Neural Network World, 13(3), 235-245, 2003

# Lexical acquisition and developing semantic map

Igor Farkaš*

**Abstract**

In this paper, we describe a self-organizing neural network model that addresses the process of early lexical acquisition in young children. The growing lexicon is modeled by combined semantic word representations based on distributional statistics of words and on grounded semantic features of words. Changing semantic word representations are assumed to model the maturation of word meaning and serve as inputs to the growing semantic map. The model has been tested on a real child-directed parental language corpus and as a result, the map demonstrates the emergence and reorganization of various word categories, as quantified by two measures.

## 1 Introduction

Language acquisition as a cognitive process is driven by various maturational factors and starts with acquisition of individual words [1]. The child's vocabulary gradually grows as the child learns and later "fine-tunes" the meaning of various words, and then starts to actively use them. Computationally, the incremental lexical acquisition can be viewed as a dynamic process incorporating changing patterns – semantic word representations, that can be thought of as an analogue to child's mental word representations. To account for the process of conceptual maturation and reorganization taking place in the child's mind/brain, the model must also show some structure early on and undergo various stages of development.

In our model, semantic word representations are computed as vectors of word co-occurrences within the current lexicon, combined with semantic features based on WordNet lexical database [2]. Several authors have already shown that distributional statistics of words provide a considerable amount of information about word meaning, as could be seen in connectionist learning models [3] and statistical models [4, 5, 6]. As a matter of fact, the distributional approach did not emerge only recently: structural linguists had already applied this approach almost a century ago (e.g., [7]) but were later suppressed by the generative linguistics paradigm [8]. However, probabilistic approach to language learning has been revived recently [9], since it was shown to account for various psycholinguistic phenomena observed in human learners.

On the other hand, these probabilistic or word cooccurrence models remain the subject to criticism as being cognitively implausible, because the generated

*This work was done while the author was at Department of Psychology, University of Richmond, USA, on leave from Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia.

word representations lack grounding in the real world. Therefore, we combined the (changing) word distributional information with (unchanging) WordNet-based semantic features of words [10]. In addition, this combination enhanced clustering properties of word categories.

Semantic word representations described above serve as an input to the growing semantic map (GSM) in our model. The motivation for mapping word categories comes from its neurobiological plausibility: Cognitive neuroscientists have discovered that various areas of the brain respond differentially to different categories. These "brain centers" are found to be highly specific linguistically. For example, nouns and verbs elicit responses in different areas of the brain, as do concrete words versus abstract words, content words versus function words, and words for animals, persons, and tools (see, e.g., [11] and references therein).

In case of a growing lexicon we are facing a problem of dynamic input data. A number of growing self-organizing neural network models have been previously proposed to cope with this task. However, most of these models have arbitrary dimensionality and connectivity (see, e.g., [12] and references therein) which makes them difficult to visualize data in two dimensions. The IGG model [13] overcomes this difficulty in that it preserves a strictly 2D topology. Similarly to IGG, the recruitment of new nodes in GSM is restricted to the grid positions. Unlike IGG, however, new nodes in GSM are recruited around the existing nodes instead of the perimeter of the grid.

Using flexible rather than fixed architecture was also motivated by the fact that young mammalian brains undergo the process of synaptogenesis (sprouting connections between neurons), which in language-related areas may also be associated with early vocabulary growth (see [14] and references therein). Hence, by recruiting we mean that new nodes may not be physically added, but they become actively incorporated by sprouting their connections to neighboring nodes and allowing all inputs to connect to them.

## 2 The model

Our GSM model is a part of DevLex model [15] of lexical acquisition and is very similar to DISLEX model [16] that was recently used to model the disordered lexicon in patients. DevLex is also based on two self-organizing maps (SOM, [17]) – a semantic map and a phonological map – linked together with associative links, in order to allow for modeling word comprehension and word production. However, DevLex differs from DISLEX in that GSM part of DevLex has a dynamic architecture. In this paper, we focus on GSM and the process of emergence of semantic word categories. GSM is based on a SOM that has several appealing properties making it suitable as a model of the human lexical system [18, 16]. GSM uses SOM learning algorithm, which allows the formation of a topological map as a consequence of decreasing values of model parameters (learning rate and radius of winner neighborhood). As a specific feature of GSM, neighborhood is adjusted non-monotonically, to allow for reorganization of the growing lexicon (see below).

**Input data.** In our previous work [19], we have described GSM as a semantic memory of the growing vocabulary (see Figure 1). GSM self-organizes on word

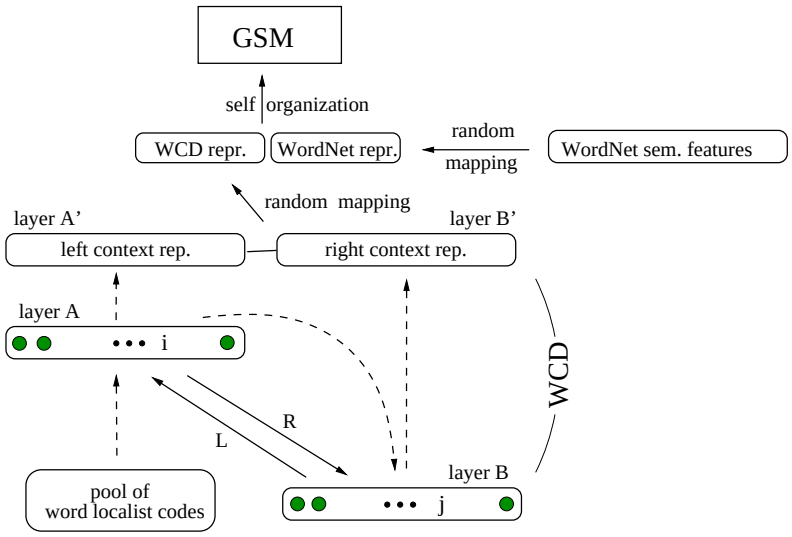


Figure 1: A diagrammatic sketch of the GSM model. The bottom half represents WCD, and the upper part represents the random mapping stage followed by GSM's self-organization. The solid links between layers of units represent activity propagation (via full connectivity weights), and dotted lines represent pattern transport (via one-to-one links).

vectors, generated off-line by word co-occurrence detector (WCD). Being a special recurrent network, WCD parses the raw input text on a word-by-word basis and transformes the local word representations (on layers A and B) to distributed representations (on layers A' and B'). It does so by learning the transition probabilities, stored in weight matrices $\mathbf{L}$ and $\mathbf{R}$ for left and right contexts respectively, for all words $i = 1, ..., n$ in the considered vocabulary (of size $n$). Word representations, represented by weights, are transferred to output unit activations (in layers A' and B') by a control mechanism described in [19]. Transitions to and from all unknown words (i.e., those not from the current lexicon) are ignored. With maximal vocabulary size denoted by $N$, the resulting word representations consist of vectors $\mathbf{q}_i \in \mathcal{R}^{2N}$ (whose last $2(N-n)$ components are zero).

Although $n$ increases (and with it, the number of potentially nonzero components), we keep the dimension of word representations constant by projecting word vectors with fixed *random mapping* matrix $\mathbf{Z}$ (of type $D \times 2N$) down to $D$ dimensions [18]). Matrix $\mathbf{Z}$ has normalized Euclidean length of columns and is not subject to adaptation. Random mapping method has a nice feature that it preserves data structure quite accurately as long as the output dimension is not too low. $D_{wcd} = 100$ we used is practically high enough. Resulting WCD word representations were then obtained as $\tilde{\mathbf{q}}_\mathbf{i} = \mathbf{Z}\mathbf{q}_i$.

WordNet-based features were obtained using a feature generation software [10] that can produce a set of binary features for words. The 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 [20, 2]. Harm extracted from WordNet relevant semantic features for nouns and verbs, but for adjectives he hand-coded the semantic features according to a taxonomy of features given in [21]. For our lexicon, the above method yielded in total a list of 459 binary features, with the number of features for any given word ranging from 1 to 12. Unlike WCD-based features, WordNet-based features are static, but for computational uniformity they were also submitted to dimension-reducing random mapping ($D_{wn} = 100$). For each word, both vectors were concatenated and formed inputs to GSM ($D = D_{wcd} + D_{wn}$).

Before the actual simulation, we needed to create the training data. In our approach, incremental learning of lexicon can be viewed as learning the data set whose patterns gradually change (due to expanding context) and whose number of patterns changes as well (due to expanding lexicon). Therefore, we first chunked the the total lexicon of $N$ words into $m$ stages with increasing increasing vocabulary size ($n_1 = N/m, n_2 = 2N/m, ..., n_m = N$). For each stage, we ran WCD for the current lexicon. That is, all other words (with indices $n_j + 1, ..., N$) were treated as noise and their presence in the data stream was not reflected in the co-occurrence matrix.

**Growing semantic map.** GSM was trained using the above data sets, one at a time. In each stage, input words were selected according the actual word token frequency, as found in the corpus. More precisely, since the word frequency distribution follows the Zipf's law even with relatively small lexicons (i.e., a few hundred words), we took the logarithm of these frequencies in order to force a more even probability of word distribution.

In each simulation, GSM was initialized with a subset of recruited nodes randomly scattered on a rectangular grid. Each node was connected with its nearest neigbors (in the map) to form a 2D structure. New nodes could be recruited in yet unoccupied grid positions. During simulation, best-matching nodes (winners) were repeatedly found for all words from the current vocabulary. The information about the number of word labels was utilized in identifying nodes with many labels (ambiguous nodes). Next to the most ambiguous node, a randomly selected neighbor became recruited (by becoming connected with its nearest existing neighbors in the map) in order to allow for a more even distribution of words onto these nodes.

We tried to use as few model parameters as possible, yet allowing the network to copy with stability-plasticity trade-off. The learning rate was kept constant throughout the simulation. The neighborhood radius was modulated in a "see-saw" fashion but its profile within each growth stage was identical: radius was always enlarged at the beginning of a growth stage and linearly decreased toward the end of it. This profile was designed to allow for greater plasticity of the map (ability to reorganize) upon seeing new words, followed by the gradual settling of a reorganized structure.

**Quantifying GSM's organization.** To monitor the development of categories in the map, we used two quantitative measures: one aimed at quantifying the map order with respect to category clusters, and the other aimed at quantifying the dynamics of change in GSM with respect to word categories. For the first measure, we used a simple, k-nearest neighbor (k-NN) classifier [22] to monitor

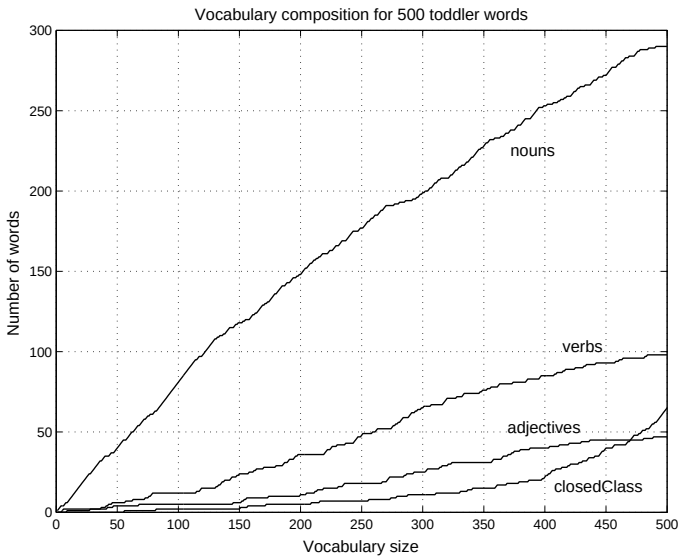


Figure 2: Vocabulary growth according to CDI order of acquisition for the four major categories.

the existence of compact category clusters in GSM. For classification of each word, the k-NN classifier was built based on all the remaining words in the considered lexicon. The label of the tested word was predicted according to the most frequent label among the $k$ nearest neighbors in the map. Ties in prediction were broken randomly. The classification rate for each word category was evaluated at the end of each growth stage.

For the second measure, we computed the amount of the map's reorganization at the end of each stage relative to the previous stage. To do this, we compared the word coordinates in the two maps at two successive stages within the same simulation. The shift of each word (as determined by its winner) was taken as the Euclidean distance of its coordinates in the two maps. Since any two maps at successive stages differ in the number of words they represent (each previous map has fewer words than the subsequent map), our comparison could only be made for words that are common in the two maps. As a reorganization measure for each category, the average word shifts were taken.

# 3 Experiment

We tested GSM on the CHILDES parental child-directed corpus [23]. We focused on a set of $N = 500$ words that are reported to be among the first ones acquired by children, according to the MacArthur Communicative Development Inventory (CDI) [24]. (CDI contains 680 words for toddlers: we excluded the homographs, word phrases and onomatopoeias.) The semantic representations for words in the lexicon were created separately for different vocabulary sizes using WCD (from 50 through 500 words, in increments of 50) yielding 10 input data sets.

From original 22 categories in CDI, we extracted 18 word categories (see [25] for rationale) that were merged into 4 major categories: (1) *nouns* (including animals, body parts, clothing, food, household items, outside things, people, rooms, toys, and vehicles), (2) *verbs*, (3) *adjectives*, and (4) *closed-class words* (including auxiliary verbs, connecting words, prepositions, pronouns, quantifiers, and question words).

The lexical composition in our simulations changed as a function of vocabulary growth (Figure 2). Nouns begin to be acquired very early, whereas closed-class words much later. Verbs and adjectives both grow steadily and their proportions remain roughly the same across words.

We ran 6 simulations with the same network and parameters. The GSM was initialized with 1,200 recruited nodes that were randomly selected over the $40 \times 50$ rectangular grid, and had randomized weight vectors. New nodes could be recruited around the "busiest" nodes every 500 iterations; hence, GSM typically ended up having around 1,500 nodes. Each stage lasted 50,000 iterations, thus amounting to 500,000 iterations per simulation. Learning rate was set to 0.02, neighborhood radius decreased from 12 to 3 within each of the 10 stages.

Figure 3 shows the classification rates of a 5-NN classifier, averaged over 6 different simulation runs, for each of the four grammatical categories.[1] In all cases except nouns, the average classification rate grows steadily toward the end. The verbs and adjectives display similar profile: both categories developed quite early during the acquisition process and formed a steady compact cluster (final classification rate 85 and 82%, respectively).[2] Closed-class words as a cluster emerged later, with similar classification rate at the end (82%). Figure 3 also suggests that nouns had formed a cluster early on and maintained their cluster-like structure throughout development (constantly above 90%). This overall high performance for nouns results from the vocabulary composition that has a strong noun bias early on – there are more nouns than all other words at any given stage. Therefore, we broke down the two major categories (into 10 noun categories and 6 closed-class categories) to see the original categories as well. As can be seen in Figure 4, even at this more detailed level at least some of the noun categories (animals, body, clothing, food, people and rooms) formed the clusters gradually, as reflected in the gradual increase of their classification rates. Within closed-class words, only prepositions and pronouns formed some cluster-like structures, albeit the former with a high degree of variability across simulations.

As a next step, we computed the amount of reorganization in the map during development. Figure 5 shows that the amount of reorganization in each category gradually decreases with the growing lexicon. Note that for the adjectives and closed-class words there were few items as input at the beginning stages (see Fig-

[1] We were aware of a potential flaw using a k-NN classifier: if a particular category consisted of separate compact clusters scattered all over the map, each of size $k + 1$, the k-NN classifier would still yield a very high accuracy for that category. However, based on visual inspection of the maps, this did not happen in most cases. We also tried higher values of $k$, but the results did not change drammatically.

[2] Earlier comparisons showed, that mainly adjectives and nouns benefited from incorporation of WordNet-based features. With only WCD-based representations used, the accuracy of adjectives would not exceed 50%, in case of nouns 70%.

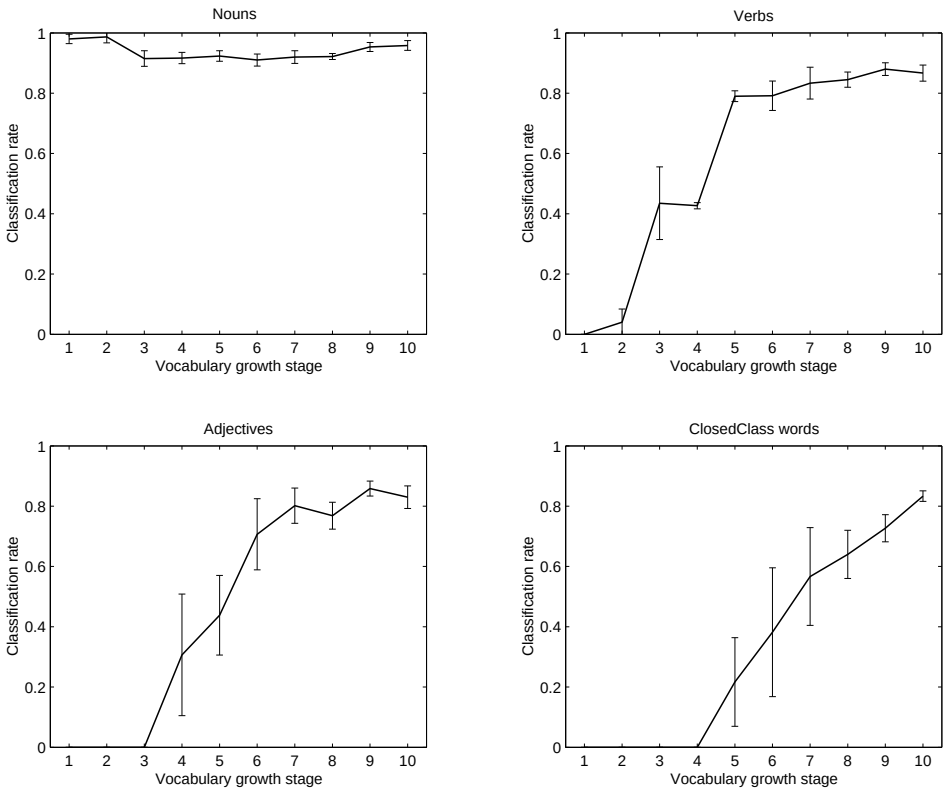


Figure 3: Classification rate with a 5-nearest-neighbor classifier as a function of vocabulary growth in GSM. The graphs present the means and standard errors from six simulations (the abscissa represents the 10 stages of growth, from 50 to 500 words).

ure 1), so reorganization does not show up initially or has a very high variance.[3] In general, these graphs indicate that, despite constant profiles of learning parameters, there is a tendency for the map to reduce the overall amount of reorganization as learning progresses. As the accuracy of the WCD-based input representation increases over time, the output space self-organizes to form an increasingly clear map of the input data structure.

Visual inspection of the map at various stages revealed that the typical development looks as follows. Nouns tend to occupy the whole map early on, because they come first and the map offers them almost all its resources. Later on, the nouns will give way to words in other categories, as they start to grow as well. Around mid-course of development, the basic organization for the major categories is already established, so that at later stages of growth there would be no radi-

[3]The noticable bump in case of nouns between stages 7 and 8 is a consequence of the WCD method and resulted from entry of the determiner "the" in the lexicon. Since this word has a very high frequency in the corpus and precedes a lof of nouns, representations of many nouns that take this determiner noticably changed.

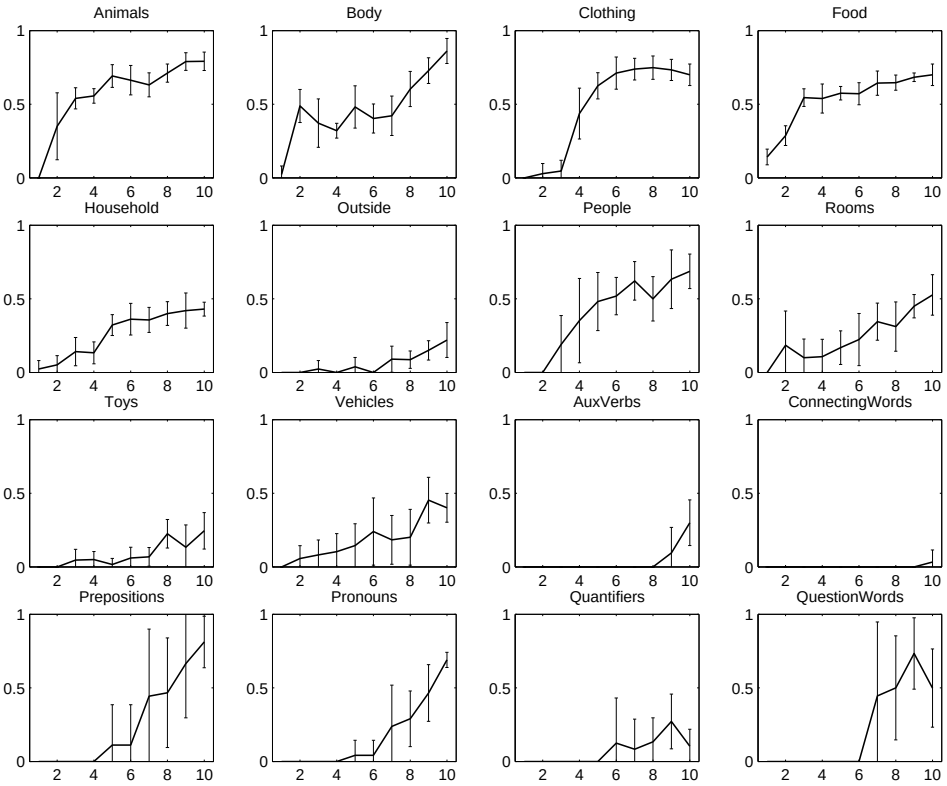


Figure 4: Classification rate with a 5-nearest-neighbor classifier as a function of vocabulary growth in GSM for 10 noun categories and 6 closed class word categories. The graphs present the means and standard errors from six simulations (the abscissa represents the 10 stages of growth, from 50 to 500 words).

cal reorganization of the major categories. The later stages mainly involve the addition of new words to established existing categories, with gradual shifting of boundaries between them, which is consistent with Figure 5. This developmental profile is consistent with recent findings [26] showing that children and adults have different patterns of neural activities in language processing. For children, the activation pattern is diffuse and unfocused, whereas for adults, the activation pattern is focused and dedicated to specific frontal regions. Our model shows that early category structures (especially nouns) are diffuse and distributed, whereas later on they become focused and localized.

# 4 Conclusion

Our attempt to provide computational account for the lexical acquisition incorporates two processes: (1) maturation of semantic representations of words, and (2) emergence and reorganization of word categories in the map. Regarding the

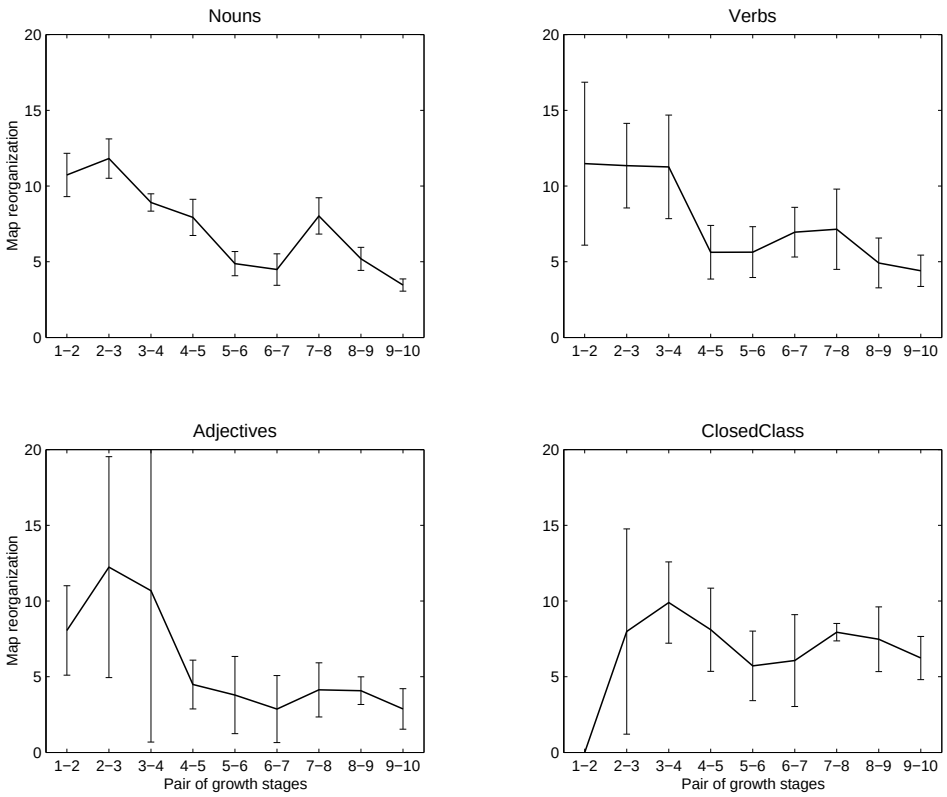


Figure 5: Map reorganization as a function of growth, computed as the amount of word shifts in the maps. Two maps are compared at a time for all words common to two successive stages. The graphs present the means and standard errors from six simulations for each pair of comparison stages.

first process, we emphasize the word distributional approach in acquisition (but we also incorporated grounded semantic features) and here, the enriching word context serves as the only maturational factor (because WordNet features are constant). As for the second process, the GSM's learning characteristics are constant with respect to each growth stage (learning rate is held constant as well as the neighborhood profile at each stage). Hence, the developmental profile observed in the map is purely input-driven. It is possible that other factors, operating either at the input level (e.g., "maturation" of grounded semantic features) or in the map (such as variations in learning rates or active neighborhoods) play a role in lexical acquisition in human, but these factors have not been examined yet.

Despite its simplified design, the GSM model can tackle the stability-plasticity trade-off: to preserve the existing structure upon seeing new words, but still being able to learn. Catastrophic interference as a common problem in neural network models trained on changing data [27] is avoided, because lexical acquisition assumes an incremental training set: new words add on rather than replace old words in the input data set.

We believe that the GSM model could be viewed as an approach attempting to provide computational account for emergence and reorganization of semantic categories in the brain "from the scratch", i.e., without having to rely on any innate predispositions on representational level [28]. Word categories are sufficiently different in the input space to allow our model to capture these differences effectively in acquiring a dynamic representation of the lexicon in the form of semantic map.

# References


[1] E.V. Clark. *The Lexicon in Acquisition*. Cambridge University Press, 1993.

[2] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[3] J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[4] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[5] C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional semantic space. *Language and Cognitive Processes*, 12(2/3):177–210, 1997.

[6] M. Redington, N. Chater, and S. Finch. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22:425–470, 1998.

[7] F. de Saussure. *Cours de linguistique générale*. Payot, Paris, 1916. (English translation: *A course in general linguistics*. New York: Philosophical Library).

[8] N. Chomsky. *Syntactic Structures*. Mouton, Hague, 1957.

[9] M.S. Seidenberg. Language acquisition and use: Learning and applying probabilistic constraints. *Science*, 275:1599–1603, 1997.

[10] M.W. Harm. Building large scale distributed semantic feature sets with WordNet. Technical Report PDP.CNS.02.1, Carnegie Mellon University, Pittsburg, PA, 2002.

[11] F. Pulvermüller. Words in the brains language. *Behavioral and Brain Sciences*, 22:253–336, 1999.

[12] F.H. Hamker. Life-long learning cell structures – continuously learning without catastrophic interference. *Neural Networks*, 14(4/5):551–573, 2001.

[13] J. Blackmore and R. Miikkulainen. Visualizing high-dimensional structure with the incremental grid growing neural network. In *Machine Learning: Proceedings of the 12th International Conference*, pages 55–63, 1995.

[14] S.R. Quartz and T.J. Sejnowski. The neural basis of cognitive development. *Behavioral and Brain Sciences*, 20:537–596, 1997.

[15] I. Farkaš and P. Li. Devlex: A self-organizing neural network model of the development of lexicon. In *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'2002)*, 2002.

[16] R. Miikkulainen. Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language*, 57:334–366, 1997.

[17] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480, 1990.

[18] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.

[19] I. Farkaš and P. Li. Modeling the development of lexicon with a growing self-organizing map. In H.J. Caulfield et al, editor, *Proceedings of the 6th Joint Conference on Information Sciences*, pages 553–556, Research Triangle Park, NC, 2002.

[20] G.A. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, pages 235–312, 1990.

[21] W. Fawley. *Linguistic Semantics.* Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1992.

[22] R. Duda, P. Hart, and D. Stork. *Pattern classification.* John Wiley and Sons, 2nd edition, 2000.

[23] P. Li, C. Burgess, and K. Lund. The acquisition of word meaning through global lexical co-occurrences. In E. Clark, editor, *Proceedings of the 30th Child Language Research Forum*, pages 167–178. Stanford, CA: Center for the Study of Language and Information, 2000.

[24] P.S. Dale and L. Fenson. Lexical development norms for young children. *Behavior Research Methods, Instruments and Computers*, 28:125–127, 1996.

[25] E. Bates et al. Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, 21:85–123, 1994.

[26] B. Schlaggar, T. Brown, H. Lugar, K. Visscher, F. Miezin, and S. Petersen. Functional neuroanatomical differences between adults and school-age children in the processing of single words. *Science*, 296:1476–1479, 2002.

[27] R.M. French. Catastrophic interference in connectionist networks: Causes, consequences and solutions. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

[28] J. Elman, E. Bates, A. Johnson, A. Karmiloff-Smith, D. Parisi, and D. Plunkett. *Rethinking innateness: A connectionist perspective on development.* MIT Press, Cambridge, MA, 1996.