Learning grammatical categories using paradigmatic representations: Substitute words for language acquisition

Mehmet Ali Yatbaz[a,*], Volkan Cirik[a], Deniz Yuret[a]

[a]*Koç University, Istanbul, Turkey*

Abstract

We study the paradigmatic representations of contexts for grammatical category acquisition in child directed speech. Learning syntactic categories is an essential task in language acquisition. Previous studies show that co-occurrence patterns or distributional knowledge could be handy for this task. These studies usually use co-occurrences of the preceding and following words to group words. However, the neighbouring words, or frames, are not able to exchange information when there is not enough overlapping between frames. In this work, we propose to use paradigmatic representations of words which are the probable substitutes of words instead of frames. Our experiments on child directed speech show that the probable substitutes are better than frames in terms of accuracy and robust in terms of the parameters on which they depend.

Keywords: Language acquisition, Grammatical categorization, Distributional information, Corpus analysis, Computational modeling, Paradigmatic approach

1. Introduction

Grammatical rules apply not to individual words (e.g. baby, talk) but to grammatical categories (e.g. noun, verb). Grammatical categories represent

*Corresponding author. Address: Department of Computer Engineering, Koç University, 34450, Istanbul, Turkey

Email addresses: myatbaz@ku.edu.tr (Mehmet Ali Yatbaz), vcirik@ku.edu.tr (Volkan Cirik), dyuret@ku.edu.tr (Deniz Yuret)

the group of words that can be substituted for one another without altering the grammatical correctness of a sentence. Therefore, learning grammatical categories is an important step in language acquisition. Studies on learning grammatical categories have shown that distributional information of word co-occurrences is one of the reliable cues (Mintz, 2003; St Clair, Monaghan & Christiansen, 2010; Redington, Chater & Finch, 1998). There is also evidence that lexical stress, prosodic and phonological cues are beneficial in learning grammatical categories (Monaghan & Mattock, 2012; Saffran, Newport & Aslin, 1996b,b).

The distributional representation of word co-occurrences can be grouped into two: syntagmatic and paradigmatic. The syntagmatic representation relates words according to the co-occurrences with the neighboring words while the paradigmatic representation relates the words that can be substituted for one another in a given context.

In this paper we hypothesize that infants represent contexts as substitute word distributions and form grammatical categories accordingly. The following two examples[1] illustrate the advantage of paradigmatic representations in uncovering similarities where no overt similarity that can be captured by a syntagmatic representation exists. The word "you" from the first sentence and the word "I" from the second sentence have no common neighbors no matter how large the context is. The paradigmatic representation captures the similarity of these words by suggesting the similar top substitutes for both (the numbers in parentheses give substitute probabilities):

(1) *"they fall out when* **you** *put it in the box ."*
you: you(.8188), I(.1027), they(.0408), we(.0146) ...

(2) *"what have* **I** *got here ?"*
I: we(.8074), you(.1213), I(.0638), they(.0073) ...

Note that substitute word distribution of a context (*"fall out when _ put it in"*) is independent of the actual word (i.e., *"you"*). The high probability substitutes reflect both semantic and grammatical properties of the context. Top substitutes for "I" and "you" are not only pronouns, but specifically pronouns compatible with the semantic context. Top substitutes for the word

[1]These examples are extracted from the Anne corpus and substitute word probabilities are calculated as described in Section 3.

"fall" in the first example consist of words that are also verbs: come(.7875), go(.0305), fall(.0232), were(.0187) ....

These examples show that the paradigmatic representation relates words according to the substitute word distribution of their context even when the surface forms of the contexts do not have any common words. Thus, this makes the paradigmatic representation more robust to the data sparsity compared to the syntagmatic representation.

In *the syntagmatic representation* the context is defined with the neighboring words, typically co-occurrences with a single word on the left or a single word on the right word called a "frame" (e.g., **the** *dog* **is**; **the** *cat* **is**) (Schütze & Pedersen, 1993; Redington, Chater & Finch, 1998; Mintz, 2003; St Clair, Monaghan & Christiansen, 2010; Lamar, Maron, Johnson & Bienenstock, 2010; Maron, Lamar & Bienenstock, 2010). A common limitation of syntagmatic representation is that it is not possible to exchange information between the words without common neighbors. To increase the chance of having common neighbors the frame size can be increased however this will introduce the data sparsity due to the low re-occurrence frequency of large frames (Manning & Schütze, 1999).

Mintz (2003) showed that non-adjacent high frequent bi-gram frames, "aXb" where a and b are the left and the right bigrams, respectively, are very informative to language learners on grammatical categorization of the middle tokens, X. The main limitation of this approach is that using only top-N *frequent frames* introduces a coverage problem while using all of the frames introduces frame sparsity. St Clair et al. (2010) overcame the limitations of frequent frames (aXb) by introducing *flexible frames* ($aX + Xb$) which represent left and right frames separately. They report improvements over aXb in terms of accuracy and coverage.

In *the paradigmatic representation* the context is defined as the distribution of the substitute words in that context (Schütze & Pedersen, 1993; Schütze, 1995; Yatbaz, Sert & Yuret, 2012). In part-of-speech induction literature Schütze (1995) incorporated paradigmatic information by concatenating the left and the right co-occurrence vectors of the right and left neighbors, respectively and grouped the words that have similar vectors. The limitation of this study is that Schutze uses the bi-gram information and suffers from sparsity as the context size gets larger. Yatbaz et al. (2012) took a more formal approach and calculate the most probable substitutes of a given context using a 4-gram statistical language model. Their model achieves the state-of-the-art result in the part-of-speech induction literature. To the best of our knowledge, no paradigmatic representation based child language acquisition models exists.

In this paper, we introduce a novel representation and hypothesize that children represent context by constructing substitute word distributions. To accomplish this, we adopt the learning framework of (St Clair, Monaghan & Christiansen, 2010) and compare our new representation with the former syntagmatic based representations.

The rest of the paper is organized as follows. In the following section, we explain other distributional approaches to grammatical category acquisition. Secondly, we provide a detailed explanation of our calculations of substitute words. In the experiments section, first we introduce the experimental setup we used and give the details of the experiments to contrast the paradigmatic approach with the syntagmatic approach. Lastly, we provide a general discussion on the findings of our results.

2. Related Work

In this section we present distributional approaches to the grammatical category acquisition. First we introduce the distributional hypothesis and the measure of success in distributional approaches. Then, we introduce previous distributional studies.

Distributional hypothesis or knowledge is a statistical approach to natural language. The distributional hypothesis suggests that words occurring in similar contexts tend to have similar meaning and grammatical properties (Harris, 1954). The two success criteria defined for distributional approaches to syntactic category acquisition are accuracy and completeness. Accuracy measures how accurate the predictions were at grouping the words into the same grammatical categoty together. It is defined as the total number of correct category predictions over total number of predictions. Completeness, on the other hand, measures how well a given category is predicted. The completeness is equal to the number of correct predictions for a category divided by number of correct predictions summed with number misses in that category.

Redington et al. (1998) defines the context of a word as the previous and following words. With this definition, they construct context vectors of target words for clustering. Using average link clustering with a threshold maximizing accuracy and completeness, target words are separated into categories. Although the categorizations are generally accurate, the method lacks of completeness. In addition, the underlying process to determine the threshold for infants is not clear (Ambridge & Lieven, 2011). Cartwright & Brent (1997) introduce an incremental learning framework for syntactic category acquisition. Their results are similar to Redington et al. (1998), high in accuracy but low in completeness. Mintz (2003) proposes frequent frames. A frequent frame consists of two jointly appearing words with one word in the middle cooccurring frequently. Experiments on child directed speech reveal that frequent frames have the ability to assign word categories with high accuracy. Though the accuracy is high, it suffers from completeness. As St Clair et al. (2010) point out, frequent frames suffer from coverage. St Clair et al. (2010) combine the bigram's coverage power (Redington, Chater & Finch, 1998) and (Monaghan & Christiansen, 2008) and accuracy of frequent frames (Mintz, 2003). The experiments demonstrate that infants make use of both bigram and trigram sources. As St Clair et al. (2010) pointed out, language learners may even use higher-order relationships between words.

Freudenthal et al. (2005) identify a complication of distributional methods for constructing syntactic categories. Distributional methods suggest that words occurring in a similar context can be used interchangebly. They claim the evaluation methods used in studies like (Redington, Chater & Finch, 1998; Monaghan & Christiansen, 2008; Mintz, 2003) could be misleading. Specifically, if a word is substituted with another one in its category, the resulting sentences could be erroneous in a way that they are not observed in infants' speech. As a success criteria, they argue that the proposed categorization should generate plausible sentences. They introduce a chunking mechanism merging words that are seen frequently. The mechanism is successful in generating meaningful sentences, still, the proposed solution is computationally complex to disclose the learning mechanism in infants.

More recently, Alishahi & Chrupała (2012) propose an incremental learning scheme inducing soft word categories while learning the meaning of words. Thothathiri et al. (2012) examine the role of prosody on infants' distributional learning of syntactic categories and concludes that the prosody shows little influence. Reeder et al. (2013) discuss to answer the use of distributional knowledge when the evidence on the possible context of a word is not enough. Furthermore, they explain how and when language users form new categories depending on the overlaps between the context words.

3. Substitute Words

In this study, we predict the syntactic category of a word in a given context based on its most likely substitute words. St Clair et al. (2010) demonstrated that learning left and right bigrams together was much more effective than learning them individually. Thus it is best to use both the left and the right context when estimating the probabilities for potential lexical substitutes. For example, in *"He lived in San Francisco suburbs."*, the token *San* would be difficult to guess from the left context but it is almost certain looking at the right context.

We define the context c_w of a given word w as the $2n-1$ word window centered around the position of $w : w_{-n+1} \ldots w \ldots w_{n-1}$. The probability of a substitute word w in a given context c_w is estimated as:

$$\begin{aligned} P(w_0 = w|c_w) &\propto P(w_{-n+1} \ldots w_0 \ldots w_{n-1}) & (1) \\ &= P(w_{-n+1})P(w_{-n+2}|w_{-n+1}) \ldots P(w_{n-1}|w_{-n+1}^{n-2}) & (2) \\ &\approx P(w_0|w_{-n+1}^{-1})P(w_1|w_{-n+2}^{0}) \ldots P(w_{n-1}|w^{n-2}) & (3) \end{aligned}$$

where w_i^j represents the sequence of words $w_i w_{i+1} \ldots w_j$. In Equation 1, $P(w|c_w)$ is proportional to $P(w_{-n+1} \ldots w_0 \ldots w_{n+1})$ because the words of the context are fixed. Terms without w_0 are identical for each substitute in Equation 2 therefore they have been dropped in Equation 3. Finally, because of the Markov property of n-gram language model, only the closest $n-1$ words are used in the experiments. Note that the substitute word distribution is a function of the context only and is indifferent to the target word.

Near the sentence boundaries the appropriate terms were truncated in Equation 3. Specifically, at the beginning of the sentence shorter n-gram contexts were used and at the end of the sentence terms beyond the end-of-sentence utterance were dropped.

4. Experimental Setup

In this section we explain the experimental setup we used. First we demonstrate how we process the input corpora. Secondly, we present the parameters used to train a language model and calculate substitute word probabilities. Lastly, we clarify the grammatical categories used for evaluation.

4.1. Input Corpora

In order to obtain comparable results with St Clair et al. (2010) and Mintz (2003), we use the same six corpora of child-directed speech from the CHILDES[2] corpus (MacWhinney, 2000): Anne and Aran (Theakston, Lieven, Pine & Rowland, 2001), Eve (Crystal, 1974), Naomi (Sachs, 1983), Nina (Suppes, 1974), Peter (Bloom, Hood & Lightbown, 1974; Bloom, Lightbown, Hood, Bowerman, Maratsos & Maratsos, 1975). Following Mintz (2003) we only analyze the adult utterances in sessions where the target child is 2.6 years old or younger. Prior to the analysis, we perform the data preprocessing detailed in Appendix A.

Table 1: Summary of the total number of tokens, utterances and types in each child corpus together with the number of utterances and types that are obserd as target word in aXb.

Corpus	Tokens	Utterances	Utterances Categorized		Types	Types Categorized	
			Count	%		Count	%
Anne	121726	93371	42789	45.82	2623	1846	70.37
Aran	129823	104997	54768	52.16	3256	2595	79.69
Eve	78778	59095	27315	46.22	2184	1465	67.07
Naomi	38302	28793	13002	45.15	1883	1194	63.40
Nina	89957	72879	39335	53.97	2036	1580	77.60
Peter	94521	72834	34997	48.05	2145	1472	68.62

Word sequences that consist of three words and do not contain any utterance boundaries are extracted from each child corpus separately (Mintz, 2003). The first and the third words of sequences are treated as frame elements while the middle utterance is the target word that is categorized. Table 1 summarizes the number of target word tokens and types in each corpus.

To calculate substitutes we extracted the 4-gram left and right contexts of each target word when they are available [3].

[2]Specifically, CHILDES version 2.0.1 is used in experiments.

[3]Lower order n-gram contexts are extracted when the 4-gram left or right context is not available.

4.1.1. Language Modeling for Substitute Words

We extracted training data of approximately 6.8 million tokens[4] of child-directed speech data from CHILDES following the steps defined in Appendix A. To calculate substitute word probabilities, we train a 4-gram language model with Kneser-Ney discounting on the training data using SRILM (Stolcke, 2002). Words that were observed less than 2 times in the language model training data were replaced with an unknown word tag <UNK>, which gave us a vocabulary size of 21734.

4.1.2. Grammatical categories and Evaluation

The grammatical category of words in CHILDES are extracted by first applying the MOR parser (MacWhinney, 2000) and then using the POST disambiguator (Sagae, MacWhinney & Lavie, 2004). The accuracy of CHILDES grammatical categories is approximately 95% (Parisse et al., 2000) and is encoded in the MOR line of the CHILDES corpus.

To evaluate classification accuracy we use the standard labeling (Mintz, 2003)[5] that categorizes target words as: nouns (including pronouns), verbs (including copula and auxiliaries forms), prepositions, adjectives, adverbs, determiners, conjunctions, wh-words, negation (i.e., "not") and interjections.

4.2. Computational Modeling Algorithm

St Clair et al. (2010) used a feed-forward connectionist model to compare the effect of distributional cues from various frame types on the grammatical category learning. We adopt their framework to compare the paradigmatic representation (substitute words) with the best performing syntagmatic representation (i.e., flexible frames).

A prototypical connectionist model consists of input, hidden and output layers. Input and output layers are connected to each other through the hidden layer. The behavior of the output units are determined by the activity of the hidden layers which is triggered by the input layer.

We train separate connectionist models to compare flexible frames ($aX + Xb$) to the substitute words ($a{*}b$). For each model we input the distributional information to the feed-forward connectionist model in the following way:

[4] Anne, Aran, Eve, Naomi and Peter corpora are excluded.

[5] Mintz (2003) also defined an expanded labeling in which pro-nouns, auxiliaries and copula forms have their own categories.

- $aX+Xb$ **model:** The first and second half of the input units correspond to the preceding bigram (a) and the succeeding bigram (b), respectively. Thus two input units are activated for each target word.

- $a*b$ **model:** Each input unit represents a distinct substitute and input units that correspond to the substitutes of the target word are set to the number of their occurrences in the sampled substitute set.

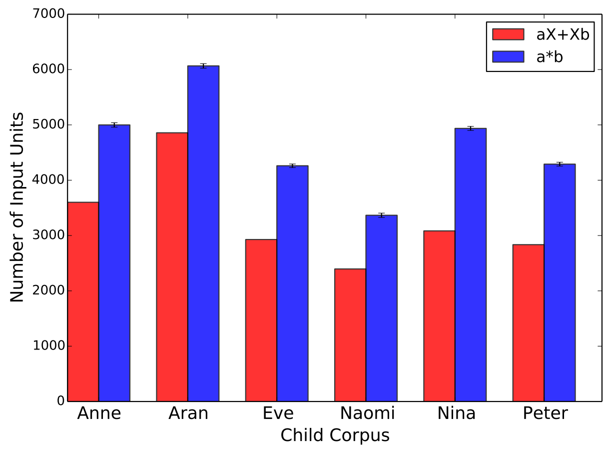


Figure 1: Number of input layer units of the flexible frame ($aX+Xb$) and the substitute based model($a*b$) are summarized. $a*b$ samples 16 substitutes per target word. Standard errors are reported with error bars.

Table 1 presents the number of input layer units of syntagmatic and paradigmatic representation based models on each child corpus separately. The number of distinct frames is fixed for any given corpus while the number of distinct substitutes varies due to the random sampling.

Each output unit represents a distinct grammatical category; therefore, the models are expected to produce only one active (non-zero) output unit for each target word. If there are more than one active units present in the output layer[6], the target word is assigned to the corresponding grammatical category of the output unit with the largest value.

Both models have 10 output units due to the standard labeling (Mintz, 2003).

[6]why neural network produces more than one active unit.

Unless stated otherwise, all connectionist models in this paper use the following parameters: (1)number of hidden units is set to 200 and initialized randomly for each model. (2) The backpropagation parameter is set to 0.1. The learning function is sigmoid and the learning rate is 0.2(!!).

4.3. Training and Testing

We measure and compare the classification accuracy of models by applying 10-fold cross validation on the union of six child corpora. To perform 10-fold cross validation we randomly split each child corpus into 10 folds. At each iteration a single fold from each child corpus is kept as the test data while the union of the remaining 9 folds of each child corpus are used as the training data. We repeat this process until all folds are used exactly once as the test data and report the average accuracy of 10 runs on each child corpus separately. The main advantage of the cross validation is that all sentences are eventually used both for testing and training. [!! citation]

To compare the effects of paradigmatic representation ($a{*}b$) with the syntagmatic one ($aX + Xb$) we train and test both models using the identical 10-fold cross validation split. Thus every model in this paper is exposed to the identical sequence of training and testing patterns. Unless stated otherwise, in the rest of this paper, we stopped the training phase of feed-forward connectionist model on each corpus after $100K$ input patterns, used the standard labeling to evaluate model accuracies, calculated substitute distributions with with the language model defined in Section 4.1.1 and sampled 16 substitutes per target word in models using the paradigmatic representation.

In the next section we replicate the corpus analysis of Mintz (2003) and St Clair et al. (2010). Section 5 compares the classification accuracies of syntagmatic and paradigmatic representation based models. The effects of the number of substitutes and the language model n-gram order on the paradigmatic model performance are inspected in Section 6 and 7, respectively.

5. Experiment 1: Syntagmatic vs Paradigmatic

In order to compare the distributional information of syntagmatic and paradigmatic representations we train separate feed-forward connectionist models for each child corpus based on these representations. St Clair et al. (2010) showed that flexible frames have richer distributional information than other frame types both in terms of classification accuracy and coverage . Thus we only report results of the models based on substitute words ($a * b$) and flexible frames ($aX + Xb$)[7].

5.1. Method

All models are trained and evaluated according to steps summarized in Section 4.3. Similar to the analysis in (St Clair, Monaghan & Christiansen, 2010), we split the training phase of each model into two as short and long training phases in which we stop and evaluate the models on the corresponding test sets after presenting identical 10K and 100K training patterns, respectively.

5.2. Results of Short Training Phase

Table 2 gives the overall classification accuracies of $aX + Xb$ and $a * b$ models on each child corpus. The accuracy of $a * b$ model significantly outperforms the $aX + Xb$ model on each child corpora even with a limited amount of training patters. Lambdas of the $a*b$ model are significantly closer to the perfect association than lambdas of the $aX + Xb$ model. Lambdas of both models are significantly different from zero association.

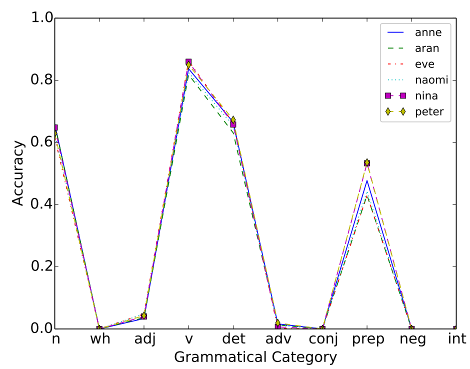
Table 2: 10-fold cross-validation classification accuracies of models based on flexible frames ($aX + Xb$) and substitutes ($a * b$) on each child corpus after 10K training patterns are summarized. Standard errors are reported in parentheses. Lambdas of $aX + Xb$ and $a * b$ are both tested against each other and zero association by using two tailed z-test. All tests have $p < .001$.

Corpus	$aX + Xb$		$a * b$	
	Accuracy	λ	Accuracy	λ
Anne	.6252 (.0231)	.4323 (.0352)	.7970 (.0069)	.6925 (.0111)
Aran	.5968 (.0218)	.3908 (.0327)	.7783 (.0083)	.6653 (.0123)
Eve	.6193 (.0192)	.4248 (.0306)	.8091 (.0100)	.7116 (.0141)
Naomi	.6054 (.0236)	.3960 (.0395)	.7771 (.0100)	.6598 (.0178)
Nina	.6438 (.0216)	.4521 (.0362)	.8146 (.0096)	.7150 (.0162)
Peter	.6255 (.0246)	.4402 (.0372)	.8086 (.0088)	.7140 (.0130)

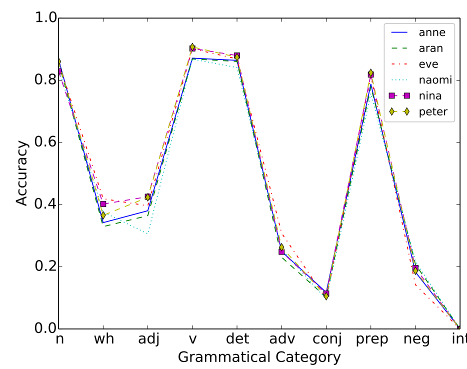
To further investigate the accuracy gap between $aX + Xb$ and $a * b$ models, we plot the classification accuracies of each grammatical category in

[7]We can put the comparison with other frames in Appendix.

the standard labeling for both models in Figure 2. Even after 10K training patterns both models are able to achieve relatively higher accuracies on nouns(n), verbs(v), determiners(det) and prepositions($prep$) than the rest of the grammatical categories. The $a * b$ model is more successful than the $aX + Xb$ model in learning grammatical categories such as wh-words(wh), adjectives(adj), adverbs(adv), conjunctions($conj$) and negations(neg).



(a) $aX + Xb$



(b) $a * b$

Figure 2: Individual tag accuracies of $aX + Xb$ and $a * b$ on each child corpus after $10K$ training patterns are presented.

Finally, with limited amount of training patterns, the $a * b$ model is able to categorize nine out of ten grammatical categories in each child corpus with different levels of accuracies. On the other hand, the $aX + Xb$ model performs poorly on *wh*, *conj*, *adv*, *neg* and *int* and can not correctly classify any members of these grammatical groups in at least one of the child corpora.

5.3. Results of Long Training Phase

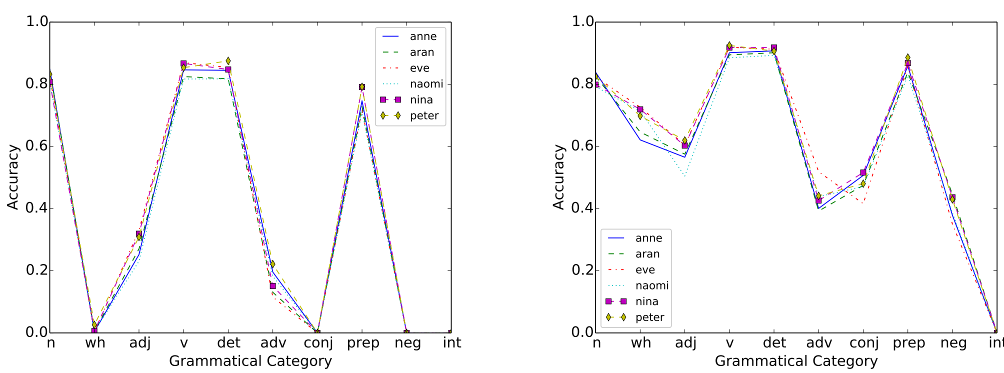
Previous section shows that the $a * b$ model is more accurate than the $aX + Xb$ model on learning grammatical categories with limited amount of language exposure. In this section each model is trained with 100K input patterns to observe the effect of extensive language exposure on learning.

Table 3 summarizes the overall classification accuracies of $aX + Xb$ and $a*b$ models on each child corpus. Although differences between corresponding accuracies and lambda values of $aX + Xb$ and $a * b$ models are less than 10K experiments, the $a * b$ model is still significantly more accurate than the $aX + Xb$ model on all child corpora. The $a * b$ model benefit less from the

Table 3: 10-fold cross-validation classification accuracies of models based on flexible frames ($aX + Xb$) and substitutes ($a * b$) on each child corpus after 100K training patterns are summarized. Standard errors are reported in parentheses. Lambdas of $aX + Xb$ and $a * b$ are both tested against each other and zero association by using two tailed z-test. All tests have $p < .001$.

Corpus	$aX + Xb$		$a * b$	
	Accuracy	λ	Accuracy	λ
Anne	.7628 (.0075)	.6407 (.0124)	.8311 (.0068)	.7442 (.0109)
Aran	.7337 (.0059)	.5977 (.0081)	.8139 (.0073)	.7189 (.0108)
Eve	.7580 (.0068)	.6351 (.0083)	.8396 (.0107)	.7576 (.0160)
Naomi	.7316 (.0086)	.5892 (.0113)	.8041 (.0090)	.7000 (.0169)
Nina	.7755 (.0040)	.6547 (.0075)	.8389 (.0097)	.7523 (.0165)
Peter	.7670 (.0071)	.6518 (.0088)	.8379 (.0073)	.7579 (.0112)

extensive training than the $aX + Xb$ model. One possible explanation for this behavior is that the number of input units of the $a * b$ model on each child corpus is significantly higher than the $aX + Xb$ (see Figure 1) while the number of hidden units is fixed to 200 for both models. As St Clair et al. (2010) did, we experiment with the number of hidden units such that the ratio between the number of input units and the number of hidden units is the same for both models. We do not observe significant changes on the result.



(a) $aX + Xb$ (b) $a * b$

Figure 3: Individual tag accuracies of $aX + Xb$ and $a * b$ on each child corpus after $100K$ training patterns are presented.

In contrast to the 100K results, $aX + Xb$ model performs poorly only on *conj* and *int* as shown in Figure 3. Both models accurately learn the noun, verb, determiner and preposition groups. However, $a * b$ models still significantly accurate on adjectives, conjunctions and negations.

6. Experiment 2: Number of Substitutes

In this experiment we analyze the effects of number of substitutes both on the number of input units and the model classification accuracies. Aside from the effect on classification accuracies, the number of sampled substitutes also varies the number of active and non-active units in the input layer.

6.1. Method

We used the same experimental settings except that the number of substitutes per target word is varied between 1 and 64[8].

6.2. Results and discussion

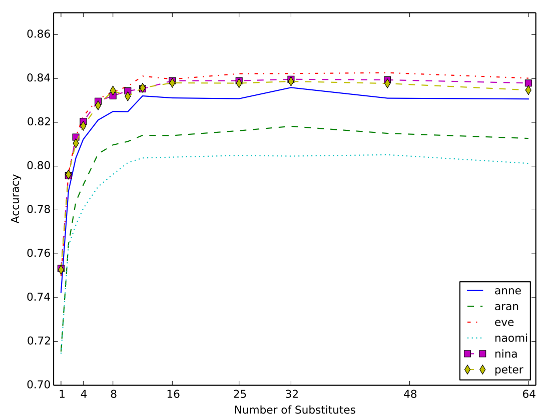


Figure 4: 10-fold cross validation accuracy of each child corpus for different number of substitutes.

Figure 4 plots the model classification accuracy of each child corpus versus the number of substitutes. The classification accuracy dramatically increases

[8]We do not observe any significant difference on model classification accuracies for the number of substitutes that are more than 64.

on each child corpus until the number of substitutes reaches 16. After 16 substitutes, increasing the number of substitutes does not significantly change the classification accuracy. Thus, the model is fairly robust to the number of substitutes as long as the model can observe at least 16 substitutes per target word.

One possible problem of these models is that the number of input units increases with the increasing number of substitutes meanwhile the number of hidden units is fixed to 200. St Clair et al. (2010) discussed this problem while comparing flexible frames with other frames and solved it by setting the number of hidden units such that the ratio between the number of hidden and input units is fixed for each model. Although they reported slight improvements over the versions with fixed number of hidden units, the classification accuracy ranking of the models did not change.

In the next experiment we analyze the effect of substitute word quality on the classification accuracy of the paradigmatic model by experimenting with the n-gram order of the language model.

7. Experiment 3: Language Model N-gram Order

In this set of experiments, we test the paradigmatic model by changing the n-gram order of the language model that are used to sample substitutes. A language model defines probabilities for the sequences of strings in a language. The n-gram order of language model determines the number of preceding items taken into account while determining the probability of the upcoming word. The previous studies show that young children are sensitive to statistical properties of language (Saffran, Aslin & Newport, 1996a) and are able to store 4-word sequences (Bannard & Matthews, 2008). Experiments with adults also suggest that the language users are sensitive to co-occurrence patterns beyond bigram (Arnon & Snider, 2010).

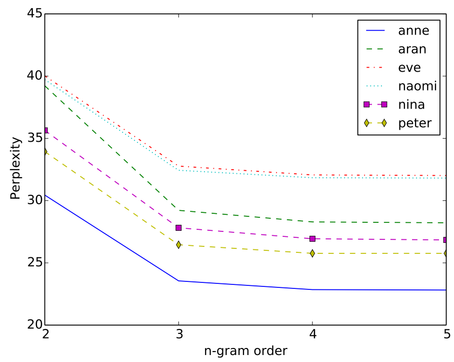
The perplexity of the language model is a measurement of the number of words that can be observed in a given n-gram context window and determined by n-gram order of the language model. Therefore, as the n-gram order increases the model assigns more relevant substitutes to the context[9].

[9]Goodman (2001) showed that the perplexity plateaued when the order is higher than 5.

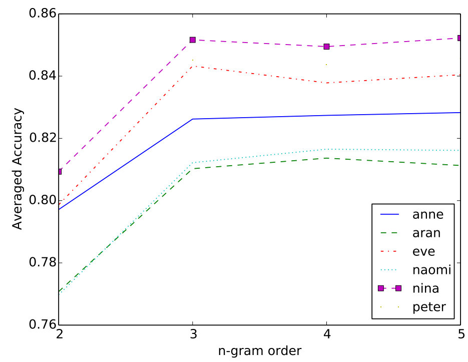
7.1. Method

We used the same experimental settings except that the n-gram order of the language model that is used to sample substitutes is varied from 2 to 5.

7.2. Results and discussion



(a)



(b)

Figure 5: Language Model perplexities on each child corpus for different n-gram orders are presented on the left figure while 10-fold cross validation accuracies calculated based on these models are presented on the right.

The perplexity of each child corpus is dramatically improved when the n-gram order of the language model is increased from 2 to 3 and varies slightly for orders higher than 3. Figure 5(a) plots the perplexity versus the n-gram order. As shown in Figure 5(b), the model classification accuracies on each child corpus are slightly improved for orders higher than 3 which is in fact parallel to the perplexity trends in Figure 5(a). Overall, the classification accuracy of paradigmatic model is highly correlated with the perplexity of the language model that is used to sample substitutes.

8. General Discussion

This study proposes paradigmatic representations of context as opposed to syntagmatic representations for syntactic category acquisition. The paradigmatic approach suggests using probable substitutes of word $(a * b)$. On the other hand, the syntagmatic approach proposes using the preceding bigram and the succeeding bigram, whichever is fruitful $(aX + Xb)$.

To contrast these two representations we replicate the experimental setup of St Clair et al. (2010). Experiments show that when the models exposed to limited amount of training patterns the $a * b$ is significantly more accurate than $aX + Xb$. Results of the long training phase show the same pattern, however, the gap between these approaches decreases.

We investigate the dependency of the model to the number of substitutes. In this experimental setup the number of substitutes varies from 1 to 64. The results show that the accuracy of the model dramatically increases up to 16. After 16 substitutes, no significant improvement in accuracy is observed. We conclude that the model is robust as long as 16 substitutes are observed.

We explore the effect of the n-gram order of language model to the accuracy of the model. While determining the probability of the next word in a sequence of words, n-gram order determines how many preceding word should be used. We hypothesise that the order of n-gram determines how accurate the substitutes of a target word. Thus, it should affect the $(a * b)$ model's accuracy. Figure 5(a) and Figure 5(a) show that the model's performance dependens on the n-gram order of the language model.

Appendix A. Preprocessing

We apply following preprocessing steps initial to our analysis:

- All punctuation, pause, trailing off and interruption marks are treated as utterance boundaries.
- Repetitions of a word are kept in the text and their grammatical categories are automatically set to the grammatical category of the original word.
- Words that are grammatically necessary but not spoken are deleted (grammatical omissions).
- Shortenings, dropping sounds out of words, are ignored and converted to the corresponding actual word forms.
- Untranscribed words such as *xxx* or *yyy* are removed.
- Assimilations, complex sound changes of words or word phrases, are not converted to the actual form.

Appendix B. Frequent Frames

In this section we replicated St Clair et al. (2010) to compare the amount of categorical information provided by the top-45 fixed and bi-gram frames.

Table B.4: Summary of the total number of utterances and types in each child corpus. For the sake of space, we only report the percentages of analyzed utterances(types) in the top-45 aXb, aX and Xb.

Corpus	Corpus Utterances(Types)	Analyzed Utterances(Types)		
		aXb	aX	Xb
Anne	93371(2623)	.0462(.1357)	.3994(.8147)	.3619(.6465)
Aran	104997(3256)	.0537(.1901)	.4383(.8353)	.4026(.6670)
Eve	59095(2184)	.0595(.1735)	.4097(.7770)	.3505(.5650)
Naomi	28793(1883)	.0572(.1603)	.3988(.7785)	.3455(.5586)
Nina	72879(2036)	.0842(.2249)	.4805(.8560)	.4028(.7062)
Peter	72834(2145)	.0671(.1762)	.4318(.8027)	.3770(.6317)

Table B.5: aXb

Corpus	Token Accuracy	Type Accuracy	Token Completeness	Type Completeness
Anne	.9693(.3909)	.8870(.4209)	.0756(.0221)	.0864(.0221)
Aran	.9527(.4166)	.8582(.4096)	.0794(.0221)	.0819(.0226)
Eve	.9731(.4935)	.8973(.4895)	.0645(.0222)	.0681(.0226)
Naomi	.9496(.4858)	.8910(.4983)	.0650(.0219)	.0630(.0224)
Nina	.9615(.4782)	.8855(.4616)	.0787(.0221)	.0902(.0219)
Peter	.9468(.4600)	.8615(.5249)	.0586(.0222)	.0739(.0217)

Appendix C. Experiment 4: Corpus analysis

Table B.6: aX

Corpus	Token Accuracy	Type Accuracy	Token Completeness	Type Completeness
Anne	.6348(.2654)	.5667(.3104)	.0850(.0219)	.0694(.0221)
Aran	.5939(.2582)	.5472(.3000)	.0791(.0220)	.0727(.0220)
Eve	.6775(.2735)	.5954(.2966)	.1002(.0221)	.0752(.0221)
Naomi	.6509(.2754)	.5996(.3082)	.1017(.0220)	.0821(.0222)
Nina	.6809(.2877)	.6287(.3525)	.1073(.0220)	.0745(.0221)
Peter	.6527(.2618)	.5043(.2715)	.1103(.0220)	.0715(.0221)

Table B.7: Xb

Corpus	Token Accuracy	Type Accuracy	Token Completeness	Type Completeness
Anne	.4462(.2613)	.4048(.2920)	.0651(.0651)	.0426(.0426)
Aran	.4758(.2755)	.4142(.3045)	.0733(.0733)	.0443(.0443)
Eve	.4492(.2601)	.3960(.2851)	.0676(.0676)	.0470(.0470)
Naomi	.4532(.2602)	.3717(.2764)	.0740(.0740)	.0438(.0438)
Nina	.4837(.2650)	.4530(.3375)	.0867(.0867)	.0458(.0458)
Peter	.4368(.2617)	.3500(.2702)	.0744(.0744)	.0417(.0417)

Table C.8: 10-fold cross-validation classification accuracies of models based on flexible frames ($aX + Xb$) and substitutes ($a * b$) on each child corpus after 100K training patterns are summarized. Standard errors are reported in parentheses. Lambdas of $aX + Xb$ and $a * b$ are both tested against each other and zero association by using two tailed z-test. All tests have $p < .001$.

Corpus	aXb		$aX + Xb$	
	Accuracy	λ	Accuracy	λ
Anne	.5416 (.0224)	.3099 (.0255)	.7628 (.0075)	.6407 (.0124)
Aran	.5156 (.0215)	.2837 (.0120)	.7337 (.0059)	.5977 (.0081)
Eve	.5370 (.0258)	.3209 (.0130)	.7580 (.0068)	.6351 (.0083)
Naomi	.5229 (.0244)	.2840 (.0220)	.7316 (.0086)	.5892 (.0113)
Nina	.5636 (.0113)	.3287 (.0183)	.7755 (.0040)	.6547 (.0075)
Peter	.5661 (.0180)	.3541 (.0206)	.7670 (.0071)	.6518 (.0088)

References

Alishahi, A., & Chrupała, G. (2012). Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 643–654). Association for Computational Linguistics.

Ambridge, B., & Lieven, E. V. (2011). *Child Language Acquisition: Contrasting Theoretical Approaches*. Cambridge Univ Pr.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, *6*, 380 – 420. doi:10.1016/0010-0285(74)90018-8.

Bloom, L., Lightbown, P., Hood, L., Bowerman, M., Maratsos, M., & Maratsos, M. P. (1975). Structure and variation in child language. *Monographs of the society for Research in Child Development*, (pp. 1–97).

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *63*, 121–170.

Crystal, D. (1974). Roger brown, a first language: the early stages. cambridge, mass.: Harvard university press, 1973. pp. xi + 437. *Journal of Child Language*, *1*, 289–307. doi:10.1017/S030500090000074X.

Freudenthal, D., Pine, J. M., & Gobet, F. (2005). On the resolution of ambiguities in the extraction of syntactic categories through chunking. *Cognitive Systems Research*, *6*, 17–25.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*, 403 – 434. doi:http://dx.doi.org/10.1006/csla.2001.0174.

Harris, Z. S. (1954). Word. *Distributional Structure*, *10*, 146–162.

Lamar, M., Maron, Y., Johnson, M., & Bienenstock, E. (2010). Svd and clustering for unsupervised pos tagging. In *Proceedings of the ACL 2010 Conference Short Papers* (pp. 215–219). Uppsala, Sweden: Association for Computational Linguistics. URL: `http://www.aclweb.org/anthology/P10-2040`.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database* volume 2. Lawrence Erlbaum.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Maron, Y., Lamar, M., & Bienenstock, E. (2010). Sphere embedding: An application to part-of-speech induction. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 1567–1575).

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91 – 117. doi:10.1016/S0010-0277(03)00140-9.

Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. *Corpora in language acquisition research: History, methods, perspectives*, (pp. 139–164).

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word–referent mappings. *Cognition*, *123*, 133–143.

Parisse, C. et al. (2000). Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers*, *32*, 468–481.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *COGNITIVE SCIENCE*, *22*, 425–469.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*, 30–54.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Childrens language*, *4*.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606–621.

Sagae, K., MacWhinney, B., & Lavie, A. (2004). Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments, & Computers*, *36*, 113–126.

Schütze, H. (1995). Distributional part-of-speech tagging. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics* EACL '95 (pp. 141–148). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. URL: http://dx.doi.org/10.3115/976973.976994. doi:http://dx.doi.org/10.3115/976973.976994.

Schütze, H., & Pedersen, J. (1993). A Vector Model for syntagmatic and paradigmatic relatedness. In *Proceedings of the 9th Annual Conference of the University of Waterloo Centre for the New OED and Text Research*. Oxford, England.

St Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, *116*, 341–60.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing* (pp. 257–286).

Suppes, P. (1974). The semantics of childrens language. *American psychologist*, *29*, 103.

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.

Thothathiri, M., Snedeker, J., & Hannon, E. (2012). The effect of prosody on distributional learning in 12-to 13-month-old infants. *Infant and Child Development*, *21*, 135–145.

Yatbaz, M. A., Sert, E., & Yuret, D. (2012). Learning syntactic categories using paradigmatic representations of word context. In *EMNLP-CoNLL* (pp. 940–951).