Learning grammatical categories from distributional cues:

Flexible frames for language acquisition

Michelle C. St. Clair

University of Manchester, Manchester, UK

Padraic Monaghan

Lancaster University, Lancaster, UK

Morten H. Christiansen

Cornell University, Ithaca, USA

Word count: 15694

Abstract word count: 147

Corresponding author:

Michelle C. St. Clair
Division of Human Communication and Deafness
School of Psychological Sciences
Ellen Wilkinson Building
University of Manchester
Oxford Road
Manchester
M13 9PL
+44 (0)161 275 8677
Michelle.StClair@manchester.ac.uk

Abstract

Numerous distributional cues in the child's environment may potentially assist in language learning, but what cues are useful to the child and when are these cues utilised? We propose that the most useful source of distributional cue is a flexible frame surrounding the word, where the language learner integrates information from the preceding and the succeeding word for grammatical categorisation. In corpus analyses of child-directed speech together with computational models of category acquisition, we show that these flexible frames are computationally advantageous for language learning, as they benefit from the coverage of bigram information across a large proportion of the language environment as well as exploiting the enhanced accuracy of trigram information. Flexible frames are also consistent with the developmental trajectory of children's sensitivity to different sources of distributional information, and they are therefore a useful and useable information source for supporting the acquisition of grammatical categories.

Introduction

During the language acquisition process, the infant is sensitive to multiple, interacting cues that assist in determining the language structure (e.g., Kelly, 1992; Monaghan, Christiansen, & Chater, 2007; Saffran, 2001; Vallabha, McClelland, Pons, Werker, & Amano, 2007). A focus of research in child language acquisition is to determine what these cues are, and how they are used by the child. In experimental and corpus-based studies on speech segmentation, for instance, many cues have been isolated that affect language learner's judgments about where words begin and end. Transitional probabilities between syllables (Saffran, Aslin, & Newport, 1996), lengthening of final syllables (Saffran, Newport, & Aslin, 1996), prosodic cues in terms of stress placement (Curtin, Mintz, & Christiansen, 2005; Theissen & Saffran, 2003), as well as phonotactic information about which phonemes occur word medially and which occur only across word boundaries (Hockema, 2006; Mattys, White, & Melhorn, 2005) all contribute to decisions about word boundaries, and interact in intriguing ways (e.g., Johnson & Jusczyk, 2001). Yet, the study of grammatical category learning has had substantially less scrutiny, but is likely to be at least as complicated a process.

There have, nonetheless, been notable exceptions that have provided crucial insight into the processes of grammatical category acquisition. Some of this work has explored how information from the word itself, including lexical stress (e.g., Kelly & Bock, 1988) and phonological patterning (e.g., Cassidy & Kelly, 2001; Durieux & Gillis, 2001; Monaghan et al., 2007), can provide cues to grammatical categories. Most of this research, though, has focused on the usefulness of *distributional* information for grammatical category learning (e.g., Finch & Chater, 1992; Mintz, 2003; Mintz,

Newport, & Bever, 1995; Redington, Chater, & Finch, 1998; Valian & Coulson, 1988). The distributional patterns in which a word tends to occur have thus been shown to be quite informative about its potential grammatical category. Indeed, distributional information may be the most reliable source for categorisation, especially for high frequency words (Monaghan, Chater, & Christiansen, 2005). However, to create psychological accounts of how learners might use distributional information for grammatical category acquisition, researchers have thus far faced a difficult choice between accuracy and coverage of this information. On the one hand, focusing on relatively detailed distributional patterns, or "frequent frames" (such as trigrams), can result in very accurate grammatical classification of words but at the cost of only covering a small part of the child's language exposure, because highly specific patterns tend to be sparse in the input (Mintz, 2003). On the other hand, more coarse distributional patterns (such as bigrams) can provide a broad coverage of the input but with a lower accuracy in terms of the grammatical classification of individual words (Monaghan & Christiansen, 2008).

In this paper, we propose a novel account to overcome the accuracy versus coverage problem in the distributional learning of grammatical categories: We hypothesise that children construct accurate higher-order "flexible frames" (trigrams) on the fly from lower-order distributional patterns (bigrams) with broad coverage. In what follows, we first discuss key research on the use of distributional information in grammatical category learning, highlighting the work on distributional cues by Mintz (2002, 2003) employing fixed, frequent frames. Results from corpus analyses and connectionist modelling are then reported replicating the original results of Mintz (2003),

while demonstrating the advantage of flexible frames over fixed frames as a means of distributional learning of grammatical categories. We conclude by considering the implications of our flexible frames approach to distributional learning for research on language acquisition.

Distributional Approaches to Grammatical Category Learning

Extending earlier work by Finch and Chater (1992), Redington et al. (1998) demonstrated that distributional information from child-directed speech provided accurate information about the grammatical categories of words, and showed that such information could form the basis of self-organising clusters of words into categories. In their analyses, they assessed co-occurrences for each word based on the previous two words, and the succeeding two words. This study was foundational in demonstrating the potential for grammatical category learning on the basis only of distributional information of words in text, yet the plausibility of all this information being used by the child acquiring her language would require ascribing a vast memory for co-occurrences between thousands of words. A consequent question, then, is what sources of distributional information may be computationally tractable to the child determining the pattern of grammatical categories within the language?

Maratsos and Chalkley (1980) proposed that local distributional information in speech may form the basis of grammatical categories, and, hence, may be a vital starting point for language learning. Essentially, grammatical categories were hypothesized to be constructed based on the overlap between frequently occurring phrases – if words X, Y and Z are heard within the same position in constructions A, B and C, then X, Y and Z

will be abducted into the same category. This allows generalisations such that if X is heard in syntactic construction D, then Y and Z should be allowed in the same position in syntactic construction D, even if they had never been heard within that particular sentential environment before. Cartwright and Brent (1997) implemented a model of this approach by examining "minimal pairs" of phrases in child-directed speech. Thus, when "the dog sat" and "the cat sat" both occurred in speech, the frame "the _ sat" and the category set {dog, cat} were extracted. Cartwright and Brent's (1997) model was effective in demonstrating that local information was available to generate sets of words that often corresponded to grammatical categories, but, as with Redington et al.'s (1998) analysis, it was computationally intensive, and additionally it resulted in an only partial coverage of the words in child-directed speech.

Building upon these previous studies, Mintz (2003) proposed that local, high frequency co-occurrences in child-directed speech could form the basis for the derivation of grammatical categories. He suggested that very frequently occurring non-adjacent pairs of words in speech would not only be available but also *useful* to the language learner as a scaffold for grouping words of the same grammatical category. For instance, when the words "to _ to" co-occur separated by one word ("__" indicates another word), then the words that can intervene tend to be of the same grammatical category – in this case, they tend to be verbs. In corpus analyses of child-directed speech, Mintz (2003) found that the 45 most frequent frames were extremely accurate in reflecting the grammatical categories of the words that occurred within them. Thus, words from the same grammatical category tended to occur in the same frame.

In a related study, Mintz (2002) tested whether such non-adjacent frames could be used by adult learners to categorise words in an artificial language learning (ALL) study. In this language, non-adjacent trigram frames were available to categorise a set of medial words together. The results of this study showed that participants had grouped the medial words together as a category, which indicated that they were able to utilise the trigram frame information in order to create the category.

Mintz' (2002; 2003) corpus analysis and ALL study together demonstrate that non-adjacent frames are not only useful but are also potentially useable by the language learner: they do not require intensive memory or computational resources, and so are presumably tractable to the child in early stages of development. Additionally, it was demonstrated that participants respond to the co-occurrences when learning an artificial language. Such analyses are a critical first step in understanding the potential sources of distributional information available to the child for discovering the grammatical categories of words. The next step, however, is to determine precisely what sources of information are used by the child in grammatical category learning. Though Mintz (2002; 2003) has demonstrated that a language containing frequent frames can be learned, the frequent frame structure contained other sources of distributional information to which participants may have been sensitive during learning, and frequent frames are therefore just one candidate for the sort of information that assists in grammatical categorisation. In particular, frequent frames also contain two bigrams – distributional information about the word's category from the preceding as well as the succeeding word. These sources may operate separately, in learning from bigram structure, or jointly, in terms of frequent frames. So, for the frequent frame "to _ to", the two bigram frames "to _" and "_ to" may

each contribute to the coherence of the category of words that appears in these trigrams (i.e., verbs tend to succeed the word "to", and also tend to precede the word "to"). So, perhaps bigram information is the key source of distributional information in grammatical category learning?

Bigram information, on its own, has been shown to be useful for category learning in ALL studies. Valian and Coulson (1988) found that bigram cues could induce word categorisation within an artificial language learning paradigm, provided the bigram frame words were much more frequent than the words being categorised. Using a similar paradigm, Monaghan et al. (2005) and St. Clair and Monaghan (2005) have also demonstrated grammatical category acquisition from bigrams (see also Smith, 1966; St. Clair, Monaghan, & Ramscar, 2009). Thus, it is not yet clear whether trigram or bigram information, or even some other candidate, such as the more complex co-occurrence statistics explored in Redington et al.'s (1998) corpus analyses, may direct the child's category learning.

Yet, there are several reasons to suspect that *fixed* trigram frames, as proposed by Mintz (2002), are unlikely candidates for the initial cues exploited by children in learning grammatical categories. First, frequent frames present the problem of sparsity, as we will confirm in the corpus analyses. In Mintz' (2003) corpus analyses, the 45 most frequent frames classified only a small fraction of the corpus. This kind of sparse data problem has received considerable attention in computational linguistics, because of similar issues to those applicable in language learning, concerning the trade-off between a highly enriched context providing accurate categorisation (or parsing) and the reduced frequency of that specific context and its low probability of reoccurrence (Manning & Schütze, 1999).

Hence, most natural language processing approaches have rejected trigrams as a basis for parsing due to their sparsity. For similar reasons, various forms of bigrams rather than trigram are typically favoured for modelling visual word recognition (e.g., Grainger & Whitney, 2004). We therefore predict that trigrams are not likely to be as useful a distributional cue as bigram information for a child learning the categories of the language. Our first experiment presents corpus analyses of child-directed speech, testing whether the sparsity problem is evident in child-directed speech just as it is in broader language corpora (Manning & Schütze, 1999), and also whether this sparsity problem is resolved by computations based on flexible, interacting bigram distributional information.

A second reason why fixed frames may be an unlikely source of information for categorisation is due to their difficulty of use. There is bountiful evidence of learning adjacent (i.e., bigram) statistics by young infants (Saffran, Newport et al., 1996), yet non-adjacent dependencies, such as in fixed frequent frames, are only available at later stages of development (Gómez & Maye, 2005), and even then they are difficult to learn (Endress, Dehaene-Lambertz, & Mehler, 2007; Onnis, Monaghan, Richmond, & Chater, 2005). Not only do frequent frames offer sparse coverage, they are also only available to use by the language learner under certain specific circumstances, even when they provide perfect categorisation information (e.g., Gómez, 2002).

A third limitation of fixed frequent frames is that they can only categorise words that are regularly surrounded by other frequent words. This applies well in English for grammatical categories that may occur surrounded by a limited number of function words, such as verbs, which are often preceded by a pronoun and succeeded by an article

(e.g., “you _ the”), or nouns which are often preceded by an article, and succeeded by an auxiliary or a preposition (e.g., “the _ can”, “a _ on”), but it is unlikely to provide information about membership for other categories, such as adverbs or adjectives, which are generally adjacent to a content word of lower frequency. Indeed, fixed frames appear to provide a rather limited cue to grammatical categories in German, which have a relatively large number of functions words e.g., because articles are marked for case, number and gender (Stumper, Bannard, Lieven, & Tomasello, 2010), and also seem similarly inappropriate for Dutch (Erkelens, 2009). Equally, function words are unlikely to be classified accurately due to the relatively low frequencies of the words that surround them. Thus, there are strong constraints imposed by the overall distributional structure of the language that limit which categories can be formed on the basis of fixed frames, generally restricting learning to noun and verb categories in English and, by extension, other Indo-European languages.

Trigram information, such as in Mintz’ (2003) analyses clearly captures a great deal of information about grammatical categories in children’s early language exposure, and so is likely to have some influence on learning. Yet, there are alternative ways in which trigram information may be exploited for classification, at the same time avoiding the problems of sparsity associated with fixed frames. We propose instead that children make use of *flexible* frames in categorisation – where the preceding and the succeeding word provide converging and integrative information about the category of the intervening word. Such flexible frame information is already present in the fixed frames of Mintz’ (2002; 2003) experiment and analyses, but is not maximally exploited by the learner if the frequent frames are treated as compositional wholes. We show that

exploiting all the distributional information present in the trigram provides both accuracy and coverage for categorising the words in child-directed speech. First, we compare the extent to which frequent frame trigram information and frequent bigrams support the development of knowledge about the grammatical categories of words in child-directed speech. We then show, by assessing computational models trained on different sources of distributional information, that flexible frames provide the optimal basis for category learning.

Experiment 1: Corpus Analysis

The corpus analyses were designed to first replicate Mintz' (2003) demonstration that frequent frame trigrams provide highly accurate information about grammatical categories in child-directed speech. The second aim was to test whether frequent bigrams can also provide accurate information about grammatical categories, and to directly compare categorisations based on trigrams and bigrams.

Method

Input corpora. We selected the same six corpora of child-directed speech from the CHILDES corpus (MacWhinney, 2000) used by Mintz (2003): Anne and Aran (Theakson, Lieven, Pine, & Rowland, 2001), Eve (Brown, 1973), Naomi (Sachs, 1983), Nina (Suppes, 1974), and Peter (Bloom, Hood, & Lightbrown, 1974; Bloom, Lightbrown, & Hood, 1975). As in Mintz (2003), only the sessions in which the child was 2;6 years or younger were analysed. All utterances from children were excluded, leaving only adult speech spoken in the presence of the child.

The Aran corpus contained aran1a to aran20b, with the exception of 14a and 14b. The Nina corpus contained nina01 to nina23, with the exception of nina8. The remaining corpora are identical to those in Mintz's (2003) analyses (Anne corpus: anne01a to anne23b; Eve corpus: eve01 to eve20; Naomi corpus: n01 to n58; Peter corpus: peter01 to peter12). The actual corpora used were slightly different to those used by Mintz (2003), as some of the subcorpora are no longer available in CHILDES.

The analysis was performed on the CHILDES MOR line, which coded the grammatical category of each word. This categorisation has an accuracy of approximately 95% correct (Sagae, MacWhinney, & Lavie, 2004). Before the corpora were analysed, all punctuation, pause marking, trailing off and interruption markings were replaced with an utterance boundary marker, as all of these markings either signalled the end of a sentence or a break in the utterance. Coded on the MOR line in some of the corpora were words that were grammatically necessary but not actually said (grammatical omissions). As these words were not spoken they were deleted from the analysis. In the CHILDES database, any repetitions of words that were marked with "[/]" to indicate the repetition (e.g., "duck [/] duck") were fully transcribed on the normal transcription line, but only one version of the repetition was transcribed on the corresponding grammatical category line (MOR line). All repetitions were inserted into the MOR grammatical category line by hand.

Analysis Procedure. Each corpus was analysed separately. The procedure for the fixed trigram, frequent frames analysis will be covered in detail; a similar process was applied for the bigram analyses. A list of all consecutive three word phrases was compiled.

None of these three word phrases crossed utterances boundaries. The 45 most frequent frames within each corpus were then selected with all the words that occurred within them, we denote these frames aXb, where a__b refers to the non-adjacent co-occurrence frame, and X refers to the set of words that occur in this context. For the fixed frame analysis, then, the most frequent a_b non-adjacent co-occurrences were selected and the words that intervened between the a and the b word were grouped together. For the preceding bigram analysis, denoted aX, the most frequent 45 words were selected, and all the words that occurred immediately following one of these most frequent 45 words were grouped together. For the succeeding bigram analysis – Xb – the words that occurred immediately before each of the most frequent 45 words were grouped. Both token (every word occurrence was counted) and type (only distinct words were counted) analyses were performed.

For the objective grammatical category labels, Mintz (2003) used two classifications, termed standard and expanded labelling. There were ten categories of words in the standard labelling: nouns (including pronouns), verbs (including auxiliaries and copula forms), adjectives, prepositions, adverbs, determiner, wh-words, "not", conjunctions, and interjections. Expanded labelling simply divided the noun and pronoun distinctions and the verb, auxiliary and copula distinctions into separate categories. As the differences in the results between standard and expanded labelling were small in both Mintz' study and our own analyses, only standard labelling is reported.

In order to determine how well the frequent frames were able to categorise the words, accuracy and completeness measures were computed so as to be comparable to Mintz (2003). The accuracy measure assessed categorisation success by looking at pairs

of medial words within the frames. All the words that occurred within one of the frames contributed to measures of accuracy and completeness, but, crucially, this did not include all the words that occurred in the child's corpus. In order to determine accuracy, the number of hits (when two words occurring in the same frame were of the same grammatical category) was divided by the number of hits plus the number of false alarms (when two words occurring in the same frame were from a different grammatical category) (accuracy = hits/(hits + false alarms)). Accuracy gave an overall measure of how successful the distributional cues were at grouping words of the same grammatical category together.

Completeness measured how well the distributional cues grouped all words from one grammatical category together in the same distributional cue grouping. Completeness was the number of hits divided by the number of hits plus the number of misses (when two words of the same category occurred in different frames) (completeness = hits/(hits + misses)).

Both accuracy and completeness had values in the range [0,1]. A value of 1 for accuracy meant that the distributional category contained only one type of grammatical category (e.g., comprised only adjectives). A completeness score of 1 meant that the frame contained all of the words from a particular grammatical category, e.g., a distributional category that contained all adjectives (but potentially other grammatical categories as well).

To establish a random baseline, all of the words that were categorised in the analysis were randomly assigned across the 45 frequent frames to create a random analysis. Each frequent frame category contained the same number of random tokens

that were found in the frequent frame based analysis. This baseline provides a measure of accuracy and completeness that would be expected if the frequent frames did not aid in grouping words of the same grammatical category together. These random baseline values are shown in parentheses in the relevant tables.

Results

All multiple *t* tests and pairwise comparisons were Bonferonni corrected. Each individual corpus was entered as a separate "subject" in the statistical analyses.

aXb analysis. The 45 most frequent aXb fixed frames for the Aran corpus are shown in Appendix I. Table 1 displays the data for the total number of word tokens and types within each of the six child-directed speech corpora and the number of tokens and types categorised in the aXb frames. It was found that on average 6.3% of the total word tokens and 19.4% of the total word types were categorised in this analysis.

Table 1. Summary of the total number of word tokens and types in the corpora and the number and percentage of tokens and types included in the aXb analysis.

Corpus	*Corpus tokens*	*Corpus types*	*Tokens categorized*		*Types categorized*	
			N	*%*	*n*	*%*
Anne	95255	2602	4870	5.1	388	14.9
Aran	106931	3249	6041	5.6	745	22.9
Eve	60929	2125	3430	5.6	386	18.2
Naomi	28979	1877	1725	5.9	315	16.8
Nina	70867	1968	6252	8.8	463	23.5
Peter	74170	2127	5204	7.0	429	20.2
Mean	**72855**	**2325**	**4587**	**6.3**	**454**	**19.4**

Table 2 displays the accuracy and completeness measures for the token and type analyses. The aXb fixed frames analysis had much higher accuracy than the random baseline for both token and type analyses, $t(5) = 36.47$, $p < .001$ and $t(5) = 58.21$, $p < .001$, respectively. The completeness measure was also significantly higher than the random baseline for both token and type analyses, $t(5) = 15.49$, $p < .001$ and $t(5) = 13.56$, $p < .001$, respectively.

Table 2. Token and type accuracy and completeness measures for the aXb corpora, random baseline values are in parentheses.

Corpus	*Token Accuracy*	*Type Accuracy*	*Token Completeness*	*Type Completeness*
Anne	.94 (.28)	.82 (.30)	.07 (.03)	.08 (.04)
Aran	.88 (.27)	.76 (.25)	.08 (.03)	.09 (.04)
Eve	.95 (.32)	.80 (.32)	.06 (.03)	.07 (.03)
Naomi	.94 (.32)	.87 (.34)	.07 (.03)	.06 (.04)
Nina	.96 (.32)	.84 (.30)	.08 (.04)	.10 (.05)
Peter	.92 (.28)	.82 (.32)	.07 (.03)	.08 (.04)
Mean	**.93 (.28)**	**.82 (.31)**	**.07 (.03)**	**.08 (.04)**

aX analysis. The 45 most frequent aX frames from the Aran corpus are shown in Appendix I. Table 3 shows a summary of the total word types and tokens categorised in the aX frames. An average of 42.9% of the tokens and 85.6% of the word types were analysed, which was substantially higher than in the aXb analyses (compare Table 1). For the aX frames, accuracy was higher than the random baseline for token and type analyses, $t(5) = 23.81$, $p < .001$ and $t(5) = 18.05$, $p < .001$, respectively. Completeness was also higher than the random baseline, $t(5) = 14.00$, $p < .001$ and $t(5) = 10.25$, $p < .001$ for token and type analyses, respectively (see Table 4).

Xb analysis. Appendix I reports the 45 most frequent Xb frames from the Aran corpus. Table 5 shows the data for the number of word tokens and types categorised in the Xb succeeding word analysis from each corpus.

Table 3 Summary of the total number of word tokens and types in the corpora and the number and percentage of tokens and types included in the aX analysis.

Corpus	*Corpus tokens*	*Corpus types*	*Tokens categorized*		*Types categorized*	
			n	*%*	*n*	*%*
Anne	95255	2602	39071	41.0	2235	85.9
Aran	106931	3249	47822	44.7	2810	86.5
Eve	60929	2125	23890	39.2	1776	83.6
Naomi	28979	1877	11598	40.0	1503	80.1
Nina	70867	1968	34402	48.5	1811	92.0
Peter	74170	2127	32384	43.7	1813	85.2
Mean	**72855**	**2325**	**31528**	**42.9**	**1991**	**85.6**

Table 4 Token and type accuracy and completeness measures for the aX corpora, random baseline are in parentheses.

Corpus	*Token Accuracy*	*Type Accuracy*	*Token Completeness*	*Type Completeness*
Anne	.48 (.17)	.41 (.20)	.08 (.04)	.07 (.04)
Aran	.43 (.16)	.38 (.18)	.07 (.03)	.07 (.04)
Eve	.51 (.17)	.42 (.19)	.09 (.04)	.07 (.04)
Naomi	.52 (.18)	.47 (.19)	.10 (.04)	.08 (.04)
Nina	.55 (.19)	.47 (.23)	.10 (.05)	.08 (.04)
Peter	.48 (.17)	.36 (.16)	.08 (.04)	.07 (.04)
Mean	**.50 (.17)**	**.42 (.19)**	**.09 (.04)**	**.07 (.04)**

Table 5 Summary of the total number of word tokens and types in the corpora and the number and percentage of tokens and types included in the Xb analysis.

Corpus	*Corpus tokens*	*Corpus types*	*Tokens categorized*		*Types categorized*	
			n	*%*	*n*	*%*
Anne	95255	2602	36101	37.9	1843	70.8
Aran	106931	3249	45006	42.1	2469	76.0
Eve	60929	2125	20807	34.1	1268	59.7
Naomi	28979	1877	10156	35.0	1082	57.6
Nina	70867	1968	28955	40.9	1446	73.5
Peter	74170	2127	28661	38.6	1400	65.8
Mean	**72855**	**2325**	**28281**	**38.1**	**1585**	**67.2**

Word groupings based solely on the succeeding word were more accurate than expected by chance for both the token and type analyses, as can be seen in Table 6, $t(5) = 20.05$, $p < .001$ and $t(5) = 12.65$, $p < .001$, for tokens and types, respectively. Completeness was again higher than chance, $t(5) = 13.15$, $p < .001$ and $t(5) = 6.71$, $p < .005$, for tokens and types, respectively.

Table 6 Token and type accuracy and completeness measures for the Xb corpora, random baseline are in parentheses.

Corpus	*Token Accuracy*	*Type Accuracy*	*Token Completeness*	*Type Completeness*
Anne	.31 (.17)	.25 (.18)	.06 (.04)	.04 (.03)
Aran	.32 (.17)	.27 (.19)	.07 (.04)	.04 (.03)
Eve	.29 (.16)	.25 (.18)	.07 (.04)	.05 (.03)
Naomi	.31 (.16)	.24 (.17)	.07 (.04)	.04 (.03)
Nina	.34 (.17)	.30 (.21)	.08 (.05)	.05 (.03)
Peter	.29 (.17)	.22 (.17)	.07 (.05)	.04 (.03)
Mean	**.31 (.17)**	**.26 (.18)**	**.07 (.04)**	**.04 (.03)**

Comparing the aXb, aX and Xb analyses. The categorisations based on the three types of frames were directly compared using three one-way ANOVAs, on the number of words categorised, and on accuracy and completeness. When the results did not differ qualitatively for the token and type analyses, only the token ANOVA results are reported. To determine which analyses categorised the largest portion of the whole corpus, the dependent variable was the number of word tokens or types categorised from the whole corpus. There was a significant main effect of distributional frame, $F(2,15) = 12.74$, $p = .001$, $\eta_p^2 = .63$. Pairwise comparisons indicated that the aXb analysis categorised fewer tokens than both the aX and Xb analyses, both $p < .01$, but the aX and Xb analyses did not differ ($p = 1.0$).

For accuracy of token categorisation, there was a significant main effect, $F(2, 15) = 635.38$, $p < .001$, $\eta_p^2 = .99$. Pairwise comparisons indicated that the aXb frames were more accurate than the aX frames, which were more accurate than the Xb frames, all $p < .001$.

For completeness, there was a significant main effect for word tokens, $F(2, 15) = 6.23$, $p < .05$, $\eta_p^2 = .45$; pairwise comparisons indicated that there was a significant difference between the aX and aXb frames, $p < .05$, and between the aX and Xb frames, $p < .05$, but there was no difference between the Xb and aXb frames ($p = 1.0$). This result demonstrated that the aX analysis produced distributional categories that captured more total words of the same grammatical category than the aXb and Xb analysis. For completeness in the type analyses, there was a significant main effect, $F(2, 15) = 29.01$, $p < .001$, $\eta_p^2 = .80$, with pairwise comparisons revealing that the aXb and aX frames were

significantly higher than the Xb frames, both $p < .001$, but did not differ from one another, $p = .64$.

Discussion

We were successful in replicating the general findings of Mintz's (2003) analyses of the aXb frequent frames. Mintz (2003) found that for token and type analyses, accuracy was extremely high, with means of .93 and .91, respectively. For our aXb frame analyses, we also found very high accuracy, with mean for type analyses of .93 and .82 for token analyses[1]. For completeness, Mintz (2003) found that the aXb frames were significantly greater than chance, .08 and .10 for token and type analyses, respectively. Our analyses were similar, with .07 and .08 for token and type analyses, respectively. Though significantly greater than chance, this indicates that approximately eleven out of every twelve occurrences of each word occurred in different frames, meaning the categorisation of words was somewhat diffuse (though see the frame-joining mechanism proposed in Mintz, 2003, for addressing this problem).

As anticipated, the aX and Xb frames were less accurate than the aXb analysis, due to their reduced specificity, but aX frames were also found to be more accurate than Xb frames, indicating that high frequency preceding words were more effective at classifying target words than succeeding words. This supports the hypothesis that frequent frames are highly accurate in categorising words. However, in the completeness scores, it was aX frames which resulted in the highest values for the token analyses, greater than both aXb and Xb frames, indicating that the aX frames resulted in large,

[1] The current results differ slightly from those reported in Mintz (2003), which can be explained by slight differences in the input corpora and small differences in occurrences of certain trigrams. St. Clair (2007) provides a full account of the slight differences between Mintz (2003) and the current results.

generic categories where each grammatical category was more likely to be represented by a few frames. Note that the completeness scores are for the set of words categorised, so for the aX frames in the type analysis, completeness is .09 of 42.9% of the whole corpora, and for the aXb analysis it is .07 of 6.3% of the corpora, resulting in an 8-fold difference across all corpora. Thus frequent frames are highly accurate, but initial bigrams have a much wider coverage of the corpora than frequent frames.

As we have indicated above, both these measures of accuracy and completeness were only across the words classified within the frames. The aX and Xb frames contained more types (more than three times as many) and tokens (more than six times as many) than the aXb frames, and so direct comparisons of accuracy and completeness between these types of distributional information are biased in favour of analyses containing a smaller number of words. Mintz (2003) points out that the number of types categorised in his analyses is double the words categorised in studies by Cartwright and Brent (1997) and reports that "the types constituting half of the tokens in each corpus were contained in the 45 most frequent frames" (Mintz, 2003, p.98). This is certainly the case, but because few of the token instances actually occurred within these frames it remains to be seen whether the category of a frequent word can be determined when just a few of its occurrences are within a frequent frame but the majority of the time occurring in other contexts. Equally, it remains speculative to claim that it cannot.

The corpus analyses indicated that the child may exploit either sparse but accurate frequent frame contexts, or bountiful but more approximate bigram information for categorisation. However, we propose that flexible frames present a third way which exploits the merits of *both* trigram accuracy and bigram coverage. We hypothesise that

the best learning situation is when the frequent frames are decomposed into their constituent bigrams. So, for each word that occurs in an aXb frame, the learner can be informed as to the grammatical category from information about the preceding *and* the succeeding word simultaneously. Thus, flexible frames function as a committee of experts, whereby each bigram generates a hypothesis about the grammatical category of the intervening word, and this combination allows incorrect categorisations based on only one bigram to be remedied by the other bigram source.

We next present a series of simulations in order to determine which source of distributional information may be most conducive to category learning for a statistical learning system which utilises child-directed speech corpora to determine grammatical categories. We compare connectionist models that learn to map between different types of distributional information and the grammatical category of each word in the corpus. We compare the learning from fixed frames (aXb), either preceding (aX) or succeeding (Xb) bigrams, and flexible frames (aX+Xb), in order to determine the usefulness of each of these methods as a basis for categorising the language to which a child is exposed. If the model learns categories more effectively from the aXb frames compared to the aX or Xb frames then this indicates that a general purpose learning system finds accuracy is most beneficial to learning at the expense of coverage of the language, whereas if the model learns better from the bigram frames this indicates that broad coverage is better for learning. However, if, as hypothesised, the flexible frame (aX+Xb) model shows the best learning then this indicates that decomposing the trigram information is best for learning categories of words.

Experiment 2: Computational Modelling of Trigrams and Bigrams

In order to test how effective learning could be based on the aXb fixed frames compared to the aX, Xb, and aX+Xb flexible frame information, we trained a feedforward connectionist model to learn to map between the distributional information of the frame and the grammatical category of the word within the frame (whether that was the fixed frame, or the flexible frame). Feedforward connectionist models are computationally similar to multiple regression techniques, and the model is intended to reflect general purpose learning mechanism responding only to the statistics of the distributional information in the child-directed speech. The model we employ had a hidden layer between the input distributional information and the output grammatical category, meaning that potential interactions between different sources of input information could occur, and so is equivalent to a multiple regression model where all terms can interact. However, we used a connectionist model instead of a statistical model as the connectionist model enables learning over time to be observed and not only a snapshot of the final result of the ideal learner. In this respect the connectionist model provides insight into the plausibility of learning grammatical categories from the language environment from the various types of distributional information.

Method

Architecture. The model comprised an input layer, which was connected to a hidden layer of 200 units, which was connected in turn to an output layer. We trained four different versions of the model corresponding to each of the four types of distributional information. In the aXb model, each unit in the input layer represented one of the aXb

frames in the corpus. So, one particular unit in the input was activated when a particular frame was inputted from the language. For the aX+Xb model, there was a set of units for the word preceding the target word and another set for the word following the target word. So, when a frame was inputted from the language, one unit was active to represent the preceding word and another was active to represent the following word. The aXb and aX+Xb models are shown in Figure 1 when inputting the sequence “you go to”. “go” specifies that the verb unit in the output should be active in both models. For the aXb model, the unit in the input representing “you_to” is active, and for the aX+Xb model, one unit representing the preceding word “you” and another unit representing the succeeding word “to” are active. In the aX model (not illustrated), only the preceding word unit was active, and in the Xb model, only the following word unit was active.

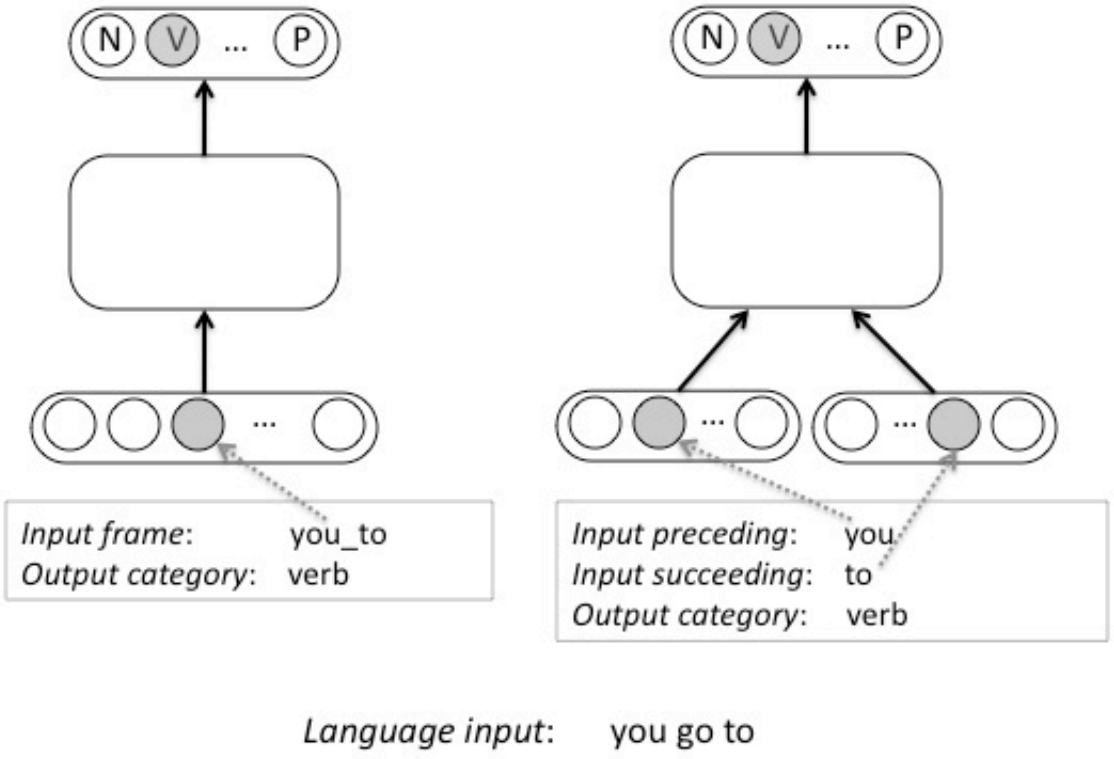

Figure 1. The aXb (left) and aX+Xb (right) models processing the input "you go to". Solid line arrows indicate fully connected layers of units in the model, dashed line arrows indicate how the input translated into units activity in the model.

The output units represented each of the grammatical categories in either the standard or expanded labelling, as described in the corpus analyses. The results section reports only the standard labelling models, as the expanded labelling resulted in similar performance and distinctions between the different distributional information types. All frames, and not only the 45 most frequent, were included in the model's input to determine whether additional information about category would be available from all frames with the particular distributional characteristics.

Training and testing. The model was trained to produce an activity of 1 on the output unit corresponding to the grammatical category of the word within the frame. So, when the sequence "you go to" was selected the input units corresponding to "you_to" were activated, and the model had to learn to activate the verb unit at the output. Units in the hidden and output layers of the model were activated in the range [0,1] and activation was a sigmoid function of the sum of the units connecting to the unit multiplied by the weights on the connections. The model learned to solve the task by adjusting the weights on the connections between units to get closer to the target activation at the output, using the backpropagation learning algorithm with learning rate 0.1.

Input-output patterns were selected randomly according to the frequency with which they occurred in each child's corpus. We tested the model's performance at two stages of training, first after 10,000 words had been presented to the model, to test whether categorisation was effective after limited input to the model, and we also tested the model's performance after more extensive training up to 100,000 patterns presented. These two stages of performance will indicate whether aXb frames provide more accurate information initially, or whether bigrams appear to be the best early foundation of learning, and also provide data on whether, after greater exposure to the language, one type of distributional information is more effective as a basis for grammatical categorisation. At each stage, the model's performance was tested by presenting the whole of the child's corpus and determining whether the model produced an output activation closer to the target grammatical category than to any other unit in the set of output units. In these analyses separate computation of accuracy and completeness is not appropriate as the model's ability to classify the entire corpus is calculated, and therefore

the results are analysed in terms of the model's ability to correctly predict the category of each word in the whole corpus.

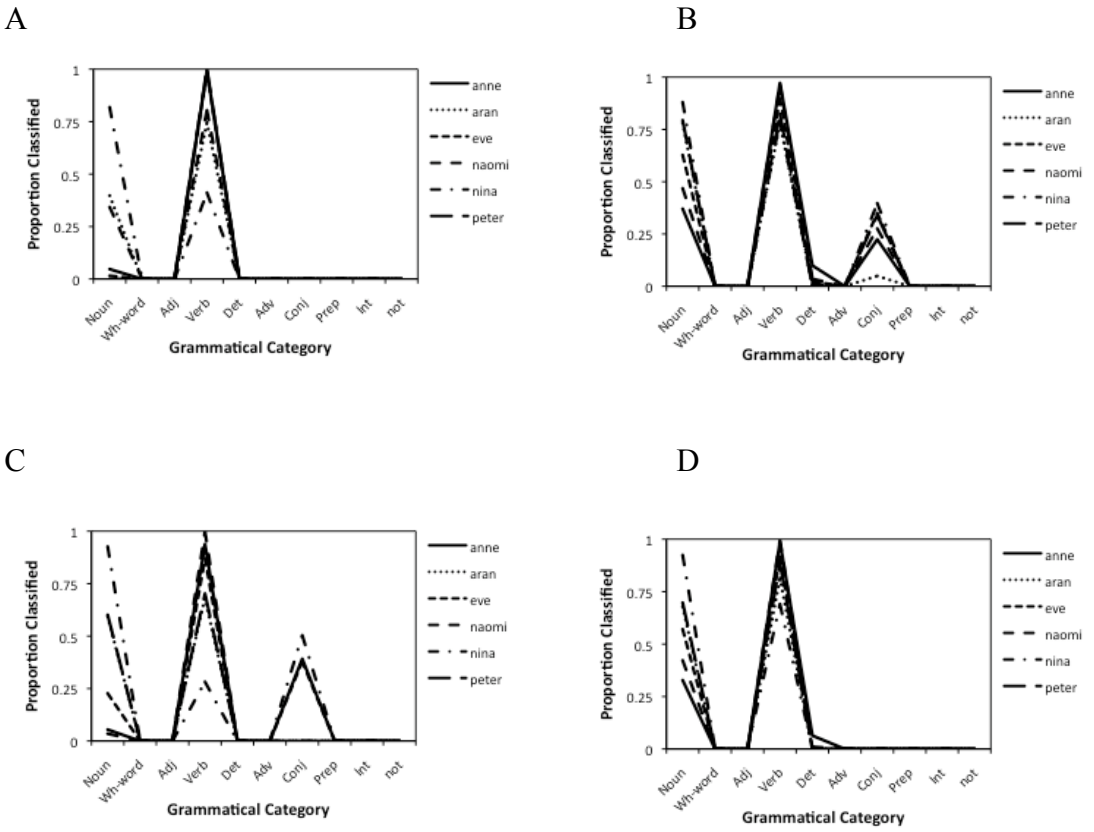

Figure 2. Proportion correctly classified for each grammatical category in each child-directed speech corpus after 10,000 word presentations, for (A) aXb fixed frames, (B) aX+Xb flexible frames, (C) aX frames, and (D) Xb frames. Categories are for standard labelling.

Results

Table 7 shows the number of distinct aXb, aX, and Xb frames in each child corpus. These values indicate that each preceding bigram word occurs with an average of between five and nine succeeding bigram words to produce the trigram frequent frames. We analysed

the output of the computational model to determine whether the precise pairing of particular preceding and succeeding words as in the aXb frames is beneficial for learning, or whether the independent contribution of these words better benefits learning grammatical categories within the language.

Table 7. Number of distinct aXb, aX and Xb frames in each child's corpus.

Corpus	Distinct aXb frames	Distinct aX frames	Distinct Xb frames
Anne	15984	1874	2132
Aran	23499	2710	3100
Eve	9224	1310	1496
Naomi	5293	980	1202
Nina	11555	1494	1714
Peter	10939	1318	1617

Early training stage. We tested the model's performance early in training, after 10,000 word tokens had been presented. Figure 2 shows the results of the model's classification accuracy for each grammatical category in the standard labelling, for each type of distributional information. For each child, we determined the accuracy of the overall classification of the model by determining how many words of each category were classified into each of the 10 categories in the standard labelling. We then calculated the asymmetric lambda value (Goodman & Kruskal, 1954) on the ability of the model to predict the categories of the words. Lambda is a measure of association between two classifications, and is particularly appropriate when one classification precedes the other temporally or causally, as then the extent to which the second classification system predicts variance in the first can be determined, and thus applies in the case where we measure the extent to which the distributional information classifications can reflect the

given grammatical categories. Lambda provides a parametric measure of the predictive accuracy of the model from a set of non-parametric categorisations (the lambda value indicates the reduction in prediction error of each word's category given the model's classifications), and provides a value in the range [0,1], where 0 indicates zero predictability and therefore no association between predicted and actual categories, and 1 indicates a perfect association. Consequently, and unlike other non-parametric tests of association such as χ^2 or Cramer's V, the lambda values for each classification can be compared to each other by converting the difference between the values into a z-score. The lambda value is also advantageous over other measures of association as it is asymmetric, and so determines the extent to which the model's judgments approximate the actual categories, so the lambda value is thus unaffected by the ability of the actual categories to predict the model's judgments. The formula for the asymmetric lambda value of association is:

$$\lambda_b = \frac{\sum_{j=1}^{k} n_{Mj} - \max(R_i)}{N - \max(R_i)}$$

where the table of cross-classifications has r rows and k columns, n_{Mj} is the highest frequency in the jth column of the table, max(R_i) is the largest row total, and N is the total number of observations (Goodman & Kruskal, 1954). The formula for variance of lambda is:

$$var(\lambda_b) = \frac{\left(N - \sum_{j=1}^{k} n_{Mj}\right)\left(\sum_{j=1}^{k} n_{Mj} + \max(R_i) - 2\sum' n_{Mj}\right)}{[N - \max(R_i)]^3}$$

where $\sum' n_{Mj}$ is the sum of all the maximum frequencies for each column that are in the row *i* (Siegel & Castellan, 1988).

In the following analyses, we used a conservative test of the difference in lambda values by taking the higher of the two standard deviation values for each of the two lambdas in calculating the z-score of the difference. Table 8 reports the mean accuracy across all the words in each child corpus in the standard labelling, as well as the lambda values, and the z-score for the differences between each of the classifications.

Table 8. Classification accuracy and lambda values for the standard labelling after 10,000 words of training.

Corpus	aXb		aX+Xb		aX		Xb		Difference in λ values as z-scores					
	Acc	λ	Acc	λ	Acc	λ	Acc	λ	aX+Xb-aXb	aX - aXb	Xb - aXb	aX+Xb - aX	aX+Xb - Xb	aX - Xb
Anne	.33	.00	.52	.26***	.47	.17***	.40	.06***	78.76***	67.03***	25.58***	26.69***	60.87***	43.99***
Aran	.36	.05***	.56	.33***	.53	.29***	.45	.17***	92.10***	76.50***	44.10***	12.85***	52.00***	37.60***
Eve	.39	.01***	.53	.22***	.46	.11***	.41	.03***	49.86***	33.27***	6.13***	26.75***	45.72***	27.31***
Naomi	.41	.05***	.59	.33***	.54	.26***	.48	.15***	32.78***	24.33***	11.92***	8.92***	20.76***	12.33***
Nina	.38	.00	.63	.41***	.57	.30***	.48	.15***	93.92***	70.12***	31.63***	24.12***	53.60***	31.71***
Peter	.42	.06***	.58	.31***	.54	.25***	.46	.13***	46.74***	33.36***	12.09***	11.47***	33.98***	21.59***
All	.37	.02***	.56	.32***	.52	.24***	.44	.13***	119.57***	89.84***	42.22***	41.60***	101.57***	62.53***

Note. Tests for lambda values are against zero association, *** indicates $p < .001$.

Though the corpus sizes are very different, and the number of distinct fixed frames and bigrams are highly variable, there is remarkable consistency in the classification accuracy of each model for each child's language environment, both in terms of overall accuracy as well as the lambda values of the association between predicted and actual grammatical categories. Over all corpora and all grammatical categories, the aX+Xb flexible frame model resulted in the highest accuracy, with the aXb frames classifying words more poorly even than the aX and Xb frame models. The aXb model resulted in accuracy of 37-41% correct classification for each of the corpora, yet this was due to the model classifying nearly all words as verbs or all words as nouns (see Figure 2A), and so the overall accuracy of the classification is better reflected in the lambda values, which for the aXb model were small, and not significantly different from zero for two of the six corpora.

In contrast, the aX+Xb model achieved a high degree of accuracy, even after this small quantity of training, resulting in more than half of the words correctly classified in each corpus, and lambda values were highly significantly different from a zero association. aX+Xb frames were effective at classifying a substantial proportion of nouns, the verbs, and conjunctions (see Figure 2B). The comparison between lambda values for the aXb model and the aX+Xb model indicated a huge difference in classification accuracy, with z-scores of the difference ranging from 31.45 to 69.98 across the corpora.

The Xb model was less accurate than the aX model, indicating that the word preceding each target word resulted in greater accuracy of classification than the succeeding word. The aX model was effective in classifying either the nouns or the verbs

as well as some of the conjunctions, the Xb model effectively classified a generally high proportion of both the nouns and the verbs. However, both the single bigram models attained a level of classification accuracy higher than that of the fixed frame model, and both were less accurate than the flexible frame model, indicating that the greater frequency of occurrence of bigrams in the corpus resulted in better categorisation performance, but that combining two bigrams resulted in better performance than using single bigrams. The z-scores of the differences were highly significant in each case.

After a small amount of exposure to each corpus, the bigram models were able to categorise accurately several of the grammatical categories in each corpus, whereas the aXb model was poorer at learning the objective categories of the words.

Later training stage. Though we were interested in determining whether each type of distributional information is an effective contributor to grammatical categorisation early in language development, therefore after a small quantity of exposure to the language, it may be that the aXb fixed frame model, due to the rarity of occurrences of each input frame, would demonstrate a learning advantage after more extensive training. Figure 3 shows the results of the model's classification accuracy for each type of distributional information in the standard labelling after the model had been exposed to 100,000 words from each corpus. Table 9 shows the categorisation accuracy and predictive strength (lambda) of each model for the standard labelling. The results are qualitatively similar to those for the earlier training stage, with the flexible aX+Xb frames resulting in the most accurate classification, followed by the aX frame, then by the Xb frame, and least accurately by the fixed aXb frames. This was the case for each of the corpora, which

showed a remarkable consistency in terms of how accurately words could be classified into their respective grammatical categories.

The principal effect of additional training was to increase the accuracy of classification for all the models. Figure 3 shows that the aXb model classifies nouns and verbs with a degree of accuracy, and shows some correct classification of conjunctions. However, the improvement in classification of words in the aX+Xb model is more rapid than that of the aXb model, with the difference in z-scores between the aXb and the aX+Xb models' classifications increasing compared to the early training stage. Figure 3 shows that, for the aX+Xb model, performance is accurate for nouns, verbs, determiners, and conjunctions, and that some occurrences of wh-words, adjectives and "not" were also correctly classified.

Table 9. Classification accuracy and lambda values for the standard labelling after 100,000 words of training.

Corpus	aXb		aX+Xb		aX		Xb		Difference in λ values as z-scores					
	Acc	λ	Acc	λ	Acc	λ	Acc	λ	aX+Xb-aXb	aX - aXb	Xb - aXb	aX+Xb - aX	aX+Xb - Xb	aX – Xb
Anne	.51	.23***	.71	.55***	.62	.41***	.55	.30***	86.68***	45.26***	14.42***	34.72***	50.77***	22.47***
Aran	.47	.20***	.68	.52***	.59	.39***	.51	.26***	99.43***	53.87***	15.11***	37.70***	62.05***	30.28***
Eve	.56	.28***	.75	.60***	.66	.44***	.54	.24***	61.02***	29.60***	-5.56***	27.77***	47.48***	27.19***
Naomi	.61	.36***	.78	.64***	.67	.47***	.59	.34***	37.07***	14.10***	-2.53**	22.42***	30.73***	13.42***
Nina	.59	.33***	.79	.67***	.68	.49***	.60	.35***	71.27***	33.59***	3.70***	44.25***	65.30***	28.82***
Peter	.57	.30***	.75	.60***	.66	.45***	.58	.32***	67.32***	29.75***	2.72**	31.25***	48.43***	22.23***
All	.53	.26***	.73	.58***	.64	.43***	.55	.43***	165.72***	88.06***	16.38***	84.42***	136.45***	64.83***

Note. Tests for lambda values are against zero association, *** indicates $p < .001$, ** indicates $p < .01$.

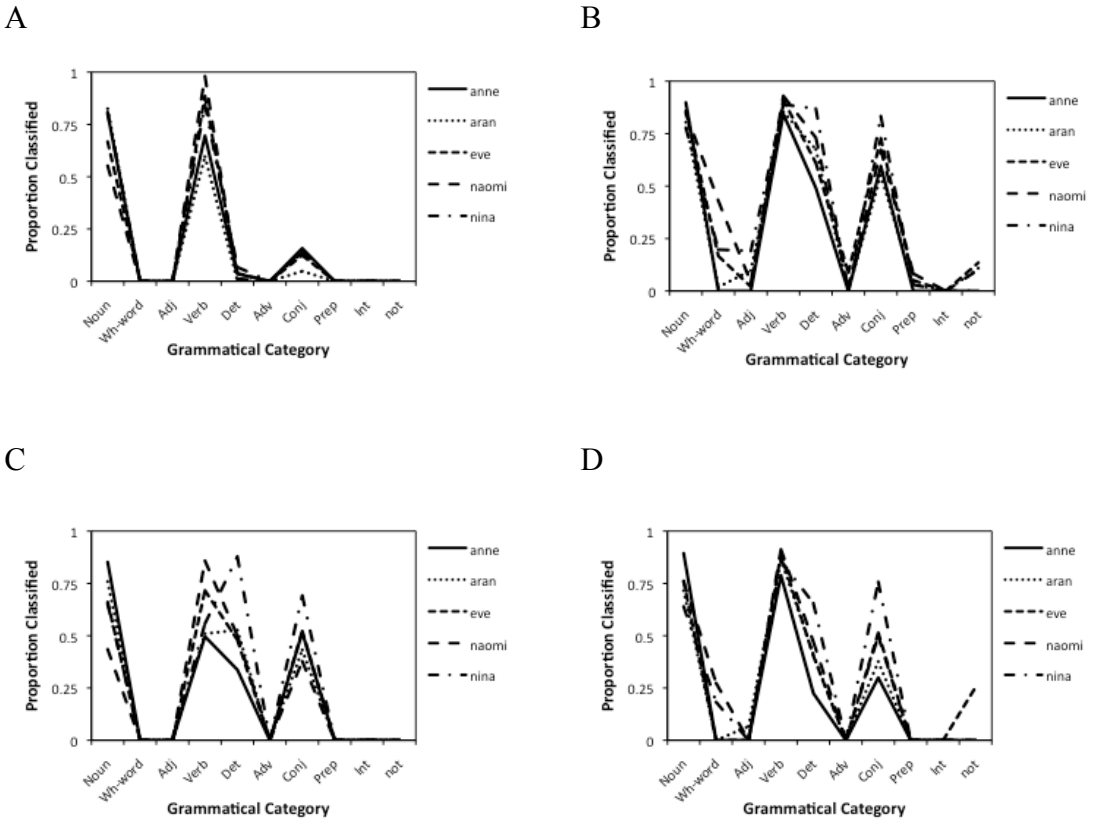

Figure 3. Proportion correctly classified for each grammatical category in each child-directed speech corpus after 100,000 word presentations for (A) aXb fixed frames, (B) aX+Xb flexible frames, (C) aX frames, and (D) Xb frames. Categories are for standard labelling.

It is possible that the aXb fixed frame model may have performed more poorly than the other models because this model has an order of magnitude more input units, meaning that each hidden unit in this model has to integrate information over more input

units. Indeed, the computational power of connectionist models is a function of the number of units in the model (including hidden units), and the number of connections. Keeping the number of hidden units stable for all the simulations has the advantage of equating the ratio of input to hidden layer connections relative to the number of input units (or number of distinct frames) for each model. However, this had the consequence that the hidden layer in the aXb model was performing a greater compression of the inputs than the aX, the Xb or the aX+Xb models, due to the large difference in the ratio of connections to each hidden unit between these models. The next Experiment tests the possibility that differences in the required compression may have influenced the degree of categorisation possible.

Experiment 3: Computational Model of Fixed Frames with Increased Resources

In order to test whether greater compression for the aXb fixed frame model resulted in reduced accuracy, we repeated the aXb fixed frame simulations but increased the numbers of hidden units to be the same ratio to input units as for the aX+Xb flexible frame model, thereby equalising the amount of compression required[2].

Method

Architecture, training and testing. The simulations were identical to those reported for the aXb fixed frame model, except for the increase in hidden units. The model was assessed as before by determining the asymmetric lambda values, and these new classification results were compared to the other simulations of Experiment 2.

[2] We are grateful to two anonymous reviewers for prompting the design and inclusion of Experiments 3 to 6.

Results and discussion

Table 11 shows the results for the aXb model with increased hidden units, with the number of hidden units used for modelling each child's corpus. The effect of increasing the number of hidden units resulted in a slight increase in classification accuracy and lambda values, however, the aXb model continued to be significantly worse at classifying the category of intervening words than the aX+Xb and the aX models. Overall, the aXb model with more hidden units did improve classification to be significantly better than the Xb model for most of the corpora (marked with $^{>}$ in the Table).

Table 11. Classification accuracy and lambda values for the standard labelling after 10,000 and 100,000 words of training for the aXb model with increased hidden units.

Corpus	# Hidden Units	aXb +hiddens 10,000		aXb +hiddens 100,000	
		Acc	λ	Acc	λ
Anne	334	.37	$.02^{***}$	.54	$.28^{***}$
Aran	351	.38	$.07^{***}$	.48	$.22^{***}$
Eve	271	$.43^{>}$	$.06^{***}$	$.60^{>}$	$.34^{***}$
Naomi	168	.40	$.03^{***}$	$.62^{>}$	$.38^{***}$
Nina	304	.41	$.05^{***}$	$.62^{>}$	$.39^{***}$
Peter	304	.42	$.06^{***}$	$.61^{>}$	$.36^{***}$
All		.40	$.05^{***}$	$.56^{>}$	$.31^{***}$

Note. Tests for lambda values are against zero association, *** indicates $p < .001$. > indicates lambda value is greater than that for the xB model.

Although the size of the hidden unit layer relative to the number of input units does not seem to explain the poor performance of the aXb fixed frame model, the large number of input units may nonetheless affect learning negatively as each unit is only activated a few times. Our next experiment therefore addressed this issue by reducing the number of input units, and thus the number of fixed frames to consider.

Experiment 4: Computational Modelling of the Most Frequent Frames

For each of the computational modelling experiments, the aXb fixed frame resulted in the lowest categorisation accuracy, principally due to the poor coverage of the entire corpus from such specific information. However, it remains a possibility that this may be an artefact of using the entire set of frames from each corpus for determining the grammatical categorisation. For the aXb fixed frames, this results in a large number of fixed frames, and the sheer size of this set may have led to poorer performance compared to the other models which had a smaller set of frames and consequent smaller set of input units. To test whether performance was due to this artefact, we repeated each of the simulations of Experiment 2 but used only the most frequent 45 frames in each case, to make the computational results comparable to the corpus analyses in Experiment 1 and to Mintz' (2003) study.

Method

Architecture, training, and testing. The models were identical to those used in Experiment 2 except that the input layers contained 45 units – one for each of the most frequent 45 frames – plus one additional dummy unit that was active for all other frames. Therefore, there were 45 frequent units and one dummy unit, which was activated for the

less frequent frames. For training and testing, if the frame was one of the most frequent 45 frames then the corresponding unit in the input was activated, and for all other frames the dummy unit was activated. For the aX+Xb model, the same principle applied – if the aX bigram was one of the most frequent 45 preceding bigrams then the corresponding unit was active, and the dummy unit otherwise, and if the Xb bigram was one of the most frequent 45 succeeding bigrams then one of the frame units was active, and the dummy unit otherwise. The model was tested in the same way as for Experiment 2 except that only the 45 most frequent frames activated the frame input units, and categorisation accuracy was assessed using the asymmetric lambda value.

Results and discussion

Table 12 shows the results after 100,000 words of training. Performance after 10,000 words of training showed a similar pattern of accuracy. Compared to the models trained on all frames in Experiment 2, the models with the top 45 frames were worse at classifying the corpus, but, critically, the lower classification accuracy of the aXb fixed frames was also observed in the current results. Consistent with the other computational models, the aXb fixed frames were significantly worse at classifying words from the corpus than the preceding and succeeding bigram models, which in turn were significantly worse than the aX+Xb flexible frames. The benefit of flexible frames over fixed frames, then, is thus not an artefact of the large number of units in the input layer for the aXb fixed frame model.

Table 12. Classification accuracy and lambda values for the 45 most frequent frame computational models for standard labelling after 100,000 words of training.

Corpus	aXb		aX+Xb		aX		Xb		Difference in λ values as z-scores					
	Acc	λ	Acc	λ	Acc	λ	Acc	λ	aX+Xb-aXb	aX - aXb	Xb - aXb	aX+Xb - aX	aX+Xb - Xb	aX - X
Anne	.40	.06***	.63	.42***	.53	.26***	.58	.34***	84.04***	61.76***	59.34***	36.40***	17.94***	-15.53
Aran	.38	.08***	.59	.39***	.52	.28***	.53	.29***	104.34***	73.94***	64.87***	37.32***	29.23***	-4.43
Eve	.40	.02	.65	.42***	.53	.23***	.54	.24***	44.51***	24.09***	25.22***	31.23***	29.50***	-2.11
Naomi	.43	.08***	.64	.41***	.55	.27***	.60	.35***	38.02***	28.96***	28.01***	16.24***	6.61***	-8.18
Nina	.45	.10***	.67	.46***	.56	.28***	.61	.36***	76.34***	33.14***	50.92***	34.27***	19.92***	-15.31
Peter	.42	.05***	.64	.42***	.55	.27***	.57	.30***	49.81***	29.81***	34.12***	25.58***	20.12***	-5.51
All	.41	.07***	.63	.42***	.54	.27***	.56	.31***	159.16***	91.89***	111.63***	75.10***	49.64***	-20.62

Note. Tests for lambda values are against zero association, * indicates $p < .05$, *** indicates $p < .001$.

The computational studies thus far have shown that decomposing the fixed frame into two bigrams results in the best classification of the child-directed corpora. However, the flexible frames also contain fixed frame information, and it may be that the advantage of the flexible frames over the bigram information is due to the fixed frames that they contain. The next Experiment tests the role of fixed frames in the flexible frame model.

Experiment 5: Separating Fixed Frame from Flexible Frame Information

The aX+Xb flexible frame model showed an advantage over the aX and Xb bigram models (as well as a large advantage over the aXb fixed frame model). However, it could be that this advantage over the bigram models was due to the aX+Xb model using fixed-frame information in addition to bigram information to assist in classification. As both

the preceding and succeeding word were presented simultaneously, the model may have picked up on the co-occurrence of particular preceding and succeeding words to boost categorisation performance. To test this hypothesis, we compared the aX+Xb model's performance to a model where during training only the preceding *or* the succeeding bigram was presented, but at test the patterns were identical to the aX+Xb model. We call this the aX+Xb-separated model. For this model, the co-occurring fixed frame is not available, though the preceding and succeeding bigram are both available, and combine in their classification during the testing stage. If the aX+Xb-separated model performs at a level similar to the aX+Xb model then this suggests that fixed frames do not contribute to the aX+Xb model's performance advantage over the bigram models. However, if the aX+Xb-separated model is close in performance to the aX or the Xb model then this suggests that the aX+Xb model's improved performance was due in part to the aXb fixed frame information.

Method

Architecture, training and testing. The aX+Xb-separated model was identical in architecture to the aX+Xb model of Experiment 2. For each training trial, an aX+Xb flexible frame was selected randomly from the same training set as used for the aX+Xb model. However, the aX+Xb-separated model was first presented only with the preceding bigram information (aX), the weights were then adjusted according to the backpropagation of error, and then the Xb bigram information from the frame was presented to the model, with weights again adjusted. After training, the aX+Xb-separated model was tested in exactly the same way as the aX+Xb model: both aX and Xb information was simultaneously available to the model. Note that this enables the overlap

between the preceding and succeeding bigram information to influence categorisation, but it does not permit the model to extract the fixed frame information as this was never simultaneously presented during training. The aX+Xb-separated model was assessed for classification accuracy using asymmetric lambda and compared to the other models' classifications in Experiment 2.

Results

Tables 13 and 14 show the aX+Xb-separated model's performance, for 10,000 words and 100,000 words, respectively. At both early and later training stages, the aX+Xb-separated model performed significantly better than the aX, the Xb and the aXb models, as with the aX+Xb-simultaneous model. However, for most of the corpora there was a slight advantage for the aX+Xb-simultaneous model over the aX+Xb-separated model. This shows that there is a small amount of variance accounted for by the model's use of the fixed frame information in the aX+Xb-simultaneous model. This is, however, an extremely small amount of information when compared to the disadvantage of learning only from fixed frames, as in the aXb model.

Table 13. Classification accuracy and lambda values for the aX+Xb-separated model for standard labelling after 10,000 words of training.

Corpus	aX+Xb-sep		Difference in λ values as z-scores			
	Acc	λ	aX+Xb – aX+Xb-sep	aX+Xb-sep - aXb	aX+Xb-sep - aX	aX+Xb-sep - Xb
Anne	.49	.20***	22.17***	61.54***	9.46***	43.64***
Aran	.55	.33***	3.81***	88.17***	9.04***	48.07***
Eve	.46	.22***	38.58***	23.06***	-.06	18.91***
Naomi	.58	.32***	.82	31.95***	8.05***	19.93***
Nina	.60	.41***	10.10***	82.79***	12.99***	43.50***
Peter	.56	.31***	4.84***	41.69***	6.63***	29.01***
All	.54	.27***	21.19***	103.44***	19.03***	80.38***

Note. Tests for lambda values are against zero association, *** indicates $p < .001$.

Table 14. Classification accuracy and lambda values for the aX+Xb-separated model for standard labelling after 100,000 words of training.

Corpus	aX+Xb-sep		Difference in λ values as z-scores			
	Acc	λ	aX+Xb – aX+Xb-sep	aX+Xb-sep - aXb	aX+Xb-sep - aX	aX+Xb-sep - Xb
Anne	.70	.53***	4.00***	81.37***	29.81***	46.77***
Aran	.64	.47***	12.37***	83.49***	23.02***	49.68***
Eve	.74	.57***	3.10***	56.51***	23.53***	44.39***
Naomi	.76	.62***	2.16	34.22***	19.61***	28.57***
Nina	.78	.64***	6.13***	64.94***	36.82***	59.17***
Peter	.74	.58***	3.85***	62.25***	26.66***	44.58***
All	.71	.55***	15.53***	148.88***	66.12***	120.92***

Note. Tests for lambda values are against zero association, *** indicates $p < .001$.

The advantage of the aX+Xb model over the other sources of distributional information arises from several sources. First, the combination of aX and Xb bigrams enables two distributional cues to be simultaneously applied to the classification of words. The corpus analysis of the aX and Xb bigrams in Experiment 1 revealed that categories based just on a bigram were too broad and contained many false positive classifications. However, a false positive due to the aX bigram can be corrected by the information from the Xb bigram.

Take as an example the aXb frame "you___to". In the aXb frame analysis of the Naomi corpus, 119 verb tokens are classified, 1 adjective, 2 adverbs, and 1 preposition. For the preceding bigram analysis on the same corpus, the word "you___" precedes 883 verbs, 20 adverbs, 1 adjective, 7 conjunctions, 12 nouns, 20 prepositions, 1 negative particle, and 3 pronouns. For the succeeding bigram analysis, the word "___to" succeeds 325 verbs, 13 adverbs, 26 adjectives, 12 prepositions, 38 nouns, 73 pronouns, 9 negative particles, and 6 determiners. Note that in each of these cases the bigrams have occurred in the corpus much more frequently than the aXb fixed frame. The combination of information in the flexible frame, then, contributes to discovering that the overlap in the most frequent category for the aX and the Xb frame is the verb token. For each corpus, the aX bigrams misclassified a mean of 44.3% of words. Of these, 31.5% were reclassified correctly by the aX+Xb frame, with just 4.2% of the words correctly classified by the aX model reclassified incorrectly by the aX+Xb frame. Of the 31.5% correctly reclassified by the aX+Xb model, the Xb model also classified them correctly in 79.7% of the cases.

Figure 4 shows that this combined advantage is general across the different grammatical categories. The Figure shows the mean percentage of each corpus classified correctly by each of the aX, Xb, and aX+Xb models, and their overlap. The relevant portions of each bar indicates the correct classifications for the aX+Xb model, with the relative proportions that are also correctly classified by the aX model, the xB model, or both. In very few instances does one of the bigram models correctly classify a word and the aX+Xb flexible frame misclassify. For the noun category, for instance, almost 70% of each corpus was correctly classified by the aX, the Xb and the aX+Xb model, as indicated by the black portion of the bar. The aX+Xb and aX model correctly classified approximately 12% of the nouns that the Xb model misclassified. The aX+Xb and the Xb model correctly classified 8% that the aX model misclassified, and the aX+Xb model correctly classified a further 2% that both the aX and the Xb models misclassified. There was just a small proportion of nouns that the aX or Xb model correctly classified but the aX+Xb model misclassified, reinforcing a view of the aX+Xb model as a committee of bigram experts.

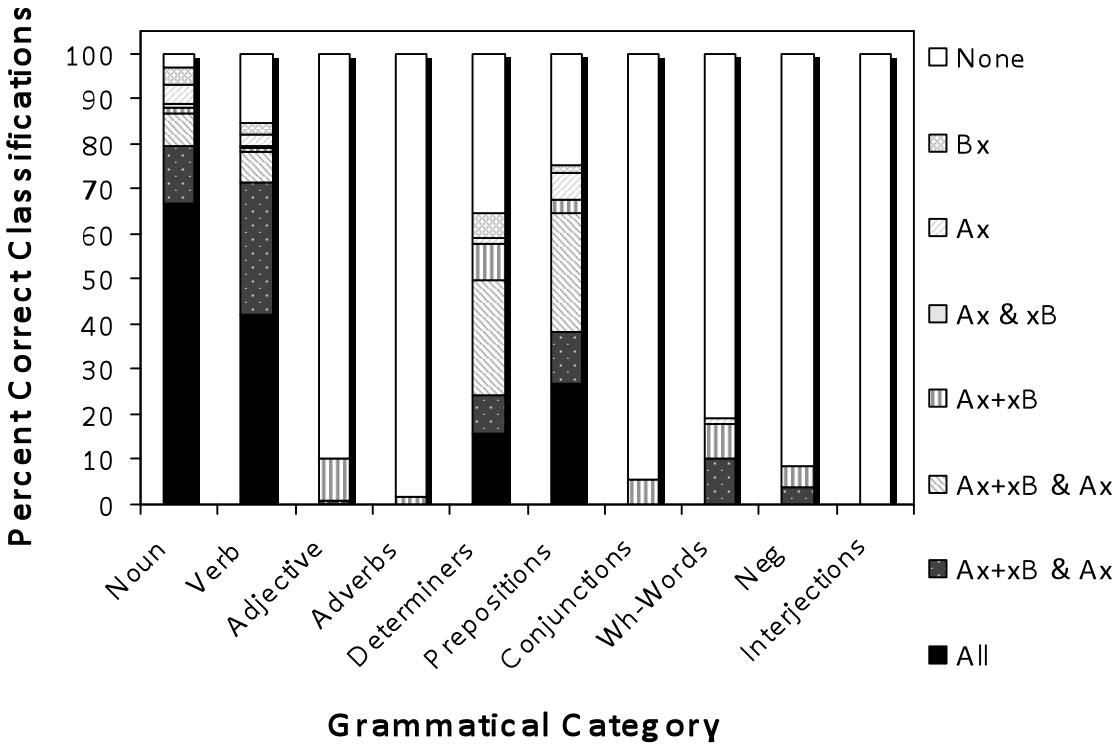

Figure 4. Classifications for each grammatical category across the 6 corpora indicating proportions of each category where the aX, Xb, and aX+XB models make same and different categorisations.

A second advantage of the aX+Xb flexible frame distributional information is that rarer fixed frames may still contain at least one frequent bigram, thus the frequent bigram can be used to guide categorisation, based on previous learning situations. Take the example "you___with", which occurs 27 times in the Naomi corpus, and is not one of the 45 most frequent frames in the corpus. All 27 tokens that occur in the fixed frame are verbs. The majority of words that occur after "you___" are verbs (431 from 495 tokens), and the words that occur before "___with" are also a majority of verbs, with double the occurrence of the fixed frame: 59 verbs, 40 nouns, 7 pronouns, 1 adjective, 4 adverbs,

and 6 prepositions. The overlap of frequent information can focus the learner on the verb, but can simultaneously mean that the 40 occurrences of nouns preceding "with" do not influence categorisation of the word in the "you___with" frame.

In all the models we have presented thus far, the models have been provided with the target grammatical category for the word in each distributional context. However, this information is not available to the child learning her language from the outset, and this assumption that the categories are provided to the learner and the only task is then to match lexical items to these categories is a far less difficult problem than having to discover the categories in the first place. In the next Experiment, we test the extent to which an unsupervised analysis can reflect the grammatical categories of all the words in the child's exposure, based on the learning of just a very few, high frequent content words from the language. Though this model still requires the categories to be specified on the basis of the first few words, it does indicate the extent to which generalisations based on the different types of distributional information we have considered provide potential for reflecting category membership and also defining the learner's category structure itself.

Experiment 6: Unsupervised generalisation from a subset of early acquired words

In this final model, we test the extent to which unsupervised learning based on a small set of early-acquired words can successfully generalise to the whole corpus. In these analyses we restricted the comparison of different sources of distributional information to generalisations based on either flexible or fixed frames. For each child's corpus, we have assumed that the category of a small set of high frequency words can be ascertained on

the basis of information external to the distributional contexts in which words occur (e.g., semantic, distributional, phonological, etc.). We next determine the extent to which the distributional contexts in which these high frequency (hence early-acquired, Kauschke & Hofmeister, 2002) words occur reflect the grammatical categories of the whole corpus. Thus, we tested whether the contexts in which a few learned words occur can lead to accurate reflection of the grammatical categories found in the whole corpus. We restricted our analyses to the noun/verb distinction, which categories represent the largest proportion of word types in the language (Monaghan et al., 2005), and a distinction that is critical for early comprehension and production ability. We predicted that the contexts of a few early acquired words would be able to accurately reflect the grammatical categories of a large sample of the language. Furthermore, we predicted that the flexible frames contexts would provide a better generalisation than fixed frames because the co-occurrence of bigrams with a small set of high frequency words would provide better coverage, and sufficient accuracy, for the whole corpus than the fixed frames in which these words occurred.

Method

Corpus preparation. We used the same child-directed speech corpora as in the previous Experiments. For each child's corpus, we selected the nouns and verbs from a set of the most frequent words. For each word in this set, we then determined the distributional contexts in which the frequent word occurred, either in terms of the bigram information in the flexible frames, or the trigrams in the fixed frames. These frames were then labelled as predicting the grammatical category of the high frequency word that occurred

within this context. So, for the high frequency word “think”, in the Peter corpus it occurred in the context “I think it’s” 53 times. So, in the aX+Xb flexible frames model, each time the frame “I ___” occurred in the corpus, the model generalised the context of the following word to predict a verb category. Similarly, each time the frame “____ it’s” occurred, again the model generalised the context of the preceding word to a verb category. For the aXb fixed frames model, each time the “I ___ it’s” trigram context occurred in the corpus the intervening word was classified as a verb. If more than one word occurred in the same distributional context, then the grammatical category label for that context was assigned as the most frequent occurrence of the category. As an example from the Peter corpus, the frame “can think of” occurred once, but because the flexible frame preceding context “can ____” co-occurred more frequently with nouns and pronouns than with verbs (245 times preceding nouns and pronouns, 220 times preceding a verb), this preceding context was labelled as predicting a noun.

We selected the 10 most frequent nouns and verbs from the corpus, and assigned the target grammatical category for contexts in which these words occurred according to the category of the frequent word (noun or verb). For each context in the corpus in which a high frequent word did not occur, the model was not provided with a target grammatical category, and so no learning took place for these contexts. Pilot testing with a greater number of nouns and verbs did not show improved generalisation over the smaller word set.

Architecture, training and testing. The architecture, training and testing was the same as in the previous Experiments, and as with the previous simulations we compared the model trained on fixed frames to a model trained on flexible frames. We stopped training

after 100,000 words from the corpus had been presented to the model (including words with no target grammatical category provided from the high frequent words). The models were then tested on their ability to predict the grammatical category of the word occurring in each frame with the grammatical category taken from the full corpus. The training and testing target grammatical categories were therefore distinct – during training they were generalisations from the high frequency words, during testing they were the actual grammatical category for each word as indicated by the MOR line in CHILDES, similar to the previous experiments.

Results and Discussion

The categorisation results based on learning just the most frequent 10 nouns and verbs from each corpus are shown in Table 15. The values indicate the classifications based on the whole corpus excepting the items used to form the categories (so omitting the 10 most frequent nouns and verbs from the final analysis). As in previous simulations, we report lambda values as well as overall accuracy of classifications, though in the current Experiment these values are computed just for nouns and verbs.

For the aX+Xb analyses, the categorisation of the remaining corpus based on the distributional information generated from just a small set of frequent words was significantly better than chance for all corpora except the Peter corpus. For the remaining five corpora, and for the combined analysis of all six corpora, lambda values were significantly above chance, all $p < .001$. This suggests that, if the child has learned the grammatical category of a small set of high frequency words then, based on the flexible frame distributional contexts in which these words occur, generalisation to discovering

the grammatical category of a large proportion of the other nouns and verbs in the corpus can generally be achieved. So, if the child can learn the grammatical category of these words from sources other than distributional information then bootstrapping from this set can provide accurate categorisation. The failure of the classification based on the Peter corpus was due to most words being classified as verbs, and so the prediction of the categories based on the distributional information was at chance level. In order for the lambda statistic to be significant for a two-category analysis, accuracy of classification of both categories has to be greater than 50%, as otherwise it is not possible to determine from the distributional information alone whether a word is classified as a noun or a verb. It is thus important to consider both the accuracy level and the lambda statistic when interpreting the results.

A striking feature of the results is that accurate generalisation can occur even when the learner knows less than a dozen words. For the aX+Xb flexible frames model, based on just the 10 most frequent words, the child had access to distributional information that could correctly categorise a further 70% of the nouns and verbs in the child's language environment. For the Aran corpus, for instance, this meant that from the contexts of just 10 nouns and verbs, 69% of the remaining 2528 nouns and verbs could be correctly classified. The potential of this information for generalisation from the aX+Xb flexible frames analysis emerges early in the child's language acquisition, and indeed, generalising from 25 or 50 words did not substantially improve categorisation accuracy for the flexible frames.

For the aXb fixed frames analyses, the model was not able to generalise from the contexts in which a small subset of words occurred. In all cases, the model predicted that

most of the words based on their fixed frame distributional contexts were nouns, and so the association from the distributional information was at chance. The classification accuracy in each case was close to .5, which is the random baseline for categorising into two groups. This qualitative difference in the generalisation effects between the aX+Xb and the aXB contexts was due to the greater specificity of the fixed frame context. The high frequent words occurred in particular contexts, and these were often distinct from the contexts in which other nouns and verbs tended to occur.

Table 15. Classification accuracy and lambda values for the unsupervised modeling results, based on the contexts of the 10 most frequent nouns and verbs in each corpus.

Corpus	10 Most Frequent Nouns/Verbs			
	aXb		aX+Xb	
	Acc	λ	Acc	λ
Anne	.50	.00	.75	.36***
Aran	.52	.00	.69	.24***
Eve	.52	.00	.73	.31***
Naomi	.53	.00	.69	.22***
Nina	.57	.00	.64	.11***
Peter	.49	.00	.68	.00
All	.52	.00	.70	.22***

Note. Tests for lambda values are against zero association, *** indicates $p < .001$.

However, assuming the child can generalise from the contexts of a few words in turn presumes that the child has already determined the category of these words from other, non-distributional information. Whilst there is general agreement that children can learn a few nouns from semantic information outside the distributional contexts in which these words occur, there is debate over whether any verbs can be learned without distributional information (Gleitman, 1990; Gleitman, Cassidy, Papafragou, Nappa, & Trueswell, 2005). However, assuming that children can acquire some knowledge about only nouns before distributional information can become useful allows sufficient information to promote learning of the noun category and an "other" category. Generalisation of contextual information from only the top 10 nouns in each child corpus resulted in highly accurate categorisation of nouns as distinct from categories in which nouns did not occur. Across the six corpora, 48.7% of nouns could be correctly categorised in this way. In terms of words in the "other" or "non-noun" category, a mean of 92.7% of the words in this category were verbs, providing the possibility that verbs could be learned on the basis of categorisation formed from the contexts of a few nouns. Though this was not as accurate as the aX+Xb information based on both nouns and verbs, it does show the potential for bootstrapping of grammatical categories based on generalising categories from only a few words from a single grammatical category.

General Discussion

The purpose of this paper was to test a new approach to distributional learning of grammatical categories based on the hypothesis that children construct on-line trigram-based flexible frames (aX+Xb) from coarser bigram-based information (aX and Xb). We

therefore first quantified the relative usefulness of different sources of distributional information as cues to grammatical categories in language. Our corpus analyses replicated the original results from Mintz (2003), indicating that frequently occurring aXb fixed frames of non-adjacent words provided a highly accurate context for the grouping together of words of the same grammatical category. However, as in the sparsity problem of corpus linguistics (Manning & Schütze, 1999), there is a trade-off between accuracy and coverage of the language. The aXb fixed frames tended to be highly specific and so words of the same category were often not grouped together. In contrast, the child may utilise information about co-occurrences between words at a coarser grain size than trigrams. We determined the extent to which bigrams could provide information about grammatical categories, and, though there was a reduction in accuracy compared to aXb trigrams, there was an enormous increase in the extent to which the whole corpus could be categorised.

The subsequent computational modelling demonstrated that, for a general purpose learning mechanism, bigrams resulted in more effective learning of the grammatical categories of words in child-directed speech than did the trigrams. Furthermore, the computational modelling demonstrated that the convergence of two general bigrams – aX and Xb – within a flexible frame was much more effective for learning than single bigrams (either aX or Xb), which were in turn more effective than the fixed frame aXb trigrams.

The effectiveness of these flexible frames was because they harness the advantage of coverage from bigrams that occur frequently throughout the corpus, together with the increased accuracy that follows from greater specificity of the context. Flexible frames

inherit the strengths of both the bigram and the fixed frame statistics, but avoid the weaknesses of both methods. Flexible frames are also consistent with the developmental trajectory of children's sensitivity to different sources of distributional information for learning language: Adjacent, bigram information can be used before children become sensitive to non-adjacent dependencies (Gómez & Maye, 2005); succeeding bigram information can be used as well as preceding information as cues for categorisation (e.g., Frigo & McDonald, 1998; St. Clair et al., 2009); and adjacent bigram information remains easier to use for determining language structure even after considerable language exposure (Onnis et al., 2005).

The results of Experiment 5 also indicate that the accuracy of classification from the flexible frames is partially due to being able to harness the value of fixed frames – when the fixed frame information was removed from the combination of bigrams, performance reduced slightly. This indicates that flexible frames are able to exploit both low-level bigram as well as higher-order trigram information to assist in categorisation. We suggest that the developmental trajectory of the use of the trigram is likely to follow the initial use of bigrams, given that bigram information is available to infants at an earlier age than non-adjacent dependencies in learning (Gómez & Maye, 2005; Saffran, Newport et al., 1996).

Though we have focused in this paper on the distinction between bigram and trigram distributional information as the basis of grammatical categorisation, the point we wish to make is about discovering the cues in the child's language environment that provide useful and useable information for grammatical categorisation. We have highlighted the relative merits and disadvantages of two potential sources of information:

bigrams which provide large coverage but low accuracy, and trigrams that are highly accurate but also highly specific. Yet, the sources of distributional information that the child uses are likely to transcend such distinctions between particular sources of information. We have focused on the bigram/trigram distinction to make the more general point about the computational usefulness of alternative sources of information. We contend that the child will exploit any sources of information that prove useful to categorisation, and some combination of highly frequent and useful trigrams, bigrams, and even higher-order co-occurrences (e.g., Bannard & Matthews, 2008), according to their usefulness (and constrained by their availability) is likely to constitute the child's repertoire of distributional cues. Cartwright and Brent's (1997) model of categorisation, for instance, indicates that frames with different amounts of specificity may be useful, but that with more extensive training, categories increase in generality, and become defined by more general distributional information. The bigram/trigram distinction provides a test case, then, to highlight the distinction between alternate cues in terms of their usefulness for language acquisition.

We have focused in our analyses on how distinct grammatical categories may be distinguished based on distributional information within child-directed speech utterances. For the majority of the models utilising distributional information, the training was "supervised", in that the target grammatical category for each word was provided to the learner. This is evidently not the case for the child learning her first language, and these supervised models ignore the question of how such categories are discovered in the first place. However, our final study focused on what category information can be bootstrapped on the basis of learning the category of a very few, highly frequent words. If

the child can learn the category of these words then we have shown that, using the same principles as the supervised analyses that reveal the potential distributional information available in the child's language environment, the categories of the majority of words in the language can follow based on the same distributional information linked to these high frequency words. As in the supervised analyses, the generalisation from a small set of words was better accomplished based on the flexible frames distributional information, and in fact, the fixed frames information resulted in classifications close to chance level. It is a moot point about how many words, and how many categories can be learned by the child in this way, but even with the very conservative assumption that only a few nouns can be learned, our Experiment 6 has shown accurate generalisation from flexible frames to almost half of all nouns, distinct from almost 90% of verbs. Throughout our analyses, we have shown that flexible frames provide the best quality of information for grammatical categorisation, so if verbs are to be discovered from distributional information alone (see, e.g., Gleitman et al., 2005) then we contend that flexible frames are most fit for this purpose.

Even without assuming the child can determine the grammatical category of a small set of words to begin the process of categorisation of the whole language, it is possible that our measures of the statistics useful for reflecting the grammatical categories within the child's language exposure may equally apply to determining what those categories are in the first place. Mintz (2003) indicated how clustering of words in frequent frames may give rise to hypotheses within the learner about the category structure of those words (see also Harris, 1954; Maratsos & Chalkley, 1980) – words that occur in similar categories can be clustered together (Redington et al., 1998) – and a

similar process for discovering the categories, and not only the category membership of words can apply to the analyses we have presented here. We have shown that combinations of preceding and succeeding bigram information best reflect the objective grammatical category structure of child-directed speech, and that clustering based one these flexible frames will lead to the most accurate hypotheses about the categories that are present in the language, as well as the membership of those categories. We predict that a repeat of the unsupervised methods of Redington et al. (1998) that enable flexible combination of clusters based on preceding and clusters based on succeeding information will enable the generation of categories that most closely resemble those of the objective classification of the language. For instance, a category based on words succeeding the word "the" would be clustered together, but then this category would be further subdivided in terms of whether these words precede the word "goes", for instance. In this way, nouns (which can both succeed articles and precede verbs) will be distinguished from adjectives (which can only succeed articles), resulting in increasing accuracy of the clustering. The benefit of flexible frames is that words that succeed "the" and precede "goes" can be clustered together, even if they never occur in the frame "the ____ goes", but do occur in more general frames "the ___" and "___ goes".

Yet, the demonstrations that distributional regularities can provide highly accurate grammatical categorisations of the language do not necessarily indicate that such word co-occurrence information alone drives the child's knowledge and construction of the grammatical categories themselves. Our previous work has indicated the importance of phonological and prosodic cues to grammatical categories, for instance (Monaghan et al., 2005; Monaghan et al., 2007), and other language-external information, such as gesture,

attention, or social cues (Gleitman, 1990; Monaghan & Christiansen, 2008; Siskind, 1996; Tomasello, 2003), which we envisage as also being of critical importance for guiding the discovery of the categories.

For accounts of language acquisition that posit innate grammatical categories (e.g., Chomsky, 1981) or innate semantic features to which distributional categories attach (e.g., Pinker, 1999), the language environment also needs to provide a structured reflection of the categories in the language to facilitate learning to map particular words onto the innate categories. Accounts of language acquisition that instead claim that the language environment is sufficient alone to generate the grammatical categories also lay claim to the same structural properties of the language environment as the basis of learning. The studies we have presented here do not decide between these theoretical accounts of grammatical category acquisition, but they do assist in generating hypotheses about the sources of information that assist in creating either the mapping onto the category or the category itself. Establishing which sources of distributional information are used by the child provides an important precursor to future tests of the nativist/empiricist debate. For instance, determining which cues are most useful to a general purpose learning system forming categories in the language enables hypotheses about the sorts of categories that may be formed on the basis of surface features of the language – in terms of distributional and phonological cues, for instance – to be made. Such a view would be consistent with accounts of grammatical processing that propose many exceptions to a model of language with very general categories (e.g., Culicover, 1999; Goldberg, 2006). It would also be consistent with data from child production where inconsistencies with adult grammatical structure can be shown to reflect surface features

of distributional information in child-directed speech (Freudenthal, Pine, Aguado-Orea, & Gobet, 2007).

To conclude, this research provides support for a new view of the distributional information that the child uses to determine the grammatical categories within the language. We have shown that trigram information may be helpful for this process, as discovered by Mintz' (2003) seminal study. Yet, we have shown that to be effective for classifying the language, the sparsity of fixed trigram frames is a poor candidate for learning, and instead the trigram frames used by children must be flexible, in terms of permitting preceding and succeeding bigram information to be combined "on the fly" for categorisation. Furthermore, learning from flexible frames is consistent with the statistical information available to children early in the language acquisition process. We concur with a wide-range of studies of multiple cues in language acquisition that suggest children are likely to employ useful information in the environment for determining language structure (e.g., Colunga & Smith, 2005), and we have demonstrated here that flexible frames provide a particularly useful source of distributional information for the learning of grammatical categories from child-directed speech.

Acknowledgements

We would like to thank Stefan Frank and two anonymous reviewers for their helpful comments on a previous version of this paper. An earlier version of some of the computational modelling work was presented at the Cognitive Science Society conference, August 2006. The corpus analysis constituted a part of the first author's PhD thesis.

References

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

Bloom, L., Hood, L., & Lightbrown, P. (1974). Imitation in language development: If, when and why. *Cognitive Psychology, 6*, 380-420.

Bloom, L., Lightbrown, P., & Hood, L. (1975). Structure and variation in child language. *Monographs of the Society for Research in Child Development, 40*, Serial No. 160.

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

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

Cassidy, K. W., & Kelly, M. H. (2001). Children's use of phonology to infer grammatical class in vocabulary learning. *Psychonomic Bulletin & Review, 8*, 519-523.

Chomsky, N. (1981). *Lectures on Government and Binding: The Pisa lectures*. Dordrecht: Foris Publications.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review, 112*, 347-382.

Culicover, P. W. (1999). *Syntactic Nuts*. Oxford: Oxford University Press.

Curtin, S., Mintz, T. H., & Christiansen, M. H. (2005). Stress changes the representational landscape: Evidence from word segmentation. *Cognition, 96*, 233-262.

Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test. In J. Weissenborn & B. Höhle (Eds.), *Approaches to Bootstrapping* (pp. 189-229). Amsterdam: John Benjamins.

Endress, A., D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition, 105*, 577-614.

Erkelens, M. A. (2009). *Learning to categorize verbs and nouns.* Unpublished PhD Thesis, Universiteit van Amsterdam, Amsterdam.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories using statistical methods. In D. Daelemans & W. Powers (Eds.), *Proceedings of the 1st SHOE Workshop on Statistical Methods in Natural Language* (pp. 229-235). Tilburg, Netherlands: ITK.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSIAC. *Cognitive Science, 31*, 311-341.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39*, 218-245.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1*, 1-55.

Gleitman, L., Cassidy, K., Papafragou, A., Nappa, R., & Trueswell, J. T. (2005). Hard Words. *Journal of Language Learning and Development, 1*, 23-64.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Gómez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*, 183-206.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732-764.

Grainger, J., & Whitney, C. (2004). Does the human mind read words as a whole? *Trends in Cognitive Sciences, 8*, 58-59.

Harris, Z. S. (1954). Distributional structure. *Word, 10*, 146-162.

Hockema, S. A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development, 2*, 119-146.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548-567.

Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language, 29*, 735-757.

Kelly, M. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*, 349-363.

Kelly, M., & Bock, J. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception & Performance, 14*, 389-403.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk, Vol 2: The database (3rd ed.).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

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, pp. 125-214). New York: Gardner Press.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134*, 477-500.

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.

Mintz, T. H., Newport, E. L., & Bever, T. G. (1995). Distributional regularities of grammatical categories in speech to infants. In J. Beckman (Ed.), *Proceedings of the 25th Annual Meeting of the North Eastern Linguistics Society*. Amhert, Mass.: GLSA.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96*, 143-182.

Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (pp. 139-164). Amsterdam: John Benjamins.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The Phonological Distributional Coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*(55), 259-305.

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language, 53*, 225-237.

Pinker, S. (1999). *Words and Rules: The Ingredients of Language*. New York Basic Books.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 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* (4th ed., pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Saffran, J. R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition, 81*, 149-169.

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language, 35*, 606-621.

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

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39-61.

Smith, K. H. (1966). Grammatical intrusions in recall of structured letter pairs - Mediated transfer of position learning. *Journal of Experimental Psychology, 72*, 580-588.

St. Clair, M. C. (2007). *Language structure and language acquisition: Grammatical categorization using phonological and distributional information.* Unpublished Thesis, University of York, York, UK.

St. Clair, M. C., & Monaghan, P. (2005). Categorizing grammar: Differential effects of succeeding and preceding contextual cues. In *Proceedings from the 27th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

St. Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science, 33*, 1317-1329.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2010). Frequent frames in German child-directed speech: A limited cue to grammatical categories. . *Poster presented at the 24th Annual Meeting of the CUNY Conference on Human Sentence Processing, New York University, NY*.

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

Theakson, 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.

Theissen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to work boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706-716.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard University Press.

Valian, V., & Coulson, S. (1988). Anchor Points in Language Learning: The Role of Marker Frequency. *Journal of Memory and Language, 27*, 71-86.

Vallabha, G., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 13273-13278.

Appendix I: 45 most frequent frames in the Aran corpus (ordered in terms of token frequency for each individual cue).

aXb	***aX***	*Xb*
you_to	the	You
are_going	you	The
what_you	a	It
you_it	to	To
to_it	it	A
you_the	is	On
there_are	that	That
to_the	and	We
a_isn't	are	Going
the_one	we	In
it_the	oh	Is
to_with	what	Are
what_it	going	There
the_isn't	on	isn't
is_a	come	Your
i_think	in	He
do_want	well	Got
you_a	isn't	One
it_you	your	This
put_on	do	And
we_to	i	With
you_that	that's	Do
put_in	have	Have
you're_to	got	Of
the_in	put	Put
to_a	this	don't
a_of	with	Not
the_of	don't	Then
is_going	it's	For
here_are	there	What
have_got	can	Think
the_and	of	Can
the_on	not	They
the_to	he	Go
what_we	all	Want
do_think	what's	didn't
a_on	one	All
you_me	for	At
to_to	did	Up
want_to	go	I
you've_to	there's	Aran
have_look	think	She
it_to	you've	Some

and_the	didn't	Out
is_sure	some	Like

Appendix II: 10 highest frequency nouns and verbs for the unsupervised models in Experiment 6. Words are reported in order of decreasing frequency in each corpus.

Corpus	*Highest Frequency Nouns and Verbs*
Anne	**you, going, we, it, think, have do, want, got, that**
Aran	**you, it, going, we, that, got, put, your, have, think**
Eve	**you, it, have, don't, do, your, put, I, are, want**
Naomi	**you, it, want, are, don't, your, put, can, have, is**
Nina	**you, is, want, are, put, going, we, it, did, do**
Peter	**you, it, put, want, going, don't, that, think, me, I**