as compared with words in sparsely populated areas. Interestingly, word confusion occurs more often for nouns than for the other major word categories, especially in the early stages, because nouns are more densely populated in the GMAP initially, due perhaps to the early noun-bias in the CDI vocabulary. Other categories, especially the closed-class words, might have been more robustly represented in the GMAP due to low density and are thus less susceptible to confusion.

[7] In our discussion of lexical density we are concerned with the compactness of words in terms of semantic representations of the lexicon. This contrasts with the tradition of spoken word recognition research, where lexical neighborhoods and lexical density are defined by phonological similarities of words; for example, *bat*, *cat*, *hat*, *mat* and *pat* form a dense neighborhood of words, as opposed to cup and pup that share few similarly sounding words.

The higher the word density is, the more strongly words are to compete with each other during lexical retrieval in word production. This result would fit in well with the arguments made by Gerskoff-Stowe and Smith (1997). However, word density alone cannot explain the whole story. When collapsed across categories, word density for our training vocabulary was generally on the rise until it plateaued at around 300 words (see Fig. 8 in Section 3.3). Yet empirical data (and our Fig. 6) show that word confusions are only temporary early on and would decrease quickly. Thus, other factors needed to be considered in this picture. One argument made by Bowerman (1978) regarding children's word confusions at a somewhat later stage is that the recognition of shared meaning components of words, or semantic relatedness, plays an important role (see discussion in Section 1). Thus, we evaluated semantic relatedness of confused words on a node-by-node basis. For each ambiguous node we checked whether there existed a dominant category, and found that more than half of the category labels among the ambiguous nodes were identical. When semantic relatedness was considered with respect to the original 18 CDI sub-categories (see Section 2.4), we observed that about 75% of all confused words were semantically related at the beginning, going up to about 90% toward the end of vocabulary growth.

In terms of the model, the increasing profile of semantic relatedness in word confusion is a direct consequence of the topographic organization of the GMAP that attempts to map similar words close to each other in the map. This type of semantically motivated organization explains the child's inability to differentiate between two similar concepts, in which case the child uses two word labels interchangeably (hence the naming errors). The more related words children have to learn within a given period, the more likely they will experience representational confusion in an overloaded lexical memory. Thus, in this view, semantically related, but not unrelated, pairs of words are the sources of confusion in lexical retrieval during children's word production.

3.3. Age-of-acquisition (AoA) effects

Previous models using feed-forward neural networks have shown that age-of-acquisition effects can arise naturally in connectionist learning. AoA effects impose significant constraints on connectionist models because of their requirement on both plasticity and stability in learning and representation, as discussed earlier. A connectionist model has to, minimally, overcome catastrophic interference to show AoA effects. The current architecture of DevLex lends itself nicely to the modeling of AoA, precisely because of its ability to deal with the plasticity-stability problem, due to its use of the SOM and ART modes.

Two conditions must be satisfied in our model before a word is considered acquired: the word must have a resource allocated in both GMAPs, and there must exist an unambiguous link between form and meaning of that word. Resource allocation occurs immediately for all new words (ART was designed to allow fast assignment of resources to new items), as long as GMAP has a sufficiently low threshold distance (high vigilance parameter in ART). Hence, the earlier an input word is presented to the GMAP, the earlier it will have its representative unit in the GMAP, due to the relatively lower threshold distance early on. Because higher threshold distance leads to coarser clustering (the same unit responds to more, mutually similar words), the new pattern must be more dissimilar from all existing patterns to be considered new and hence assigned a new unit. As mentioned earlier, the threshold distance was set to increase linearly with a growing lexicon in the first half and then remained constant.

Whereas the resource allocation condition could be satisfied immediately, the second condition requires more learning time. An unambiguous link between the meaning and the form (word production) requires training with Hebbian learning on the units that are initially all connected with one another across maps, and the gradual weight changes by small amounts can be time-consuming. However, the earlier a word enters the lexicon, the earlier the link between its form and meaning may be established, resulting in earlier acquisition of this word (i.e. ability to produce the word correctly). Such gradual fine-tuning of associative links between form and meaning could be compared to a situation in realistic language learning whereby children, having acquired some relevant lexical semantic representations, have not firmly established the links between forms and meanings (e.g. imaging a situation in which the child calls both dog and cat with *doggie*). Such lack of strong associative connections may result in retrieval difficulties during the naming of objects (recall our early discussion on lexical confusion).

Fig. 7 presents the picture of the time at which words are acquired, as a function of the current vocabulary size. Acquisition times were expressed in terms of current vocabulary size, and were based on word production rates (semantic-to-phonological association links, given that most AoA studies have focused on production). It can be seen that although the overall shape of word development appeared curvilinear in Fig. 7, the data set could be approximated by a positive regression line after the vocabulary size reached 150 words. This means that for most words after the 150-word mark, the earlier a word entered the vocabulary learning pool, the earlier it was likely to be acquired. Note that the 150-word mark corresponds to the end of the SOM mode, where lexical reorganization was more vibrant than the later periods of learning in the ART mode.

Our model differs from previous connectionist networks of AoA effects in several respects: (1) instead of using arbitrary patterns of random bits, we used realistic words that have phonological and semantic information as our input patterns; (2) instead of dividing words into early

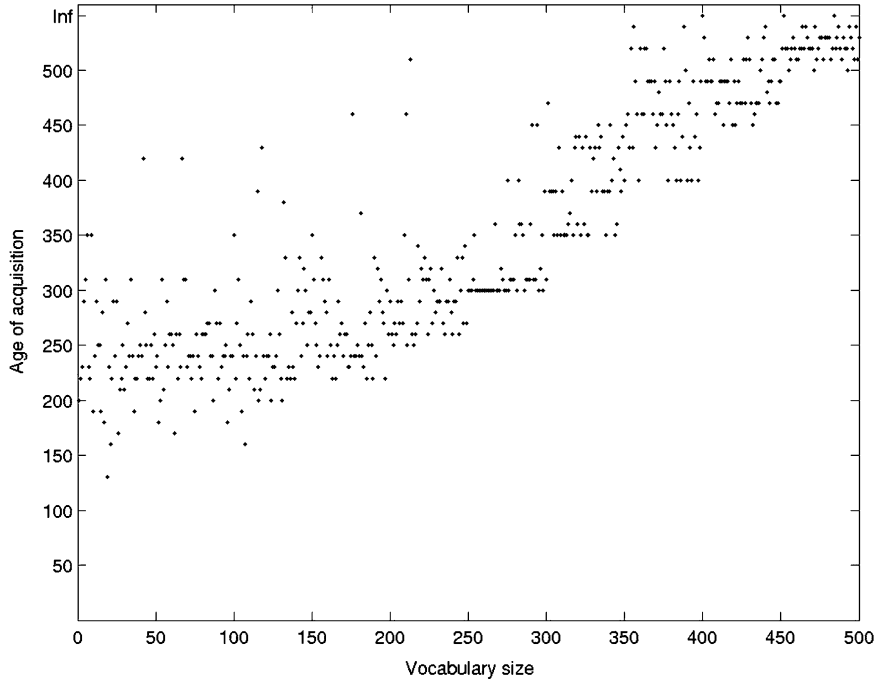

Fig. 7. Age of acquisition in the growing vocabulary as a function of the current vocabulary size, evaluated every 50 epochs.

versus late learned groups, we examine the natural order of learning in which words are acquired. In our model, the time at which words enter training is naturally different for different words (according to CDI); and (3) instead of modeling AoA effects as the network's ability to reconstruct earlier presented patterns, we treat AoA as a natural correlation of the time at which words are acquired and the dynamic change that occurs to the vocabulary as a whole. Thus, although all of the words in our training vocabulary would probably be regarded as early learned words in empirical studies, we think that there might be changes in the structural properties of the vocabulary that differentiate between words acquired at different times, which our model can pick up as a function of learning the vocabulary. Word density seems to be a good candidate in this respect.

Fig. 8 shows AoA effects as a function of the density of the words across stages of learning in our model. Word density here is calculated as the number of words mapped to the target node and all surrounding nodes (within radius 1; see also discussion of word density in Section 3.2). It can be seen that word density gradually increases as learning progresses: in general, words were less densely populated in the map early on, but they became more densely populated with increasing vocabulary size, a natural result of learning (although word density could vary from category to category, as discussed in Section 3.2 in connection with lexical confusions). An interesting prediction of this pattern is that words that are learned earlier should be more resistant to confusion or noise/damage, because of the relatively weaker competition among lexical items in the nearest neighborhood. Such a prediction may be empirically or computationally tested in future research.

4. General discussion

DevLex is the first full-scale SOM-based developmental model of language acquisition. Our goal has been to provide a cognitively plausible, linguistically scalable model to account for lexical development in children. We wanted a model that can capture important insights into mechanisms underlying lexical development. We succeeded in

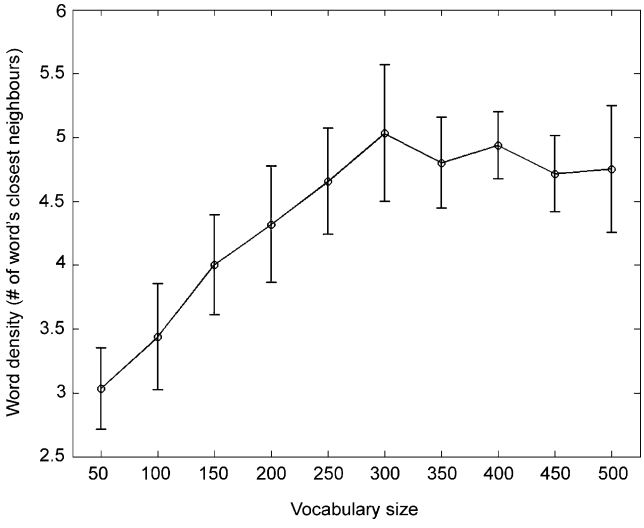

Fig. 8. Average word density as a reflection of AoA, evaluated every 50 epochs on groups of most recent 50 words. Error bars represent the standard errors of density for the given group.

constructing a model that allows for the representation of a dynamically changing linguistic environment in language acquisition. The model develops topographically organized representations for linguistic categories over time, displays lexical confusion as a function of word density and semantic similarity, and shows AoA effects in the course of learning a growing lexicon.

First, with regard to topographically organized representations, our model has direct implications for understanding the representation of linguistic category in the brain. Cognitive neuroscientists have identified various 'brain centers' of language for nouns, verbs, tools, fruits, animals, and so on. Their studies have often operated under the assumption that the brain is a highly modularized system, with different cognitive functions localized to different cerebral regions, perhaps from the beginning (the classical 'modularity of mind' hypothesis, Fodor, 1983). Our DevLex model, however, speaks to a process of 'emergent organization' through which localized brain centers can arise as a function of the developmental processes during ontogenesis (Elman et al., 1996; MacWhinney, 2001b). Our simulation results show how the organization of local maps can give rise to emergent categories across stages of learning, with substantial early plasticity and early competition for the organization and reorganization of category members.

Second, with regard to lexical confusion in early vocabulary learning, our simulations have allowed us to identify a number of crucial factors that lead to lexical confusions observed in early child language. The CDI inventory includes the earliest words that children produce and comprehend, and modeling with this vocabulary permits us to see the early developmental characteristics of the lexicon. For example, lexical confusion in DevLex is directly related to how densely words are populated in the GMAP, measured as the amount of words mapped onto the nearest neighborhood of the target node. Interestingly, lexical confusion occurs more often for nouns than for other word categories, because nouns are more densely populated in the GMAP. There is empirical evidence that the early child English vocabulary is highly biased toward nouns (Gentner, 1982), and hence nouns have a higher chance to get confused with each other.

Not only the sheer number of nouns, but also the high semantic similarities that hold between the words can cause lexical confusions. Semantically motivated lexical confusions and substitutions have been documented in empirical research, as discussed earlier. The majority of confused words in our model are those that are semantically related. Thus, our model simulates word confusion as a function of both word density and semantic similarity. In addition, our model indicates that the sources of confusion cannot be uniquely assigned to either lexical representations or associations between form and meaning: it is reasonable to think that both representations and associations contribute to the early stages of lexical confusion, given that semantically related words may be more difficult to distinguish in the representation, and may also be more difficult to retrieve as they compete for retrieval in particular speech contexts.

Finally, with regard to the age-of-acquisition effects, our model displays both plasticity in learning and stability in representation: plasticity because of the use of SOM in allowing for the organization and reorganization of new words along with learned words, and stability because of the adaptive learning of ART in preventing the learned structure from being disrupted by new learning. The combination of SOM and ART modes in our model represents two modes of the learning dynamics that we think fit the realistic language-learning scenario. Thus, age of acquisition in our network shows up as a smooth function of the correlation between the increasing vocabulary size and the learning time. The effects of age of acquisition are reflected in adult lexical processing speed in empirical studies and in reconstruction rates in some connectionist networks, but in our simulations this is reflected in the density of words across stages of learning.

The ability of our model to capture lexical organization, word confusion, and age of acquisition in formal mechanisms attests to the utility of self-organizing neural networks as psychologically and biologically plausible models of language acquisition. The mental representation and acquisition of words has been a focus of much psycholinguistic research, and our model demonstrates the emergence and organization of the lexicon as a self-organizing process. The model's ability to capture empirical phenomena also lends itself to many interesting predictions that may be evaluated against future empirical findings, such as the relationship between lexical density and lexical confusion (the more densely words are populated in the representation, the more likely they are to be confused by children), and lexical density and resistance to damage (early learned words have lower density and therefore are less susceptible to damage or noise).

In a review article on brain plasticity and language development, Bates (1999) laid out a proposal for early language development in which she highlighted three important features of the developmental process: early plasticity, early competition, and experience-dependent synaptic changes. Our model can be seen as an implementation of these features in a computationally concrete form. Our model displays significant early plasticity in the emergent organization of linguistic categories, and significant early competition in lexical representation and retrieval, especially for words with high-density neighbors. Moreover, the model's ability to learn a growing lexicon and show age-of-acquisition effects is a direct function of the experience-dependent (input-dependent) synaptic changes that are part of the learning dynamics of the self-organizing system. The learning dynamics of DevLex, in particular, early plasticity and late stability, match up well with what we know about general principles of cognitive

development and language development (Bates, 1999; Bates & Elman, 1993; Elman et al., 1996). And finally, the topographically organized maps as used in SOM and our DevLex model may have significant neural underpinnings for the representation and development of the lexicon in the brain (Miikkulainen, 1993, 1997; Spitzer, 1999).

Acknowledgements

This research was supported by grants from the National Science Foundation (BCS-9975249; BCS-0131829) to PL. We thank Risto Miikkulainen for helpful discussions on various aspects of the model. Igor Farkas was with the University of Richmond while the research project was carried out in the university's Cognitive Science Laboratory. He is also in part with the Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia.
Please address correspondence to Ping Li, Department of Psychology, University of Richmond, Richmond, VA 23173, USA. E-mail: pli@richmond.edu.

Appendix A. Word co-occurrence detector (WCD)

Assume that at time t the current word is $i(i=1,\ldots,n)$ and is represented by a localist vector $\mathbf{o}=[o1,o2,\ldots,oN]$ in layer A (see Fig. A1). The previous word j is represented by a localist vector $\mathbf{c}=[c_1, c_2,\ldots,c_N]$ in layer B. Link l_{ij} learns to approximate $P(j^{t-1}|i^t)$, i.e. the probability that the word j precedes word i. Likewise, link r_{ji} learns to approximate $P(i^t|j^{t-1})$, i.e. the probability that the word i follows j. Learning follows the Hebbian rule of the form:

$$\Delta l_{ij}^t = \beta o_i^t (c_j^t - l_{ij}^t) \text{ and } \Delta r_{ji}^t = \beta c_j^t (o_i^t - r_{ji}^t) \quad \text{(A1)}$$

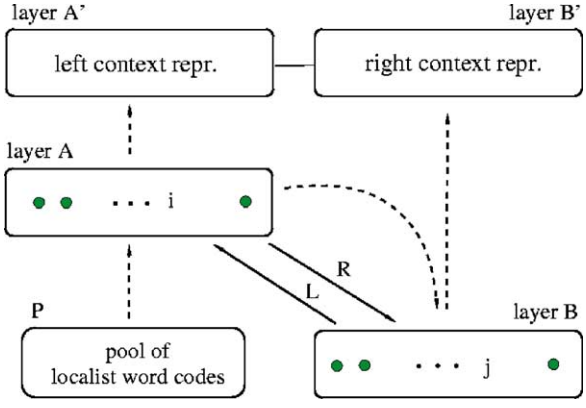

Fig. A1. A diagrammatic sketch of WCD. The solid links between layers of nodes represent activity propagation (via full connectivity), and dotted lines represent pattern transport (via one-to-one links). Layers A and B hold localist representations, whereas layers A′ and B′ hold distributed representations.

where $0<\beta<1$. Word i is then characterized by a concatenation of vectors

$$\mathbf{l}_i = [l_{i1}, l_{i2}, \ldots, l_{iN}] \text{ and } \mathbf{r}_i = [r_{1i}, r_{2i}, \ldots, r_{Ni}]$$

As these values are stored in weight matrices $\mathbf{L}$ and $\mathbf{R}$ between layers A and B, a sequence of operations is required at each iteration to transport weights to unit activations $\mathbf{q}_i^L$ and $\mathbf{q}_i^R$ in layers A′ and B′, respectively, to make them available for further processing. Each operation falls into one of the three categories: pattern transport (denoted by an arrow), activity propagation, and weight adaptation. All units have linear activation functions. The whole sequence consists of the following steps: (1) transport previous word w^{t-1}: A→B, (2) pick up (transport) a new word w^t from the pool: P→A, (3) adapt l and r links (Eq. (A1)), (4) A→B, (5) propagate $\mathbf{o}^t=\mathbf{L}\mathbf{c}^t$ to layer A, (6) A→A′ yielding $\mathbf{q}_i^L$, (7) pick up w^t again: P→A, (8) propagate $\mathbf{c}^t=\mathbf{R}\mathbf{o}^t$ to layer B, (9) B→B′ yielding $\mathbf{q}_i^R$, (10) propagate (and process) $\mathbf{q}_i=[\mathbf{q}_i^L, \mathbf{q}_i^R]$ further up (random mapping, etc.), (11) go to step 1.

Fig. A1 illustrates WCD diagrammatically.

Appendix B. Random mapping

Random mapping reduces the vectors to lower, fixed dimensions ($D=100$ in our case) and can preserve the data structure with high accuracy (as long as D is sufficiently high). Random mapping as a linear transformation is very useful because (1) it does not need to be learned (the mapping coefficients are fixed), and (2) it allows a consistent and on-the-fly transformation with dimensionality reduction ($2N\rightarrow D$).

Generally, in the random mapping method the original data vector, $\mathbf{x}\in\Re^N$ is multiplied by a random mapping matrix $\mathbf{Z}$, resulting in a dimension-reduced vector, $\mathbf{x}'=\mathbf{Z}\mathbf{x}\in\Re^D$. $\mathbf{Z}$ consists of random values and the Euclidean length of each column has been normalized to unity. If we denote the Euclidean norm in k-dimensional space as $||\cdot||_k$, then it can be proven (see Ritter & Kohonen, 1989) that

$$\langle(\|\mathbf{x}'-\mathbf{y}'\|_D^2 - \|\mathbf{x}-\mathbf{y}\|_N^2)^2\rangle \leq \frac{2}{D}\|\mathbf{x}-\mathbf{y}\|_N^4$$

where $\langle\,\rangle$ stands for the average over all possible pairs of vectors $\mathbf{x}$ and $\mathbf{y}$. In other words, the relative distortion of vectors $\mathbf{x}$ and $\mathbf{y}$ under mapping $\mathbf{Z}$ will be small, if the output dimension D is large enough. Hence, we can expect that $\mathbf{Z}$ will preserve the data structure with higher accuracy for higher D.

Appendix C. The DRP procedure

The Distance Ratio Preservation (DRP) procedure is used to maintain topographic 2D order of nodes throughout

GMAP growth. Its design originated from the following considerations. If the input is considered to be a new word (i.e. it is not sufficiently similar to any existing node, compared to the current threshold), the new word's winner (a new node) will be recruited in the GMAP. It will be the node that maximally preserves the ratio of distances of the new word with three nearest neighbors in the input space. The intuitive motivation behind this procedure is to transfer spatial relations between a new word and its nearest neighbors from input (feature) space to output (GMAP) space. The DRP procedure is actually the results of solving the set of three nonlinear equations, and it works well in most cases, except when the positions of three nearest neighbors in the GMAP are highly linearly dependent. The new node is connected by links to the 3 nearest neighbors to become incorporated into the graph. Also, the strength of all edges in GMAP is decreased by a constant amount. Edges with negative weight are deleted.

Algorithmically, the DRP procedure involves the following steps:

1. Find 3 nearest neighbors c_1, c_2 and c_3 to $\mathbf{x}$ such that $d_{c1} \le d_{c2} \le d_{c3}$, where $d_c = ||\mathbf{x} - \mathbf{m}_c||$ and for all $k \in A \backslash \{c_1, c_2, c_3\}$: $d_{c3} \le d_k$.
2. If c_1, c_2 and c_3 are highly linearly dependent (>0.95), skip DRP procedure.
3. Given the 2D map coordinates $\mathbf{r}_{ci}$ of nodes c_i, find the map coordinates $\mathbf{r}_q$ for the new node q that will yield an optimal solution (Levenberg–Marquardt algorithm is used) for the set of 3 equations of the form $||\mathbf{r}_q - \mathbf{r}_c|| = z \cdot d_c$, for $c \in \{c_1, c_2, c_3\}$.
4. Round $\mathbf{r}_q$ to integers. If this position is already taken, randomly choose one position from among the 8 neighbors in the grid. If all positions are taken, skip the DRP procedure.
5. Connect q with c_i: $C = C \cup \{(c_1, q), (c_2, q), (c_3, q)\}$, and also connect nearest neighbors with each other: $C = C \cup \{(c_1, c_2), (c_1, c_3), (c_2, c_3)\}$.
6. Set $\mathbf{m}_q = \mathbf{x}$.
7. Decrease the strength of all connections in C by constant amount $\Delta c = 1/(100n)$, where n is the current lexicon size. If there are connections whose strength falls below zero, delete them. Division coefficient 100 was found heuristically.

Appendix D. A run-down Pseudocode for DevLex

Initialization

Initialize P-GMAP and S-GMAP with 300 nodes each, randomly scattered over a 50×50 grid. Connect nodes within each map to make a 2D graph. Connect both GMAPs with unidirectional links.

Training

For each stage S
  Get current word meanings to the pool
  Set NBHD radius, $q_{sem}(S)$, $q_{pho}(S)$
  For each substage
    Set SOM mode
    /* within each substage, parameter k linearly runs from 0 to 1 */
    For each iteration
      If $k >$ GMAP-mode-coef(S)
        Set ART mode
      End If
      Choose a word from the pool (represented by vectors $\mathbf{x}_{pho}$, $\mathbf{x}_{sem}$)
      Find best matching units (bmu) in both GMAPs
      If SOM mode
        adapt weights within eighborhood radius in both GMAPs
        adapt associative weights between GMAPs
      End If
      If ART mode
        If $||\mathbf{x}_{pho} - \mathbf{w}_{pho-bmu}|| > q_{pho}(S)$
          Apply DRP procedure for phonological map
        End If
        If $||\mathbf{x}_{sem} - \mathbf{w}_{sem-bmu}|| > q_{sem}(S)$
          Apply DRP procedure for semantic map
        End If
        Decrease the strength of all edges in both GMAPs.
        Remove edges with non-positive strength.
      End If
      Update associative links
    End For
  End For
End For

References

Bates, E. (1999). Plasticity, localization and language development. In S. Broman, & J. M. Fletcher (Eds.), *The changing nervous system: Neurobehavioral consequences of early brain disorders* (pp. 214–253). New York: Oxford University Press.

Bates, E., & Carnevale, G. (1993). New directions in research on language development. *Developmental Review*, *13*, 436–470.

Bates, E., & Elman, J. (1993). Connectionism and the study of change. In M. Johnson (Ed.), *Brain development and cognition: A reader* (pp. 623–642). Oxford: Blackwell Publishers.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P. S., Reznick, J. S., Reilly, J. S., & Hartung, J. P. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*, 85–123.

Blackmore, J., & Miikkulainen, R. (1995). Visualizing high-dimensional structure with the incremental grid growing neural network.

In A. Prieditis, & S. Russell (Eds.), *Machine learning: Proceedings of the 12th international conference (ICML'95, Tahoe City, CA)* (pp. 55–63). San Francisco: Kaufmann.

Bloom, P. (2000). *How children learning the meanings of words*. Cambridge, MA: MIT Press.

Bowerman, M. (1978). Systematizing semantic knowledge: Changes over time in the child's organization of word meaning. *Child Development*, *49*, 977–987.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner, & L. Gleitman (Eds.), *Language acquisition: The state of the art*. Cambridge: Cambridge University Press.

Burgess, C., & Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 1–34.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, *349*, 788–790.

Carpenter, G., & Grossberg, S. (1987). ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics: Special Issue on Neural Networks*, *26*, 4919–4930.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighborhoods of words in young children's lexicons. *Journal of Child Language*, *17*, 205–215.

Chomsky, N. (1975). *Reflections on language*. New York: Parthenon Press.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, and Computers*, *28*, 125–127. CDI is available at http://www.sci.sdsu.edu/cdi/.

Damasio, H., Grabowski, T., Tranel, D., Hichwa, R., & Damasio, A. (1996). A neural basis for lexical retrieval. *Nature*, *380*, 499–505.

Dromi, E. (1987). *Early lexical development*. Cambridge, UK: Cambridge University Press.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification* (2nd ed.). New York: Wiley.

Ellis, A. W., & Lambon-Ralph, M. A. (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.

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 515–523.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Science*, *8*, 301–306.

Elman, J., Bates, A., Johnson, A., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Farkas, I., & Li, P. (2001). A self-organizing neural network model of the acquisition of word meaning. In E. M. Altmann, A. Cleeremans, C. D. Schunn, & W. D. Gray (Eds.), *Proceedings of the fourth international conference on cognitive modeling* (pp. 67–72). Mahwah, NJ: Lawrence Erlbaum.

Farkas, I., & Li, P. (2002a). Modeling the development of lexicon with a growing self-organizing map. In H. J. Caulfield, et al. (Ed.), *Proceedings of the sixth joint conference on information sciences* (pp. 553–556). Durham, NC: JCIS/Association for Intelligent Machinery, Inc.

Farkas, I., & Li, P. (2002b). DevLex: A self-organizing neural network model of the development of lexicon. In L. Wang, et al. (Ed.), *Proceedings of the ninth international conference on neural information processing, #1514, CCE*. Singapore: Nanyang Technology University.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D., & Pethick, S. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59, No. 5, Serial 242*.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Frawley, W. (1992). *Linguistic semantics*. Hillsdale, NJ: Lawrence Erlbaum Associates.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*, 128–135.

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

Gasser, M., & Smith, L. B. (1998). Learning nouns and adjectives: A connectionist account. *Language and Cognitive Processes*, *13*, 269–306.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. A. Kuczaj, *Language development. Language, thought and culture* (Vol. 2) (pp. 301–334). Hillsdale, NJ: Lawrence Erlbaum.

Gerskoff-Stowe, L., & Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, *34*, 37–71.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*, 171–183.

Gould, E., Tanapat, P., Hastings, N. B., & Shors, T. J. (1999). Neurogenesis in adulthood: A possible role in learning. *Trends in Cognitive Sciences*, *3*, 186–191.

Harm, M. (2002). Building large scale distributed semantic feature sets with WordNet. *Technical Report PDP-CNS-02-1*, Carnegie Mellon University.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

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

Kohonen, T. (2001). *The self-organizing maps* (3rd ed.). Berlin: Springer.

Ladefoged, P. (1982). *A course in phonetics*. San Diego, CA: Harcourt Brace.

Li, P. (1999). Generalization, representation, and recovery in a self-organizing feature-map model of language acquisition. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the twenty-first annual conference of the cognitive science society* (pp. 308–313). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (2000). The acquisition of lexical and grammatical aspect in a self-organizing feature map model. In L. R. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the twenty-second annual conference of the cognitive science society* (pp. 304–309). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (2003). Language acquisition in a self-organizing neural network model. In P. Quinlan (Ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp. 115–149). New York: Psychology Press.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In E. V. Clark (Ed.), *Proceedings of the thirtieth stanford child language research forum* (pp. 167–178). Stanford, CA: Center for the Study of Language and Information.

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia, & J. Altarriba (Eds.), *Bilingual sentence processing* (pp. 59–85). North-Holland: Elsevier Science.

Li, P., Jin, Z., & Tan, L. (2004). Neural representations of nouns and verbs in Chinese: an fMRI study. *NeuroImage*, *21*, 1533–1541.

Li, P., & MacWhinney, B. (1996). Cryptotype, overgeneralization, and competition: a connectionist model of the learning of English reversive prefixes. *Connection Science*, *8*, 3–30.

Li, P., & MacWhinney, B. (2002). PatPho: a phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers*, *34*, 408–415.

Li, P., & Zhao, X. (2004). From avalanche to vocabulary spurt: Dynamics in self-organization and children's word learning. Manuscript under review.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*, 199–227.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.

MacWhinney, B. (2001a). Lexicalist connectionism. In P. Broeder, & J. M. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches* (pp. 9–32). Oxford, UK: Oxford University Press.

MacWhinney, B. (2001b). Language emergence. In P. Burmeister, P. Thorsten, & A. Rohde (Eds.), *An integrated view of language development* (pp. 17–42). Trier: Wissenschaftlicher Verlag.

Marsland, S., Shapiro, J., & Nehmzow, U. (2002). A self-organizing network that grows when required. *Neural Networks*, *15*, 1041–1058.

Menn, L., & Stoel-Gammon, C. (1995). Phonological development. In P. Fletcher, & B. MacWhinney (Eds.), *Handbook of child language* (pp. 335–359). Oxford: Blackwell.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. Cambridge, MA: MIT Press.

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

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235–312.

Morrison, C. M., & Ellis, A. W. (1995). The roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 116–133.

Pinker, S. (1994). *The language instinct: How the mind creates language*. Boulder, CO: William Morrow and Co.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, *4*, 293–312.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*, 253–336.

Quartz, S. R., & Sejnowski, T. J. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences*, *20*, 537–596.

Quinlan, P. (2003). Modeling human development: In brief. In P. Quinlan (Ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp. 1–12). New York: Psychology Press.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, *61*, 241–254.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland, D. Rumelhart, & PDP research group, *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. II) (pp. 216–271). Cambridge, MA: MIT Press.

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

Smith, M. A., Cottrell, G. W., & Anderson, K. (2001). The early word catches the weights. In T. K. Leen, T. G. Dietterich, & V. Tresp, *Advances in Neural Information Processing Systems* (13) (pp. 52–58). Cambridge, MA: MIT Press.

Spitzer, M. (1999). *The mind within the net: Models of learning, thinking, and acting*. Cambridge, MA: MIT Press.