Frequent frames as a cue for grammatical categories in child directed speech

Toben H. Mintz*

Departments of Psychology and Linguistics, and Program in Neuroscience, University of Southern California, Los Angeles, CA, USA

Abstract

This paper introduces the notion of frequent frames, distributional patterns based on co-occurrence patterns of words in sentences, then investigates the usefulness of this information in grammatical categorization. A frame is defined as two jointly occurring words with one word intervening. Qualitative and quantitative results from distributional analyses of six different corpora of child directed speech are presented in two experiments. In the analyses, words that were surrounded by the same frequent frame were categorized together. The results show that frequent frames yield very accurate categories. Furthermore, evidence from behavioral studies suggests that infants and adults are sensitive to frame-like units, and that adults use them to categorize words. This evidence, along with the success of frames in categorizing words, provides support for frames as a basis for the acquisition of grammatical categories.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Language acquisition; Grammatical categories; Distributional analysis; Corpus analysis

1. Introduction

Grammatical categories (e.g. noun verb, etc.) are fundamental building blocks of grammar, yet it is not fully known how child language learners initially categorize words. There has been recent interest in the idea that distributional information carried by the co-occurrence patterns of words in sentences could provide a great deal of information relevant to grammatical categories. For example, words in position X in sentences

* Department of Psychology, SGM 501, University of Southern California, Los Angeles, CA 90089-1061, USA. Tel.: +1-213-740-2253; fax: +1-213-746-9082.
 E-mail address: tmintz@usc.edu (T.H. Mintz).

containing the English fragment in (1) are likely to belong to the same grammatical category, verb.

(1)   … to X to …

Building on early ideas from structural linguistics and related proposals for acquisition (Harris, 1951; Maratsos & Chalkley, 1980), several recent studies have investigated whether distributional patterns in English-learning children's input could be a reliable source of information for learning the category structure of the language (Cartwright & Brent, 1997; Mintz, Newport, & Bever, 1995, 2002; Redington, Chater, & Finch, 1998). These studies demonstrated that distributional patterns were informative and potentially viable bases for an initial categorization of words, and provided evidence against speculative yet influential claims that distributional information would be extremely unreliable. As an example of the potential problems faced by distributional approaches, Pinker (1987) argued that, given sentences in (2a,b), a distributional learner would incorrectly categorize *fish* and *rabbits* together and, hearing (2c), would incorrectly assume that (2d) is also permissible.

(2)   a. John ate fish.
      b. John ate rabbits.
      c. John can fish.
      d. *John can rabbits.

The crux of the problem exemplified in (2) is that a given word form (in this case, *fish*) can belong to multiple categories and thus occur in different syntactic contexts (e.g. as a noun in 2a or a verb in 2c), potentially providing misleading category information. Pinker argued that the resulting erroneous generalizations would be common, and would render a distributional approach to categorization untenable.

   Another potential difficulty is that important distributional regularities are not always local, as in (1), but can occur over a variable distance, as in (3) (Chomsky, 1965; Pinker, 1987).

(3)   … *to* hurriedly and effortfully *X to* …

Here, the informative verb environment for X in (1) (… *to X to* …) spans many words. The fundamental issue is that lexical adjacency patterns are variable: in any particular utterance, the words in a specific position relative to a target is somewhat accidental, and a learner that categorized only from fixed positions could be led to make erroneous generalizations. Thus, another question is how the learner is to know *which* environments are important and which should be ignored. Distributional analyses that consider all the possible relations among words in a corpus of sentences would be computationally unmanageable at best, and impossible at worst.

   Hence, it was an important advance when results from recent empirical investigations into the viability of distributional approaches to categorization revealed that, in children's actual input, these potential problems do not significantly undermine the informativeness

of distributional patterns. For example, Mintz et al. (1995, 2002) and Redington et al. (1998) showed that local contexts restricted to immediately adjacent words can be informative as to a word's category membership. Such findings suggested that, although problematic environments may exist, there is nonetheless enough "signal" in the distributional patterns compared to the noise created by the problematic environments that categorization from distributional patterns is not intractable. Moreover, by showing that local contexts are informative, these findings suggested a solution to the problem of there being too many possible environments to keep track of: focusing on local contexts might be sufficient.

Although prior studies have shown distributional information to be useful for categorizing words, many open questions remain and much research is still needed to determine what type of distributional information is especially informative, and what kinds of distributional cues infants and young children are sensitive to and use in categorizing words. Questions also remain as to how distributionally defined categories could be integrated into a grammatical system. The studies presented here contribute to the understanding of these issues by proposing and testing a novel analysis procedure based on a previously unstudied kind of distributional pattern: frequent frames. Frequent frames are defined as ordered pairs of words that frequently co-occur with exactly one word position intervening (occupied by any word). Any words that occur as the intervening word inside a given frequent frame are categorized together. The motivation for investigating the informativeness of frequent frame contexts as a basis for category learning was twofold: (1) to study a distributional unit for which there is evidence that infants and adults attend to: the logical problem of which distributional context(s) to attend to can be circumvented if one can demonstrate that the distributional contexts that learners *do* attend to can support categorization; and (2) to study a procedure that categorizes only words that occur in contexts that are likely to be informative: frequent frames, as defined here, might provide reliable category information because requiring the joint occurrence of contextual elements eliminates many accidental contexts from the analysis. A brief overview of these two points follows in a discussion of relevant psycholinguistic evidence, and in comparing the distributional patterns investigated here with those investigated in other studies.

1.1. Psycholinguistic evidence relevant to frames

A number of recent studies suggest that frame-like units might be relevant representational units for children, and could plausibly be used to categorize words. In sequence-learning studies, Gómez (2002) has shown that adults and 18-month-old infants can track (slightly) long distance dependency relationships similar to those that define frames here (see also Santelmann & Jusczyk, 1998). Related findings with adults have been reported (Peña, Bonatti, Nespor, & Mehler, 2002, Experiments 3 and 5) using structures similar to the frames defined here. Finally, in a categorization study, Mintz (2002) has shown that adults spontaneously form categories of novel words based on the frames in which they appear.

In word-learning research, Childers and Tomasello (2001) have shown that children more easily acquire novel verb meanings when the verbs occur in lexical frames that occur frequently in the children's input. In addition, more abstract linguistic frames, for example

a verb's argument structure, have been shown to be informative for verb learning (Fisher, Gleitman, & Gleitman, 1991; Gillette, Gleitman, Gleitman, & Lederer, 1999; Gleitman, 1990; Landau & Gleitman, 1985; Naigles & Hoff-Ginsberg, 1998).

Taken together, these studies provide compelling support for the idea that frames, construed in a general sense, are structural units that young language learners pay attention to. Moreover, they provide evidence that the specific kind of frame investigated here is psycholinguistically relevant for language learners.

1.2. Comparisons with previous distributional approaches

The approach taken here differs in many significant ways with the other recent investigations of category-relevant distributional information. Although space prohibits a detailed comparison, one unique feature of the present approach that might be favorable for extracting category information is highlighted below.

In the present approach the word 'W' in the environment '… X W Y …' is stored as "jointly following X and preceding Y", but such would not be the case if W occurred after X and before Y on independent occasions. In contrast, the Mintz et al. (2002) and Redington et al. (1998) studies use "bigram" contexts, which record only independent co-occurrence patterns (e.g. "following X", "preceding Y"). There are potentially important consequences of using frame contexts as opposed to bigram contexts. In particular, the property of joint co-occurrence in the frame contexts involves an additional relationship between the context elements themselves, as well as between context and target word. Hence, it is reasonable to assume a priori that if a given frame occurs *frequently* in a corpus of natural language, it is likely to be caused by some systematic aspect of the language, rather than by accident. Therefore, the target words that occur inside each instance of the frequent frame are likely to have some linguistically pertinent relationship, such as grammatical category membership. This is not a necessary outcome, but is an arguably likely consequence of using frequent frames as contexts. To capitalize on this possibility, not only are frames used as contexts in the present approach, but only words that occur in the most frequent frames are categorized (as opposed to the Mintz et al. (2002) and Redington et al. (1998) studies in which target words were selected based on the frequency of the target words themselves). Another important difference between frame and bigram contexts is that, as mentioned above, adults will categorize words in an artificial language based on their occurrence within frames (Mintz, 2002), whereas bigram regularity alone has failed to produce categorization in artificial grammar experiments, without additional cues (Braine, 1987; Gerken, Gomez, & Nurmsoo, 1999; Smith, 1966; Wilson, Gerken, & Nicol, 2000).

The procedure used in Cartwright and Brent's (1997) study differs considerably from both the current approach and the ones just discussed. Nevertheless, as in the frequent frames approach taken here but in contrast to the bigram procedures, joint co-occurrence is a property of the context items in Cartwright and Brent's procedure as well. However, for them, the entire utterance is used as a word's context, hence the scope and complexity of the joint contexts are much greater than in the present approach, and the algorithms involved in computing category membership are very different. In general, Cartwright and

Brent's learning model has many more parameters and specialized computations than the approach taken here.

Finally, the present procedure differs from Redington et al. (1998), and some manipulations in Cartwright and Brent (1997), in that in those studies the input was pooled across many children, whereas here, input corpora to individual children were analyzed separately, resulting in much smaller samples.[1] Important questions can be asked by analyzing the distributional properties of input pooled across many children, however analyzing individual input corpora allows one to evaluate the informativeness of patterns in the input to individuals, which is ultimately the database from which individual children learn.

The goal of the work described here was not to provide a model of grammatical categorization by learners (cf. Cartwright & Brent, 1997), but to examine, based on evidence from human cognition, what assumptions would be reasonable to build into such a model. Specifically, the goal was to formulate a unit to which there is some evidence that children and adults attend, and with which adults have been shown to categorize, and examine how predictive it is of grammatical category membership.

2. Experiment 1

2.1. Method

2.1.1. Input corpora

Six corpora from the CHILDES database (MacWhinney, 2000) served as input for the analysis procedure: Eve (Brown, 1973), Peter (Bloom, Hood, & Lightbown, 1974; Bloom, Lightbown, & Hood, 1975), Naomi (Sachs, 1983), Nina (Suppes, 1974), Anne (Theakston, Lieven, Pine, & Rowland, 2001), and Aran (Theakston et al., 2001). Only sessions of each corpus in which the target child was 2;6 or younger were analyzed, and only utterances of the adults were analyzed. The session range for each corpus analyzed is given in Table 1.

The analyzed utterances were minimally treated before the distributional analysis procedure was performed. All punctuation was removed and all special CHILDES transcription codes were removed.[2]

2.1.2. Distributional analysis procedure

The following procedure was carried out separately on each corpus. First, an exhaustive tally was made of all the frames – where a frame is an ordered pair of words with any word intervening – and the number of times each frame occurred in the corpus. Utterance

---

[1] For example, Redington et al.'s results were based on analyzing a pooled corpus consisting of millions of words of speech from almost 6000 speakers, whereas the largest corpus analyzed here has approximately 107,000 word tokens and the smallest under 30,000. The samples analyzed here are thus substantially smaller than Redington et al.'s, although they were larger than Cartwright and Brent's.

[2] This approach differs from Cartwright and Brent (1997) in that there, declaratives, imperatives, and questions were analyzed separately. They also collapsed certain orthographically and phonologically distinct words, like "dada" and "daddy". Here, utterance types were not differentiated, and differently transcribed words were kept distinct.

Table 1
Experiment 1 session ranges for analyzed corpora, number of utterances, number of tokens and types categorized, percentage of corpus (tokens) accounted for by categorized types, and percentage of corpus (tokens) analyzed

Child	CHILDES sessions	# of utterances	Tokens categorized	Types categorized	Percentage of corpus accounted for	Percentage of corpus analyzed
Peter	peter01-peter12	19846	5690	446	48%	6%
Eve	eve01-eve20	14922	3513	400	46%	5%
Nina	nina01-nina23	14417	6265	469	51%	8%
Naomi	n01-n58	6950	1617	297	38%	5%
Anne	anne01a-anne23b	26199	4389	405	54%	4%
Aran	aran01a-aran20b	20857	5628	620	61%	5%
Mean			4517	439.5	50%	6%

boundaries were not treated as framing elements, nor could frames cross utterance boundaries. Next, a subset of these frames was selected as the set of frequent frames. The principles guiding inclusion in the set of frequent frames were that frames should occur frequently enough to be noticeable, and that they should also occur enough to include a variety of intervening words to be categorized together. While these criteria were not operationalized in the present experiment, a pilot analysis with a randomly chosen corpus, Peter, determined that the 45 most frequent frames satisfied these goals and provided good categorization. Hence, the frames analyzed for each corpus were the 45 most frequent frames for that corpus.

Next, each instance of a given frequent frame was located in the corpus, and the intervening word was stored together in a group with the other intervening words for that frame, creating a frame-based category. The number of times each word occurred in a frame was also recorded. One can therefore distinguish between the number of word types that occur in a frequent frame, and the number of word tokens. (A "word type" is a particular word, for example, "dog", whereas a "word token" is a specific instance of the type in the corpus.)

2.1.3. Quantitative evaluation procedure

In order to obtain a standard measure of categorization success, comparable across corpora and to a control condition (described below), a quantitative measure of categorization called accuracy was calculated for each corpus. Accuracy is a standard metric for measuring categorization success, and was used for reporting results by Cartwright and Brent (1997) and Redington et al. (1998). To compute accuracy, for each frame-based category all possible pairs of word tokens in the category were compared. Each pair was then classified as either a Hit or a False Alarm. A Hit was recorded when two items were from the same grammatical category (i.e. they were correctly grouped together), and a False Alarm was recorded when two items were from different grammatical categories (i.e. they were incorrectly grouped together). Accuracy measures the proportion of Hits to the number of Hits plus False Alarms (i.e. the proportion of all words grouped together that were correctly grouped together), maximum accuracy being 1.

The equation for accuracy is given as Eq. (1).

$$\text{Accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}} \tag{1}$$

In order to calculate Hits and False Alarms, the categorized tokens were first labeled with their true grammatical category. Two different labeling protocols were used. In Standard Labeling, each categorized token was labeled as noun (nouns and pronouns), verb (verbs, auxiliaries, and copula forms), adjective, preposition, adverb, determiner, wh-word, "not", conjunction, or interjection. In Expanded Labeling, nouns and pronouns were labeled as distinct categories, as were verbs, auxiliaries, and the copula. In situations where the grammatical category of the word was ambiguous (for example, if it was unclear whether "walk" was used as a noun or a verb) the corpus was consulted to disambiguate and appropriately label the word.

Accuracy was the primary outcome measure in this analysis, however, a second measure, completeness, assessed the degree to which the analysis grouped in the same distributional category words that belong to the same grammatical category. Completeness measures the proportion of the number of Hits to the number of Hits plus Misses (Eq. (2)). Misses are computed by comparing all possible pairs of word tokens that were categorized. A pair is counted as a Miss if the members belong to the same grammatical category but were not categorized together by the analysis. Thus, whereas accuracy is penalized when the analysis groups together words belonging to different grammatical categories, completeness is penalized when the analysis fails to group together words that belong to the same grammatical category.[3] As with accuracy, maximum completeness is 1.

$$\text{Completeness} = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{2}$$

The calculation of Hits, Misses, and False Alarms described above was based on comparing tokens to tokens. In evaluating the analysis outcomes, these measures were also calculated by comparing *types*. To illustrate how token measures might differ from type measures, consider a hypothetical category that contained ten different word types: five noun types consisting of 500 tokens all together, and five verbs types consisting of 50 tokens all together. Calculating token accuracy, there would be 125,975 Hits (500*499/2 for nouns plus 50*49/2 for verbs) and 25,000 False Alarms (500*50), yielding an accuracy of about 0.83. Calculating type accuracy would yield 20 Hits (5*4/2 for nouns and verbs each) and 25 False Alarms, resulting in an accuracy of about 0.44. Basically, token-based measures are frequency weighted, whereas type measures are not. Which measure is most relevant depends on how the distributional categories are further processed (for example, if lower frequency members are later filtered out, then token accuracy might be the more indicative rating of the resulting categories). In any case, both token and type accuracy and completeness measures are reported below and, as will be seen, they are numerically very close for these corpora.

---

[3] It should be noted that this calculation of completeness is evaluated against the words that the procedure actually categorizes (the tokens occurring in frequent frames), rather than against *all* words in the corpus.

2.1.4. Computing chance categorization

As a baseline control against which to compare the accuracy and completeness of the frame-based categories, chance categories were created for each corpus. Chance categories were designed to match in number and in size with the frame-based categories created by the analysis procedure for a given corpus. The content of the chance categories was determined by selecting the word tokens in the frame-based categories and randomly distributing them among the chance categories. Token and type accuracy and completeness were computed (for both Standard and Expanded Labeling) on the chance categories for each corpus to yield baseline measures. The baseline indicates the accuracy and completeness that could be achieved given the category structure that resulted from analyzing the corpus – number and size of categories – but without considering the distributional structure of the corpus.

2.2. Results

2.2.1. Overview

The mean number of word types categorized across input corpora was 440 ($M = 439.5$, SD = 106.3). This is more than twice the number of words categorized in Mintz et al. (2002) and in Cartwright and Brent (1997) (although Cartwright and Brent did not report this information for all of their analyses). Table 1 shows the number of word types and tokens categorized for each corpus, and the percentage of each corpus that consisted of tokens of the categorized types ($M = 50.0\%$, SD = 7.9%). On average, the types constituting half of the tokens in each corpus were contained in the 45 most frequent frames.

2.2.2. Sample frame-based categories

Frequent frames contained words from a range of categories, including nouns, verbs, adjectives, pronouns, adverbs, and auxiliaries. Table 2 provides representative examples of the resulting frame-based categories. As the table shows, the words contained in each frame-based category were almost exclusively from one grammatical category.

2.2.3. Accuracy and completeness

Mean token accuracy across corpora for Standard and Expanded Labeling was 0.98 and 0.91, respectively. These scores were significantly higher than baseline ($M = 0.46$, $t(5) = 25.0$, $P < 0.0001$, Standard Labeling; $M = 0.27$, $t(5) = 35.2$, $P < 0.0001$, Expanded Labeling). Mean type accuracy was 0.93 and 0.91 for Standard and Expanded Labeling, respectively. These scores were significantly higher than baseline ($M = 0.47$, $t(5) = 25.8$, $P < 0.0001$, Standard Labeling; $M = 0.38$, $t(5) = 25.6$, $P < 0.0001$, Expanded Labeling). The difference between token and type accuracy was significant for Standard Labeling ($t(5) = 5.8$, $P < 0.01$), however accuracy was clearly quite high in both cases. Table 3 shows token and type accuracy with Standard and Expanded Labeling for the frame-based categories and for the baseline random categorization.

Mean token completeness across corpora for Standard and Expanded Labeling was 0.07 and 0.12, respectively. These scores were significantly higher than baseline ($M = 0.03$, $t(5) = 11.0$, $P < 0.0001$, Standard Labeling; $M = 0.03$, $t(5) = 15.8$,

Table 2
Samples of representative categories from several corpora. The number of tokens categorized for each type is in parentheses

---

Peter

Frame you__it

put (52), see (28), do (27), did (25), want (23), fix (13), turned (12), get (12), got (11), turn (10), throw (10), closed (10), think (9), leave (9), take (8), open (8), find (8), bring (8), took (7), like (6), knocked (6), putting (5), pull (5), found (5), make (4), have (4), fixed (4), finish (4), try (3), swallow (3), opened (3), need (3), move (3), hold (3), give (3), fixing (3), drive (3), close (3), catch (3), threw (2), taking (2), screw (2), say (2), ride (2), pushing (2), hit (2), hiding (2), had (2), eat (2), carry (2), build (2), brought (2), write (1), wiping (1), wipe (1), wind (1), unzipped (1), underneath (1), turning (1), touching (1), tore (1), tie (1), tear (1), swallowed (1), squeeze (1), showing (1), show (1), said (1), rip (1), read (1), reach (1), pushed (1), push (1), play (1), pick (1), parking (1), made (1), love (1), left (1), knock (1), knew (1), hid (1), flush (1), finished (1), expected (1), dropped (1), drop (1), draw (1), covered (1), closing (1), call (1), broke (1), blow (1)

Frame I__it

see (18), put (12), think (9), got (8), thought (5), have (5), found (5), do (4), take (3), open (3), fix (3), did (3), closed (3), use (2), tie (2), tear (2), need (2), know (2), hear (2), guess (2), give (2), doubt (2), wear (1), took (1), throw (1), threw (1), saw (1), read (1), pushed (1), pick (1), move (1), leave (1), knock (1), knew (1), get (1), fixed (1), finished (1), close (1), build (1), bet (1)

the__one

other (21), red (11), yellow (8), green (8), orange (6), big (6), blue (5), right (4), small (3), little (3), wrong (1), top (1), round (1), only (1), light (1), empty (1), black (1)

Aran

Frame you__it

put (28), want (15), do (10), see (7), take (6), turn (5), taking (5), said (5), sure (4), lost (4), like (4), leave (4), got (4), find (4), throw (3), threw (3), think (3), sing (3), reach (3), picked (3), get (3), dropped (3), seen (2), lose (2), know (2), knocked (2), hold (2), help (2), had (2), gave (2), found (2), fit (2), enjoy (2), eat (2), chose (2), catch (2), with (1), wind (1), wear (1), use (1), took (1), told (1), throwing (1), stick (1), share (1), sang (1), roll (1), ride (1), recognize (1), reading (1), ran (1), pulled (1), pull (1), press (1), pouring (1), pick (1), on (1), need (1), move (1), manage (1), make (1), load (1), liked (1), lift (1), licking (1), let (1), left (1), hit (1), hear (1), give (1), flapped (1), fix (1), finished (1), drop (1), driving (1), done (1), did (1), cut (1), crashed (1), change (1), calling (1), bring (1), break (1), because (1), banged (1)

Frame the__and

tractor (5), horse (4), shark (3), back (3), zoo (2), top (2), tiger (2), roof (2), leg (2), grass (2), garage (2), window (1), wellingtons (1), water (1), video (1), train (1), sun (1), station (1), stars (1), shop (1), shirt (1), sand (1), round (1), rain (1), pussycat (1), postbox (1), panda (1), nuts (1), mother (1), monkey (1), lion (1), kite (1), ignition (1), hut (1), holes (1), hippo (1), hens (1), ham (1), giraffe (1), floor (1), fire + engine (1), eye (1), entrance (1), elephant (1), dolly (1), doctor (1), cups (1), cows (1), controls (1), carts (1), carpark (1), cake (1), bus (1), bull (1), brush (1), box (1), bottom (1), book (1), blue (1), bits (1), bank (1), bananas (1), animals (1), air (1)

Frame put__in

it (49), them (14), him (11), things (6), that (5), those (4), teddy (2), dolly (2), yourself (1), you (1), what (1), this (1), these (1), some (1), panda (1), her (1), Pingu (1)

Naomi

Frame you__it

like (11), put (8), want (6), throw (4), think (3), see (3), eat (3), did (3), take (2), open (2), got (2), turning (1), turn (1), touched (1), throwed (1), spit (1), spill (1), snapped (1), shaking (1), say (1), rubbing (1), pull (1), pour (1), pick (1), left (1), hurt (1), how (1 wh), holding (1), have (1), guessed (1), give (1), finish (1), find (1), enjoy (1), eating (1), dropped (1), distorted (1), discovered (1), cutting (1), coloring (1), closed (1), cleaning (1), call (1), ate (1)

Frame the__is

moon (6), sun (5), truck (3), smoke (2), kitty (2), fish (2), dog (2), baby (2), tray (1), radio (1), powder (1), paper (1), man (1), lock (1), lipstick (1), lamb (1), kangaroo (1), juice (1), ice (1), flower (1), elbow (1), egg (1), door (1), donkey (1), doggie (1), crumb (1), cord (1), clip (1), chicken (1), bug (1), brush (1), book (1), blanket (1), Mommy (1)

Table 3
Experiment 1 token and type accuracy for Standard and Expanded Labeling including baseline accuracy of random categories

Corpus	Token accuracy (Standard)		Token accuracy (Expanded)		Type accuracy (Standard)		Type accuracy (Expanded)	
	Analysis	Random	Analysis	Random	Analysis	Random	Analysis	Random
Peter	0.98	0.49	0.97	0.32	0.96	0.55	0.95	0.49
Eve	0.98	0.51	0.91	0.25	0.92	0.50	0.89	0.40
Nina	0.98	0.48	0.98	0.29	0.95	0.46	0.94	0.36
Naomi	0.97	0.48	0.96	0.30	0.94	0.49	0.93	0.41
Anne	0.98	0.37	0.84	0.24	0.94	0.41	0.90	0.31
Aran	0.97	0.44	0.80	0.23	0.89	0.42	0.87	0.33
Mean	0.98	0.46	0.91	0.27	0.93	0.47	0.91	0.38

$P < 0.0001$, Expanded Labeling). Mean type completeness was 0.08 and 0.10 for Standard and Expanded Labeling, respectively. These scores were significantly higher than baseline ($M = 0.04$, $t(5) = 10.2$, $P < 0.0001$, Standard Labeling; $M = 0.04$, $t(5) = 9.0$, $P < 0.001$, Expanded Labeling). Token and type completeness scores were not significantly different for either labeling protocol. Table 4 shows token and type completeness with Standard and Expanded Labeling for the frame-based categories and for the baseline random categorization.

2.2.4. Cross-corpus consistency

In order to assess how consistent the most frequent frames were across corpora, an analysis was performed to determine the extent to which corpora overlapped in their frequent frames. Fig. 1 shows the average percentages of frequent frames per corpus that occurred in all, only five, only four, only three, only two, and only one corpus.

Table 4
Experiment 1 token and type completeness for Standard and Expanded Labeling including baseline accuracy of random categories

| Corpus	Token completeness (Standard)		Token completeness (Expanded)		Type completeness (Standard)		Type completeness (Expanded)	
	Analysis	Random	Analysis	Random	Analysis	Random	Analysis	Random
Peter	0.06	0.03	0.09	0.03	0.07	0.04	0.08	0.04
Eve	0.06	0.03	0.12	0.03	0.07	0.04	0.09	0.04
Nina	0.08	0.04	0.13	0.04	0.10	0.05	0.12	0.05
Naomi	0.07	0.03	0.11	0.04	0.07	0.03	0.08	0.04
Anne	0.08	0.03	0.11	0.03	0.09	0.04	0.12	0.04
Aran	0.08	0.04	0.13	0.04	0.09	0.04	0.10	0.04
Mean	0.07	0.03	0.12	0.03	0.08	0.04	0.10	0.04
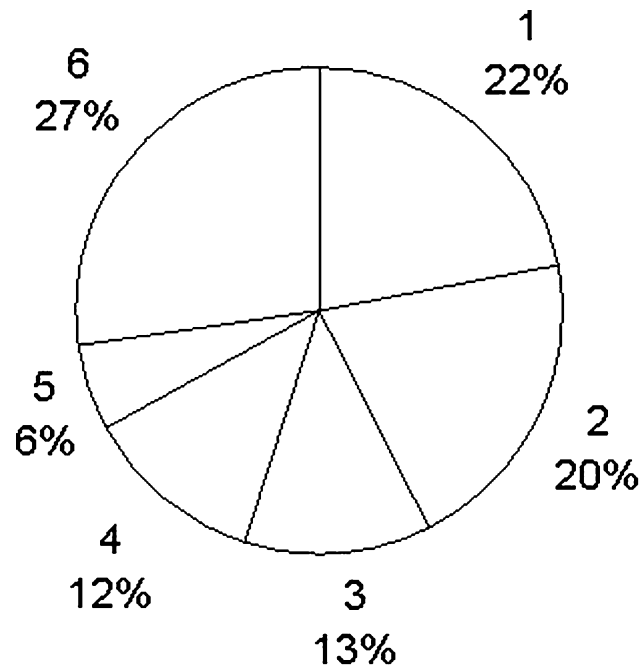
Fig. 1. Percentage breakdown of frequent frames in an average corpus that occur only in that corpus (1), in exactly one other (2), up to all six corpora.

As the figure shows, on average 45% of the frequent frames of a given corpus were frequent frames for at least three other corpora, indicating that many informative distributional contexts are shared from corpus to corpus. Table 5 lists all the frequent frames that occurred in at least two corpora, organized by the number of corpora in which they occurred.

2.3. Discussion

The major finding was that frequent frames are extremely effective at categorizing words. Token accuracy scores were 0.97 or more for all corpora under the Standard Labeling protocol, and averaged over 0.90 under the Expanded Labeling protocol; type accuracy was comparable, although slightly lower (0.93 average) under Expanded Labeling. The difference in Standard and Expanded scores indicates that some nouns and pronouns were categorized together, and some auxiliaries, main verbs, and/or copula forms were categorized together. The similarity in token and type scores indicates that word types in a given frame-based category largely belonged to the same grammatical category, regardless of the frequency of those types within the frame. The accuracy obtained here was on par with, or higher than results in other studies (Cartwright & Brent, 1997; Mintz et al., 2002; Redington et al., 1998), although the force of this comparison is limited due to differences in the number of words categorized and the number of resulting categories. It is clear from both the quantitative and qualitative results that this method

Table 5
Frames that were frequent frames in at least two corpora, organized by number of corpora in which each occurred

6	5	4	3	2
do__want	a__of	I__think	I__know	I__it
put__on	put__in	I__you	are__doing	a__on
the__in	to__the	are__going	did__do	can__it
the__on		is__a	go__the	do__think
to__it		it__the	the__of	don't__it
want__to		to__a	the__one	have__got
what__you		would__like	there__is	have__look
you__a		you__that	to__to	here__are
you__it			what__it	in__box
you__me			why__you	put__back
you__the			you__with	shall__put
you__to				the__and
				the__is
				there__are
				to__on
				to__with
				we__a
				we__to
				what're__doing
				what__is
				what__that
				what__we
				you're__to
				you__have
				you__some
				you__what
				you__your

produces extremely accurate categories. These results are especially impressive when one considers the restricted distributional contexts used here – the 45 most frequent frames – and they are compelling given the evidence that infants and adults naturally attend to this type of unit.

Although the categories formed were unquestionably accurate, there were often several noun categories and several verb categories (all very accurate), rather than one category of all the nouns, one of all the verbs, etc.[4] This outcome is reflected in the comparatively low completeness scores. Nevertheless, it is clear from Table 2 that the categories, in general, are relatively large (by token or type counts); thus, it was not the case that low completeness was due to numerous accurate categories with only a few members each. It is not surprising that there should be multiple distributional categories that correspond to the same grammatical category, since the distributional analysis procedure dictates that any given frequent frame, in essence, constitutes a category. Of course, it would be desirable if distributional information could be used to make categories that are comprehensive as well

---

[4] This property also holds of the other recent studies discussed above.

as accurate. Plausible methods for combining categories to make them more complete are taken up in Section 4.

As Table 5 shows, the frames themselves largely consisted of closed class items, including determiners, prepositions, auxiliary verbs, and pronouns. However, some open class items made up framing elements as well: for example, the verbs 'put' and 'want' made up frequent frames in each corpus, 'think' and 'like' occur in four, 'know' in three, and 'look' in two. Nouns also made up several framing contexts, although not to the same degree as verbs: for example, the noun 'box' was a framing element in two corpora, and frames idiosyncratic to an individual corpus included 'bag' and 'table'. The potential significance of this finding is addressed further in Section 4.

The distributional information provided by frequent frames was robust. The word types that were categorized constituted, on average, 50% of the tokens in a given corpus. This coverage was achieved by analyzing only about 6% of the tokens and their contexts. That is, the tokens of the categorized types making up the 6% contained in frequent frames constituted half the tokens in a given corpus. Frequent frames can thus focus a learner on a relatively small number of contexts that can have broad impact on how words in the input are categorized. The efficiency and accuracy provided by frequent frames could thus be very useful to young language learners, who have limited memory and processing resources.

Although for the most part, words in a given frame-based category belonged to the same grammatical category, there were some categorization errors. As mentioned above, differences in accuracy between Standard and Expanded Labeling indicate that nouns and pronouns were occasionally grouped together, as were auxiliaries and main verbs. In addition, in some cases prepositions and verbs were grouped together. For example, the frame it__the was a frequent frame for prepositions in four of the corpora, however in three of those corpora the frame contained some verbs as well. Similarly, go__the occurred in four corpora as a frequent frame containing prepositions, but contained some verbs as well in three of those corpora. In the it__the case, the verbs that occurred were infrequent (occurring once or twice in the analyzed samples). Thus, one way to circumvent the erroneous classifications such as these would be to filter out extremely low frequency targets. Of course, this might have the undesirable effect of eliminating some correctly classified words as well. Erroneous classifications and their significance to a frame-based approach to early word categorization will be taken up further in Section 4.

A limitation in the present experiment is that the set of frequent frames was selected by the same absolute threshold for all corpora (the 45 most frequent frames). It would be desirable to analyze the corpora using a frequency threshold for each corpus that is based on a relativized frequency criterion, as the salience of frequent frames to human learners is more likely to be a factor of relative frequency than absolute number. Experiment 2 addresses this issue. In addition, a contributing factor to the high accuracy in Experiment 1 might have been the several frame-based categories that had only one or two member types (for example, the frame want__put in the Nina corpus contained 98 tokens of *to*, only). If members of these miniscule categories belonged to the same grammatical category, they would contribute to raising accuracy (especially token accuracy, if the token frequency for the few members within a frame were high) without necessarily

contributing to qualitatively good categorization. Experiment 2 was designed to mitigate this effect, in addition to implementing a relativized criterion for frequent frames.

3. Experiment 2

The purpose of Experiment 2 was to examine the categorization outcome when a frame selection method is used that is sensitive to the frequency of frames relative to the total number of frames in a corpus. An additional goal was to ensure that high accuracy scores in Experiment 1 were not due to very small categories with only a few member types, as such categorization, although accurate, is not linguistically interesting.

3.1. Method

As in Experiment 1, within each corpus, all frames were tallied and ranked by frequency. The set of frequent frames was then selected to include all frames whose frequency in proportion to the total number of frames in the corpus surpassed a predetermined threshold of 0.13%. That is, a given frame in a corpus was included as a frequent frame just in case its frequency was at least 0.13% of the total number of frames in the corpus. This specific threshold was determined based on the frequent frames for each corpus in Experiment 1. In particular, the frequent frames in Experiment 1 were analyzed, corpus by corpus, by tallying the frequency of the least frequent member of the set of frequent frames, and expressing that frequency as a proportion of the total number of frames for that corpus, yielding a different proportional threshold for each corpus. These thresholds were then averaged, yielding 0.0013, or 0.13%, and this was the threshold used for all corpora in Experiment 2. Thus, the frequent frame selection method for Experiment 2 provided a kind of normalization of the method used in Experiment 1.

The other difference in the selection of frequent frames was that in Experiment 2, frequent frames consisting of only one or two word types were removed from the set of frequent frames. This modification helped to guard against minimal frame-based categories contributing to high accuracy. In addition, it brings the frame selection criterion slightly more in line with the research on infants' ability to pick up on frame-like units: work by Gómez (2002) suggests that non-adjacent dependencies are more likely to be noticed when the intervening material varies. At the same time, since some grammatical categories, such as pronouns and determiners, are indeed relatively small, it is desirable that the cutoff for the minimum number of types be low enough to include frames that capture these smaller grammatical categories. These factors guided the decision to restrict frequent frames to those having three or more different word types. On average, four frames per corpus were excluded for this reason.[5]

All other aspects of this experiment were equivalent to those in Experiment 1.

---

[5] An additional analysis was run in which these frames were not removed from the set of frequent frames. The results were numerically identical to the reported outcomes for Experiment 2.

Table 6
Experiment 2 session ranges for analyzed corpora, number of utterances, number of tokens and types categorized, percentage of corpus (tokens) accounted for, and percentage of corpus (tokens) analyzed

Child	CHILDES sessions	# of utterances	Tokens categorized	Types categorized	Percentage of corpus accounted for	Percentage of corpus analyzed
Peter	peter01-peter12	19846	5086	437	47%	5%
Eve	eve01-eve20	14922	3380	398	43%	4%
Nina	nina01-nina23	14417	4309	387	42%	6%
Naomi	n01-n58	6950	1319	294	34%	4%
Anne	anne01a-anne23b	26199	4839	512	60%	5%
Aran	aran01a-aran20b	20857	6172	676	66%	6%
Mean			4184.2	450.7	49%	5%

3.2. Results and discussion

Using the normalized frame selection criterion, the mean number of word types categorized across input corpora was 432 (SD = 96.3). Table 6 shows the number of word types and tokens categorized for each corpus and the percentage of each corpus that consisted of tokens of the categorized types ($M = 49.0\%$, SD = 8.2%). On average, the types constituting almost half of the tokens in each corpus were contained in the frequent frames. Thus, the outcome of this experiment resembles that of Experiment 1 on these general properties.

Mean token accuracy across corpora for Standard and Expanded Labeling was 0.98 and 0.91, respectively. These scores were significantly higher than baseline ($M = 0.49$, $t(5) = 15.9$, $P < 0.0001$, Standard Labeling; $M = 0.29$, $t(5) = 38.2$, $P < 0.0001$, Expanded Labeling). Mean type accuracy was 0.94 and 0.92 for Standard and Expanded Labeling, respectively. These scores were significantly higher than baseline ($M = 0.50$, $t(5) = 18.3$, $P < 0.0001$, Standard Labeling; $M = 0.39$, $t(5) = 22.3$, $P < 0.0001$, Expanded Labeling). The difference between token and type accuracy was significant for Standard Labeling ($t(5) = 5.7$, $P < 0.01$), however accuracy was quite high in both cases. Table 7 shows token and type accuracy with Standard and Expanded Labeling. Comparing Table 7 with Table 3, it is clear that the accuracy outcomes for Experiments 1 and 2 were extremely similar. Thus, frequent frames yielded high category accuracy across corpora under relativized frame selection criteria, as well as when frames were selected based on absolute frequency (Experiment 1). In addition, the high accuracy scores did not appear to be inflated by very small categories (in word types), as the scores remained high even when the minimal categories were removed in this experiment.

Mean token completeness across corpora for Standard and Expanded Labeling was 0.08 and 0.13, respectively. These scores were significantly higher than baseline ($M = 0.04$, $t(5) = 6.9$, $P < 0.001$, Standard Labeling; $M = 0.04$, $t(5) = 9.4$, $P < 0.001$, Expanded Labeling). Mean type completeness was 0.08 and 0.10 for Standard and Expanded Labeling, respectively. These scores were significantly higher than baseline ($M = 0.04$, $t(5) = 6.8$, $P < 0.01$, Standard Labeling; $M = 0.04$, $t(5) = 7.2$, $P < 0.001$,

Table 7
Experiment 2 token and type accuracy for Standard and Expanded Labeling including baseline accuracy of random categories

Corpus	Token accuracy (Standard)		Token accuracy (Expanded)		Type accuracy (Standard)		Type accuracy (Expanded)	
	Analysis	Random	Analysis	Random	Analysis	Random	Analysis	Random
Peter	0.98	0.51	0.97	0.32	0.95	0.59	0.95	0.53
Eve	0.98	0.56	0.91	0.27	0.92	0.52	0.89	0.38
Nina	0.98	0.52	0.97	0.32	0.95	0.48	0.94	0.36
Naomi	0.96	0.56	0.96	0.39	0.94	0.51	0.93	0.42
Anne	0.98	0.37	0.82	0.23	0.94	0.40	0.90	0.34
Aran	0.97	0.45	0.80	0.22	0.91	0.42	0.88	0.33
Mean	0.98	0.49	0.91	0.29	0.94	0.50	0.92	0.39

Expanded Labeling). The difference between token and type completeness was significant for Expanded Labeling ($t(5) = 4.4$, $P < 0.01$). Table 8 shows token and type completeness with Standard and Expanded Labeling for the frame-based categories and for the baseline random categorization. Comparing Table 4 with Table 8 clearly shows that, like accuracy outcomes, completeness outcomes for Experiments 1 and 2 were extremely similar.

As in Experiment 1, frequent frames resulted in extremely accurate categories, with accuracy scores that were very similar in both analyses. The sample categories given in Table 2 for Experiment 1 were produced in Experiment 2 as well (simply because the associated frames were frequent frames in both analyses), and are equally representative of the kinds of categories resulting from the present analysis. Likewise, cross-corpus consistency in Experiment 2 resembled that of Experiment 1 (shown in Table 5 and Fig. 1),

Table 8
Experiment 2 token and type completeness for Standard and Expanded Labeling including baseline accuracy of random categories

Corpus	Token completeness (Standard)		Token completeness (Expanded)		Type completeness (Standard)		Type completeness (Expanded)	
	Analysis	Random	Analysis	Random	Analysis	Random	Analysis	Random
Peter	0.06	0.03	0.08	0.03	0.06	0.04	0.07	0.04
Eve	0.07	0.04	0.13	0.04	0.07	0.04	0.09	0.04
Nina	0.08	0.04	0.13	0.04	0.08	0.04	0.10	0.04
Naomi	0.07	0.04	0.11	0.05	0.07	0.04	0.08	0.04
Anne	0.10	0.04	0.13	0.04	0.10	0.04	0.13	0.05
Aran	0.10	0.05	0.17	0.05	0.11	0.05	0.13	0.05
Mean	0.08	0.04	0.13	0.04	0.08	0.04	0.10	0.04

as the frequent frames were very similar across the two experiments. The average coverage of each corpus from the categorized types was virtually the same in each experiment as well (49% here compared to 50% in Experiment 1). Thus, in all respects the outcomes from Experiment 1 were replicated here with relativized criteria for selecting frequent frames, and with filtering out minimal categories.

4. General discussion

To summarize, categorizing words in child directed speech on the basis of their distribution within frequent frames produces extremely accurate categories, and frequent frames provided categorization accuracy that was equal to or surpassed the accuracy of distributional methods investigated in prior studies (Cartwright & Brent, 1997; Mintz et al., 2002; Redington et al., 1998). Moreover, the information provided by frequent frames was robust, in that by analyzing only a small portion of each corpus (6% in Experiment 1, 5% in Experiment 2), the types that constituted about half of the corpus were categorized. The efficiency of frequent frames, and the relative simplicity of the computations and representations involved in making use of them for categorization make them compelling candidates for learners with limited resources. In addition, although there was variability across corpora in the frames that occurred, there was also considerable consistency, suggesting that categorization mechanisms that use this kind of information would be able to fruitfully analyze linguistic input from different sources without having to constantly re-evaluate what frames to pay attention to. Finally, results were stable when different criteria were used for selecting frequent frames. These findings, along with independent evidence that human learners, including infants, attend to frame-like units (Gómez, 2002; Mintz, 2002; Peña et al., 2002; Santelmann & Jusczyk, 1998), suggest that this kind of distributional information is a viable basis for young children's early word categorization.

However, as was mentioned in the discussion of Experiment 1 (Section 2.3), although the resulting categories for both analyses were very accurate, there was often more than one frame-based category for a given grammatical category. For example, most corpora yielded several noun categories, verb categories, etc., and this was reflected in the relatively low completeness scores. This is not surprising given that each frame defines a distinct distributional category, and there was generally more than one frame that contained members of a given grammatical category. Of course, it would be desirable if these frame-based categories could be made more comprehensive. There is at least one simple way to unify distinct frame-based categories that contain words from the same grammatical category. It is a prevalent characteristic of these frame-based categories that there is considerable overlap in the words they contain. For example, the verb categories defined by frames you__to, she__to, you__the, etc., will generally have a number of member words in common because many of the same verbs can appear in each environment. Hence, two frame-based categories could be unified if they surpass a threshold of lexical overlap. This possibility was tested on the results from one of the corpora, Peter, using a criterion of 20% overlap. The outcome was that 17 different verb categories were joined to form one category of 261 word types, 99.3% of which were verbs. The non-verb items were from disparate grammatical categories and only occurred

once or twice in the frames that constituted the composite category.[6] Accuracy was not adversely affected by the unification of categories, remaining at 0.90 or above (for all combinations of type/token and Standard/Expanded Labeling), indicating that the unification procedure did not join together frame-based categories containing words from different grammatical categories. Furthermore, type completeness reached 0.91 (compared to 0.07 before unification!), indicating that, as expected, the distributional categories that had been fragments of grammatical categories were merged by the unification procedure. Although further research is needed to better understand the effectiveness and limitations of this technique, it appears that a very simple conglomeration procedure based on lexical overlap could be used to join accurate smaller categories together into a more complete category.

4.1. Mechanisms for identifying informative frames

Of course, despite the robust informativeness of the restricted set of frequent frame contexts, corpus-wide processing is necessary at least at a superficial level to identify the frequent frame contexts to begin with. However, as was mentioned earlier, there is evidence that co-occurrence information of the type required here is tracked by infants (Gómez, 2002; Santelmann & Jusczyk, 1998) as well as adults (Gómez, 2002; Mintz, 2002; Peña et al., 2002), that adults categorize words based on frame contexts (Mintz, 2002), and that humans' sensitivity to co-occurrence patterns in speech occurs very early on (e.g. Saffran, Aslin, & Newport, 1996). Furthermore, the capacity to notice co-occurrences in speech appears to be available to non-human primates as well (Hauser, Newport, & Aslin, 2001). Given this broad range of evidence, it is not unreasonable to expect that identifying the most frequently occurring frames would be an achievable, and perhaps natural task for the young language learner.

That said, it is worth noting that another avenue potentially exists for arriving at a set of informative frames that does not rely on corpus-wide tracking of frame frequencies. Recall that the framing elements in the frequent frames predominantly consisted of closed-class words. The prevalence of closed-class framing elements is not terribly surprising: since the closed-class items have the highest frequencies, they should be the most likely overall to co-occur in an utterance in the domain defined by frames. But that they provide such accurate categorization environments is potentially significant. Neonates have been shown to differentiate English closed-class words from open-class words on prosodic grounds (Shi, Werker, & Morgan, 1999). Other researchers have proposed that closed-class items might act as "anchors" that establish a starting point for a distributional analysis around the anchoring item (Valian & Coulson, 1988), and as a cue for phrase marking (Gerken, Landau, & Remez, 1990; Kimball, 1973; Mintz et al., 2002; Morgan, Meier, & Newport, 1987). Thus, the unique perceptual and constitutional properties of these items – a small

---

[6] Verbs were the most fragmented category, so the conglomeration effects are most dramatic for verbs. Nevertheless, in the same pilot test of this procedure, the two groups containing all but one (type and token) of the nouns were joined to form one category of 88 types (217 tokens), all but one of which were nouns and the remaining word was a pronoun. Similarly, the two adjective categories in the Eve corpus were joined to make one group of 25 types (107 tokens) consisting entirely of adjectives except for one instance of the determiner *another*.

set of short, unstressed, and vocalically reduced words – might guide the learner to attend to these potential frame environments (more than, say, ones defined by open class words), and this would result in a set of frames greatly resembling the frequent frames analyzed here. Hence, although frequent frames might be naturally salient to young learners, learners could additionally rely on prosodic cues correlated with closed-class words to arrive at a set of grammatically informative frames.[7]

4.2. Errors

As was mentioned in the discussion of Experiment 1 (Section 2.3), despite the excellent overall categorization yielded by frequent frames, there were some frames that grouped together words belonging to different grammatical categories. For example, the frame it__the was a frequent frame in four of the corpora, containing mostly prepositions, but also some verbs in three of those four corpora. It was noted, however, that within the frame, the frequency of the few verbs was very low, each occurring only once or twice. Just as the distributional analysis procedure itself is sensitive to frame frequency, perhaps the frequency of target words within a frame must be considered as well: one can imagine a mechanism that makes categorization commitments only for words that surpass some frequency threshold within a frame. In most frames where there are a small number of low frequency words that do not conform to the category of the majority of words – tokens and types – this would filter out the non-conforming elements. Of course, such a mechanism would filter out low frequency words inside a frame that do conform to the majority category as well. Might such a mechanism reduce the overall coverage of a frame-based categorization procedure? Perhaps not: as discussed in the preceding section, there is much lexical overlap of target words across frames. A conglomeration procedure like the one discussed above joins frames that have a certain degree of lexical overlap. In addition to the much improved completeness of the resulting category, another potential effect is that a correctly grouped word that might be below the frequency threshold in one or more individual frames would be above the threshold in the joined category, since its frequency across the joined frame categories would be summed. The variety of environments represented by the joined categories would help ensure that the erroneously grouped target words would remain below the threshold after conglomeration, because a given misclassified word is unlikely to be one of the overlapping words, and thus would not gain a frequency boost in the conglomeration process.

However, not all situations of misclassification can be corrected by such a procedure. For example, in the go__the frame in the Peter corpus, the verb *get* is almost as frequent as the preposition *to*. There are, perhaps, other distributional facts that could correct this kind of misclassification as well. For example, *get* occurs five

---

[7] An additional set of analyses was run in which the frequent frames in Experiment 2 were restricted to closed-class elements only, to see what categories would result if the learner only had access to frames defined by closed-class words. The following scores were obtained: under Standard Labeling, token and type accuracy were 0.98 and 0.94, respectively, and token and type completeness were 0.10 and 0.11, respectively; under Expanded Labeling, token and type accuracy were 0.90 and 0.93, respectively, and token and type completeness were 0.15 and 0.12, respectively. Thus, the results when only frequent closed-class frames were used were on par with those when framing elements were not restricted to closed-class words.

times as often in the conglomerated verb category as it does in the mostly preposition-filled go__the frame. Understanding the degree to which this type of information would be helpful must be left to future research.

It should be noted that the distributional methods investigated in previous studies resulted in some misclassifications as well (Cartwright & Brent, 1997; Mintz et al., 1995, 2002; Redington et al., 1998). In particular, the bigram methods (but not the current method) often resulted in one or two large categories that were linguistically incoherent (Mintz et al., 1995, 2002; Redington et al., 1998). Thus, there do appear to be limits on the extent to which a purely distributional analysis procedure could provide a full categorization of words, and most studies in this area discuss how other sources of information could work in conjunction with distributional information (the section that follows sketches out part of such a scenario, see also Mintz et al., 2002). What is remarkable given the potential pitfalls for distributional analyses is that with frequent frames, the misclassifications are quite rare, as indicated by the high accuracy scores and the representative categories shown in Table 2.[8]

4.3. From distributional to grammatical categories

This study investigated a method by which words from the same grammatical category can be grouped together on distributional grounds. However, true grammatical categories are more than clusters of words. They embody information about the kinds of structural relationships category members can enter into with words from other grammatical categories. For example, being a verb means that a word can enter into specific kinds of relationships with one or more noun phrases. Clearly the categories derived here do not directly contain information about syntactic privileges. However, there are plausible ways that the necessary grammatical facts could be linked to distribution-based categories. One possibility is briefly sketched below.

One way to view the utility of the distributional information described here is as a way of bootstrapping into a parameterized universal grammar that contains category distinctions, such as noun and verb, specifications of whether the category is a phrasal head, specifications of options of ways in which phrases could be combined, etc. Fitting the distributional categories into the grammar would then amount to labeling the distributional categories as noun, verb, adjective, etc. One might call such a procedure Distributional Bootstrapping, as this brief sketch resembles in some respects Pinker's (1984) Semantic Bootstrapping proposal. However, for Pinker the foundational "boot-strap" categories were derived from semantic information rather than distributional information (see also Grimshaw, 1981; MacNamara, 1972). According to Pinker's proposal, learners identify the semantic category of a word (e.g. action-word) and then

---

[8] It may be tempting to inquire whether the occasional misclassifications in frequent frames are attested in children's productions. To my knowledge, these kinds of errors (e.g. treating verbs as prepositions) are not attested. However, the absence of such errors should not be taken as evidence against frames as a basis for categorization. As was just stated and as is further developed below, the proposal put forward is that frame-based analyses might support an initial bootstrapping into categorization, and that other sources of information could in subsequent refinements. It is thus difficult to make direct predictions from the distributional analysis results to child errors until these factors and how they interact are further specified.

innate linking rules classify the word (e.g. as a verb). The newly categorized word can then be fit into the developing grammar, and at early stages might be used as a source of information for determining certain language-specific aspects of the grammar (e.g. head branching direction, case marking, etc.). The primary difference between semantic bootstrapping and a distributional approach is that in the latter, the bootstrap categories are defined distributionally rather than on semantic grounds. Given the similarity with semantic bootstrapping of some of the machinery in the scenario briefly sketched here, it is appropriate to evaluate what advantages distributional bootstrapping might have.

One problematic aspect of semantic bootstrapping that has been discussed extensively in the literature has to do with the difficulty of identifying the meaning, and thus the semantic category, of unknown words, especially in the case of verbs (see for example, Gillette et al., 1999; Gleitman, 1990). The problem is not that the meaning cannot be recovered, clearly it can since children learn word meanings, but the argument is that, especially for verbs, structural information about the carrier utterance is necessary to help focus the learner on the relevant aspects of the world, and the related concepts that the word refers to. But on a semantic bootstrapping account, that structural information would not yet be available, as that is precisely what is hypothesized to be ultimately deduced once the category of the word is identified. These arguments have led many scholars to believe that forming an initial semantic classification as a bootstrap might not be possible. Another problem for approaches that take semantic categories as an initial categorization basis is that the links between semantic and grammatical categories are not one-to-one, but many-to-many. One aspect of the problem is that there are words for which the semantic antecedent conditions (action → verb, object → noun) do not apply. For example, 'know' is not an action and 'justice' is not a physical object; these facts produce a many-to-one mapping situation. In addition, as Maratsos and Chalkley (1980) discuss in detail, semantic-to-syntactic linking rules are subject to one-to-many mappings as well (i.e. one semantic type associated with several syntactic types). For example, 'action' and 'noisy' are not verbs, but, Maratsos and Chalkley argue, they would be mapped to the verb category based on their action-like semantics. Thus, even if mapping the to-be-categorized word to the correct semantic type could be reliably achieved, many incorrect grammatical assignments could be made based on purely semantic-to-grammatical linking rules.

Distributional bootstrapping does not run into these problems. Since the initial categorization is not dependent on accessing word meanings, the problems of finding a word's referent in the world do not come up. Nor are semantic factors directly involved in linking the bootstrap (distributional) categories to grammatical categories, so problems concerning the many-to-many mappings between semantic and syntactic categories are avoided as well. However, the question still remains as to what could guide the linking between distributional categories and grammatical categories: what are the mechanisms that could provide the appropriate grammatical labels for distributional categories? A possible mechanism is sketched out below with respect to nouns and verbs (since being able to identify nouns and verbs in order to determine verb-argument structure is arguably one of the more important tasks the learner has to accomplish). Before providing an account of how this might be achieved, it will be useful to lay out some background assumptions.

First, although it is questionable that verb referents can be identified by learners without access to sentential structural information (Gleitman, 1990), the referents of concrete nouns have been argued to be recoverable from observations of the circumstances in which they are used (Gillette et al., 1999; see also discussions in Fisher, Hall, Rakowitz, & Gleitman, 1994). If this is so, then the distributional category that contains nouns could be readily identified based on the concrete nouns that are its members. Note that using a semantic-to-syntactic generalization to label an independently derived category avoids the one-to-many mapping problem encountered when attempting to derive syntactic categories from semantic ones, since the semantic information is simply used to determine a general tendency of a group of words that is independently coherent.[9] Once the noun category (or categories) is labeled, identifying the distributional class which contains verbs becomes much more straightforward, and is perhaps achievable without recourse to additional semantic information: as it turns out, an effective procedure for the corpora analyzed here would be to apply the verb label to the distributional category (or categories) that is the largest category not labeled as noun. However, such a simplistic procedure for identifying the verb categories might not turn out to be viable cross-linguistically. A universally viable approach might label as verb the distributional category whose members satisfy one of a predetermined set of possible relationships with already identified nouns, specifically, the category whose members take the nouns as arguments. The information and representations involved in such a process are similar in some respects to the information involved in syntactic bootstrapping, but much less elaborated. A coarse representation of the argument structure of a set of utterances – the position of the nouns and a limited set of possible verb positions – would be sufficient to determine which distributionally defined word class is the verb category. Thus, initially words would be clustered distributionally, and the nouns would be labeled as such based on semantic correspondences. The location of nouns in utterances would then be used by syntactically constrained mechanisms to guide the labeling of the verb category. More sophisticated syntactic mechanisms could then be used to help determine the meanings of the verbs (Fisher et al., 1994; Gleitman, 1990; Landau & Gleitman, 1985).

The preceding paragraphs outline a method by which learners might plausibly link distributionally defined categories to a pre-given set of syntactic category labels. On this account, distributional information provides a bootstrap into a pre-existing (albeit under-specified) grammatical system. It is also conceivable, although the mechanisms are less clear, that properly constrained distributional information, perhaps in concert with prosodic information (e.g. Fisher & Tokura, 1996), could be used to induce higher order grammatical relations such as phrasal constituency and hierarchical structure that are pre-given in most bootstrapping accounts. The frequent frames investigated here capture some local information that might be relevant for positing higher order structural relationships (for example, argument structure for verbs with pronominal arguments: I__you, etc.). Thus, frames might be the seeds for growing higher order trees that would effectively make distributional categories syntactic. The question would then be whether the mechanisms required to motivate the construction of higher order generalizations, and to

---

[9] This combined use of distributional and semantic information is along the lines of Maratsos and Chalkley's (1980) proposal.

constrain the induction mechanisms to focus on the right kinds of distributional facts, would be theoretically discernable from the kinds of innate knowledge generally assumed under most bootstrapping accounts. At this point the success of such an approach is only speculative (but see Finch & Chater, 1994, for a model that attempts to construct grammatical phrases from distributional information).

A final comment regarding semantic bootstrapping: although the problems associated with reliably identifying a novel word's semantic class might render semantic bootstrapping intractable, the other problems discussed are not fatal. Pinker allows for categorizing words such as 'know' and 'justice' by *distributional* mechanisms that would register these words with already-classified verbs and nouns, respectively (what he calls "structure-dependent distributional learning"). Indeed, virtually *all* acquisition theories include distributional analyses at some point. What the present study and related research suggests is that categorization processes can operate on distributional information from the outset, thus making for a more economical theory, and avoiding some of the pitfalls inherent in semantically based categorization proposals.

4.4. Cross-linguistic applicability

All accounts of categorization are challenged to some degree by cross-linguistic differences. As an example, difficulties arise for semantically based proposals because cross-linguistic variability gives rise to different semantic-to-syntactic correspondences. For instance, all languages express the semantic types that English expresses with adjectives (e.g. PHYSICAL PROPERTY, HUMAN PROPENSITY, AGE, DIMENSION, etc.), but some languages express these types (or subsets of them) with nouns and verbs (Dixon, 1982). Hence, mechanisms that rely on a word's meaning to recover its syntactic type would have to independently determine what the language-specific correspondence rules are. Although this type of cross-linguistic variability would not adversely affect mechanisms that relied on distributional information, some typological differences might. For instance, the success of the frame-based analysis in English might be due to the fact that English has a relatively fixed word order. One might expect that a frame-based categorization mechanism would not be successful with languages that have freer word order. Two points about this issue are worth noting. First, in a language in which word order is relatively free – grammatical relations being marked by inflectional morphology – it may turn out that there is nevertheless enough consistency in word orders that informative frequent frames would result. This is clearly an empirical question, however data cited by Slobin and Bever (1982) concerning canonical word order in free word order languages (e.g. Turkish) show that both children and adults have preferred orders for sequencing nouns and verbs. Perhaps these canonical patterns will turn out to yield informative frequent frames. Second, even if this turns out not to be the case, the fundamental notion behind frequent frames might nevertheless be relevant for categorization in languages in which grammatical relations correlate more with morphology than with word adjacency. The fundamental notion is that a relatively local context *defined by frequently co-occurring units* can reveal a target word's category. In the procedures explored here, the units were words and the frame contexts were defined by words that frequently co-occur. In other languages, a failure to find frequent word frames

could trigger an analysis of co-occurrence patterns at a different level of granularity, for example, at the level of sub-lexical morphemes. The frequently co-occurring units in these languages are likely to be the inflectional morphemes which are limited in number and are extremely frequent. The details of how the morphological patterns could then be used is also an empirical question: would frequent frames made out of bound morphemes be informative, or would some other type of distributional analysis be better? Further research into typologically different languages is necessary to determine the practical universal applicability of the frame-based approach. But with straightforward modifications of the type just described, the approach is amenable, at least in principle, to categorization in typologically different languages.[10]

4.5. Frequent frames in acquisition

This paper discusses information inherent in child directed speech, given empirically motivated sensitivities and computational abilities on the part of the learner. Thus, these findings are an important first step in positing, on empirical and logical grounds, a specific type of information that could be useful in early acquisition. Additional studies will be needed in order to determine whether children actually make use of frame-like information in categorizing words. One way to do this is by examining children's productions, and examining whether an important factor influencing children's first production of a word is whether it previously occurred in frequent frames in the child's input. The success of frequent frames at predicting children's productive vocabulary would then have to be compared to other influences, like simple lexical frequency in the input. In addition, further analysis as to the difference in the frequent frames across corpora could be carried out to see if having different frequent frames in their input has consequences in children's productions. While such analyses are possible, a danger is that the relatively small samples and sample density in these corpora make these kinds of correlations difficult to detect (Tomasello, 2002). As new corpora become available that have greater density of both adult and child utterances, these questions can start to be addressed with analyses of this type (Lieven, Behrens, & Tomasello, 2001). On the other hand, more direct evidence is obtainable through controlled laboratory experimentation. Experiments like the one carried out by Mintz (2002) on adults could be carried out on young children, to investigate whether they, too, form categories based on words' occurrences within frames.

---

[10] Of course, the possibility of using bound morphemes to categorize words might motivate one to question whether morphology would be a useful source of information in a language like English, as well. If the ideas behind this approach are correct, then the categorization mechanism should be flexible enough to seek out the appropriate level, or levels, of analysis for the language at hand. Here the idea that the context itself is based on co-occurrence patterns is crucial, because in satisfying the co-occurrence requirement on the context elements themselves, the mechanism will settle on the contexts that are informative for that language. The search space of possibilities could be quite limited – words and sub-lexical morphemes being the most obvious candidates – and the cognitive/perceptual mechanisms of the type described by Gómez (2002) would be sufficient for discovering the appropriate environments. On this account, categorization by frequent word frames would be a specific case of a more general frame-based distributional mechanism that first discovers which of a limited set of local frame contexts are appropriate for the language at hand, and then categorizes words based on these contexts.

4.6. Summary and conclusion

In summary, frequent frames have been shown here to be an extremely effective and efficient source of information for categorizing words in children's input. Thus, as with earlier studies (Cartwright & Brent, 1997; Mintz et al., 2002; Redington et al., 1998), these findings demonstrate that the theoretical problems with distributional approaches turn out not to be problematic when actual corpora are analyzed. Moreover, the present experiments showed that by examining a relatively small number of frame contexts, a relatively large number of words were accurately categorized. It was argued that the condition on joint co-occurrence between the framing elements themselves provides a potentially more informative context than the bigram contexts investigated in prior studies, and also provides a cross-linguistically viable framework for focusing on the level of analysis (e.g. sub-lexical morpheme, word) that might be most informative for a particular language. Moreover, evidence from other studies suggests that the representations and processes involved in making use of frame-like information are plausibly within the scope of young children's capacities, as is the information needed to link distributional categories to syntactic ones. Thus, although the present approach does not yet include an explicit model of category acquisition, the crucial components seem well supported empirically. Overall, the sum of these facts and findings offers evidence in support of frame-like units as a basis for children's initial categorization of words, and suggests that further investigation of frequent frames will be a fruitful approach to advance our understanding of children's early grammatical knowledge.

Acknowledgements

References

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

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

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

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

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

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology*, *37*, 739–748.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Dixon, R. M. W. (1982). *Where have all the adjectives gone? and other essays in semantics and syntax*. Berlin: Mouton.

Finch, S. P., & Chater, N. (1994). Distributional bootstrapping: from word class to proto-sentences. *Proceedings of the 16th annual meeting of the Cognitive Science Society of America*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Fisher, C., Gleitman, H., & Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, *23*, 331–392.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: syntactic and conceptual constraints on vocabulary growth. In L. Gleitman, & B. Landau (Eds.), *The acquisition of the lexicon*. Cambridge, MA: MIT Press.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: cross-linguistic evidence. *Child Development*, *67*, 3192–3218.

Gerken, L. A., Gomez, R. L., & Nurmsoo, E (1999, April). *The role of meaning and form in the formation of syntactic categories*. Paper presented at the Society for Research in Child Development, Albuquerque, NM.

Gerken, L., Landau, B., & Remez, R. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology*, *26*, 204–216.

Gillette, J., Gleitman, L., Gleitman, H., & Lederer, A. (1999). Human simulations of lexical acquisition. *Cognition*, *73*, 135–176.

Gleitman, L. R. (1990). The structural sources of verb meaning. *Language Acquisition*, *1*, 3–55.

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

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. L. Baker, & J. McCarthy (Eds.), *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

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

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, *78*, B53–B64.

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition*, *2*, 15–47.

Landau, B., & Gleitman, L. R. (1985). *Language and experience*. Cambridge, MA: Harvard University Press.

Lieven, E. V. M., Behrens, H., & Tomasello, M (2001, November). *Corpus-based studies of children's development of verb-argument structures*. Paper presented at the 26th annual Boston University Conference on Language Development, Boston, MA.

MacNamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

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

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, *19*, 498–550.

Naigles, L., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120.

Peña, M., Bonatti, L. L., Nespor, N., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*, 604–607.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 435–469.

Sachs, J. (1983). Talking about the there and then: the emergence of displaced reference in parent-child discourse. In K. E. Nelson (Ed.), (*4*). *Children's language*, Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Santelmann, L., & Jusczyk, P. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition*, *69*, 105–134.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, *72*, B11–B21.

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: a crosslinguistic study of word order and inflections. *Cognition*, *12*, 229–265.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology*, *72*, 580–588.

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

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

Tomasello, M (2002, July). *Issues in the developmental analysis of spontaneous speech corpora*. Paper presented at the IX International Congress for the Study of Child Language, Madison, WI.

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

Wilson, R., Gerken, L. A., & Nicol, J (2000, November). *Artificial grammar research extends to natural language: implicit learning of categories*. Paper presented at the annual meeting of the Psychonomics Society, New Orleans, LA.