Cross-situational Learning of Low Frequency Words: The Role of Context Familiarity and Age of Exposure

Afsaneh Fazly, Fatemeh Ahmadi-Fakhr
Computer Sciences and Engineering
Shiraz University
Shiraz, Iran
{fazly,ahmadifakhr}@cse.shirazu.ac.ir

Afra Alishahi
Computational Linguistics and Phonetics
Saarland University
Saarbrücken, Germany
afra@coli.uni-saarland.de

Suzanne Stevenson
Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

Higher frequency has been shown to have a positive effect on the acquisition of words and other linguistic items in children. An important question that needs to be answered then is how children learn low frequency items. In this study, we investigate the acquisition of meanings for low frequency words through computational modeling. We suggest that for such words, the familiarity of the context they appear in has an important effect on their acquisition. We note that context familiarity is confounded with another factor, namely the age of exposure to a word, and hence examine the independent role of each of the two factors on word learning.

Cross-situational Word Learning

Learning the meaning of words is a challenging task for young children, especially given that most words are learned from noisy and ambiguous contexts. Many specific word learning biases and constraints, as well as general learning mechanisms, have been suggested to be at work in the course of child lexical development. In particular, the learning of word–meaning mappings has been suggested to be based on cross-situational observation (Quine, 1960; Pinker, 1989) — that is, the meaning of a word can be learned by detecting the common set of meaning elements across all situations in which the word occurs. Psychological experiments on adults and children show that they are capable of learning word–referent mappings from their co-occurrences over time, even when each single occurrence of a word–referent pairing is ambiguous (Yu & Smith, 2007; Smith & Yu, 2007). The learning process seems to be sensitive to the statistical properties of the input, such as word frequency, the degree of ambiguity of presentation (i.e., how many words appear together), and the familiarity of the context. However, the relevant properties and their precise impact on learning word meanings are not well understood.

For example, higher frequency has been shown to have a positive effect on learning many linguistic constructions (Huttenlocher et al., 1991; Naigles & Hoff-Ginsberg, 1998). But frequency must be investigated carefully, since in many learning situations — including word learning — it may be confounded with other factors, such as the diversity of the context that a word can appear in (see, e.g., Kachergis, Yu, & Shiffrin, 2009). Moreover, many experimental studies have shown that children can acquire the meaning of a novel word with only one or a few exposures (i.e., when the word has very low frequency), especially if it is presented in a familiar context, a phenomenon known as *fast mapping* (Carey & Bartlett, 1978; Gershkoff-Stowe & Hahn, 2007; Alishahi et al., 2008). These observations about the interactions of various factors with frequency are especially important given that words in the input children receive have a Zipfian distribution (Zipf, 1949) — that is, a large proportion of words have a very low frequency of occurrence, yet generally they are successfully learned. Therefore, the relevant statistical features of the input data and their independent effect on learning need to be carefully investigated.

The experimental study of Kachergis et al. (2009) on adult subjects is one such attempt to identify and study the role of some of the important statistical properties of input on word learning. By varying the frequency of co-occurrence of different word–referent pairs, Kachergis et al. examine the independent and differential role of frequency, contextual diversity (i.e., diversity in the co-occurring words across usages of a target word), and within-trial ambiguity (i.e., the number of co-occurring words in each sentence) in cross-situational word learning. In particular, Kachergis et al. suggest that high contextual diversity and low within-trial ambiguity can boost the acquisition of low frequency words. However, some of the experimental results in their study cannot be explained by the factors they propose. Also, contextual diversity cannot explain children's ability to easily learn the referent of a novel word from a single exposure to that word, since it captures a property of the input across multiple exposures to a word. These observations suggest that other factors may be at play, especially for the acquisition of low frequency words.

Our goal is to investigate what other factors may have an effect on learning the meanings of words in general, and on the acquisition of low frequency words in particular. We use an existing computational model of cross-situational word learning to simulate the experiments of Kachergis et al. (2009), and to examine the effect of two additional factors in word learning: the familiarity of the context that a word appears in (i.e., how well the model/learner knows the other words in the sentence), and the age of exposure to a word. The computational simulations of our model show a matching behavioural pattern with that of adult word learners in the experiments of Kachergis et al. Moreover, our results suggest that for low frequency words, it is not the contextual diversity that helps learn their meaning, but the degree of familiarity of their context. We further test this claim by applying our model to a large corpus of child-directed speech, and examine the role of the proposed factors in the learning performance of the model.

Statistical Properties of the Input

Contextual Diversity and Within-trial Ambiguity

Kachergis et al. (2009) report a series of studies on adult subjects learning word–referent mappings from ambiguous utterance–scene pairs, where the utterance contains a bag of words, and the scene is the set of their referents. They investigate the effect of frequency by having some words and referents appear more often than others. They also investigate the interaction between word–referent frequencies and the diversity of the contexts that a word appears in, to examine the independent effect of each of the two factors on word learning. In these experiments, contextual diversity is varied by manipulating either the overall rate of co-occurrence among words, or the number of co-occurring words within a trial. More precisely, Kachergis et al. study the interactions among the following three factors:

- Word frequency:
  F(w) = total #occurrences of w in the input
- Contextual diversity:
  CD(w) = total #words co-occurring with w across all usages of w in the input
- Within-trial ambiguity:
  WA(w) = mean #words co-occurring with w in each utterance

Their results show that a higher F often leads to better learning of a word, and to boosting the learning of other words. However, F is usually confounded with CD, which can be seen as an alternative explanation for learning facilitation. When F is controlled for, a higher CD improves learning (i.e., more word–meaning pairs are learned), whereas a higher WA harms learning. Similarly when CD is controlled for, a higher F improves learning. Most interestingly, when a higher CD is achieved by interleaving high frequency words in the presentation of low frequency words, the learning of the low frequency words is improved.

Age of Exposure and Context Familiarity

As described in the previous section, the results of Kachergis et al. (2009) suggest that contextual diversity (CD) is particularly important for the acquisition of low frequency words. However, some of their results show a boost in the acquisition of low frequency words where there is no notable difference in CD. In an attempt to explain these results, and inspired by the well-studied fast mapping effect (Carey & Bartlett, 1978), we study two additional statistical factors that might play a role in cross-situational learning:

- Age of exposure:
  AE(w) = time at which w first appears in the input
- Context familiarity:
  CF(w) = mean *familiarity* of words co-occurring with w, averaged across all usages of w in the input

where *familiarity* of a word is determined by its frequency of occurrence prior to its current appearance.

Computational Analysis

We investigate the role of each of the above factors in cross-situational learning through two sets of experiments. First, we replicate the results of Kachergis et al. (2009) using the computational model of Fazly et al. (2008) (briefly explained in the next section), and examine the impact of our proposed factors as well as the ones proposed by Kachergis et al. on learning. Second, we apply our model on a larger corpus of actual child-directed speech to better understand how the model learns the meaning of low frequency words in a more naturalistic situation, and to study the impact of the statistical factors and their interaction during the course of learning.

Overview of the Computational Model

We use an incremental probabilistic word learning algorithm, explained in full detail in Fazly et al. (n.d.). Here we repeat a brief explanation of how the model works.

Utterance and Meaning Representations

The input to our word learning model consists of a set of utterance–scene pairs that link an observed scene (what the child perceives) to the utterance that describes it (what the child hears). We represent each utterance as a set of words, and the corresponding scene as a set of meaning symbols.

Utterance: { *Joe*, *rolled*, *the*, *ball* }
Scene: {`joe`, `roll`, `the`, `ball`}

Given a corpus of such utterance–scene pairs, our model learns the meaning of each word w as a probability distribution, $p(.|w)$, over the semantic symbols appearing in the corpus. In this representation, $p(m|w)$ is the probability of a symbol m being the meaning of a word w. We assume that in the absence of any prior knowledge, all symbols are equally likely to be the meaning of a word. Hence, prior to receiving any usages of a given word, the model assumes a uniform distribution over all semantic symbols as its meaning.

Meaning Probabilities

Our model combines probabilistic interpretations of cross-situational learning (Quine, 1960) and a variation of the principle of contrast (Clark, 1990), through an interaction between two types of probabilistic knowledge acquired and refined over time. Given an utterance–scene pair received at time t, i.e., $(\mathrm{U}^{(t)}, \mathrm{S}^{(t)})$, the model first calculates an alignment probability a for each $w \in \mathrm{U}^{(t)}$ and each $m \in \mathrm{S}^{(t)}$, using the meaning probabilities $p(.|w)$ of all the words in the utterance prior to this time (Step 1 below). The model then revises the meaning of the words in $\mathrm{U}^{(t)}$ by incorporating the alignment probabilities for the current input pair (Step 2). This process is repeated for all the input pairs, one at a time.

Step 1: Calculating the alignment probabilities. We estimate the alignment probabilities of words and meaning symbols based on a localized version of the principle of contrast: that a meaning symbol in a scene is likely to be highly associated with only one of the words *in the corresponding utterance*. For a symbol $m \in \mathrm{S}^{(t)}$ and a word $w \in \mathrm{U}^{(t)}$, the

higher the probability of m being the meaning of w (according to $p(m|w)$), the more likely it is that m is aligned with w in the current input. In other words, $a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$ is proportional to $p^{(t-1)}(m|w)$. In addition, if there is strong evidence that m is the meaning of another word in $\mathrm{U}^{(t)}$ — i.e., if $p^{(t-1)}(m|w')$ is high for some $w' \in \mathrm{U}^{(t)}$ other than w — the likelihood of aligning m to w should decrease. Combining these two requirements:

$$a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)}) = \frac{p^{(t-1)}(m|w)}{\sum_{w' \in \mathrm{U}^{(t)}} p^{(t-1)}(m|w')} \quad (1)$$

Step 2: Updating the word meanings. We need to update the probabilities $p(.|w)$ for all words $w \in \mathrm{U}^{(t)}$, based on the evidence from the current input pair reflected in the alignment probabilities. We thus add the current alignment probabilities for w and the symbols $m \in \mathrm{S}^{(t)}$ to the accumulated evidence from prior co-occurrences of w and m. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w, m) = \mathrm{assoc}^{(t-1)}(w, m) + a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)}) \quad (2)$$

where $\mathrm{assoc}^{(t-1)}(w, m)$ is zero if w and m have not co-occurred before. The association score of a word and a symbol is basically a weighted sum of their co-occurrence counts.

The model then uses these association scores to update the meaning of the words in the current input:

$$p^{(t)}(m|w) = \frac{\mathrm{assoc}^{(t)}(m, w)}{\sum_{m_j \in \mathcal{M}} \mathrm{assoc}^{(t)}(m_j, w)} \quad (3)$$

where $\mathcal{M}$ is the set of all symbols encountered prior to or at time t. We use a smoothed version of the above formula, as described in (Fazly et al., n.d.).

Word Comprehension Score

Our model updates the meaning of a word every time it is heard in an utterance. The strength of learning of a word at time t is reflected in $p^{(t)}(m = m_w|w)$, where m_w is the "correct" meaning of w according to a gold-standard lexicon. We refer to $p^{(t)}(m_w|w)$ as the comprehension score (Comp) of word w at time t. Ideally, a word is accurately learned when the probability distribution $p(.|w)$ is highly skewed towards the correct meaning m_w. In our experiments reported in the following sections, we first train our model on a number of utterance–scene pairs, and then examine the comprehension scores of words as an indirect way of measuring the performance of our model in selecting referents of words.

Analysis of Artificial Word Learning Data

Here we report the results of our simulations on artificially-generated data similar to that of Kachergis et al. (2009). Their (human) experiments examine the effect of three factors on word learning: frequency (F), contextual diversity (CD), and within-trial ambiguity (WA), as defined on page 2. The artificial input data set used in our simulations is explained next, and then the results of the experiments are presented.

Input Data

The artificial data set consists of randomly-generated sequences of utterances in the form of an unordered bag of novel words, each paired with a set of novel meaning symbols. In the artificial data, one of the three factors under study is changed while the other factors are kept constant, in order to better understand the role each plays in learning, as well as the interactions among the different factors. We use nine sets of artificial data (each containing 18 word–meaning pairs), one set for each experimental condition of Kachergis et al. (2009). The first experiment investigates the role of F: one condition divides words into two frequency groups (F=3,9), the other into three frequency groups (F=3,6,9). The second experiment examines the role of context by manipulating either CD or WA, while keeping F constant. One condition manipulates CD by dividing words into two unequal-sized groups (with 6 and 12 words, respectively), and allowing words in each group to co-occur only with other words from the same group. In two other conditions, a word appears with either 2 or 3 other words in each trial (WA=3 and WA=4, respectively). The third experiment studies the interaction between F and CD by controlling the co-occurrence among words from three frequency groups (F=3,6,9), resulting in four conditions: In Low CD condition, words from each frequency group co-occur only with other words from the same group. In Med CD conditions, low frequency words (F=3) are allowed to either co-occur with words in F=6 (Med CD-3&6), or with those in F=9 (Med CD-3&9). In High CD condition, there is no restriction on the co-occurrence of words from different frequency groups. For each experimental condition, we randomly generate 30 different artificial input. Results presented here are averages over 30 different simulations, each using a different input.

Modeling Effects of Frequency and CD

Figures 1 to 3 present the performance of our model on the artificially-generated input in three experiments analogous to those of Kachergis et al. (2009).

Our findings in Experiment 1 (Figure 1) are generally in line with those of Kachergis et al. (2009): that higher frequency does not seem to have a consistently positive effect on word learning. As noted by Kachergis et al., frequency might be conflated with other factors, and thus we cannot make a decisive conclusion only on the basis of this experiment.

Figure 2 (left half) shows that contextual diversity (CD) has a significant positive effect on word learning ($p \ll .001$).[1] Figure 2 (right half) shows that an increased WA has an adverse effect on word learning, even though it also increases CD (difference is significant; $p \ll .001$).

Recall that in Experiment 3 the interaction between CD and F is examined by looking at the learning performance of low (F=3), medium (F=6), and high (F=9) frequency words in

[1] All statistical significance tests reported in this paper are for paired t-tests with a 95% confidence interval, and are performed using the R statistics package (`http://www.r-project.org`).

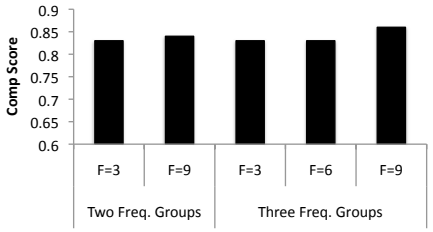


Figure 1: Average Comp scores for words from different frequency ranges (Experiment 1).

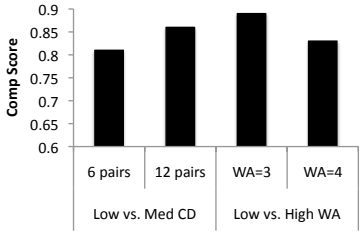


Figure 2: Average Comp scores for words with different contextual diversity (CD), or within-trial ambiguity (WA) (Experiment 2).

conditions with varying degrees of CD. The overall pattern of our results across the four conditions of this experiment, presented in Figure 3, matches those reported by Kachergis et al. for humans. Specifically, we note that, as for human subjects, the overall learning is significantly greater for our model in the High CD condition ($p \ll .001$). Moreover, we observe that, similar to human behaviour, the acquisition of low frequency words in our model is better when they are allowed to co-occur with higher frequency words (Conditions: High CD, Med CD-3&6, and Med CD-3&9); differences between each of these three conditions and the Low CD condition are statistically significant ($p \ll .001$). Whereas Kachergis et al. attribute this behaviour to an increased CD, we suggest that there is another factor (namely context familiarity), which is responsible for this boost of performance in the acquisition of low frequency words.

Context Familiarity as the Explanatory Factor

As discussed above, Kachergis et al. (2009) suggest that contextual diversity is especially important for the acquisition of low frequency words. However, there are cases (in our experiments and in those of Kachergis et al.) where we see a boost in the acquisition of low frequency words, with no notable difference in CD. Instead, as we show now, differences in context familiarity (CF) can explain the pattern of results.

Consider again the results of Experiment 3 shown in Figure 3. We also summarize some properties of the input in the four conditions of that experiment in Figure 4. For each condition we select one simulation such that the overall pattern of results (e.g., with respect to the learning of low frequency words) for these simulations match that of the average performance given in Figure 3. For each input used in the selected simulations, we then calculate the average CD and CF values for words in each of three frequency groups (i.e., F=3,6,9). To

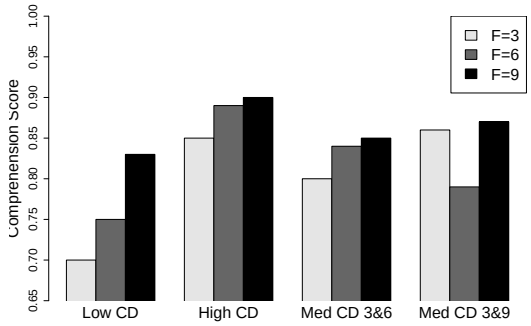


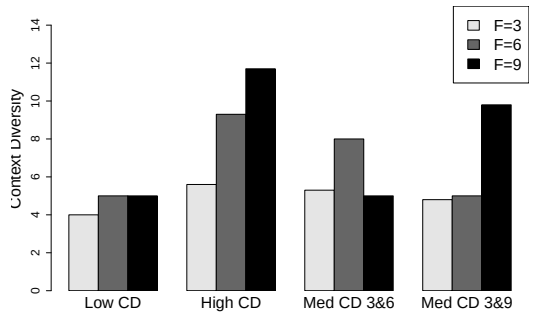
Figure 3: Average Comp scores for low, medium, and high frequency words (left to right bars) within each of the four conditions of contextual diversity (Experiment 3).

calculate CF for a word in an utterance, we set the familiarity of each co-occurring word to its frequency of occurrence prior to the current appearance.

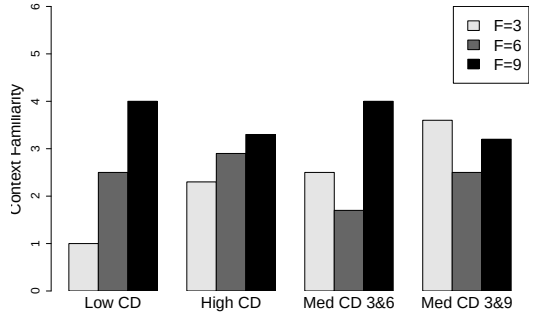
Figure 3 shows that the highest performance in learning of low frequency words is achieved when these words are allowed to co-occur with words with F=9 (Condition: Med CD-3&9). Kachergis et al. attribute this high learning performance to an increased CD. There is indeed *overall* a higher CD value for the High and Med CD conditions. However, when they are separated by the frequency of the items as in Figure 4(a), we observe that the CD values for the low frequency words do not substantially change across the four conditions. (The overall pattern of CD values in Figure 4(a) match very closely those presented in Table 3 on page 5 of Kachergis et al. CD values range from 4 to 5.6 in our experiments, and from 4 to 5.5 in those of Kachergis et al.)

Interestingly, however, if we look at the pattern of the CF values in Figure 4(b), we see that it conforms to the learning performance of our model (and those of humans) on low frequency words (compare the lightest bars in each of the two figures across the four conditions). We can explain this effect of CF in the learning of our model as follows: When low frequency words are allowed to co-occur with words with F=9, we expect the contexts of the low frequency words to be, on average, more familiar than in other conditions. Since our model is expected to have learned something about the possible meanings of a familiar word, this in turn decreases the degree of ambiguity in an utterance–scene pair, making the acquisition of a novel low frequency word easier. This result is a direct consequence of the interactions between the two sets of probabilities accumulated over time in our model, namely the alignment and the meaning probabilities. When aligning a word in an utterance to a referent/meaning in the corresponding scene, our model uses its acquired knowledge about the meaning of the co-occurring words (according to the meaning probabilities). The more familiar the co-occurring words are, the more reliable the meaning probabilities for these words will be, and this in turn makes it easier for the model to align the target word to its correct referent.

For higher frequency words (F=6 and F=9), we can see a



(a) Average CD for different frequency ranges across different conditions



(b) Average CF for different frequency ranges across different conditions

Figure 4: Average CD and CF values for low, medium, and high frequency words (left to right bars) within each of the four conditions of contextual diversity (Experiment 3).

clear effect of CD (compare darkest bars in Figure 3 and Figure 4(a)). These results together suggest that CD positively affects the learning of medium and high frequency words (as also noted by Kachergis et al.), but for low frequency words it is the familiarity of the context that is the key factor in their acquisition (in contrast to the suggestion of Kachergis et al.).

Analysis of Naturalistic Child-directed Speech

As noted previously, children are exposed to a great number of low frequency words in the input they receive (due to the Zipfian distribution of words in a language). We thus further investigate the effects of context familiarity (CF) on the acquisition of low frequency words in a more naturalistic setting, by performing experiments on a large corpus of actual child-directed speech. The child-directed corpus and the details of the experiments are explained below.

Input Data

The child-directed corpus consists of 10,000 utterance–meaning pairs, where the utterances are taken from the Manchester corpus (Theakston et al., 2001) in the CHILDES database (MacWhinney, 2000), and the corresponding meanings are artificially generated by including a distinct meaning symbol for each word in the utterance. Our focus in the following experiments is on F, CD, CF, and another factor usually confounded with CF, namely age of exposure to a word or AE. We control for the effect of within-trial ambiguity (WA) in our experiments by considering only those utterances whose length is between 5 and 7 (inclusive).

Table 1: Average frequency (F), CD, Comp, CF, and AE values for two groups of low frequency words: High Comp vs. Low Comp. Number of words in each group is given in parenthesis.

	High Comp Comp ≥ 0.9 (877)	Low Comp Comp < 0.9 (258)
F	1.50 ± 0.73	1.49 ± 0.73
CD	6.61 ± 2.82	6.70 ± 2.77
CF	4.64 ± 0.40	3.58 ± 0.62
AE	9.39 ± 5.55	6.36 ± 6.19
Comp	0.93 ± 0.02	0.52 ± 0.15

We measure the factors CD, CF and AE for each word according to the definitions on page 2. Here we measure the *familiarity* of a word slightly differently from on the artificial data. Since the frequency of words in the child-directed corpus is on a different scale and varies a lot, we set familiarity of a word to a value between 0 and 5 according to the frequency range it belongs to. The mappings between familiarity values and frequency ranges are: 0 (0), 1 (1), 2 (2–4), 3 (5–9), 4 (10–29), and 5 (≥ 30), where the numbers in parentheses specify frequency ranges. Similarly, we re-scale AE for a word to be the sequence number of the utterance in which the word is encountered for the first time, divided by 500 (e.g., all words in utterances 1 to 499 will have an AE of 0).

Modeling Effects of Context Familiarity

AE has been identified as an important factor in word learning (Carey & Bartlett, 1978; Gershkoff-Stowe & Hahn, 2007). However, it is usually confounded with CF since a later AE entails that there are generally more familiar words in the input. It is thus important to examine the independent role of CF and AE on word learning, as we see below.

After training our model on the 10,000 utterances in our child-directed corpus, we divide low frequency words (those with F < 4) into two groups according to how well they are learned: one group with a high comprehension score (Comp ≥ 0.9), and another group with a lower comprehension score (Comp < 0.9). Table 1 summarizes the averages of the different factors for the two groups. Interestingly, although F and CD are similar for both groups, we observe a substantial difference in the average Comp scores (0.93 vs. 0.52), suggesting that a factor other than F and CD must be responsible for this difference in learning. Looking at CF and AE, we can see an effect for both: words that have a high Comp score also tend to have higher CF and AE. That is, the words that are learned more confidently are those that have occurred in contexts with greater familiarity and that are first seen at a later age (i.e., when more words have been learned).

We now examine the independent effects of CF and AE on the acquisition of low frequency words, by holding one fac-

Table 2: Average AE, CF, and Comp for two groups of low frequency words: High CF vs. Low CF when AE is held constant (top part); and High AE vs. Low AE, when CF is held constant (bottom part). Number of words in each group is given in parenthesis.

	High CF	Low CF
	CF $\geq$ 4.5 (313)	CF $<$ 4.5 (160)
AE	9.90 $\pm$ 2.60	9.68 $\pm$ 2.49
CF	4.84 $\pm$ 0.17	3.98 $\pm$ 0.38
Comp	0.93 $\pm$ 0.02	0.77 $\pm$ 0.22

	High AE	Low AE
	AE $\geq$ 9 (78)	AE $<$ 9 (143)
CF	3.50 $\pm$ 0.38	3.43 $\pm$ 0.41
AE	13.62 $\pm$ 2.95	2.15 $\pm$ 2.22
Comp	0.60 $\pm$ 0.20	0.62 $\pm$ 0.21

tor constant (fixed within a range), and looking at the effect of the other factor. First, we consider low frequency words with AE values within a fixed range (here $5 < \text{AE} < 15$), and divide them into two groups based on their CF (Table 2: top part). Second, we hold CF constant within a fixed range ($2 < \text{CF} < 4$), and divide words into two groups with high and low AE (Table 2: bottom part). (Note that F and CD are the same for the two groups in both conditions.) We find that words that have occurred with differing CF values (top of Table 2) show an effect on their Comp score, with much better learning when the context familiarity is higher. On the other hand, words that have occurred with differing AE values (but with similar CF; bottom of Table 2) show no difference in learning at the different ages of exposure. These results show that CF has an independent and positive effect on the acquisition of low frequency words, whereas AE does not. We suggest that the effect we previously observed for AE (Table 1) is mostly through its effect on CF: since the model/learner learns more and more words over time, words encountered later (with higher AE) are in general more likely to appear with other familiar words, and thus to have a higher CF.

Conclusions

We have used an incremental probabilistic model of cross-situational word learning to study the effects of various statistical properties of the input on the acquisition of low frequency words. This is especially important since a large proportion of words in the input children receive have a very low frequency of occurrence. Replicating the results of a set of psychological experiments on artificial word learning (Kachergis et al., 2009), we argue that different factors affect the acquisition of high and low frequency words. These results and our findings through further experiments on natural child-directed utterances suggest that, for medium and high frequency words, the diversity in the context has a positive effect on learning (as also noted by Kachergis et al.), whereas for low frequency words it is the familiarity of the context that greatly impacts their acquisition.

These effects can be explained as a natural consequence of the interactions between two sets of probabilities that our model acquires over time. Through these interactions, our model draws on its own acquired knowledge of word meanings to boost the learning of other (novel) words. Thus, the acquisition of a set of high frequency words helps learn low frequency words by increasing their context familiarity. Generally, our model learns word meanings by drawing on the statistical regularities found in the input, and without incorporating any specific word learning biases or constraints, thus making the model appropriate for conducting studies on the relation between input properties and word learning.

References

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word learning: What probabilities tell us. In *Proceedings of CoNLL'08* (pp. 57–64).

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on Child Lang. Dev.*, *15*, 17–29.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*, 417–431.

Fazly, A., Alishahi, A., & Stevenson, S. (n.d.). A probabilistic computational model of cross-situational word learning. *Cognitive Science: An Interdisciplinary Journal*.

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proceedings of CogSci'08*.

Gershkoff-Stowe, L., & Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*, *50*, 682–697.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psych.*, *27*(2), 236–248.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In *Proceedings of CogSci'09*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database).

Naigles, L., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Quine, W. (1960). *Word and object*. The MIT Press.

Smith, L. B., & Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In *Proceedings of CogSci'07*.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.