From linear sequences to abstract structures: Distributional information in infant-direct speech

Hao Wang[1], Toben H. Mintz[1, 2]
1 Department of Psychology, University of Southern California
2 Department of Linguistics and Program in Neuroscience, University of Southern California

Introduction

Syntactic categories (such as noun and verb) are the basic units of grammar, as grammatical rules are defined over syntactic categories rather than words. Word categorization is thus a prerequisite for acquiring a full-fledged grammar. A challenge for infants is that category information is not explicitly marked in the input. Rather, learners must compute category information by analyzing cues in the input and environment. The question is, then, how children learn word categories and which source of information plays a primary role in the process.

Distributional information has been proposed by some scholars as a crucial source of lexical category information. The proposal originated in the approach of structural linguistics (e.g., Harris, 1951), in which grammatical categories were defined by lexical co-occurrence patterns. For example, if two words occur after *the* and *a*, and share other contexts in a sample of speech, they were designated as belonging to the same category. Maratsos & Chalkley (1980) developed some of the concepts from structural linguists into a theory of language acquisition, proposing that learners consider distributional information in determining which words belong together in a category.

Some early critiques of this approach to grammatical categorization argued that distributional information, without being structurally constrained, would lead to many erroneous generalizations. For example, Pinker (1987) pointed out that "generalizations based on simple distributional commonalities can do more harm than good". He gave an example that *John ate fish*, *John ate rabbits*, and *John can fish* would lead a distributional learner to accept *John can rabbits* as a grammatical sentence. In his example, both the words *fish* and *rabbits* appeared in the same context—after *John ate*—but extending the classification of *fish* with *rabbits* to a different environment where *fish* occurs is unwarranted. So, a simple distributional learner who does not have access to underlying syntactic structures would group *fish* and *rabbits* together and erroneously generalize to the sentence *John can rabbits*. On the other hand, if learners had access to the fact that *fish* occurs in different syntactic contexts in the two cases, then they would be able to restrict generalizations to the relevant contexts. This is what Pinker called "structure dependent distributional learning."

However, despite this and other potential pitfalls, a number of studies have examined a variety of kinds of distributional environments and have shown that distributional information is quite robust (Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Martin Redington, Crater, & Finch, 1998). Among all the distributional environments studied, one of them - frequent frames - is extremely good at categorizing words. Mintz (2003) defined a frame as "two jointly occurring words with one word intervening". In one analysis, frequent frames were defined as the most frequent 45 frames in a corpus. For example, *you___it* is a frequent frame in the input to many children. Words occurring in the target position of the frame (between *you* and *it*)

are predominantly verbs. In his first experiment, the data analyzed is speech to children under 2;6 from six corpora in CHILDES database (MacWhinney, 2000). The results were measured with accuracy of categorization, which ranges from 0 to 1 where 1 means all the words grouped together belong to the same linguistic category. The mean accuracy of the six corpora was 0.98, which is very close to the maximum of 1. Hence near perfect word categorization was achieved with frequent frames. Related studies have shown that frames can lead to categorization in adults and infants (Mintz, 2002, 2006, 2007), and that a learner with limited memory resources would nonetheless achieve very high accuracy by categorizing words using frames (Wang & Mintz, 2007). There is also evidence that frequent frames are a robust context for categorization in French (Chemla, Mintz, Bernal, & Christophe, 2009), and our ongoing analyses of German and Turkish child-directed speech show similar results.

Thus, evidence from different kinds of studies—corpus, behavioral and computational modeling—showed that frequent frames are good environments for categorizing words. In this study we begin to investigate why this is so. Why do the simple linear relations involved in frequent frames capture abstract syntactic information? Syntactic grammars that have been developed over the past 50 years account for structural regularities within and across languages by proposing hierarchical organization of words. Yet frequent frames are defined over simple linear sequences. What is special about the syntactic structures that frequent frames select, as opposed to other linear sequences?

Our hypothesis is that the frequent non-adjacent co-occurrence of words that defines frequent frames comes about when the context positions are structurally "close" (if not in the same constituent), and that the underlying syntactic structures are similar from instance to instance of a frequent frame. This consistency, in turn, constrains the grammatical category of the words in the intermediate (target) position across instances of a frequent frame.

To examine the syntactic structures associated with frequent frames, two analyses were conducted on several syntactically annotated child-directed speech corpora. The first analysis investigated the relation between frequent frames and syntactic structures by examining the range of syntactic structures associated with frequent frames, bigrams and two other trigram contexts that are minimally different from frequent frames. We predicted that syntactic structures associated with frequent frames consist of a more restricted set of structures; therefore, more strictly constraining the word categories that can occur in the target position. The second analysis compared the syntactic relatedness of targets and contexts in frequent frames and bigrams. Specifically, it examined the degree to which dependencies involving a word in a frequent frame or bigram link to other words within that frequent frame or bigram and to words that are outside that frequent frame or bigram. Our prediction was that the target and context in frequent frames are more closely related to each other than in bigrams; and this relation is more consistent across instances of frequent frames than in instances of bigrams.

Analysis 1

This analysis examined syntactic structures of frequent frames in child-directed speech by quantifying common grammatical relations associated with frequent frames. It also compared quantitatively the common structures of frequent frames, bigrams and

other two trigram-based environments. The hypothesis is that the structures associated with frequent frames are very typical. Only small variability in structures of frequent frames of the same type should be observed in child-directed speech; and instances of the same frequent frame type should have a restricted set of grammatical relations.

Data

Both Analysis 1 and 2 used the same six corpora that were analyzed in Mintz (2003) from CHILDES database (MacWhinney, 2000) for the same reason that these corpora covered the period of early syntax development. They are Eve (Brown, 1973), Peter (Bloom, Hood, & Lightbown, 1974; Bloom, Lightbown, & Hood, 1975), Naomi (Sachs, 1983), Nina (Suppes, 1974), Anne (Theakston, Lieven, Pine, & Rowland, 2001), and Aran (Theakston, et al., 2001). The utterances analyzed were speech to children before age of 2;6 which is the time period when children start to show some knowledge of linguistic categories in their production. Utterances were minimally treated. Punctuations and other annotations were removed from sentences before running any analyses.

Because there was no phrase structure annotation available on any English child-directed speech corpus at the time of the study, we used CHILDES corpora that were annotated with a dependency grammar (Sagae, 2007). The dependency structure of a sentence consists of grammatical relations (e.g., subject, object and determiner) between words in that sentence. Similar to phrase structure, it is also a representation of structural information in sentence. For each word in a sentence, there is a grammatical relation annotation associated with it in the format of a triple i|j|g, where *i* is the position of the word in the sentence, *j* the position of the word's syntactic head, and *g* is the name of the grammatical relation represented by the syntactic dependency between the i-th and j-th words. If the topmost head of the utterance is the i-th word, it is labeled as i|0|ROOT.

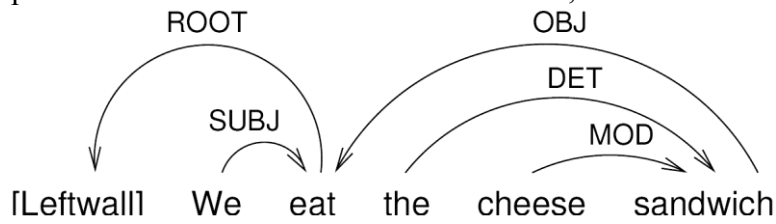


Figure 1. Dependency structure of a sentence (adapted from Sagae et al. (2005) Figure 1)

The dependency structure of an example sentence is given in Figure 1. For instance, the first word *we* is a subject (SUBJ) of *eat*, which is itself the root of the utterance. The root of a sentence is usually the main verb. A list of grammatical relations used in tagging the corpora can be found in Sagae (2005) Figure 2.

Method

We use a special notation to refer to positions in frames, bigrams or any other distributional contexts. F_1 is the first context word; F_2 is the second context word (for the trigram analyses). T is the target word that is to be categorized. For example, for frame *we__the*, F_1 refers to *we*; F_2 refers to *the*; T refers to words occurred in that frame. Apparently, bigrams only have F_1 and T but no F_2.

In comparing grammatical structures, we used four combinations of grammatical relations (later referred to as structures) to represent the structure of a frame as we don't

know which one better represents the structure. They are GR_{F1} & GR_{F2}, GR_{F1} & GR_T, GR_T & GR_{F2}, and GR_{F1} & GR_T & GR_{F2} (only GR_{F1} & GR_T for bigrams). For example, GR_{F1} & GR_{F2} is a pair of grammatical relations from the first and second context words. For two tokens of a frame, they will be treated as having the same structure only if GR_{F1} of the first token matches GR_{F1} of for the second token and so for GR_{F2}. When comparing two grammatical relations, they are the same only if both the relative positions of heads and the relations are the same. It means that if two grammatical relations have the same relation but their head positions are different, they will be treated as different ones. This is a strict constraint that could introduce more variability in our result than if positional information were not counted.

The procedure is now described. For each corpus, all the tokens of the most frequent 45 frames and their related grammatical relations were selected and analyzed. For each frame type, structures were sorted according to how many tokens of that frame share the same structure. The proportions of tokens that have the four most frequent structures were computed.

Bigrams and another two trigram-based environments (__F_1F_2 and F_1F_2__) were used as control conditions. The same procedure was repeated with bigrams and __ F_1F_2 and F_1F_2__. Bigrams used here are a distributional environment similar to frequent frames, but it does not have the second context words that could provide additional constraints to the target words. An example of bigrams is *the*__, in which *the* is the categorization context and all the words immediately following *the* are categorized together. Bigrams were found to be an informative categorization context in previous studies (Mintz, et al., 2002; M. Redington, Chater, & Finch, 1998), although it is not as accurate as frequent frames. In addition to bigrams, we also included __ F_1F_2 and F_1F_2__ because they are similar to frames. They all have two context words and a target word in a trigram. But the target word positions are slightly different from frames; they are the first word and last word in a trigram, respectively. Because in these alternative environments the target word is not framed by the joint context elements, the target word is not syntactically constrained to the degree that it is in a frequent frame. We thus predicted that frequent frames would have a high proportion of tokens in a few very frequent structures compared to the other three environments. We chose to analyze the four most frequent structures for each context.

Results

The mean proportions of tokens of the most frequent four structures are show in Figure 2. For all four kinds of structure representations, the most frequent four structures in each frequent frame have covered a large number of tokens (92%, 91%, 88% and 85%). In other words, on average 92% of tokens in each frame have only four structures that are represented by the pair of grammatical relations GR_{F1} & GR_{F2}; and on average 85% of tokens in each frame have only four structures that are represented by the triple grammatical relations GR_{F1} & GR_T & GR_{F2}.

Comparing the GR_{F1} & GR_T structure of frequent frames to that of bigrams, the proportions of tokens accounted for by the most frequent four structures is significantly higher for frequent frames (M=91%) than for bigrams (M=64%), $t(5)=26.97$, $p<0.001$. The proportion of tokens accounted for by the most frequent four GR_{F1} & GR_T structures

is also significantly higher for frequent frames (M=91%) than for __F_1F_2 (M=76%), t(5)=11.43, p<0.001 or F_1F_2__ (M=81%), t(5)=13.75, p<0.001.

Frequent frames demonstrated significantly larger proportion accounted for by the most frequent four GR_{F1} & GR_{F2} structures (92%) than__F_1F_2 (M=88%), t(5)=9.55, p<0.001 or F_1F_2__ (M=86%), t(5)=8.20, p<0.001. Since the GR_{F1} & GR_{F2} structure is a combination of relations from the two context words, higher proportion in frequent frames indicates that the grammatical relations of first and second context words are more coherent and consistent.

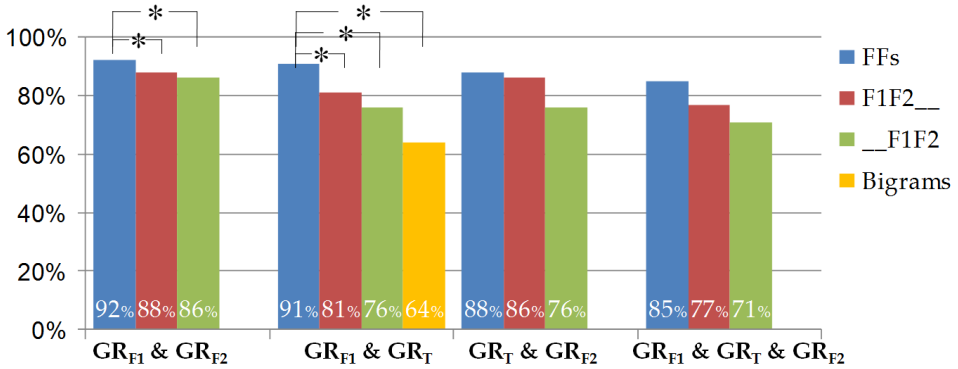
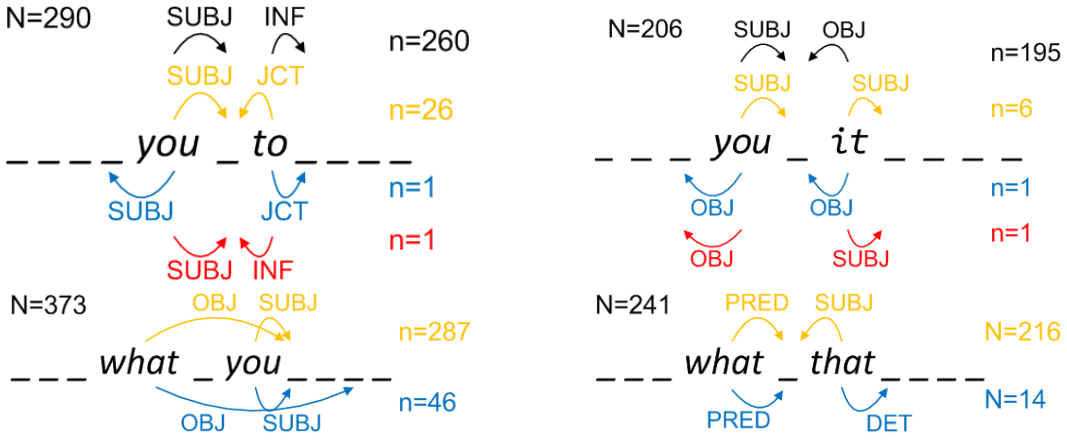


Figure 2 Mean proportions of the most frequent four structures (* p<0.001)

Figure 3 shows four frequent frames and their most frequent four GR_{F1} & GR_{F2} structures. For frames like *what__you* and *you__it*, F_1 and F_3 have the object and subject relations that give the target word very few category options; one of them is predicate. This explains how frequent frames that are linear local relations restrict the word categories of the target position through syntactic constraints.



SUBJ subject	OBJ object	INF infinitival "to"
JCT adjunct	PRED predicative	DET determiner

Figure 3 The most frequent four GR_{F1} & GR_{F2} structure of example frequent frames (N is the total number of frame tokens, n is the number of token with a specific structure)

In addition, we observed that the relations of structures for the same frame mostly remain the same while their head positions are sometimes different. For instance, the top four structures of *what__you* frame all have the same relation pair – OBJ and SUBJ. The only difference is that the object relationship is linked to different head positions.

Therefore, the proportion of tokens accounted for by the most frequent four frames would be higher if one only counts the relations and not the positions.

The exceptionally high proportions in frequent frames confirm our prediction that grammatical structures associated with each frequent frame are dominated by a restricted set of structures, which consequently constrain the possible word categories occurring in the target position of that frequent frame. The result also suggests that the two context words in frequent frames have the most coherent and consistent relations among the three trigram-based environments.

Analysis 2

The first analysis has shown that the grammatical structures selected by frequent frames are more restricted and consistent than other distributional environments. The second analysis further investigated the distribution of grammatical relations within a frequent frame and grammatical relations that cross frame boundary. In particular, we investigated the degree to which words within a frequent frame (or bigram) were grammatically related (via dependency links) to other words within the frame (or bigram) and to words outside the frame (or bigram). We predicted that words within frequent frames would be more likely to be linked to other words within the frame than words within a bigram structure. Such an outcome would be evidence that the words within a frequent frame are especially "close" syntactically.

Data

The data is the same as in Analysis 1.

Method

For each frequent frame, all the grammatical relations associated with any position in the frequent frame were tallied. These grammatical relations were then classified as either *external links* (links between a word in a frame and words outside that frame) or *internal links* (links between two words within a frame). For example, assume *we__the* in Figure 1 is a frequent frame, the subject (SUBJ) relation between *we* and *eat* is an internal link and the object (OBJ) relation between *eat* and *sandwich* is an external link. In addition, we also differentiate the direction of grammatical relations. A link always goes from a word to its head. For example, in Figure 1, the subject (SUBJ) relation goes from *we* to *the*, the object (OBJ) relation goes from *sandwich* to *eat*.

Then, the mean number of links per frame token was computed for every position/direction of each frequent frame, such as, the number of internal links that come out of F_1, the number of external links that go from F_1 to target, and so forth. Finally, the numbers were averaged over all frequent frames.

Another measure, which views a frequent frame/bigram as an entity and demonstrates its interaction with other words in the same utterance, is the total number of links going out a frame/bigram (outgoing links) and coming to a frame/bigram (incoming links). The total number of links was divided by the number of words since a frame has one more word than a bigram.

The above procedure was also applied to bigrams to serve as a control condition. Notice that bigrams do not have links to or from F_2, so analyses involving F_2 were not carried out on bigrams.

Results

The mean numbers of links per token from or to every position in a frequent frame and in a bigram are shown in Appendix Tables A1 and A2, respectively. The numbers of links for corresponding positions in frequent frames and bigrams were compared. The most interesting result emerged from comparing the links from or to F_1 as illustrated in Figure 4. For links from F_1, frequent frames have significantly more internal links (0.73) than bigrams (0.31), $t(5)=21.27$, $p<0.001$; frequent frames also have significantly fewer external links (0.17) than bigrams (0.51) $t(5)=-16.26$, $p<0.001$. For links to F_1, bigrams have significantly more external links (0.58) than frequent frames (0.23), $t(5)=10.72$, $p<0.001$. When examining specifically the links from F_1 to T, frequent frames have significantly more links (0.49) than bigrams (0.31), $t(5)=5.39$, $p<0.005$, although F_1 and T in frequent frames and bigrams have the same relationship.

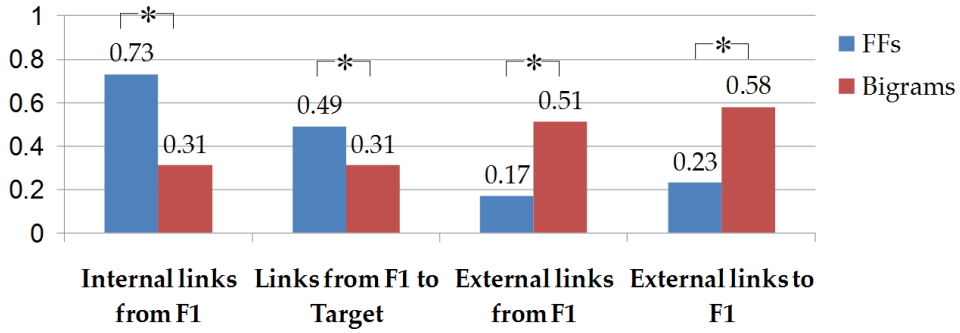


Figure 4 Links from/to W1 for FFs and bigrams (all ps<0.01)

Links associated with the target word may also give us some clues on the internal structure of frequent frames and bigrams. There are much more internal links to T for frequent frames (0.75) than for bigrams (0.31), $t(5)=8.59$, $p<0.001$, which suggests a more integral structure for frequent frames. In addition, both frequent frames and bigrams have many external links to T, 0.60 and 0.52 respectively. As for the links from T, frequent frames and bigrams present a contrary pattern. For frequent frames, the links from T mainly go to words within that frame (0.42) and some go to words outside the frame (0.34). But for bigrams, the majority of links from T go to words outside the bigram (0.63) and very few links go to F_1.

There are relatively few links from F_2 to F_1 (0.10) and T to F_1 (0.14), which could be due to the fact that English is a head-first language.

Another measurement, the total outgoing links per word is 0.32 for frequent frames and 0.57 for bigrams. The total incoming links per word is 0.45 for frequent frames and 0.55 for bigrams. It suggests that frequent frames have less interaction with outside than bigrams that makes frequent frames as a stable and self-contained entity.

The number of internal links per word is 0.49 for frequent frames and 0.25 for bigrams. Frequent frames have much more internal links than bigrams; and bigrams have more links that cross boundary than frequent frames. This indicates that frequent frames are internally very coherent and words of frequent frames are more closeness related to each other.

The above results revealed the overall pattern of grammatical relations involving frequent frames and bigrams. Since a variety of grammatical relations are involved in frequent frames and bigrams, it is not surprising to observe distinct patterns within

frequent frames or bigrams. Table A3 in Appendix lists the number of internal links and external links per token for each frame in Eve corpus. The same statistics for bigrams is reported in Appendix Table A4. Figure 5 is a scatter plot of internal links for every frequent frame and bigram type. Each data point is the mean number of links per token for the specific link of a frequent frame or bigram type. For example, 96% of all tokens of *what__that* frame have a link from F_1 to T; almost no *what__you* frame have that link; and nearly half of all tokens of the bigram *what__* have that link. It is easy to notice the difference in distributions: frequent frames are mostly gathered around the two extreme values while bigrams are more evenly distributed across the whole range. These indicate that for all tokens of a particular frame there is either no link or very consistent links between the two positions while many bigrams have a mixed pattern. This further supports that frequent frames are internally very consistent and stable.

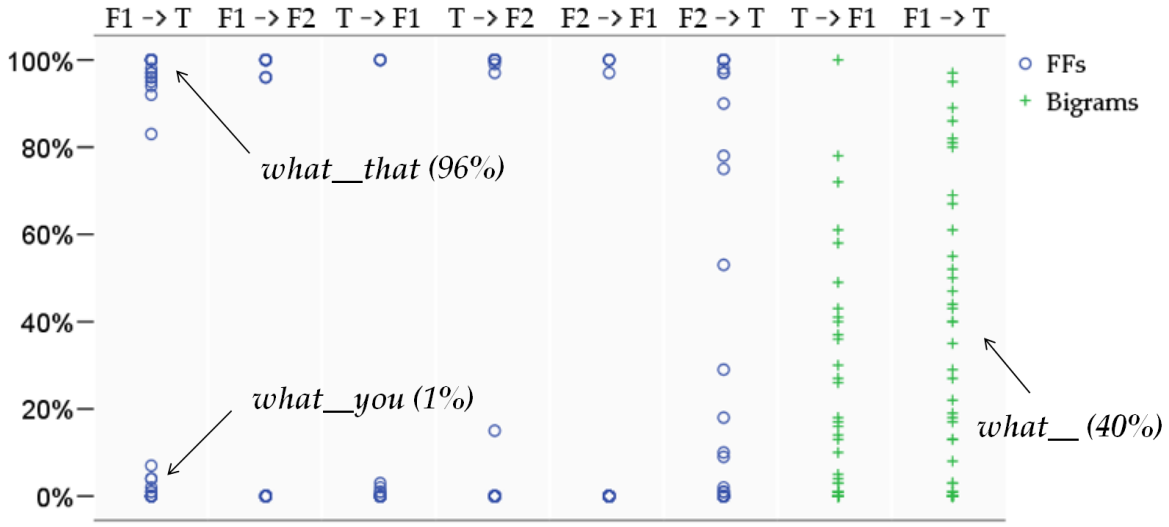


Figure 5 Proportions of internal links for every frequent frame and bigram type (Eve corpus)

Discussion and conclusion

The first analysis showed that the syntactic structures selected by frequent frames are dominated by a restricted set of structures that are highly consistent and internally coherent. This tightly constrains categories of words occurring in the target position and hence led to robust categorization. For bigrams and non-framing trigram contexts, they have lower proportions for the most frequent four structures, which introduces some variation to the word categories of the target position hence is detrimental to word categorization. We conclude from Analysis 1 that frequent frames are accurate categorizers because they identify linear sequences that are syntactically highly constrained.

The second analysis confirmed our predictions that a target and its context in a FF are more syntactically closely related to each other than in bigrams; and grammatical relations between words are more consistent in individual frequent frames than in bigrams. This provides converging evidence that frequent frames select syntactically constrained word sequences.

In summary, our analyses confirm that frequent frames are internally more coherent and consistent than bigrams and trigram environments that are not frames. In

light of these results we return to Pinker's critique of distributional analyses and his claim for the need of structure dependent distributional learning. What we have shown is that limiting distributional generalizations to structurally similar contexts is possible without requiring a prior structural analysis, at least in some cases. Frequent frames can be viewed as a proxy for structural information, and it is perhaps for this reason, in part, that it is such a robust cue to lexical categories. Furthermore, this relation between surface-level patterns and the syntactic structure could provide a cue to children for bootstrapping into phrasal syntax.

It would be interesting to continue this research in two directions. Firstly, it would be informative to run the second analysis for $\_\_F_1F_2$ and $F_1F_2\_\_$ and compare the result to frequent frames. The two environments are similar to frames except the target positions are different. Since bigrams lack of the constraints from F_2, analysis with these two environments might reveal additional insights on why frequent frames are exceptionally good at categorizing words. Secondly, it is worth to repeat the first analysis on child-directed speech annotated with phrase structure. Although the dependency structure annotation used in CHILDES database represents a kind of syntactic relations between words, it misses some information like the attachment of adjuncts. All the words attached to a head are seen at the same level, not organized as a binary syntactic tree. Therefore, it would be worth to annotate some child language corpora and look at phrase structures of frequent frames.

Acknowledgement

This research was supported in part by a grant from the National Science Foundation (BCS-0721328).

References

Bloom, L., Hood, L., & Lightbown, P. M. (1974). Imitation in language development: if, when and why. *Cognitive Development, 6*, 380-420.

Bloom, L., Lightbown, P. M., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40*(2), 1-97.

Brown, R. W. (1973). *A first language: The early stages*. Cambridge, Mass.: Harvard University Press.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*(2), 121-170.

Chemla, E., Mintz, T. H., Bernal, S., & Christophe, A. (2009). Categorizing words using frequent frames: what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science, 12*, 396-406.

Harris, Z. S. (1951). *Structural linguistics*. Chicago: University of Chicago Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. E. Nelson (Ed.), *Children's Language* (Vol. 2). New York, NY.: Gardner Press.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*, 678-686.

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

Mintz, T. H. (2006). Finding the verbs: distributional cues to categories available to young learners. In R. M. G. K. Hirsh-Pasek (Ed.), *Action Meets Word: How Children Learn Verbs* (pp. 31-63). New York: Oxford University Press.

Mintz, T. H. (2007). *Category Induction from Lexical Co-occurrence Patterns in Artificial Languages*. Paper presented at the Current Issues in Language Acquisition: Artificial & Statistical Language Learning, University of Calgary.

Mintz, T. H., Newport, E. L., & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Pinker, S. (1987). The bootstrapping problems in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* New York, NY.: Springer-Verlag.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Redington, M., Crater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K. E. Nelson (Ed.), *Children's language* (Vol. 3). Hillsdale, NJ.: Lawrence Erlbaum Associates, Inc.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B. and Wintner, S (2007). *High-accuracy annotation and parsing of CHILDES transcripts*. Paper presented at the ACL-2007 Workshop on Cognitive Aspects of Computational Language Acquisition, Prague, Czech Republic.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). *Automatic measurement of syntactic development in child language*. Paper presented at the Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics.

Suppes, P. (1974). The semantics of children's language. *American Psychologist, 29 (2)*, 103-114.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb argument structure. *Journal of Child Language, 28*(1), pp. 127-152.

Wang, H., & Mintz, T. H. (2007). *A Dynamic Learning Model for Categorizing Words Using Frames*. Paper presented at the 32nd annual Boston University Conference on Language Development, Boston, MA.

Appendix

Table A1 Mean number of links per token for frequent frames

Corpus	Token count	External links						Internal links					
		to W1	to W2	to W3	from W1	from W2	from W3	W1->W2	W1->W3	W2->W1	W2->W3	W3->W1	W3->W2
Eve	3601	0.19	0.54	0.50	0.15	0.33	0.39	0.52	0.25	0.10	0.28	0.10	0.29
Peter	4541	0.28	0.71	0.44	0.25	0.30	0.52	0.44	0.21	0.16	0.27	0.13	0.20
Nina	6709	0.19	0.46	0.71	0.15	0.32	0.40	0.48	0.29	0.09	0.37	0.07	0.23
Naomi	1447	0.20	0.77	0.46	0.13	0.36	0.52	0.60	0.17	0.13	0.21	0.07	0.24
Anne	4435	0.24	0.50	0.54	0.18	0.32	0.43	0.41	0.29	0.17	0.34	0.12	0.17
Aran	5245	0.27	0.61	0.51	0.17	0.39	0.51	0.50	0.20	0.16	0.24	0.10	0.21
Average		0.23	0.60	0.52	0.17	0.34	0.46	0.49	0.24	0.14	0.29	0.10	0.22

* ROOT is not counted in any internal or external link.

Table A2 Mean number of links per token for bigrams*

Corpus	Token count	External links				Internal links	
		to W1	to W2	from W1	from W2	W1->W2	W2->W1
Eve	28076	0.52	0.51	0.51	0.62	0.32	0.18
Peter	35723	0.65	0.48	0.53	0.66	0.28	0.20
Nina	37055	0.66	0.58	0.49	0.64	0.32	0.15
Naomi	12409	0.59	0.50	0.51	0.63	0.30	0.19
Anne	38681	0.52	0.48	0.44	0.62	0.36	0.16
Aran	49302	0.52	0.55	0.54	0.60	0.30	0.22
Average		0.58	0.52	0.51	0.63	0.31	0.19

* ROOT is not counted in any internal or external link.

Table A3 Mean number of links per token for each frequent frame in Eve corpus*

Frequent frames	Token count	External links						Internal links					
		to W1	to W2	to W3	from W1	from W2	from W3	W1->W2	W1->W3	W2->W1	W2->W3	W3->W1	W3->W2
put_X_in	76	1.54	0.00	0.79	0.29	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00
you_X_to	290	0.00	1.76	0.09	0.00	0.07	0.90	1.00	0.00	0.00	0.00	0.00	0.10
you_X_me	101	0.00	1.69	0.00	0.00	0.08	0.00	1.00	0.00	0.00	0.00	0.00	1.00
why_X_you	23	0.00	0.04	0.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
are_X_going	101	0.00	0.00	1.08	0.00	0.00	0.03	0.00	1.00	0.01	0.99	0.00	0.00
the_X_and	53	0.00	0.00	1.92	0.00	0.85	0.34	1.00	0.00	0.00	0.15	0.00	0.00
what_X_it	64	0.00	0.03	0.00	0.17	0.27	0.22	0.83	0.00	0.02	0.00	0.00	0.78
put_X_back	39	2.05	0.00	0.00	0.13	0.00	0.03	0.00	0.00	1.00	0.00	0.97	0.00
did_X_do	71	0.00	0.00	1.63	0.00	0.00	0.06	0.00	1.00	0.00	1.00	0.00	0.00
you_X_have	71	0.00	0.03	1.61	0.03	0.01	0.15	0.01	0.96	0.00	0.97	0.00	0.01
the_X_on	46	0.00	0.02	0.78	0.00	1.00	0.91	1.00	0.00	0.00	0.00	0.00	0.09
I_X_it	103	0.00	1.17	0.01	0.00	0.17	0.47	1.00	0.00	0.00	0.00	0.00	0.53
I_X_think	3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
to_X_it	115	0.00	0.56	0.00	0.00	0.99	0.00	1.00	0.00	0.00	0.00	0.00	1.00
you_X_a	133	0.00	1.95	0.00	0.05	0.20	0.99	0.95	0.00	0.00	0.00	0.00	0.00
want_X_to	62	2.53	0.00	0.00	0.02	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00
what_X_I	50	0.00	0.00	0.00	0.98	1.00	0.98	0.02	0.00	0.00	0.00	0.00	0.02
you_X_it	206	0.00	1.30	0.03	0.02	0.21	0.03	0.98	0.00	0.00	0.00	0.00	0.97
on_X_table	55	0.05	0.00	0.04	0.91	0.00	0.00	0.00	0.00	0.00	1.00	1.00	0.00
are_X_doing	110	0.00	0.00	1.43	0.00	0.00	0.01	0.00	1.00	0.00	1.00	0.00	0.00
the_X_in	49	0.00	0.02	0.92	0.00	1.00	0.82	1.00	0.00	0.00	0.00	0.00	0.18
would_X_like	92	0.00	0.00	1.29	0.00	0.00	0.01	0.00	1.00	0.00	1.00	0.00	0.00
to_X_with	63	0.00	0.46	0.78	0.03	0.97	0.03	0.97	0.00	0.03	0.00	0.00	0.97
you_X_on	49	0.00	0.84	0.96	0.00	0.27	0.00	1.00	0.00	0.00	0.00	0.00	1.00
you_X_your	85	0.00	2.01	0.00	0.06	0.34	0.99	0.94	0.00	0.00	0.00	0.00	0.01
what_X_you	373	0.00	0.01	0.00	0.99	0.99	0.98	0.01	0.00	0.00	0.00	0.00	0.01
you_X_the	111	0.00	2.31	0.00	0.04	0.16	1.00	0.96	0.00	0.00	0.00	0.00	0.00
to_X_a	75	0.00	1.31	0.00	0.00	0.95	1.00	1.00	0.00	0.00	0.00	0.00	0.00
you_X_with	35	0.00	1.63	1.00	0.03	0.11	0.00	0.97	0.00	0.00	0.00	0.00	1.00
put_X_on	46	1.17	0.00	0.83	0.39	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00
a_X_one	51	0.00	0.00	0.22	0.00	0.00	0.80	0.04	0.96	0.00	1.00	0.00	0.00
do_X_want	125	0.00	0.00	1.54	0.00	0.00	0.02	0.00	1.00	0.00	1.00	0.00	0.00
I_X_you	84	0.00	1.10	0.00	0.00	0.14	0.71	1.00	0.00	0.00	0.00	0.00	0.29
you_X_that	61	0.00	1.38	0.00	0.00	0.07	0.25	1.00	0.00	0.00	0.00	0.00	0.75
there_X_is	39	0.05	0.00	0.15	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00
it_X_the	72	0.00	0.99	0.00	0.93	0.94	1.00	0.07	0.00	0.01	0.00	0.00	0.00
you_X_in	61	0.00	0.89	0.70	0.08	0.41	0.02	0.92	0.00	0.00	0.00	0.00	0.98
a_X_of	66	0.00	0.15	1.00	0.00	0.80	0.00	1.00	0.00	0.00	0.00	0.00	1.00
what_X_that	241	0.01	0.17	0.07	0.04	0.13	0.10	0.96	0.00	0.00	0.00	0.00	0.90
the_X_one	51	0.00	0.00	0.10	0.00	0.00	0.78	0.04	0.96	0.00	1.00	0.00	0.00
Average		0.19	0.54	0.50	0.15	0.33	0.39	0.52	0.25	0.10	0.28	0.10	0.29

*The length of each blue bar represents the number of that cell.

Table A4 Mean number of links per token for each bigram in Eve corpus*

Bigrams	Token count	External links				Internal links			
		w1_in	w2_in	w1_out	w2_out	w1_in_i	w2_in_i	w1_out_i	w2_out_i
right_X	203	0.01	0.16	0.03	0.85	0.00	0.97	0.97	0.00
have_X	604	2.57	0.05	0.36	0.63	0.36	0.00	0.00	0.36
don_X	448	0.05	0.00	0.98	0.00	1.00	0.00	0.00	1.00
here_X	142	0.04	0.92	0.55	0.52	0.01	0.43	0.43	0.01
'll_X	621	0.01	2.78	0.05	0.25	0.00	0.95	0.95	0.00
what_X	1387	0.05	0.42	0.58	0.63	0.04	0.40	0.40	0.04
't_X	816	0.00	2.45	1.00	0.35	0.00	0.00	0.00	0.00
your_X	845	0.00	0.03	0.11	0.96	0.00	0.89	0.89	0.00
this_X	234	0.00	0.39	0.18	0.77	0.00	0.82	0.82	0.00
there_X	303	0.06	0.98	0.50	0.54	0.01	0.50	0.50	0.01
I_X	1327	0.00	0.97	0.48	0.56	0.00	0.52	0.52	0.00
and_X	704	1.33	0.35	0.25	0.57	0.43	0.00	0.00	0.43
no_X	407	0.00	0.10	0.81	0.87	0.00	0.19	0.19	0.00
are_X	502	0.54	0.08	0.65	0.71	0.27	0.03	0.03	0.27
one_X	181	0.66	0.57	0.57	0.71	0.10	0.35	0.35	0.10
we_X	571	0.01	0.92	0.60	0.70	0.01	0.40	0.40	0.01
in_X	784	0.85	0.05	0.87	0.80	0.18	0.00	0.00	0.18
okay_X	80	0.00	0.25	0.80	0.79	0.03	0.18	0.18	0.03
not_X	452	0.02	1.08	0.81	0.61	0.05	0.13	0.13	0.05
is_X	758	1.17	0.11	0.17	0.25	0.72	0.03	0.03	0.72
see_X	235	1.79	0.03	0.38	0.39	0.61	0.00	0.00	0.61
me_X	250	0.01	0.33	0.99	0.98	0.00	0.01	0.01	0.00
do_X	531	0.96	0.16	0.73	0.62	0.37	0.01	0.01	0.37
put_X	388	3.04	0.04	0.28	0.42	0.58	0.00	0.00	0.58
going_X	461	3.03	0.08	0.12	0.84	0.16	0.00	0.00	0.16
on_X	594	0.83	0.04	0.89	0.82	0.17	0.00	0.00	0.17
where_X	283	0.01	0.76	0.33	0.36	0.00	0.67	0.67	0.00
he_X	441	0.00	1.07	0.39	0.51	0.00	0.61	0.61	0.00
's_X	2855	1.28	0.47	0.18	0.49	0.40	0.22	0.22	0.40
're_X	481	0.64	1.27	0.19	0.32	0.26	0.55	0.55	0.26
well_X	415	0.00	0.26	0.83	0.83	0.00	0.17	0.17	0.00
did_X	351	0.20	0.20	0.81	0.78	0.14	0.08	0.08	0.14
yes_X	343	0.00	0.08	0.87	0.88	0.00	0.13	0.13	0.00
with_X	293	0.53	0.01	0.92	0.50	0.49	0.00	0.00	0.49
it_X	1383	0.00	0.91	0.53	0.62	0.01	0.47	0.47	0.01
to_X	1163	0.06	1.01	0.18	0.86	0.13	0.81	0.81	0.13
can_X	342	0.07	1.24	0.52	0.43	0.30	0.44	0.44	0.30
want_X	321	2.62	0.04	0.04	0.59	0.41	0.00	0.00	0.41
oh_X	389	0.00	0.14	0.73	0.75	0.00	0.27	0.27	0.00
for_X	299	0.27	0.01	0.88	0.21	0.78	0.01	0.01	0.78
Eve_X	315	0.03	0.48	0.69	0.75	0.03	0.29	0.29	0.03
the_X	1919	0.00	0.06	0.14	0.97	0.00	0.86	0.86	0.00
a_X	1320	0.00	0.15	0.31	0.93	0.01	0.69	0.69	0.01
that_X	1335	0.02	0.98	0.19	0.35	0.00	0.80	0.80	0.00
Average		0.52	0.51	0.51	0.62	0.18	0.32	0.32	0.18

*The length of each blue bar represents the number of that cell.