Vocabulary Development in English and Chinese: A Comparative Study with Self-Organizing Neural Networks

Xiaowei Zhao (xzhao2@richmond.edu)
Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23173 USA

Abstract

In this paper we present a self-organizing neural network model to simulate the early vocabulary development in English and Chinese. We focus on how the different lexical composition patterns in the two languages can emerge, develop and change when the learner acquires an increasing number of words. Our results suggest that certain lexical characteristics in the linguistic input (e.g., word frequency and length) play significant roles in the presence or absence of given patterns. Our study presents a dynamic developmental picture for early lexical acquisition, which is dependent on the joint contributions of mechanisms of learning and characteristics of the learning environment.

Keywords: SOM; DevLex; Vocabulary Development.

Introduction

Previous studies of children's vocabulary development have shown that there are often distinct patterns of acquisition, especially at the early stages. For example, in English it has been found that children exhibit a so-called 'noun bias' (Gentner, 1982), a pattern that shows a predominance in the number of nouns, as compared with other categories of words, in the child's early lexical composition. Some researchers argue that this noun bias might represent a universal pattern across languages (See the "*natural partitions hypothesis*" proposed by Gentner, 1982). Other investigators, however, have found that the noun bias is not present in some languages, particularly East Asian languages such as Chinese and Korean (Choi, 2000; Tardif, 1996, 2006; Tardif, et al., 1999). They argue that nouns are not always learned before verbs, and that other factors such as parental input and language-specific properties of the lexicon might determine the acquisition pattern.

Just what factors determine the early lexical composition patterns is so far unclear. Some investigators have started to look at characteristics of the linguistic input. For example, Sandhofer, Smith and Luo (2000) found that in parental speech of both English and Chinese, nouns often have a "flat distribution" in frequency: most nouns are presented with similar (and low) frequency levels; while verbs follow a "steep distribution" that a few verbs have extremely higher frequency level than others. These different distribution patterns may present learning advantages for nouns than for verbs. Second, in terms of sheer frequency, it has also been suggested that Chinese adults use more verbs than nouns when they speak to their children, while English parents use more nouns than verbs (Tardif, et al. 1999). Third, other characteristics of the lexicon and grammar may also be important. For example, Tardif (2006) suggested that verbs may be more informative in Chinese than in English, which partly contributes to the privilege of verbs in Chinese children's early vocabulary. Goodman, Dale and Li (2008) argue that syntactic complexity and ease of perception of words, can also affect the age of acquisition of these words.

The complexity of these factors in the learning environment need to be systematically investigated, but empirical studies are often limited in their abilities to put all variables to test at once. Computational work from several researchers, including our own, indicates that neural networks are ideally suited for identifying mechanisms of early vocabulary development (e.g., Li, Farkas & MacWhinney, 2004; Li, Zhao, & MacWhinney, 2007; Regier, 2005). The current study represents an effort to extend this line of research to a comparative, crosslinguistic understanding of early lexical patterns in different languages. (See also our bilingual language model in Zhao & Li, 2007). Our study attempts to examine vocabulary development in both English and Chinese with a self-organizing neural network model called DevLex-II. The model relies on simple but powerful computational principles of self-organization and Hebbian learning, and has been developed to capture the interactive developmental dynamics in language acquisition. Previously we have applied it to account for a variety of empirical patterns found in both early monolingual lexical acquisition and bilingual lexical representation (Li, et al., 2007; Zhao & Li, 2007). Here we apply this model to study how distinct patterns in two languages may emerge as a function of lexical learning in our network, and what characteristics of the learning environment (the input) may have caused the emergence of distinct patterns in early vocabulary development.

The Model

A Sketch of the Model

DevLex-II is a multi-layer self-organizing neural network model, diagrammatically depicted in Figure 1 (see Li, et al. 2007 for details). It includes three basic levels for the representation and organization of linguistic information: phonological content, semantic content, and output sequence of the lexicon. The core of the model is a two-dimensional self-organizing, topography-preserving, feature map (SOM; Kohonen, 2001), which handles lexical-semantic representations. This feature map is connected to two other feature maps, one for input (auditory) phonology,

and another for articulatory sequence of output phonology. Upon training of the network, the word meaning representations, input phonology, and output phonemic sequence of a word are presented to the network. This process can be analogous to the child's analysis of a word's semantic, phonological, and phonemic information upon hearing a word. On the semantic and phonological levels, the network forms representational patterns of activation according to standard SOM algorithm.

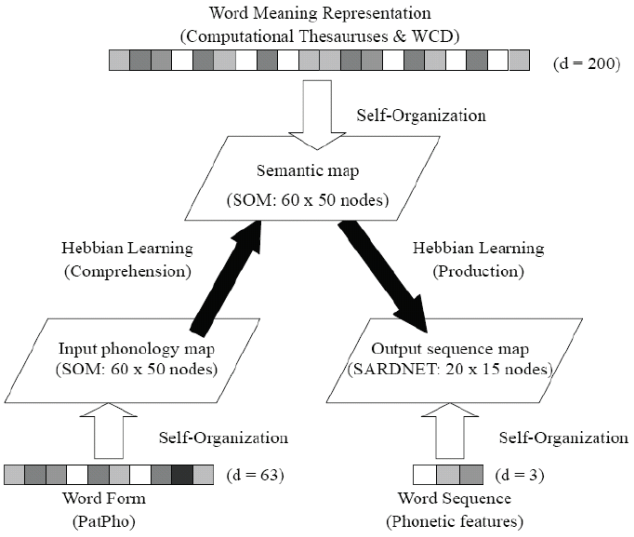

Figure 1: DevLex-II (Li, Farkas, & MacWhinney, 2007)

Here, given a stimulus **x** (the phonological or semantic information of a word), a best matching unit (BMU) on the SOM is found if its weight vector has the smallest Euclidean distances to **x.** After a BMU is identified, the weights of the nodes surrounding it in a given area (the *neighborhood*) are updated proportional to a constant learning rate α. Unlike the original DevLex model (Li et al., 2004), DevLex-II has a separate output sequence level, which is slightly different from the other two levels where standard SOM is used. The addition of this level in the model is inspired by models of word learning based on temporal sequence acquisition. It is designed to simulate the challenge that language learners face when they need to develop better articulatory control of the phonemic sequences of words. Here, the activation pattern corresponding to phonemic sequence information of a word is formed according to the algorithms of SARDNET (James & Miikkulainen, 1995), a type of temporal or sequential SOM network (see Li et al., 2007). In DevLex-II, the activation of a word form can evoke the activation of a word meaning via form-to-meaning links (to model word comprehension) and the activation of a word meaning can trigger the activation of an output sequence via meaning-to-sequence links (to model word production). Concurrent with the training of the three maps, the associative connections between maps are trained via Hebbian learning with a constant learning rate β.

In standard SOM, the radius of the neighborhood usually decreases according to a fixed training timetable. This scenario, though practically useful, is subject to the criticisms that 1) learning is tied directly and only to time (or the amount) of training, and is rather independent of the input-driven self-organizing process; and 2) the network often loses its plasticity for new inputs when neighborhood radius becomes too small. In DevLex-II, we attempt to correct these problems by using a learning process in which the neighborhood size is not totally locked with time, but is adjusted according to the network's learning outcome (experience). In particular, neighborhood function depends on the network's average quantization error on each layer, with quantization errors defined as the Euclidean distances between an input pattern and the input weight vector of its BMU (Kohonen, 2001). (see Li, et al., 2007, for details in the implementation of the self-adjustable neighborhood function.) This method gives DevLex-II certain plasticity by increasing the neighborhood size when facing new patterns (due to the increment of error)[1], while at the same time keeping a certain degree of stability. With this function the learning process will not be totally fixed *a prior*, but be more dependent on the learning experience of the network.

The Lexicons and Input Representations

For our modeling purposes we created two comparable input lexicons based on the vocabulary from the MacArthur-Bates Communicative Development Inventory, or CDI (Dale & Fenson, 1996). The English lexicon was identical to that of Li et al. (2004). The Chinese lexicon was derived from the Chinese version of the CDI (Tardif et al., 1999). Each of the two lexicons included 500 words chosen from the Toddler list of the corresponding CDI. The words were extracted roughly according to their order of acquisition by the toddlers. After excluding the homographs, word phrases, game words, words about time, words about place and onomatopoeias, the English lexicon includes 286 nouns, 98 verbs, 51 adjectives, and 65 words in other categories; and the Chinese lexicon includes 242 nouns, 145 verbs, 47 adjectives, and 66 words in other categories.

To derive the input representations of the two lexicons for our model, first, we used PatPho, a generic phonological pattern generator for neural networks (Li & MacWhinney, 2002), to construct the basic input phonological patterns of the English and Chinese words. A left-justified template with 54 dimensions was adopted. In addition, a separate group of 9 units was used to represent lexical tones in Chinese, and the values of these units were left empty for English. This method provides us a universal syllabic-template based phonological representation of the words in the examined languages. Second, there were in total 55 phonemes from the two languages (36 in Chinese; 38 in English), which we represented as vectors of articulatory features of the phonemes to the output sequence map (as in PatPho). Third, for each language, we constructed two sets

[1]Our network selected a word each time according to its frequency in the parental CHILDES corpus, so some words may be presented to the model later than others due to frequency differences.

of lexical semantic representations through two different methods, and then combined them to increase the accuracy of the lexical representation (see Li et al., 2004 for rationale). The first set was generated by WCD (the word co-occurrence detector, Li et al., 2004), a special recurrent network that learns the lexical co-occurrence constraints of words by reading through input in linguistic corpus (here it is the child-directed parental speech from CHILDES). The second set of semantic representations was generated from computational thesauruses available for each of the two languages. The similarity matrix of the 500 words in each language were calculated respectively according to their semantic features extracted from the HowNet database (for Chinese, http://www.keenage.com), and WordNet database (for English, Miller, 1990). A Random Mapping (Kohonen, 2001) method was further used to reduce the size of each set of the semantic representation to a lower dimension (from d=500 to d=100), and the two sets were then combined together to form each word's semantic vector.

Simulation Parameters

Since we were examining vocabulary development in two languages, we wanted to keep our modeling parameters to be as similar as possible across languages. In our simulations, the English and Chinese lexicons were presented to the DevLex-II model separately, with identical simulation parameters. Specifically, the phonological map and the semantic map each consisted of 60 x 50 nodes, and the output sequence map consisted of 20 x 15 nodes. During training, both learning rate α and β were kept constant (0.25 and 0.1 respectively). The radii of a winner's neighborhood on each map were adjusted automatically according to the neighborhood function discussed earlier. The neighborhood radii followed an overall tendency of decreasing from 10 to zero while the average quantization errors also decreased. The initial numbers were chosen to be large enough to discriminate between the words and phonemes, while keeping the computation tractable.

During a simulation, words from a training lexicon (English or Chinese) were presented to the network one by one. To simulate the effect of word frequency in early child lexicon, our network selected one word at a time for training roughly according to the word's frequency of occurrence in the child-directed speeches. Particularly, we extracted the Chinese-speaking parents' speeches from the East Asian database in CHILDES; the Chinese corpus included about 380,000 word tokens. For English, the data were extracted from the American English database in CHILDES. To equate the size of the two corpora, we only used portion of the American English database, and also got about 380,000 words for English parental corpus. The logarithms (base 10) of these occurrence frequencies were used to force a more even distribution of the words in the input. Such a setup has been widely used in other computational simulations based on real corpora, given that word frequency distributions follow the famous Zipf's law. Our simulation results reported below are based on the average performance from ten simulation runs that used the same training parameters of α, β, map size, and initial neighborhood radius.

Results and Discussion

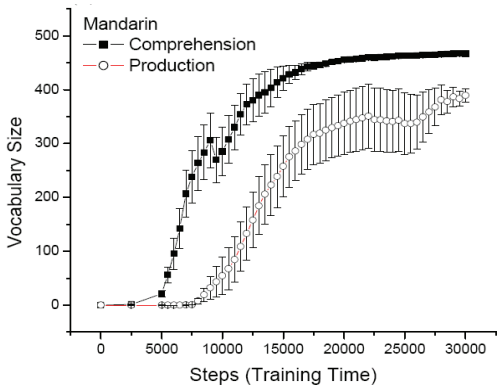

Figure 2. Vocabulary development in the learning of Chinese lexicon by DevLex-II. Results are averaged across ten simulation trials (error bars indicate standard errors).

Overall Developmental Pattern

Our simulation results capture many developmental patterns in children's early lexical acquisition. First, not surprisingly, as the training progresses, the size of the vocabulary that our network can successfully comprehend or produce increases for both English and Chinese. Figure 2 presents the vocabulary profile in the learning of Chinese, as a function of training time[2]. Second, similar to the findings for English lexical acquisition in our previous work (Li, et al, 2007), we can observe a clear vocabulary spurt, the rapid increase in the size of acquired vocabulary within a short time span, in our network for Chinese. As seen in Figure 2, on average, the model's productive vocabulary did not accelerate until the vocabulary size reached about 50 words, which was about 10,000 steps (around a third of the total training time). As we have discussed previously (Li, et al, 2007), this period of rapid increase in vocabulary size was prepared by the network's slow learning of the structured representation of phonemic sequence, word phonology, and word semantics. Once the basic structures are established on the corresponding maps, Hebbian learning based associative connections between maps reliably strengthened to reach a critical threshold, which triggers the onset of the vocabulary spurt. Third, the results also indicate that comprehension occurs earlier than production, consistent with previous empirical evidence that children's comprehension generally precedes their production (Clark & Hecht, 1983). This discrepancy can be best understood by that language learners need more cognitive efforts related to phonological rehearsal in their working memory to develop better articulatory control of the phonemic sequences of words (which can be caught by the gradually development of meaning-to-sequence links in our model).

[2] Since a similar result for English has been reported in our previous monolingual study (Figure 2 in Li et al., 2007), we are not showing here our network's vocabulary development for English.

Another common developmental pattern in our network is the increased lexical complexities across the learning of both Chinese and English. As shown in Figure 3, along with the increase of the vocabulary size of the words that our network can successfully produce, the mean length of these words (in phonemes) also gradually increased. From the figure, we can clearly observe that the average phonemic length of the words learned by the network became longer as more words entered the network's vocabulary; the average number of phonemes in the learned words increased from around one (for Chinese) or two phonemes per word (for English) to almost four phonemes per word. This result indicates that, just like the child, our network's articulatory ability to control the phonemic sequence of words develops over time. Initially, it can only handle those words that are simple and easily to pronounce, but with its improved sequences learning ability, the network can manipulate more complicated and longer words for both languages.

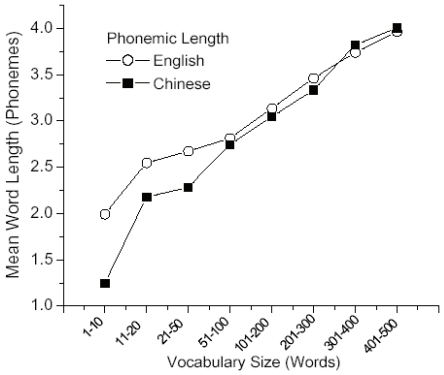

Figure 3. The increase of mean phonemic lengths of words that DevLex-II can successfully produce.

Development in Lexical Composition

The investigation of lexical composition in children's early vocabulary is a main goal of our current study. As discussed in the introduction, empirical studies have shown that discrepancies do exist for different languages in the overall percentage of words that belong to various grammatical categories in children's vocabulary (e.g., the ratio of nouns versus verbs). Here, we calculated the average number of words correctly produced by our network in three main grammatical categories (nouns, verbs, and adjectives), as a function of our network's increasing vocabulary size. The objective here is to identify the possible factors that may affect early lexical composition patterns. The results for English words and Chinese words are depicted in Figure 4(a) and 4(b) respectively. These figures show clearly both similarities and differences in the developmental trajectories for English and Chinese lexicons (similar to the real data shown in Figure 18.1 of Tardif, 2006).

First, as the training progresses, more and more words enter the list that can be correctly produced by the network in each of the grammatical categories. Second, for both English and Chinese, at most stages of vocabulary development (which are represented by the total number of learned words), there are more nouns than verbs, and more verbs than adjectives that can be produced by the network. On average, at the latest stage of vocabulary development (total vocabulary size between 400 and 500 words), there are about 250 nouns, 90 verbs and 48 adjectives that have been learned in the English lexicon; similarly, there are about 200 nouns, 131 verbs and 41 adjectives that have been learned in the Chinese lexicon[3]. This overall "noun bias" in our network's lexical acquisition for both languages is consistent with recent corpus-based studies of crosslinguistic lexical development (see Liu, 2008), but is at odds with the argument that verbs dominate over nouns in early Chinese vocabulary (see Introduction).

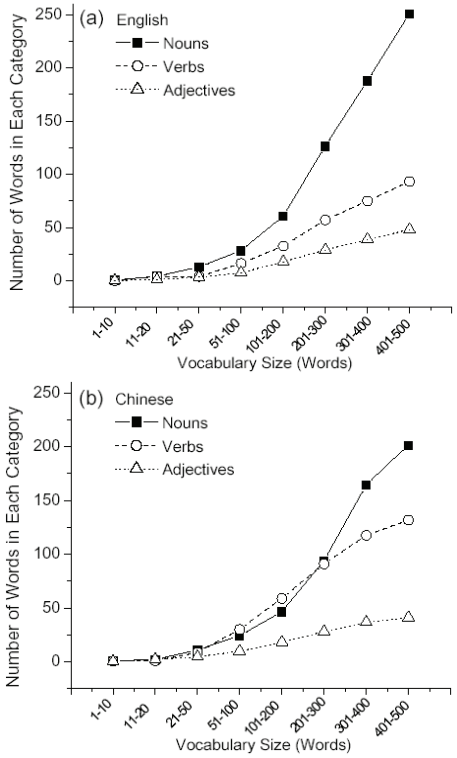

Figure 4. Mean number of nouns, verbs, and adjectives learned by DevLex-II at different developmental stages for (a) English and (b) Chinese.

However, clear differences in lexical composition can also be observed between our simulation results for English versus Chinese. Comparing Figures 4(a) and 4(b), we can see that the network generally produced more nouns in English than in Chinese, and more verbs in Chinese than in English, across most stages. For Chinese, it produced comparable numbers of nouns and verbs at the early vocabulary stages. In two cases (e.g., when the total vocabulary size was between 51-100 and between 101-200 words) there were more verbs than nouns produced by the model, but after 300 words there were more nouns than verbs. In English, nouns always dominate over verbs in number, starting from the very earliest stages. Moreover, the

[3] Nouns also occur more often than verbs at the earliest stages, but this cannot be shown clearly on the figures due to small vocabulary size.

rate of increase for nouns is also more rapid than that for verbs in English. In short, our network displayed a much stronger noun bias in English acquisition than in Chinese acquisition, consistent in general with empirical evidence on crosslinguistic differences in vocabulary growth.

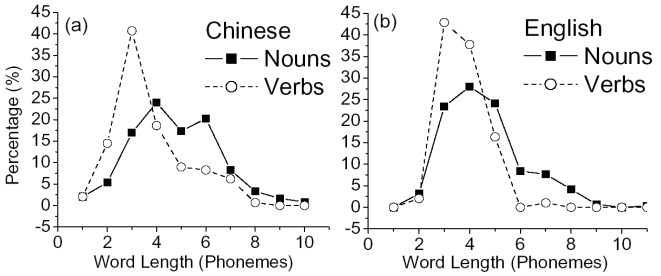

Figure 5. Distribution of nouns and verbs as a function of phonemic length of words: (a) Chinese; (b) English.

Length and Frequency Effects

Why does our network show the different patterns in lexical composition for the two languages? In the current study, since our networks learning English and Chinese were using the same set of training parameters, we can safely conclude that the cross-language differences in lexical compositions must be dependent on the different input characteristics in the two lexicons. In particular, the occurrence frequency of words in language input and their phonemic length may be two important factors that led to the differences.

As can be observed in Figure 3, short words are learned more easily and earlier than longer words in our network. Based on this finding, we believe that the differences in morphological characteristics between Chinese and English words can partly account for the difference in children's early lexical compositions. In Figure 5, we plotted the word length distributions (in phonemes) of nouns and verbs in our Chinese and English lexicons respectively. The distribution of Chinese words shown in Figure 5(a) indicates that most verbs have their phonemic length in the range from two to four phonemes. The peak of the distribution is at three phonemes, and more than 40 percent of all the verbs have such a word length. Nevertheless, the nouns are more evenly distributed in the range from three to six phonemes, the peak occurs at four phonemes, and nearly 20 percent of nouns have six phonemes. Figure 5(b) presents a different picture for English that nouns and verbs generally follow similar and overlapped distribution patterns, and most of the words ranged from 3 to 5 phonemes in length. Although the verb distribution is still somewhat left-skewed and their length is shorter than nouns on average, the disparity between nouns and verbs is not as large as in Chinese. The reason for this difference is that, in the Chinese language, a large percentage of words are made up of only one syllable (character)[4], and this feature appears more obviously in verbs than in nouns. So the proportion of monosyllabic words in nouns is much lower than the proportion of such words in verbs; consequently Chinese verbs may be generally easier to produce than nouns. In contrast, the difference of length between nouns and verbs in English is not so clear. It is thus highly conceivable that both children and our network capitalize on the word length effect in the production of early vocabulary, and therefore verbs get advantages in early vocabulary development in Chinese.

Based on the length effect, an interesting prediction can be made on Chinese children's lexical composition. As demonstrated by our model, children often need more cognitive efforts related to working memory to develop appropriate articulatory control for correctly pronouncing words with different length; thus the length effect (if any) should be more salient in production than in comprehension. In comprehension, the advantage of short verbs in Chinese might disappear. Consequently, we predict that Chinese children's lexical composition in comprehension should present stronger noun bias than that in production. It is often difficult studying children's early language comprehension, but if we can find such a pattern, it could certainly serve as a good example of word length effects.

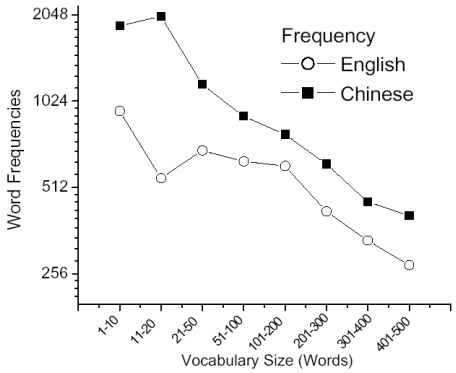

Figure 6. The development of the average frequency of words that can be correctly produced by DevLex-II.

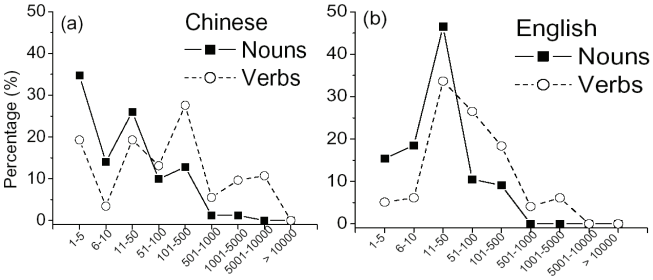

Figure 7. Distribution of nouns and verbs as a function of word frequency. The abscissa indicates the frequency level of words: (a) Chinese; (b) English.

In addition to the length effect, the effect of word frequency is also important. In Figure 6, we show the average token frequency of words learned by our network at different stages of vocabulary development[5]. It can be clearly seen that their average frequency has a general

[4] According to Wang (1994), 44 percent of 3000 most frequently speaking Chinese words are monosyllabic; and in the 8822 words used by Chinese government to evaluate student's vocabulary level, 22 percent of them are monosyllabic words (Xing, 2006).

[5] Frequency here is counted as each word's occurrence number in the corresponding parental CHILDES corpus.

tendency to decrease as more words are learned. It suggests that the high frequency words, which occur more often in children's language input, are learned more easily than low frequency words (but see the recent discussion of frequency effects in Goodman, et al. 2008). Figure 7 shows the frequency distribution of nouns and verbs in our Chinese (7a) and English lexicons (7b). Here, we calculate the percentages of the total noun (or verb) types that each frequency level is associated with. There are more verbs having higher frequency in Chinese than English: in total, about 20 percent of all the Chinese verbs fall in the frequency range of occurring more than 1000 times in the Chinese parental corpus (which includes 380,000 word tokens); however, only six percent of English verbs fall within these high frequency ranges. Since words with higher frequencies tend to be learned earlier, and Chinese-speaking children's language input has more high frequency verbs (also see Sandhofer et al., 2000), it should be clear why higher frequency verbs are more easily learned by our network at earlier stages.

Conclusion

In this study we extended DevLex-II, a self-organizing neural network model, to a comparative, crosslinguistic study of lexical development in English and Chinese. Our model can successfully capture the gradual process of development in size and complexity of children's lexicons for both of the two target languages. Additionally, our model can display important lexical compositional patterns (e.g., the presence or absence of "noun bias") found with children acquiring lexicon in different languages. Our analyses suggest that the different patterns of lexical composition in early child vocabulary should be best understood with respect to (a) the dynamical consolidation of lexical-semantic structures in representation, meaning-form association, self-organization within and between varieties of linguistic information; and (b) varied input characteristics, such as the length and frequency distributions of the words to be learned. It is important to note, however, that frequency does not operate in a straightforward fashion to impact lexical acquisition, and there are many factors need to be considered on whether and how frequency counts in determining the process of vocabulary acquisition (e.g., types of words, modality, and time of acquisition; see Goodman, et al, 2008). It will be our goal to consider these additional factors in our future studies of computational and crosslinguistic studies.

Acknowledgments

This research was supported by a grant from the National Science Foundation (BCS-0642586).

References

Choi, S. (2000). Caregiver input in English and Korean: Use of nouns and verbs in book-reading and toy-play contexts. *Journal of Child Language*, *27(1)*, 69-96.

Clark, E.V., & Hecht, B.F. (1983). Comprehension, production, and language acquisition. *Annual Review of Psychology, 34*, 325-349.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127.

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

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

Goodman, J.C., Dale, P.S. & Li, P. (2008). Does frequency count: Parental input and the acquisition of vocabulary. *Journal of Child Language, 35,* 1-17.

James, D., & Miikkulainen, R. (1995). SARDNET: A self-organizing feature map for sequences. In G. Tesauro et al., (Eds.), *Advances in neural information processing systems 7* (pp.577-584). Cambridge, MA: MIT Press.

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

Li, P., Farkas, I., & MacWhinney (2004). Early lexical development in a self-organizing neural network. *Neural Networks. 17*, 1345-1362.

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., & MacWhinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*, 581-612.

Liu, S. (2008). *Early vocabulary development in English, Mandarin, and Cantonese: A cross-linguistic study based on CHILDES.* Master thesis, University of Richmond.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science, 29*, 819.

Sandhofer, C., Smith, L., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of Child Language*, *27*, 561.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speaker's early vocabularies. *Developmental Psychology, 32*, 492-504.

Tardif, T. (2006). But are they really Verbs? Chinese words for action. In K. Hirsh-Pasek, R.M. Golinkoff (Eds.), *Action Meets Word: How Children learn Verbs*. New York: Oxford University Press.

Tardif, T., Gelman, S.A., & Xu, F. (1999). Putting the "noun bias" in context: a comparison of English and Mandarin. *Child Development, 70*, 620-635.

Wang,Y. (1994). The analysis and study of common words in Chinese. *Global Chinese Education, 2,* 58-62.

Xing, H. (2006). Statistic analysis of disyllabic words in Chinese. *Global Chinese Education, 77,* 63-71.

Zhao, X., & Li, P. (2007). Bilingual lexical representation in a self-organizing neural network. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 755-760). Nashville, TN.