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

Mehmet Ali Yatbaz[1]* **Volkan Cirik**[2]* **Aylin Küntay**[3] **Deniz Yuret**[3]
[1]Facebook Inc. 1 Hacker Way Menlo Park, CA, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]Koç University, İstanbul, Turkey
{myatbaz,vcirik,akuntay,dyuret}@ku.edu.tr

Abstract

Learning word categories is a fundamental task in language acquisition. Previous studies show that co-occurrence patterns of preceding and following words are essential to group words into categories. However, the neighboring words, or frames, are rarely repeated exactly in the data. This creates data sparsity and hampers learning for frame based models. In this work, we propose a paradigmatic representation of word context which uses probable substitutes instead of frames. Our experiments on child-directed speech show that models based on probable substitutes learn more accurate categories with fewer examples compared to models based on frames.

1 Introduction

Children abstract grammatical rules from individual words (e.g. baby, talk) and eventually apply them to word categories (e.g. noun, verb, adverb). A word category represents a group of words that can be substituted for one another without altering the grammatical appropriateness of a sentence. Learning word categories is an important step in language development.

The Distributional Hypothesis (Harris, 1954) suggests that words occurring in similar contexts tend to have similar meanings and grammatical properties. Studies on extraction of word categories have shown that distributional information of word co-occurrences is a reliable cue for the learning of word categories (Mintz, 2003; St Clair et al., 2010; Redington et al., 1998). Children need to extract word categories from incoming speech stream in order to be able to predict how words behave in various grammatical contexts and to produce words in appropriate grammatical constructions. Children tend to form word categories that group words used in similar contexts.

To judge how similar two contexts are, one can use syntagmatic or paradigmatic representations of the word context: A syntagmatic representation is based on the neighbors of the target word whereas a paradigmatic representation uses potential substitutes for the target word.

In this paper, we hypothesize that children judge context similarity using a paradigmatic representation: a context is similar to another if the same words can be substituted in both. The following two examples[1] illustrate the advantage of paradigmatic representations in uncovering latent similarities where a syntagmatic representation would fail to see any overt similarities. The word "you" from the first sentence and the word "I" from the second sentence have no common neighbors within the same sentence. The paradigmatic representation, shown below the sentences as substitute word probabilities, captures the similarity of these contexts by suggesting similar top substitutes for both:

(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) . . .

*Work was done when authors were at Koç University.

[1]These examples are extracted from the Anne corpus from CHILDES (MacWhinney, 2000). The computation of substitute word probabilities is described in Section 3.3.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 707–716, Osaka, Japan, December 11-17 2016.

The high probability substitutes reflect both semantic and grammatical properties of the context. The top substitutes for "I" and "you" are pronouns. As an additional example, the top substitutes for the word "fall" in the first sentence are other motion verbs: come(.7875), go(.0305), fall(.0232), . . .

These examples show that the paradigmatic representation can relate a pair of words according to the substitute word distribution of their contexts even when the surface forms of the contexts do not share any common neighbors. This makes the paradigmatic representations of word context more robust to the data sparsity compared to the syntagmatic representations, due to low re-occurrence frequency of large frames.

The rest of the paper is organized as follows. In the next section, we describe prior distributional approaches to word category acquisition. In Section 3, we provide a detailed explanation of our calculations of substitute words. In Section 4, first we introduce the experimental setup we used and give the details of the experiments to contrast the paradigmatic approach with the syntagmatic approach. The last section summarizes our contributions.

2 Related Work

In this section we review distributional approaches to word category acquisition and evaluate them based on two success criteria: accuracy and completeness. Accuracy measures how accurate the predictions were at grouping the words into the same word category together. Completeness, on the other hand, measures how well a given category is predicted.

Redington et al. (1998) define the context of a word as the previous and following words. They construct context vectors of target words for clustering. Using average link clustering, target words are separated into categories. Although the resulting categorizations are generally accurate, the method is weak in completeness because words that do not appear in frequent frames cannot be covered.

Mintz (2003) proposes top-N frequent frames surrounding a target word as a more fitting context to derive word categories from. A frequent frame consists of left and right neighbors that co-occur frequently. Experiments on child-directed speech reveal that frequent frames have the ability to assign word categories with high accuracy. However, this method also lacks satisfactory completeness. St Clair et al. (2010) combine the bigram's coverage power (Redington et al., 1998; Monaghan and Christiansen, 2008) and the accuracy of frequent frames (Mintz, 2003). Their experiments suggest that to match the performance of infants both bigram and trigram sources may need to be used.

Freudenthal et al. (2005) identify a complication of distributional methods for constructing word categories. Distributional methods suggest that words occurring in a similar context can be used interchangeably. They claim the evaluation methods used in studies like (Redington et al., 1998; Monaghan and 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 criterion, 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 too complex as a learning mechanism in infants.

Alishahi and 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 in infants' distributional learning of syntactic categories and concludes that the prosody shows little influence. Reeder et al. (2013) discuss the use of distributional knowledge when the evidence in 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.

In the related part-of-speech induction literature[2], Schütze (1995) incorporates paradigmatic information by concatenating the left and the right co-occurrence vectors of the right and left neighbors respectively, and groups words that have similar vectors. The limitation of this model is that it uses only bi-gram information and suffers from sparsity as the context size gets larger. Yatbaz et al. (2012) calculate the most probable substitutes of a given context using a 4-gram statistical language model. Their

[2]See (Christodoulopoulos et al., 2010) for a comprehensive review of the part-of-speech induction literature.

model achieves the state-of-the-art result in the part-of-speech induction literature. Part-of-speech induction aims to induce word-categories from large amounts of unannotated text (mostly news corpora). Our paper evaluates the substitute-based context representation by Yatbaz et al. (2012) as a possible feature for classifying words in relatively small amounts of child-directed speech.

3 Method

In this section we explain the experimental methodology we used, including how the input corpora was processed, the language model was trained, the evaluation metrics, and the computational model to learn grammatical categories.

3.1 Input Corpora

To compare results with (St Clair et al., 2010) and (Mintz, 2003), we use the same six corpora of child-directed speech from the CHILDES[3] corpus (MacWhinney, 2000): Anne and Aran (Theakston et al., 2001), Eve (Crystal, 1974), Naomi (Sachs, 1983), Nina (Suppes, 1974), Peter (Bloom et al., 1974; Bloom et al., 1975). Following (Mintz, 2003) we only analyze the adult utterances in sessions where the target child is 2.6 years old or younger. Table 1 summarizes the number of target word tokens and types in each corpus.

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 observed as target word in a three word window 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

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

3.2 Algorithm

We use supervised learning with a feed-forward connectionist model (a single hidden layer neural network) to compare the effect of distributional cues from various context representations on the word category learning. The input is a representation of the word context and the output is a word category. We evaluate two models with different input representations:

- $aX + Xb$ **model:** This is the flexible frames model of St Clair et al. (2010), the best performing syntagmatic model. Consider a five word window $a_1a_2Xb_1b_2$ where X is the target word. The input to the model consists of two one-hot vectors, one that correspond to the preceding bigram (a_1a_2) and one to the succeeding bigram (b_1b_2). Thus two input units are activated to represent the context of each target word.

- $a * b$ **model:** This is the paradigmatic model investigated in this paper. First, a small number of substitutes are sampled for the target word based on an n-gram language model as described in the next section. The input to the model consists of the counts of the substitutes in the sampled set. The length of the input vector is equal to the size of the substitute vocabulary.

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

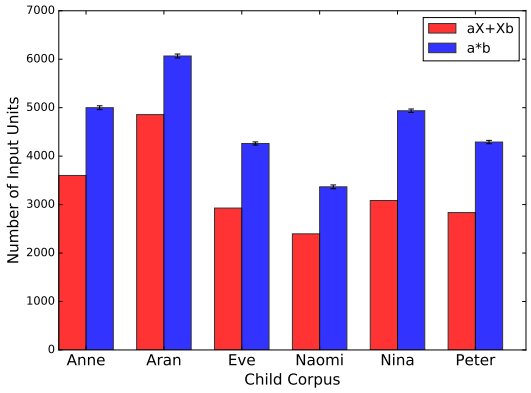


Figure 1: Number of input layer units of the flexible frame ($aX + Xb$) and the substitute based ($a * b$) models. ($a * b$) samples 16 substitutes per target word. Standard errors are reported with error bars.

Figure 1 gives the number of input layer units of syntagmatic ($aX + Xb$) and paradigmatic ($a * b$) 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. Both models have 10 output units due to the standard labeling (Mintz, 2003).

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 non-linearity is sigmoid and the learning rate is 0.2.

3.3 Substitute Words

In the paradigmatic ($a * b$) model, we predict the word category of a word in a given context based on its most likely substitute words. We measure the likelihood of substitute words using an n-gram language model. Here, we first describe how substitute probabilities can be computed using an n-gram model and give details on training the n-gram model.

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 certainly determined 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 : $w_{-n+1} \dots w \dots w_{n-1}$. The probability of a substitute word w in a given context c_w is estimated as:

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

where w_i^j represents the sequence of words $w_i w_{i+1} \dots w_j$. In Equation 1, $P(w_0 = w|c_w)$ is proportional to $P(w_{-n+1} \dots w_0 \dots w_{n+1})$ 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, 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.

To train the n-gram model, we first extracted training data of approximately 6.8 million tokens[4] of child-directed speech data from CHILDES. We trained a 4-gram language model on this data with Kneser-Ney discounting 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, which gave us a vocabulary size of 21734.

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

3.4 Evaluation

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), prepositions, adjectives, adverbs, determiners, conjunctions, wh-words, negation (i.e., "not") and interjections. Following St Clair et al. (2010), we also report the asymmetric lambda value (Goodman and Kruskal, 1979) to compare the association among the classification of grammatical categories.

3.5 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. We randomly split each child corpus into 10 folds. The main advantage of the cross validation is that all sentences are eventually used both for testing and training.

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 as described in Section 3.3, and sampled 16 substitutes per target word in models using the paradigmatic representation.

Section 4 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 5 and 6, respectively.

4 Experiment 1: Syntagmatic vs Paradigmatic

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 ($aX + Xb$).

Method. All models are trained and evaluated according to steps summarized in Section 3.5. To see the effect of training data size, similar to the analysis in (St Clair et al., 2010), we split the training of each model into 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.

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 corpus even with a limited amount of training patterns. 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.

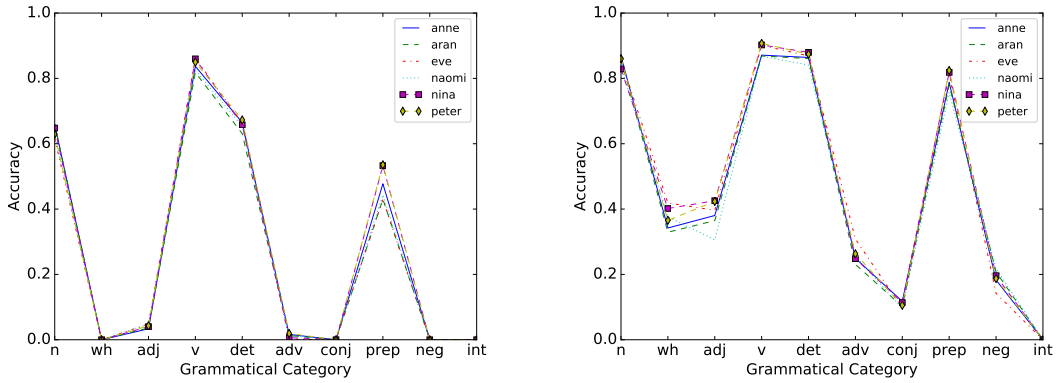
To further investigate the accuracy gap between $aX + Xb$ and $a * b$ models, we plot the classification accuracies of each grammatical category in the standard labeling for both models in Figure 2. Even after 10K training patterns both models are able to achieve relatively high accuracies on nouns (*n*),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

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 some accuracy. 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 word groups in at least one of the child corpora.

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

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)



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

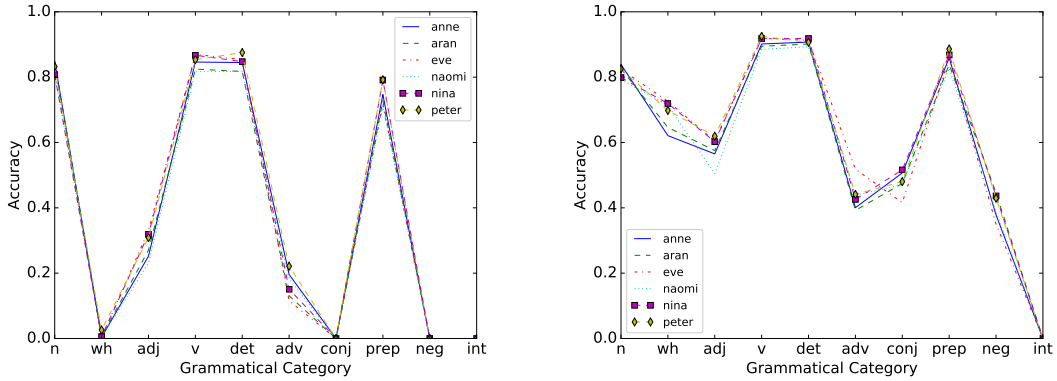
Figure 2: Individual tag accuracies of $aX + Xb$ and $a * b$ on each child corpus after $10K$ training patterns are presented.

Results of Long Training Phase. The previous section shows that the $a * b$ model is more accurate than the $aX + Xb$ model on learning word categories with limited amount of language exposure. In this section each model is trained with 100K input patterns to observe the effect of more 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 benefits less from the 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. Following St Clair et al. (2010), we experimented 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 did not observe significant changes on the result.

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 used. 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)



(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.

Similar to the 10K results, $aX + Xb$ model performs poorly on *wh*, *conj*, *neg*, and *int* as shown in Figure 3. Both models accurately learn the noun, verb, determiner, and preposition groups. However, the $a * b$ model is significantly more accurate on adjectives, conjunctions, and negation.

5 Experiment 2: Number of Substitutes

In this experiment we analyze the effect of the number of substitutes sampled per context on the classification accuracy.

Method. We used the same experimental settings except that the number of substitutes per target word is varied between 1 and 64. We did not observe any significant difference on model classification accuracies for the number of substitutes that are more than 64.

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

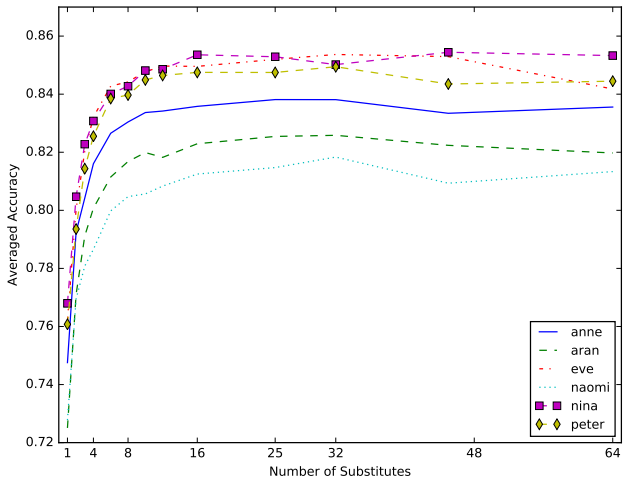


Figure 4: 10-fold cross validation accuracy of each child corpus for different number of substitutes.

6 Experiment 3: Language Model N-gram Order

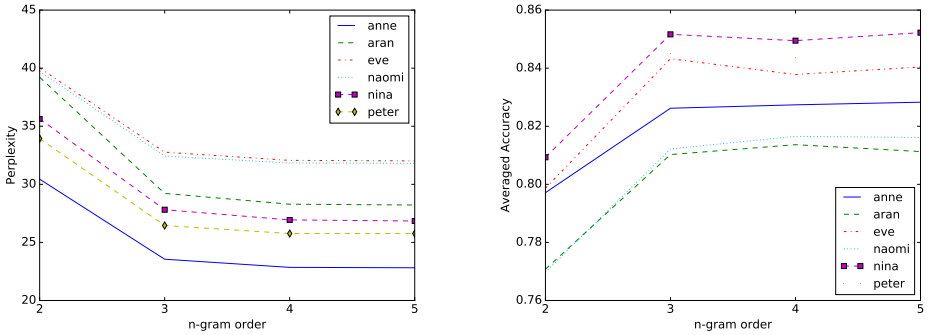
In this set of experiments, we test the paradigmatic model by changing the n-gram order of the language model that is used to sample substitutes. A language model defines probabilities for the sequences of

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 et al., 1996) and are able to store 4-word sequences (Bannard and Matthews, 2008). Experiments with adults also suggest that the language users are sensitive to co-occurrence patterns beyond bigram (Arnon and Snider, 2010).

The perplexity of the language model is a measurement of the variation of words that can be observed in a given n-gram context window and is 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[6].

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.

Results and discussion. 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. Figure 5(b) plots the model classification accuracy versus the n-gram order on each child corpus which slightly improve 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.



(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.

7 Conclusion

In this work, we proposed representing word context with substitute-based paradigmatic representations as opposed to neighbor-based syntagmatic representations for word category acquisition. The paradigmatic approach suggests using probable substitutes of word ($a * b$). On the other hand, the syntagmatic approach we used proposes using the preceding bigram and the succeeding bigram, whichever is fruitful ($aX + Xb$). Our experiments showed that paradigmatic representation of context is more accurate in learning word categories.

To contrast these two representations we replicated the experimental setup of St Clair et al. (2010). Experiments showed 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 demonstrated the same pattern, however, the gap between these approaches decreased.

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

[6](Goodman, 2001) showed that the perplexity plateaued when the order is higher than 5 for datasets of about 10^8 words.

We explored the effect of the n-gram order of the 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 words should be used. Our results demonstrated that the model's performance depends on the n-gram order of the language model up to order 3, larger contexts do not seem to improve the performance.

Acknowledgements

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) grants 114E628 and 215E201.

References

Afra Alishahi and Grzegorz Chrupała. 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*, pages 643–654. Association for Computational Linguistics.

Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62(1):67–82.

Colin Bannard and Danielle Matthews. 2008. Stored word sequences in language learning the effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19(3):241–248.

Lois Bloom, Lois Hood, and Patsy Lightbown. 1974. Imitation in language development: If, when, and why. *Cognitive Psychology*, 6(3):380 – 420.

Lois Bloom, Patsy Lightbown, Lois Hood, Melissa Bowerman, Michael Maratsos, and Michael P Maratsos. 1975. Structure and variation in child language. *Monographs of the society for Research in Child Development*, pages 1–97.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two Decades of Unsupervised POS Induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 575–584, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Crystal. 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, 10.

Daniel Freudenthal, Julian M Pine, and Fernand Gobet. 2005. On the resolution of ambiguities in the extraction of syntactic categories through chunking. *Cognitive Systems Research*, 6(1):17–25.

Leo A Goodman and William H Kruskal. 1979. Measures of association for cross classifications. In *Measures of association for cross classifications*, pages 2–34. Springer.

Joshua T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403 – 434.

Zellig Sabbetai Harris. 1954. Word. *Distributional Structure*, 10(23):146–162.

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

Toben H Mintz. 2003. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91 – 117.

Padraic Monaghan and Morten H Christiansen. 2008. Integration of multiple probabilistic cues in syntax acquisition. *Corpora in language acquisition research: History, methods, perspectives*, pages 139–164.

Christophe Parisse and M. Le Normand. 2000. Automatic disambiguation of morphosyntax in spoken language corpora. *Behavior Research Methods, Instruments, & Computers*, 32(3):468–481.

Martin Redington, Nick Chater, and Steven Finch. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *COGNITIVE SCIENCE*, 22(4):425–469.

Patricia A Reeder, Elissa L Newport, and Richard N Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, 66(1):30–54.

Jacqueline Sachs. 1983. Talking about the there and then: The emergence of displaced reference in parent-child discourse. *Children's language*, 4.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Kenji Sagae, Brian MacWhinney, and Alon Lavie. 2004. Automatic parsing of parental verbal input. *Behavior Research Methods, Instruments, & Computers*, 36(1):113–126.

Hinrich Schütze. 1995. Distributional part-of-speech tagging. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, EACL '95, pages 141–148, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Michelle C St Clair, Padraic Monaghan, and Morten H Christiansen. 2010. Learning grammatical categories from distributional cues: flexible frames for language acquisition. *Cognition*, 116(3):341–60.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*, pages 257–286, November.

Patrick Suppes. 1974. The semantics of children's language. *American psychologist*, 29(2):103.

Anna L Theakston, Elena VM Lieven, Julian M Pine, and Caroline F Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, 28(1):127–152.

Malathi Thothathiri, Jesse Snedeker, and Erin Hannon. 2012. The effect of prosody on distributional learning in 12-to 13-month-old infants. *Infant and Child Development*, 21(2):135–145.

Mehmet Ali Yatbaz, Enis Sert, and Deniz Yuret. 2012. Learning syntactic categories using paradigmatic representations of word context. In *EMNLP-CoNLL*, pages 940–951.