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

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

[a] *Koç University, Istanbul, Turkey*

Abstract

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

1. Introduction

1.1. Psycholinguistic evidence relevant to substitutes

1.2. Comparison with previous distributional approaches

2. Related Work

Previous research demonstrate that infants have a mechanism to process statistical properties of natural language(Saffran et al., 1996). Distributional knowledge,which is also a statistical approach to natural language, refers to the notion that context of a word determines grammatical properties of it. For instance, (Gomez, 2002) and (Van Heugten et al.,2010) demonstrated that distributional statistics help infants acquire non-adjacent dependencies. Distributional knowledge also plays an important part in word segmentation task(Saffran,Newport, et al., 1996).

One of the approaches of distributional information to construct syntactic categories is the work of (Redington et al., 1998). They define the context of

*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)

a word as previous and following words. With this definition, they construct context vectors of target words for clustering. Using average link clustering with a treshold maximizing accuracy and completeness, target words are seperated into categories. Although the categorizations are generally accurate, the method lacks of completeness. Furthermore, as (Ambridge and 2011) pointed out one might question what could be the underlying process to determine the treshold for infants to do such clustering.

(Cartwight et al.,1997) introduces an incremental method to make use of distributional knowledge. Intrinsic idea behind the method is that no language learner has ever a chance to process all the sentences of the language to learn syntactic categorization, thus, a generalization method such as distributional processing may help learner confine the search space for later generalization. To accomplish this, they convert the problem into Minimum Descriptive Language optimization problem. The crucial part of the method is that after a sentence is processed, it is forgotton. Results are similar with Redingtons's work, high in accuracy but low in completeness. Still, it shows that distributional knowledge is powerful for syntactic categorization even if the learner is exposed the small amount of syntactic knowledge of the language.

By extending the work on adults with artificial language inputs (Mintz,2002), (Mintz, 2003) proposes a notion to represent the context of a word. Frequent frames can be defined as two jointly appearing words with one word in the middle. Experiments on child directed speech reveal that even relatively small fraction of frames has ability to categorize the half of the corpora. Though the accuracy is impressive, the same as previous work, it suffers from completeness, in addition, it has covarege problem. As a further step, (St. Clair et al., 2010), combines the bigram's coverage power(Redington et al.,1998),(Monaghan and Christansen, 2008) and accuracy of fixed frames(Mintz,2003). Extensive experiments result that infants make use of both bigram and trigram sources. As they pointed out, they may even use higher-order relationships among words.

(Freudenthal et al.,2004) points out a different complication on distributional methods for constructing syntactic categories. One of the core ideas behind distributional methods is that words in the same context share the same syntactic categories and can be used interchangebly(REFERENCE HERE). Freudenthal claims that evaluation methods used in previous work such as (Mintz, 2003), (Cartwright, 1997) and (Redington et al., 1998) can be misleading. In those work, if a word is substituted with another one in its

category, the resulting sentences erronous in a way that are not observed in infants speech. As a success criteria, they argue that proposed categorization should generate plausible sentences. They introduce chunking mechanism to overcome this problem by merging words that are seen frequently. Results seems successful to generate meaningful sentences after substitution, still, the proposed solution is computationally complex to disclose the learning mechanism in infants.

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. Note that the substitute word distribution is a function of the context only and is indifferent to the target word.

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 c_w as the $2n-1$ word window centered around the target word position: $w_{-n+1} \dots w_0 \dots w_{n-1}$. The probability of a substitute word w in a given context c_w can be 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, $P(w|c_w)$ 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, because of the Markov property of n-gram language model, only the closest $n-1$ words are used in the experiments.

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 compute substitute probabilities we trained a language model using approximately 6.8 million tokens of child-directed speech data from the CHILDES corpus (MacWhinney, 2000) (excluding sections of [test-set]) We used SRILM (Stolcke, 2002) to build a 4-gram language model with Kneser-Ney discounting. Words that were observed less than 2 times in the LM training data were replaced by UNK tags, which gave us a vocabulary size of 21734. [What is the test data? Where should we put this?] [perplexity]

4. Experiment Setup

1. We need to clarify the tag set that is used during the experiments. May be it is better to give the whole mapping as an Appendix section.
2. Data statistics (it is common to all experiments)

4.1. Input Corpora

In order to be consistent with St Clair et al. (2010) and Mintz (2003), we use the same six corpora of child-directed speech from the CHILDES 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.

4.1.1. Preprocessing

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 it is encoded in the MOR line of the CHILDES corpus.

We apply the following pre-processing steps (St Clair, Monaghan & Christiansen, 2010) initial to our analyses:

- 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).
- **Short usages ??** boundaries.

4.1.2. Target Word

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 one is the target word that is categorized. The correct grammatical category of target words are extracted from CHILDES. [!! put number of target words table]

4.1.3. Language Modeling Substitute Words

To compute substitute probabilities of target words we trained a language model using approximately 6.8 million words[1] of child-directed speech data from CHILDES. We used SRILM (Stolcke, 2002) to build a 4-gram language model with Kneser-Ney discounting. Words that were observed less than 2 times in the LM training data were replaced with unknown word tag <UNK>, which gave us a vocabulary size of 21734.

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,

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

- $aX+Xb$ **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.

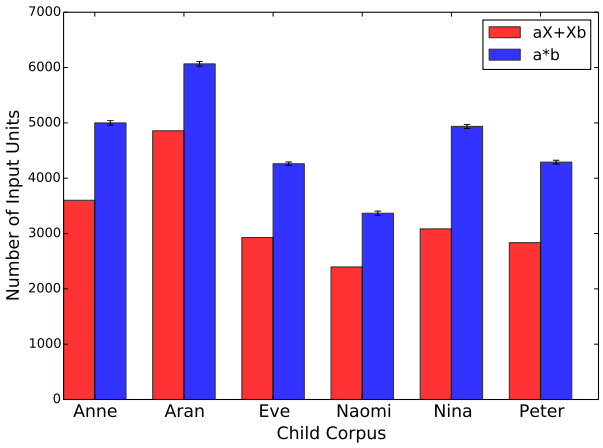


Figure 1: Number of input layer units of the flexible frame ($aX+Xb$) and the substitute based model($a*b$) are summarized. $a*b$ samples 16 substitutes per target word. Standard errors are reported with error bars.

Table 1 presents the number of input layer units of syntagmatic and paradigmatic representation based models on each child corpus seperately. 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[2], 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).

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

Unless stated otherwise, all connectionist models in this paper uses the following parameters: (1)number of hidden units is set to 200 and initialized randomly for each model, (2)**backprobagation(0.1)**, (3)**learning rate**,(4) **sigmoid...**

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 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. [!! cite]

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 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 LM defined in Section 4.1.3 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 6 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 7 and 8, respectively.

5. Experiment 1: Corpus analysis

6. Experiment 2: 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

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
Anne	121726	93371	42789	2623	1846
Aran	129823	104997	54768	3256	2595
Eve	78778	59095	27315	2184	1465
Naomi	38302	28793	13002	1883	1194
Nina	89957	72879	39335	2036	1580
Peter	94521	72834	34997	2145	1472

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 ($aX + Xb$)[3].

6.1. Method

All models are trained and evaluated according to steps summarized in Section 4.3. Similar to analysis in St Clair et al. (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.

6.2. 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.

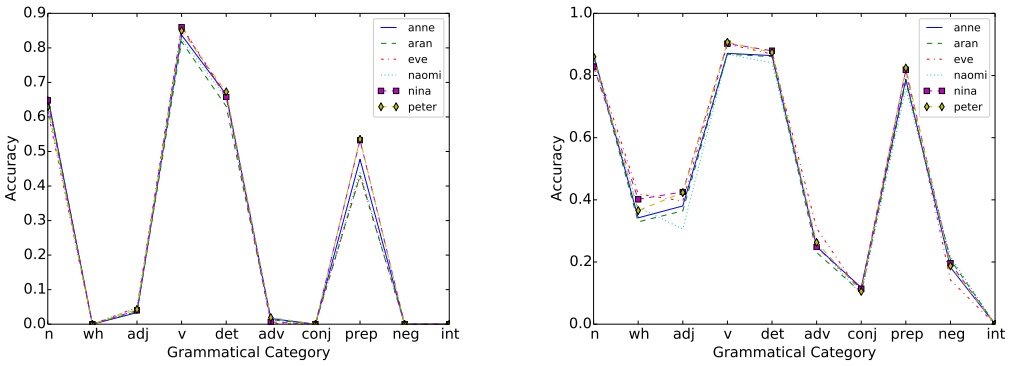
To further investigate the accuracy gap between $aX + Xb$ and $a * b$ models, we plot the classification accuracies of each grammatical category in the standard labeling for both models in Figure 2. Even after 10K training patterns both models are able to achieve relatively higher accuracies on

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

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)

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.

6.3. Results of Long Training Phase

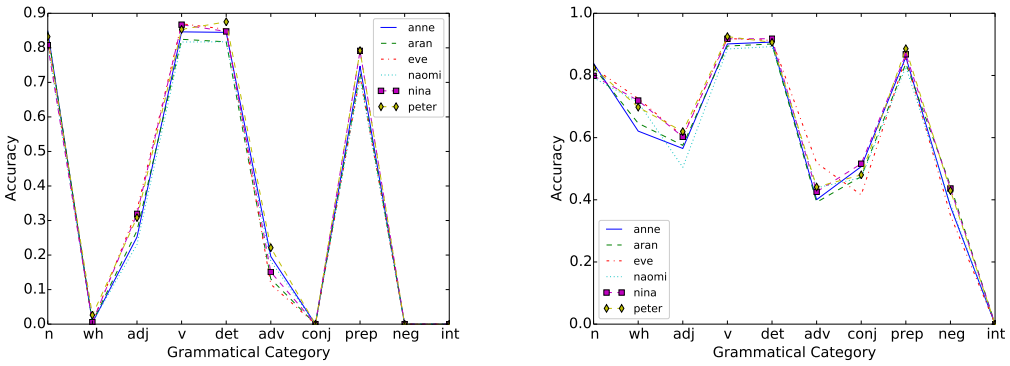
Previous section shows that the $a * b$ model is more accurate than the $aX + Xb$ model on learning grammatical categories with limited amount of language exposure. In this section each model is trained with 100K input patterns to observe the effect of extensive language exposure on learning.

Table 3: 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)

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 extensive training than the $aX + Xb$ model. One possible explanation of 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.

In contrast to the 50K 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.



(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.

7. Experiment 3: 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. A side 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.

7.1. Method

We used the same experimental settings except that the number of substitutes per target word is varied between 1 and 64[4].

7.2. 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 to 16. After 16 substitutes the effect of increasing number of substitutes resulted in very minor changes on the classification accuracies. 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.

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

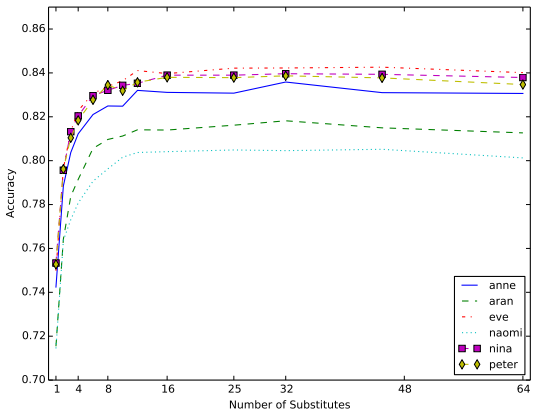


Figure 4: 10-fold cross validation accuracy of each child corpus for different number of substitutes.

Figure 4 [Put subs vs input graph] shows the increasing trend of the number of input units as the number of substitutes on each child corpus increases. 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) came up 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 was same 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.

8. Experiment 4: Language Model N-gram Order

The n-gram order effects the perplexity of the language model which is in fact a measurement of the number of words that can be observed in a given n-gram context window. Therefore one can expect that as the n-gram order increases the model assigns more relevant substitutes to the context[5]. In this

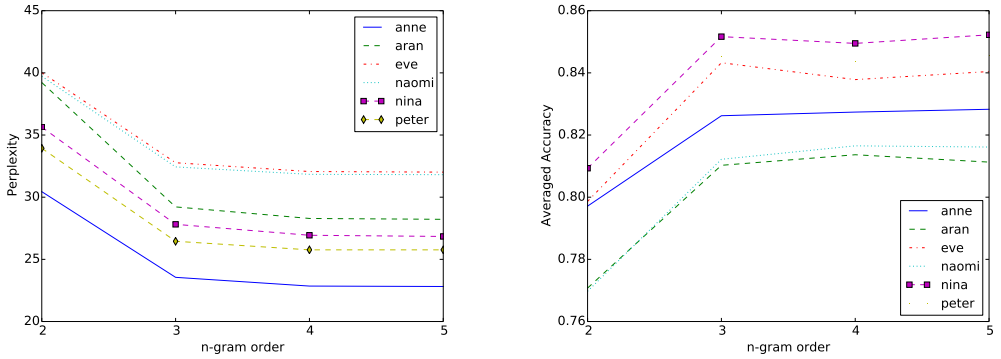
[5]Goodman (2001) showed that the perplexity plateaued when the order is higher than 5.

set of experiments we test the paradigmatic model by changing the n-gram order of the language model that are used to sample substitutes.

8.1. 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.

8.2. 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.

9. General Discussion

References

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).

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.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language*, *15*, 403 – 434. URL: http://www.sciencedirect.com/science/article/pii/S0885230801901743. doi:http://dx.doi.org/10.1006/csla.2001.0174.

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

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91 – 117. URL: http://www.sciencedirect.com/science/article/pii/S0010027703001409. doi:10.1016/S0010-0277(03)00140-9.

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

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

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

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. URL: http://www.biomedsearch.com/nih/Learning-grammatical-categories-from-distributic

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.