What distributional information is useful and usable for language acquisition?

Padraic Monaghan (pjm21@york.ac.uk)
Department of Psychology, University of York
York, YO23 1ED, UK
Morten H. Christiansen (mhc27@cornell.edu)
Department of Psychology, Cornell University
Ithaca, NY 14853, USA

Abstract

Numerous theories of language acquisition have indicated that distributional information is extremely valuable for assisting the child to learn syntactic categories, yet these theories differ over the type of information that is proposed as useful in acquisition. Mintz (2003) has proposed that children utilize the previous word and the following word (AxB frames) for acquiring categories, whereas Monaghan, Chater, and Christiansen (submitted) have suggested that information about the previous word alone provides a rich source of data for categorization. In three modeling experiments we found that bigrams were better than fixed AxB frames for learning syntactic categories in a corpus of child-directed speech. However, presentation of the preceding and succeeding words when these can be processed separately resulted in better learning than presenting the preceding word alone, and also improved performance over presenting the previous two words.

Introduction

What sort of information does the child use to develop an understanding of their language? The rational analysis approach answers this question by assessing what sort of information is *useful* for learning the language. If a particular source of information proves to be rich and reliable then a computational system (of which the child is a very special case) will exploit it. The child learns a sense of syntactic categories early in language development. In order to understand speech and relate it to the world, the child must know which part of speech refers to an action, and which to objects, and which words modify relations between objects. “Look at the cow mooing” elicits many possibilities for relations between words and the world, for example, whether the animal in question is referred to by the word “cow”, “look”, or “mooing”. Constraints *within* the language, restricting which words in the sentence can refer to objects, for example, greatly limit the number of possibilities for relating words to the world.

But what sort of information is useful for constructing syntactic categories? A variety of different types of information have been proposed as useful for categorization, including gestural, semantic, phonological, and distributional information. Combining more than one type of information has indicated improvements in categorization (Reali, Christiansen, & Monaghan, 2003), and it may indeed be the case that combining multiple sources is necessary for categorization to take place (Braine, 1987).

This paper focuses on distributional information as a cue for syntactic categorization, and questions what type of information is most useful and thus usable by the child. Theories of the use of distributional information in language acquisition have suggested different analyses of the context in which a word (category) occurs, but no empirical comparisons of these competing accounts have been made. We present a series of computational models that compare the extent to which accurate syntactic categorization of language directed to the child can be made on the basis of different sources of distributional information.

Sources of distributional information

Theories of distributional information in language acquisition have tended to focus on demonstrating that such information can contribute significantly toward categorization, rather than proposing that the particular implementation is psychologically realistic. Redington, Chater, and Finch (1998) produced context vectors based on the two preceding words and the two words following the target word from the CHILDES (MacWhinney, 2000) database of child-directed speech. The resulting vectors for the most frequent 1000 words in the database clustered together with a high correspondence to syntactic categories. Redington et al. (1998) also assessed vectors resulting from using different context words. They found that good results were also obtained for the one preceding and one following word, and also for the two preceding words, and for the two succeeding words (with better performance for preceding words than succeeding words). Yet, using only the immediately preceding word also resulted in good performance, though addition of richer contextual information improved performance.

An alternative approach is the proposal that particular sequences of words are useful for determining syntactic category. Fries (1952) produced a set of “frames” in which only words of a certain category could appear. For example, only a noun could appear in “The __ is/was/are good”. Similarly, Maratsos and Chalkley (1980) proposed that there were local constraints on the occurrence of particular word categories, such as that only a verb can occur before the inflection *–ed*.

Mintz (2003) provided an empirical test of this local source of information, by analyzing corpora of child-directed speech for the occurrence of frames of the preceding and the succeeding words. We refer to these as AxB frames, where A and B are fixed, and x indicates the intervening word. For example, for the frame “you __ to”, “go” and “have” both occur as “x” words in the frame.

Mintz selected the 45 most frequent frames involving the preceding and succeeding word, and then grouped the words that occurred within each of these frames. In the above example, "go" and "have" would be grouped together in the analysis. Accuracy was assessed by counting the number of times that words of the same category were grouped together, and dividing this by the number of pairings of all words within the groups. Completeness was determined by counting the number of pairings of words of the same category within the group, and dividing this by the number of pairings of words of the same category occurring in any of the groupings.

The 45 most frequent frames resulted in high accuracy but low completeness, indicating that these frequent AxB frames grouped together words of the same category, but that many words of the same category tended to occur in different groups. Relatedly, Mintz (2002) found that people categorized words together when they occurred in AxB frames in an artificial language learning task, and consequently claimed that such AxB frames were a source of distributional information that children used to acquire syntactic categories.

An alternative proposal is that a frame involving only the preceding word – an Ax frame – is required in order to produce effective categorization (e.g., Valian & Coulson, 1988). Monaghan, Chater, and Christiansen (submitted) found that categorizations of child-directed speech based on the association between the 20 most frequent preceding words and the target word resulted in accurate classification of words of different categories, but critically, also resulted in a large proportion of words being classified. Additionally, Monaghan et al. showed that, in an artificial language learning task, participants could group words on the basis of Ax frame information alone.

Both AxB and Ax frames can therefore be exploited in learning artificial languages, but which source of information is most useful to the child learning their language? AxB frames result in high accuracy, but low completeness, whereas Ax frames produce high completeness at the expense of some accuracy. Should a learning system select accuracy over completeness, or vice versa?

A comparison of different sources of distributional information requires that alternative methods are subjected to the same analyses. In addition, an empirical test of whether accuracy or completeness is a priority in acquisition is necessary. We now present a series of modeling experiments that test the extent to which different types of distributional information lead to successful categorization of words in child-directed language. Experiment 1 replicated Mintz's (2003) analysis of AxB frames in child-directed speech, and directly compared the resulting classification to an Ax analysis. Experiment 2 assessed whether a neural network model learned to categorise words more accurately on the basis of AxB information or Ax information alone. Finally, Experiment 3 tested a neural network model learning from AxB information when the relationship between A and x and B and x can also contribute separately towards categorization, and compared performance to a model with information about the two preceding words.

Experiment 1

Method

Corpus preparation From the CHILDES database, we selected a corpus of speech directed towards a child of age 0-2;6 years (anne01a-anne23b, Theakston, Lieven, Pine, & Rowland, 2001). This was one of the corpora used by Mintz (2003). We replaced all pauses and turn-taking with utterance boundary markers, and the resulting corpus contained 93,269 word tokens in 30,365 utterances (mean utterance length = 3.072 words). There were 2,760 word types, and the syntactic category for these words was taken from the CELEX database (Baayen, Pipenbrock, & Gulikers, 1995), according to the most frequent category usage for each word. Some interjections, alternative spellings, and proper nouns were hand-coded. There were 12 syntactic categories: noun, adjective, numeral, verb, article, pronoun, adverb, conjunction, preposition, interjection, wh-words (e.g., *why*, *who*), and proper noun.

Analysis In accordance with Mintz (2003), we selected the 45 most frequent AxB frames from the corpus, and determined the words that occurred in the x position within each frame. Each AxB frame thus resulted in a cluster of words. Accuracy and completeness were assessed in the same way as for Mintz (2003), described above. An additional method for assessing completeness was taken as the total number of word types that were classified in (at least) one frame.

For the Ax analysis, the 45 most frequent words were selected from the corpus, and co-occurrence with these frequent words formed the clusters in the bigram analysis. Accuracy and completeness were assessed in the same way as for the AxB co-occurrence analysis.

Results

As an example of the resulting classification, Table 1 shows a summary of the words that were classified into the 5 most frequent AxB and Ax frames. For these most frequent AxB frames, two frames clustered verbs together, and two clustered only pronouns. For the Ax classifications, the results are noisier, but have far higher numbers of words classified. The most frequent Ax frame – "the x" – classifies 623 nouns, and very few verbs, whereas the next most frequent Ax frame – "you x" – classifies 210 verbs, and only 26 nouns. The accuracy and completeness results are shown in Table 2, together with those from Mintz (2003)[1]. In parentheses are the random baseline values. We closely replicated Mintz's (2003) results indicating the high accuracy of the AxB frames, though, as noted in the

[1] Data are shown from Mintz's analysis of the anne corpus, with standard labeling and word-type analyses.

Table 1. Classifications based on the 5 most frequent Ax and AxB frames.

Ax						
Ax	noun	verb	pronoun	adjective	preposition	other
a	335	33	2	56	0	11
it	37	69	12	29	13	43
to	76	107	16	6	1	9
you	26	210	15	27	8	39
the	623	23	9	38	5	14

AxB						
Ax**B**	noun	verb	pronoun	adjective	preposition	other
do_think	0	0	1	0	0	0
do_want	0	0	6	0	0	0
are_going	0	0	5	0	0	0
what_you	0	10	0	0	1	0
you_to	0	19	2	1	1	1

Table 2. Completeness and accuracy of classifications for the Ax and the AxB co-occurrence models.

	CO-OCCURRENCE MODEL		
	MINTZ	AX	AXB
Accuracy	0.94 (0.41)	0.57 (0.22)	0.88 (0.26)
Completeness	0.09 (0.04)	0.07 (0.04)	0.06 (0.03)
Words classified	405, 14.7%	1930, 69.9%	394, 14.3%

Introduction, there was very low completeness for this classification. The Ax analysis also resulted in high accuracy, and slightly higher completeness according to Mintz's definition. However, a striking difference between the AxB and the Ax analyses is the overall number of words from the corpus that were categorized. Clustering based on bigrams resulted in a classification of almost 5 times as many words as the trigram analysis. The small differences in completeness between the two analyses is therefore misleading, as this only considered words that were clustered – in the AxB case, completeness was assessed over only a fraction of the corpus considered in the Ax analysis.

Discussion

We successfully replicated Mintz's (2003) demonstration that classifications of syntactic category based on occurrence within the most frequent AxB frames resulted in impressively high accuracy. However, our prediction that high accuracy could also be achieved by the smaller, less specific Ax frame was supported. The Ax analysis had the additional advantage of enabling a classification of far more words from the child's environment than was possible using AxB frames. There is a pay-off between accuracy and completeness: a specific context will result in high accuracy, but low completeness, whereas a general context will result in lower accuracy but high completeness.

This raises the question as to whether categorization is best based on information that renders highly reliable classifications of only a few words, or whether learning would benefit from using information that classifies a larger proportion of the words in the environment, but with the possibility that such classifications may contain more errors.

One way to test this issue is to train a neural network to base predictions of the syntactic category of words based on either AxB frames, or Ax frames. After training, the neural network model's error on the predicted classifications reflects the extent to which the given source of information is beneficial for learning the syntactic categories of the language. If the model trained on AxB frames has lower error then learning is more effective when based on high accuracy but low completeness, whereas if the model trained on the Ax frames has lower error then high completeness at the expense of high accuracy is a better source of information for learning.

We were concerned with how effective the frame is in predicting the category of the x word, so we trained the models to predict the category of x without entering the identity of the x word at the input. In addition, we did not preselect the frames that were input into the model: the entire corpus was used for training and not just the 45 most frequent frames, as we were interested in whether the model would be able to pick up which frames were useful for categorisation. From Mintz's (2003) analysis, it is not clear whether the AxB frames are to be interpreted as non-compositional, or whether the relationship between A and x and between x and B may also contribute to categorization. Experiment 2 tests the non-compositional interpretation, whereas Experiment 3 assesses the compositional version of the AxB frames.

Experiment 2

We trained two neural network models to learn to predict the category of the target (x) word using the same corpus of child-directed speech as in Experiment 1. We compared the learning of models that were given either Ax or AxB information. The AxB model was designed to test whether the AxB frame was useful for learning when the frame is interpreted as a whole, i.e., the "A" and the "B" do not contribute separately toward classification.

Architecture

Ax model The model was a feed-forward network with a set of input units fully-connected to a hidden layer, which was fully-connected to an output layer. The model is shown in Figure 1. Each unit in the input layer represented one word type in the child-directed speech corpus (so there were 2,760 input units), and there was also a unit representing the utterance boundary, in accordance with other connectionist models of syntax learning (e.g., Elman, 1990) that provide this additional information to the simulated child learner. There were 10 units in the hidden layer. The output layer contained units representing the syntactic category of the next word in the corpus. The model was trained on all Ax bigrams in the corpus, with the first word in the bigram occurring in the input layer, and the category of the second word in the bigram as the target at the output layer.

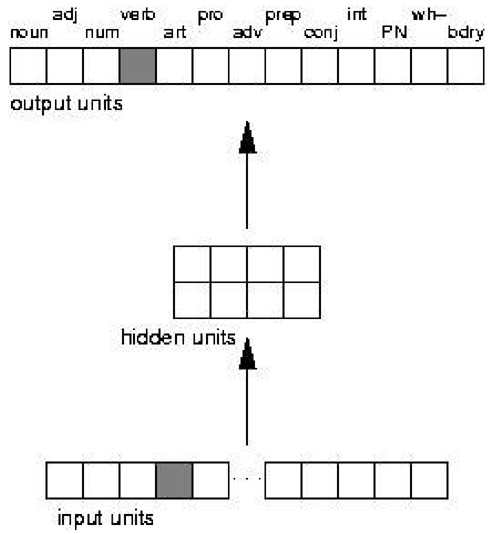

Figure 1. The feedforward neural network model of syntactic categorization. The active input unit represents either the A-word in the Ax model, or the AxB frame in the AxB model. The active output unit is the category of the x word, or the utterance boundary if x represents the end of the utterance. In the Figure, the output verb unit is active.

AxB model The AxB model was identical to that of the Ax model, except that in the input layer each unit represented one of the possible AxB frames. There were 36,607 such AxB frames, and so there were 36,607 input units in the model. The model was trained on all AxB frames in the corpus, with the A_B frame activating the appropriate unit in the input layer, and the syntactic category of the x word as the output layer target.

Training and testing

The models were trained using backpropagation with gradient descent with learning rate 0.01, and momentum 0.95. Before training, the weights between connections were randomized with mean 0 and standard deviation 0.1. We imposed a 0.1 error tolerance on the output units to prevent the development of very large weights on the connections. The models were trained on all Ax or AxB frames in the corpus, with each epoch being one pass through the corpus, and training was halted after 5 epochs, which was over 600,000 training events. As a baseline, we trained and tested the Ax model and the AxB model on a corpus where the frequency of words was maintained, but word-order was randomized. In the AxB randomized control model, there were 44,786 AxB frames and thus 44,786 input units in the model.

The models were tested after each epoch on the whole corpus, with the mean square error (MSE) across the output units taken as a measure of the ability of the model to learn to categorize words in the corpus on the basis of either the Ax or the AxB information. As an additional measure, we assessed whether the target unit – that is, the appropriate category of the x word – was the most highly activated for each pattern presentation.

Table 3. Percent correctly classified and MSE for the Ax and AxB models for each syntactic category in the corpus, with number of tokens (n) and *t*-test on MSE (all $p < 0.001$).

		% CORRECT		MSE		
CATEGORY	N	Ax	AxB	Ax	AxB	*t*
Nouns	12458	66.3	0	0.533	1.000	-116.316
Adjectives	4125	1.9	0	1.116	1.035	21.373
Numerals	1087	0	0	1.128	1.040	20.304
Verbs	23182	83.9	0	0.511	0.851	-145.602
Articles	7996	31.0	0	0.848	1.025	-52.371
Pronouns	18932	47.6	0	0.675	0.869	-71.369
Adverbs	5456	0	0	1.150	1.040	46.221
Prepositions	9491	31.3	0	0.865	1.016	-34.894
Conjunctions	1955	0	0	1.147	1.032	29.448
Interjections	3762	0	0	0.984	1.026	-24.608
Proper nouns	2104	0	0	1.149	1.032	28.642
Wh-words	3500	0	0	1.041	1.024	7.510
Boundary	30365	79.6	100	0.446	0.793	-147.391
TOTAL	**123634**	**52.4**	**22.9**	**0.680**	**0.911**	**-205.957**

Results

The Ax model performed better than the random baseline, MSE was 0.680 compared to 0.920, $t(247266) = -189.808$, $p < 0.001$. The model also classified more words correctly than the random baseline: 52.4% compared to 22.9%, $\chi^2 = 75{,}014.859$, $p < 0.001$.

The AxB model performed at a level similar to the random baseline. MSE was 0.911 which was slightly higher than the randomized version of 0.910, $t(247264) = 4.418$, $p < 0.001$. Classification was poor, with the model classifying all words as the utterance boundary, which was the single most frequent token in the input, This behavior was identical to the performance of the AxB model on the randomized corpus.

Table 3 shows the comparison between the Ax and the AxB models, for all words, and for each syntactic category. In terms of MSE, performance was better for the Ax model than the AxB model on all categories apart from adjectives, numerals, adverbs, conjunctions, proper nouns, and wh-words. However, performance was better for the large closed-class categories – pronouns and articles – and for nouns and verbs. Overall, the Ax model classified more words correctly than the AxB model, $\chi^2 = 75{,}014.011$, $p < 0.001$.

Discussion

The Ax model performed significantly better than chance in predicting the category of the x word from the preceding word. The AxB model performed at a chance level, and did not discriminate any word category. The better performance of the AxB model in terms of MSE on adjectives, numerals, adverbs, conjunctions, proper nouns and wh-words may have been due to a broader context serving these categories better: adverbs often occur after nouns in positions normally taken by verbs, and adjectives intervene between determiners and nouns. An enriched context would undoubtedly assist the categorization of these types. However, the better performance may merely have been due to a lack of discrimination between any of the word types in the AxB model.

These simulations demonstrated that categorization of a large, entire corpus of child-directed speech was best achieved using information about the preceding word, rather than information about set frames comprised of the preceding and the following word. Greater coverage of the set of words, rather than greater accuracy in categorization, resulted in better performance.

The next experiment assessed whether a compositional treatment of the AxB frame may provide better information about the syntactic category of the target x word than the Ax frame alone, and compared it to a model with information about the two preceding words.

Experiment 3

We trained neural network models to learn to predict the category of the next word from the same corpus of child-directed speech as used in Experiments 1 and 2. We compared the learning of a model that was given information about the preceding and the following word in order to predict the category of the intervening word, but could operate on this information separately and combined. We call this the AxB-compositional (AxB-c) model. We also tested a model where information was given about the two preceding words: the ABx model. Note that these models embed the bigram information from the Ax model in the input. We predicted that both models would perform better than both the Ax model and the non-compositional AxB model from Experiment 2. We also predicted that the AxB-c model would outperform the ABx model, as proximity to the target word is most informative.

Architecture and training

The AxB-c model had the same architecture as the Ax model in Experiment 2, except that it had two banks of input units. In the first bank of units the unit corresponding to the A-word was activated, and in the second bank of units the B-word unit was activated. At the output layer, the model had to learn to predict the category of the x word. The same architecture was used for the ABx model, but it had as input the two words preceding the target word.

Training and testing was identical to that for the models in Experiment 2. Baselines for learning were determined by training and testing the models on the randomized corpus.

Results

For both models, performance was better than the random baseline in terms of accurate classifications and MSE. For the AxB-c model, accuracy was 69.4% (baseline 22.9%), $\chi^2 = 82422.148$, $p < 0.001$, and MSE was 0.480 (baseline 0.920), $t(247266) = -329.487$, $p < 0.001$. For the ABx model, accuracy was 56.3% (22.9%), $\chi^2 = 60841.166$, $p < 0.001$, and MSE was 0.628 (0.920), $t(247266) = -221.728$, $p < 0.001$.

As predicted, both the AxB-c and the ABx model performed with greater accuracy than the non-compositional AxB model from Experiment 2 for all syntactic categories: overall, $t(123633) < -300$, $p < 0.001$, for each individual syntactic category, all $t < -50$, all $p < 0.001$.

Table 4. Percent correctly classified and MSE for the AxB-c and ABx models. *T*-tests are computed on MSE (all $p < 0.001$, except † $p < 0.1$).

	% CORRECT		MSE		
CATEGORY	AxB-c	ABx	AxB-c	ABx	*t*
Nouns	73.7	68.0	0.408	0.509	-43.808
Adjectives	25.8	0	0.878	1.167	-44.306
Numerals	0	0	1.185	1.149	5.969
Verbs	85.4	86.6	0.289	0.466	-77.029
Articles	67.6	38.7	0.490	0.827	-72.861
Pronouns	80.5	53.5	0.361	0.585	-81.153
Adverbs	20.8	0	0.976	1.151	-33.207
Prepositions	59.0	37.8	0.592	0.807	-50.213
Conjunctions	0.5	0	1.140	1.148	$-1.409^{\dagger}$
Interjections	80.8	0	0.671	0.957	-71.643
Proper nouns	0.1	0	1.214	1.155	11.694
Wh-words	38.6	0	0.817	1.006	-23.613
Boundary	84.7	85.8	0.283	0.350	-26.769
TOTAL	**69.4**	**56.3**	**0.480**	**0.628**	**-147.470**

Compared to the Ax model in Experiment 2, the additional word information in the AxB-c and ABx models resulted in an increase in accurate classifications. For both models, classification was more accurate ($p < 0.001$), and resulted in lower error, both $t < -300$, $p < 0.001$. For the individual syntactic categories, the AxB-c and the ABx model performed better for all syntactic categories apart from numerals, all $t < -50$, all $p < 0.001$, though the difference for conjunctions was non-significant.

Table 4 compares the AxB-c model to the ABx model, indicating that accuracy was lower and MSE higher in the ABx model. The AxB-c model performed better on all syntactic categories apart from numerals and proper nouns.

Discussion

Providing decomposable information about the preceding and following word resulted in increased accuracy of performance in the model. The AxB-c model classified words of all syntactic categories better than the non-compositional AxB and the Ax models of Experiment 2. Accuracy across all the categories was high, though classifications of adjectives and adverbs was still inaccurate – these tended to be classified as nouns/pronouns and verbs, respectively. Adding information about the two preceding words also assisted in increasingly accurate classifications, though not to the same degree as providing the preceding and succeeding word.

General Discussion

Experiment 1 demonstrated, as predicted, that AxB information provides high accuracy at the expense of completeness, whereas Ax information results in slightly lower accuracy but much higher coverage of the language.

Experiment 2 tested the extent to which a computational model could utilize AxB frame information in categorizing the intervening word. The model trained on AxB frames performed at slightly below chance level, and well below the accuracy that could be achieved from categorizing on the basis of Ax information alone. The high completeness of Ax frames resulted in significantly better learning than the high accuracy but low-coverage of AxB information.

However, when the model is able to learn on the basis of AxB information when this information is compositional, i.e., the relationship between the preceding word and the target word and between the succeeding word and the target word can be computed separately, then a different picture emerges. The AxB-c model of Experiment 3 was more accurate than the Ax model of Experiment 2. Furthermore, this provided better classification results than the two preceding words (the ABx model), though this latter model also improved performance over a non-compositional AxB frame or just the single preceding word.

The simulations presented here suggest that learning is most effective when information about the preceding word and the succeeding word is available. However, this is only the case when the AxB frame is not computed as a whole. Learning must also be based in part on the relationship between A and x and between x and B. In the experiments presented in Mintz (2002), such a distinction is not made – the learning situation resembles that of the AxB-c model, where the participant has access not only to the AxB frame, but also to the Ax and the xB bigrams. Therefore, it is not yet possible to distinguish the contribution of bigram and trigram information in adult learning situations (though see Onnis et al., 2003).

The possibility remains that the requirement for category learning depends on establishing distinctions and similarities between only a few words in the language: it is not realistic or feasible to attempt to learn the whole language simultaneously. However, performance for the most frequent 100 words was poorer in the non-compositional AxB model than the Ax model, and even taking only those words that occurred in the most frequent 45 AxB frames resulted in poorer performance than for the 45 most frequent Ax frames.

The experiments presented in this paper require the models to learn pre-ordained syntactic categories. The task facing the child is more difficult: the child must also construct the categories. Yet, both tasks concern learning about which words co-occur. When the relationship between the occurrence of certain categories in particular distributional contexts is easy to learn then this demonstrates that the category itself is more clearly defined.

We have shown that AxB frames provide poor information about categorization unless this information is componential, such that Ax information is also available. We suggest that the distributional information that a neural network model finds most useful is more likely to be used by the child in acquiring syntactic categories.

Acknowledgments

This research was supported in part by a Human Frontiers Science Program Grant (RGP0177/2001-B).

References

Baayen, R.H., Pipenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Braine, M.D.S. (1987). What is learned in acquiring word classes: A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp.65-87). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Fries, C.C. (1952). *The Structure of English: An Introduction to the Construction of English Sentences.* New York: Harcourt, Brace & Co.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, Third Edition. 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* Volume 2, pp.127-214. New York: Gardner Press.

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

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

Monaghan, P., Chater, N., & Christiansen, M.H. (submitted). The differential contribution of phonological and distributional cues in grammatical categorization.

Onnis, L., Christiansen, M.H., Chater, N., & Gómez, R. (2003). Reduction of uncertainty in human sequential learning: Evidence from artificial grammar learning. *Proceedings of the 25th Cognitive Science Society Conference* (pp. 887-891). Mahwah, NJ: Lawrence Erlbaum.

Reali, F., Christiansen, M.H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: Scaling-up the connectionist approach to multiple-cue integration. *Proceedings of the 25th Cognitive Science Society Conference* (pp. 970-975). Mahwah, NJ: Lawrence Erlbaum.

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

Theakston, A.L., Lieven, E.V.M., 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.

Valian, V. & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*, 71-86.