DEVLEX: A SELF-ORGANIZING NEURAL NETWORK MODEL OF THE DEVELOPMENT OF LEXICON

Igor Farkaš & Ping Li

Department of Psychology
University of Richmond
Richmond, VA 23173, USA

ABSTRACT

In this paper we present the DevLex model of language acquisition. DevLex consists of two self-organizing maps (a growing semantic map and a phonological map) that are connected via associative links. It simulates the early stages of lexical development in children, in particular, word confusion as evidenced in naming errors. The simulation results indicate that the rate of word confusion is modulated by developmental profile of vocabulary increase, word density of competing neighbors, and rate of lexical growth. These results match up with hypotheses from empirical research on lexical development.

1. INTRODUCTION

Connectionist modeling of language learning has made significant progresses since Rumelhart and McClelland's pioneering model [1] of the acquisition of the English past tense. However, two major limitations need to be considered for the further development of neural network models of language acquisition. First, most current models have used artificially generated input representations that are in many cases isolated from realistic language uses. In addition, these input representations are often "handcrafted" by the modeler and limited to small sets of vocabulary. Second, most previous models have used supervised learning, in particular, the back-propagation learning algorithm as their basis of network training. Although these types of networks have demonstrated success, there are serious problems concerning their biological and psychological plausibility, especially in the language learning context (see [2, 3] for arguments).

In this study, we present DevLex, a self-organizing neural network of the development of lexicon, in an attempt to overcome the limitations associated with current models. DevLex relies on corpus-based speech data to establish the sequence as well as the structure of input, using phonological and semantic representations that more closely approximate the reality of language use. Components of DevLex have been applied successfully to model the acquisition of semantics and morphology by children and bilingual learners [4, 5, 2, 6].

In the empirical literature, it has been observed that children experience a "vocabulary spurt", a sudden and rapid increase in the rate at which new words are learned (typically when the child's vocabulary reaches 150 words; [7]). Associated with this vocabulary spurt is a brief period of confusion on the use of some words, often a "naming deficit", whereby the child calls an object by the wrong name [8]. There are various explanations of children's lexical confusions (see Discussion); a prominent argument is that word confusion may be due to semantic reorganization, a process in which the child starts to recognize the shared meanings of words but not their subtle differences [9].

In this paper, we simulate early stages of lexical development with an incremental vocabulary growth profile, in order to provide insights into the mechanisms that lead to children's word confusions at various developmental stages.

2. THE MODEL

DevLex consists of two main parts: the growing semantic map (GSM) and the phonological map (PMAP) that are connected with associative pathways (Figure 1).

In our previous work [5], we have described GSM in more detail as a semantic memory of the growing vocabulary. GSM self-organizes on word vectors, generated offline 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 to distributed representations. It does so by learning the transition probabilities (for left and right contexts) for all words $i = 1, ..., n$ in the considered vocabulary (of size n). Word representations, acquired by WCD weights, are transferred to output unit activations by a control mechanism described in [5]. Transitions to and from all unknown

This research is supported by NSF (grant #BCS-9975249 to P.L.). I.F. is also with the Slovak Academy of Sciences, Bratislava, Slovakia.

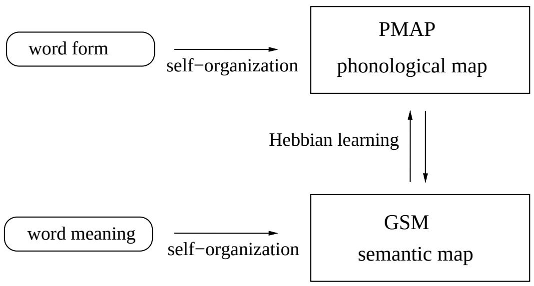

Figure 1: A sketch of the DevLex model of lexical acquisition.

words (i.e., those not from the 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, we keep the word dimension constant by projecting word vectors with fixed *random mapping* matrix **Z** (of type $D \times 2N$) down to D dimensions (we used $D = 100$) while aproximately preserving the data structure (for an underlying mathematical rationale of the method see [10]). Matrix **Z** has normalized Euclidean length of columns and is not subject to adaptation. Resulting word representations can thus directly be obtained as $\tilde{\mathbf{q}}_\mathbf{i} = \mathbf{Z}\mathbf{q}_i \in \mathcal{R}^{\mathcal{D}}$.

PMAP is a memory of the associated phonological word symbols that were created with the PatPho generator [11]. PatPho fits every word (max. 3-syllables) onto a template according to its vowel-consonant structure. We used the left-justified template with binary encoding, reduced to 54 dimensions by PCA.

2.1. Learning

Learning in DevLex is split into two major phases: (1) initialization, and (2) learning of the growing lexicon. During initialization, several steps are performed. PMAP is pretrained on the whole lexicon (550 words; see Section 3) and then kept constant. The working hypothesis behind this simplification is that children learn phonological forms much faster than semantic representations of words. In other words, given the limited repertoire of phonemes in a language, acquisition of phonological structure is considerably easier than acquisition of other linguistic components.

In contrast, GSM is pretrained on a subset of the lexicon (100 most frequent words), and the links between the two maps are learned, too. GSM is initialized with a subset of recruited units scattered randomly across the underlying rectangular grid, and connected to form a 2D structure. New units are then recruited (get connected) in the areas with the highest lexical density. The initialization of GSM serves to capture the regularities of the semantic space (initial neighborhood is large), to which new words are added during lexical growth. In order to preserve the existing map structure, the weight update during growth is local (smaller neighborhood radius at all times).

In both phases, the single iteration follows the same steps (though some learning parameters are different). For each semantic-phonological representation pair selected from the current pool, three calculations are performed: (1) map responses are computed and the winner in each map is identified; (2) weight vectors of GSM units in the winner's neighborhood are updated, and (3) the associative links connecting the winner neighborhoods are updated.

The localized output response of a unit k is computed as

$$a_k = \begin{cases} 1 - \frac{\|\mathbf{x}-\mathbf{m}_k\| - d_{\min}}{d_{\max} - d_{\min}}, & \text{if } k \in \mathcal{N}_c, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbf{x}$ is a semantic or phonological word representation vector, $\mathbf{m}_k$ the weight vector of the unit k, $\mathcal{N}_c$ the neighborhood around the winner c, $d_{\min}$ and $d_{\max}$ the smallest and largest distance of $\mathbf{x}$ to a unit's weight vector in the considered neighborhood. Weights of GSM units around the winner are updated as

$$\Delta\mathbf{m}_k(t) = \alpha_{\text{sem}}(t)\ [\tilde{\mathbf{q}}_\mathbf{i} - \mathbf{m}_k(t)], \quad \text{if } k \ \in \mathcal{N}_c.$$

The associative weights between active units in both maps are then increased proportional to their activity, using Hebbian learning

$$\Delta w_{kl} = \alpha_{\text{assoc}}(t) a_k^S a_l^D$$

where w_{kl} is the unidirectional weight leading from unit k in the source map to unit l in the destination map, and a_k^S and a_l^D are the associated unit activations in the source map and destination maps, respectively. As is common with Hebbian learning, the associative weight vectors are then normalized:

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

At every iteration, Hebbian learning is applied to units situated around the winners in both maps. As the unit neighborhood radius shrinks, it decreases the number of units whose associative links are updated. The smaller the neighborhood taken, the more focused the update is.

Due to associative links, the activity in the source map is propagated (translated) to the destination map:

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

where the activation function $g(y) = y/y_{max}$ scales down the activations in the destination map linearly within 0 and 1.

In terms of activity propagation and learning, DevLex is similar to the DISLEX model [12]. However, it differs from DISLEX in the word representations used, as well as the flexible architecture of GSM that allows to learn a growing lexicon.

2.2. Learning of the growing lexicon

During growth, words are extracted from the pool, according to their frequency in the parental CHILDES corpus [13, 14]. More precisely, since the word frequency distribution follows the Zipf's law even with as few as 550 words, we take the logarithm of these frequencies in order to force a more even distribution of words and thus a more balanced distribution of units in the map. Learning is *cumulative*, i.e., new words are added to the existing pool, rather than replacing old words. This scenario of learning avoids catastrophic interference and matches more closely with vocabulary growth in children.

The above learning scheme is enhanced by what we call *focused learning*, during which only the confused words in GSM (i.e., those mapped to the same unit) can be selected from the pool for further learning. Whenever turned on, this mechanism always helps to decrease word confusion rate.

3. EXPERIMENTS

We tested DevLex on the CHILDES parental corpus [14]. We focused on a set of 550 words that are reported to be among the first ones acquired by children, according to the MacArthur Communicative Development Inventory (CDI) [15]. (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 (from 100 through 550 words), yielding 10 sets of input data (in increments of 50 words).

The CDI words fall into 22 categories, which we merged into 4 major categories: (1) *verbs*, (2) *closed-class* words (including auxiliary verbs, connecting words, prepositions, pronouns, quantifiers, and question words) (3) *adjectives*, and (4) *nouns* (including animals, body parts, clothing, food, games and routines, household items, outside things, people, places to go, rooms, toys, and vehicles). The lexical composition in our simulations changed as a function of vocabulary growth. Figure 2 shows the lexical composition for each of the 10 data sets and Figure 3 the relative proportions of lexical composition. Comparison of the two figures shows that nouns increase in number both absolutely and relatively, especially towards the later stages. The closed-class words have the opposite pattern: most of them come at the very early stages. Verbs and adjectives both grow steadily and their proportions remain roughly the same across words.

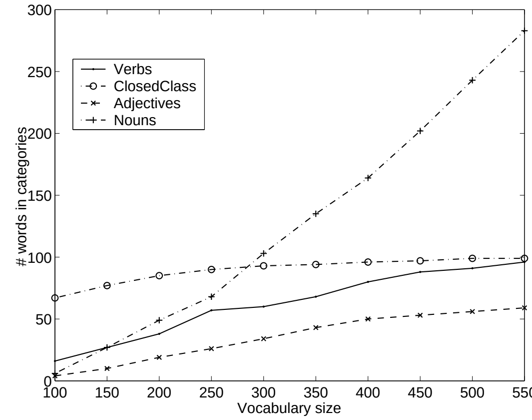

Figure 2: Absolute number of words falling into four word categories as a function of vocabulary size.

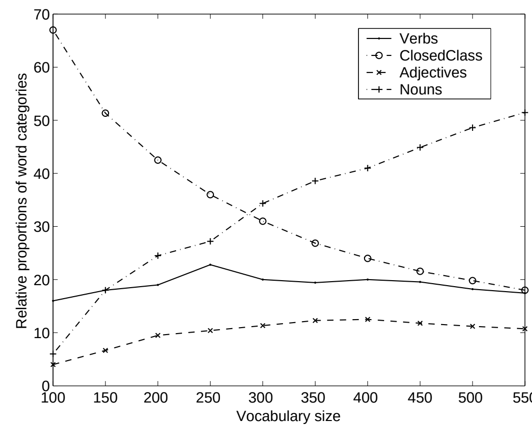

Figure 3: Lexical composition as a function of vocabulary size.

After intialization with 100 most frequent words in GSM, new words were added to the lexicon until all 550 words are seen. The whole simulation thus consists of 9 growth stages associated with corresponding data sets. Each stage included 50 substages of one-word growth, thus ensuring incremental growth. In all simulations, the rate of vocabulary growth was kept constant (i.e., each stage had the same amount of training). Initially, each GSM contained 1500 recruited units scaterred randomly over the 50×60 grid and as a result of growth, it typically ended up having around 2000 units.

3.1. Word confusion during vocabulary growth

The first set of simulations focused on modeling how vocabulary growth affects word confusion in GSM and in word production (via activity propagation from GSM to PMAP). Figures 4 and 5 display the confusion rates, averaged over

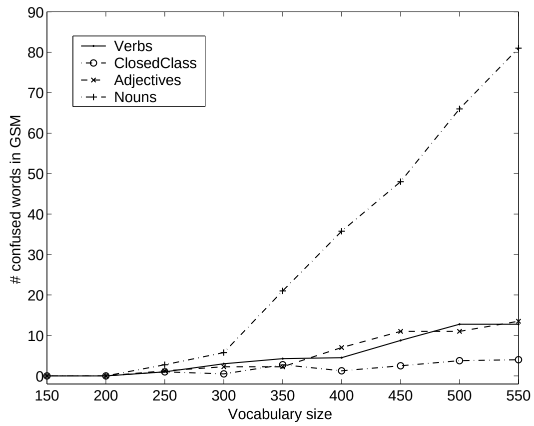

Figure 4: Number of words confused in GSM as a function of vocabulary size.

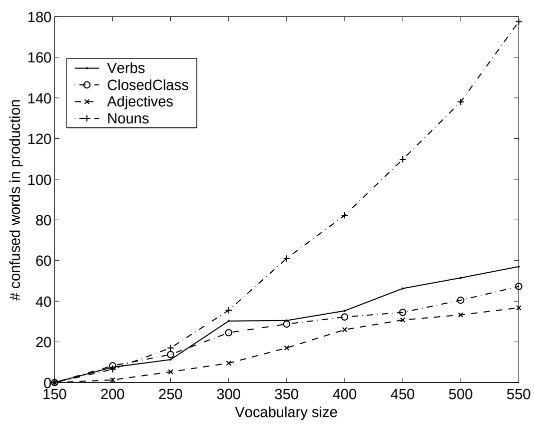

Figure 5: Number of words confused in production as a function of vocabulary size.

5 simulations (all having 50000 iterations per stage), with varying proportions of focused learning (20-80%). Word confusion in GSM was evaluated as the number of words that are represented by the same unit in GSM. Confusion in production was calculated as the number of words for which there was a mismatch between the eliciting semantic representation of the word and the evoked phonological form. It can be seen that the confusion rates in production are considerably higher (twice as high for nouns) than those within GSM. This may be due to confusion within GSM itself (that typically evokes at least two active units in PMAP) and to inaccurate associative links.

We have two interesting observations to make on these results. First, the number and the profile of the confused words are closely related to growth profile of the four categories (cf. Figure 2). The majority of the confused words are nouns because they undergo the most significant growth toward the end, and words at later stages have low token

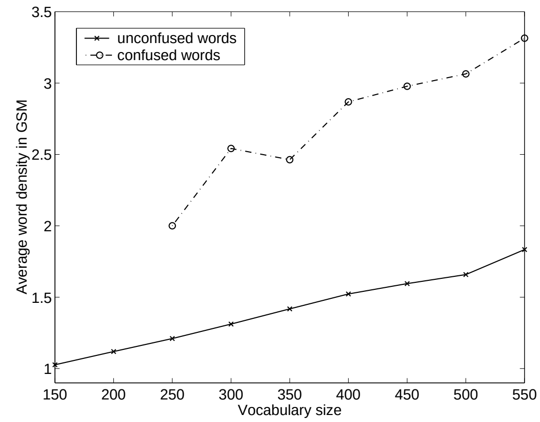

Figure 6: Average density of confused words compared to that of unconfused words. No confusion occurred before 250 words.

frequency. With more words in the pool, GSM is becoming more dense in the “noun” area which leads to higher probability of confusion. In contrast, closed-class words are less confused in GSM, because most of them appear in the initial data set (cf. Figure 2) and they tend to have high frequency of use. Therefore, they become robustly represented in GSM without being disrupted by later words.

Second, confused words mostly come from map areas with higher density, as shown in Figure 6. The density for every word was computed as the total number of words (including the target word) that were represented by the winning unit and all its nearest neigbors. This observation is quite consistent with previous predictions [8] that higher rate of naming errors is associated with high lexical density areas than with sparsely populated areas. Figure 6 also shows that the confusion rates grow with a rapidly expanding lexicon which is observed in children at an early stage of their lexical development [8], or the ”vocabulary spurt” stage.

We also observed that most of the confused words were related to each other. When relatedness was considered with respect to the 4 major categories, the number of related confused words at the end of growth was in the range of 92-100%. When relatedness was considering with respect to the original 21 CDI categories, the related confused words accounted for 53-62% of the whole vocabulary. This confirms the emergence of (though not perfectly) structured word representations in GSM and agrees with our above analysis that densely populated areas are most error-prone.

3.2. Effect of rate of vocabulary growth on word confusion

During growth, the GSM has to react to the dynamics of the changing environment. Here we modeled the rate of

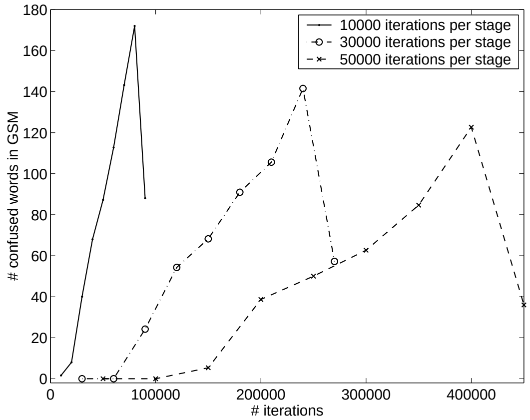

Figure 7: Number of words confused in GSM for various rates of vocabulary growth.

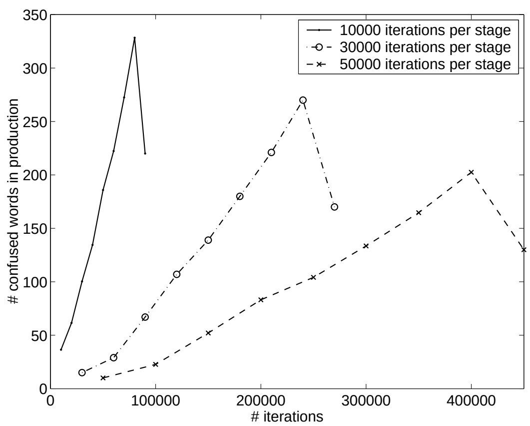

Figure 8: Number of words confused in production for various rates of vocabulary growth.

vocabulary growth by manipulating the number of iterations available per growth stage. Each of the 3 simulations was then followed by one stage aimed at the fine-tuning of GSM weights and the associative links. During fine tuning, no new words were added, and focused learning was turned on 50% of the time. Results for confusion rates in GSM and production are shown in Figures 7 and 8, respectively.

It is evident that the higher the growth rate (i.e., fewer iterations per stage, or more words added per given iterations), the earlier and the higher the confusion rates in GSM. For example, in Figure 7, after 90000 iterations the network in simulation 1 (10000 iter./stage) has been exposed to all 550 words, 170 of which are still confused. Up to that point, the network in simulation 2 (30000 iter./stage) has only seen 250 words, with some 20 words being confused. Finally, the network in simulation 3 (10000 iter./stage) is still in its second stage (190 words seen), with no confused words occurring yet.

Consistent pattern can be observed in Figure 8 for production, where the confusion rates are roughly twice as high (as in Figures 4 and 5). Again, the production accuracy depends on accuracy within GSM but also on accuracy of GSM-to-PMAP links. As the rate of vocabulary growth also affects associative links (given no extra traning), production tends to deteriorate at a higher speed than that within GSM.

In all cases, however, fine tuning for one stage only can decrease the number of confused words considerably: in the GSM by roughly 50%, and in the production less effectively (by 30%).

4. DISCUSSION

Our simulations with DevLex in this study provide important insights into mechanisms of lexical development in children. By modeling the early stages of vocabulary growth according to the CDI vocabulary, we are able to identify several factors that lead to observed lexical confusions in early child language. First, the CDI 550 words include the earliest words that children produce or comprehend, and the composition of this vocabulary is clearly biased toward nouns (a total of 51% of the vocabulary). In our simulation schedule based on frequency of use, the number of nouns gradually increases, indicating that although nouns have a high type frequency (as compared with other word categories), they are not necessarily high in token frequency right from the beginning. In contrast to nouns, the type frequency of closed-class words is low (only a total of 18% of them), but they are very high in token frequency and they enter the vocabulary from very early on. Verbs and adjectives stand somewhere in between. These kinds of word composition dynamics clearly influence the development of the lexicon, as evidenced in our simulations by the number of words confused over stages of vocabulary increase (see also [16] regarding the role of composition of word categories on lexical development).

Another important finding from our simulations is that the number of words confused in the GSM and in production is directly related to word density, measured as the amount of words mapped onto the nearest neighborhood of the target word. Interestingly, word confusion occurs more often for nouns than for other word categories, because nouns are more densely populated in the GSM map of our model (partly due to the nouns-bias discussed above). This pattern matches up with the hypothesis that nouns are more densely interrelated than verbs or closed-class words [8]. According to the ”retrieval failure” hypothesis, children may actually have the appropriate representations in lexical memory, but they fail to retrieve the appropriate words in production, perhaps due to competition among similar neighbors in densely populated regions of the lexicon [18, 8]. Our production results indicate that word retrieval in production does cause more word confusions, but the errors patterns are similar as those in the GSM map.

Finally, our results also underscore the importance of rate of vocabulary increase as a variable in modulating lexical confusion. It was suggested that naming errors in children could be a result of the sudden increase in the number and density of words in lexical memory [8]. Our simulations allow the network to "see" more or fewer words per training period, thus modeling the rate of vocabulary increase across developmental stages. At any given point in training time, the network receives different amount of words, and the rate of lexical confusion differs (cf. Figures 7 and 8). This pattern also agrees with our analysis that naming errors are due to the increase of density of the target words. 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. Our model captures both the rate and the density of lexical growth. To conclude, DevLex provides a new connectionist model that can simulate a developmental lexicon and relate to realistic language learning with self-organizing principles.

5. REFERENCES

[1] D.E. Rumelhart and J.L. McClelland. On learning the past tenses of English verbs, Chapter in *PDP: Exploration in the Microstructure of Cognition*, vol. 2, pp. 216–271. MIT, 1986.

[2] P. Li. Language acquisition in a self-organizing neural network model. In P. Quinlan (Ed.), *Connectionism and Developmental Theory*. Psychology Press, in press.

[3] B. MacWhinney. Lexicalist connectionism. In P. Broeder and J. Murre (Eds.), *Models of language acquisition : inductive and deductive approaches*. Oxford University Press, 2000.

[4] I. Farkaš and P. Li. A self-organizing neural network model of the acquisition of word meaning. In E. Altmann et al. (Eds.), *Proceedings of the 4th International Conference on Cognitive Modeling*, pp. 67–72, 2001.

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

[6] P. Li and I. Farkaš. A self-organizing connectionist model of bilingual processing. In R.R. Heredia and J. Altarriba (Eds.), *Bilingual Sentence Processing*. North Holland, in press.

[7] K. Karmiloff and A. Karmiloff-Smith. *Pathways to Language: From Fetus to Adolescent*. Harvard University Press, 2001.

[8] L. Gerskoff-Stowe and L.B. Smith. A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, 34:37–71, 1997.

[9] M. Bowerman. The acquisition of word meaning: an investigation of some current conflicts. In N. Waterson and C.E. Snow (Eds.), *The Development of Communication*, pp. 263–287. John Wiley & Sons, 1978.

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

[11] P. Li and B. McWhinney. PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments and Computers*, in press.

[12] 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.

[13] B. MacWhinney. *The CHILDES project: Tools for analyzing talk*. LEA, 2000.

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

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

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

[17] D. Gentner. On relational meaning: The acquisition of verb meaning. *Child Development*, 49(4):988–998, 1978.

[18] J. Huttenlocher. The origins of language comprehension. In R.L. Solso (Ed.), *Theories in cognitive psychology: the Loyola symposium*. LEA, 1974.