Cognitive Psychology 66 (2013) 30–54

Contents lists available at SciVerse ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes

Patricia A. Reeder [a,*], Elissa L. Newport [b], Richard N. Aslin [a]

[a] Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA
[b] Department of Neurology, Georgetown University, Washington, DC 20007, USA

ARTICLE INFO

Article history:
Accepted 17 September 2012
Available online 23 October 2012

Keywords:
Statistical learning
Category learning
Form-class categories
Artificial grammar
Language acquisition

ABSTRACT

A fundamental component of language acquisition involves organizing words into grammatical categories. Previous literature has suggested a number of ways in which this categorization task might be accomplished. Here we ask whether the patterning of the words in a corpus of linguistic input (*distributional information*) is sufficient, along with a small set of learning biases, to extract these underlying structural categories. In a series of experiments, we show that learners can acquire linguistic form-classes, generalizing from instances of the distributional contexts of individual words in the exposure set to the full range of contexts for all the words in the set. Crucially, we explore how several specific distributional variables enable learners to form a category of lexical items and generalize to novel words, yet also allow for exceptions that maintain lexical specificity. We suggest that learners are sensitive to the *contexts* of individual words, the *overlaps* among contexts across words, the non-overlap of contexts (or *systematic gaps* in information), and the *size* of the exposure set. We also ask how learners determine the category membership of a new word for which there is very sparse contextual information. We find that, when there are strong category cues and robust category learning of other words, adults readily generalize the distributional properties of the learned category to a new word that shares just one context with the other category members. However, as the distributional cues regarding the category become sparser and contain more consistent gaps, learners show more conservatism in generalizing distributional properties to the novel word. Taken together, these results show that learners are highly systematic

* Corresponding author. Address: Department of Brain & Cognitive Sciences, Meliora Hall, RC 270268, University of Rochester, Rochester, NY 14627, USA. Fax: +1 585 442 9216.
E-mail addresses: preeder@bcs.rochester.edu (P.A. Reeder), eln10@georgetown.edu (E.L. Newport), aslin@cvs.rochester.edu (R.N. Aslin).

0010-0285/$ - see front matter
http://dx.doi.org/10.1016/j.cogpsych.2012.09.001

in their use of the distributional properties of the input corpus, using them in a principled way to determine when to generalize and when to preserve lexical specificity.

1. Introduction

The ability to categorize is a powerful mechanism that learners employ to represent and interact with their environment. Categories compress information, thereby reducing demands on memory, and they allow for rapid generalizations. There are many fewer categories than exemplars, and if a particular exemplar is a member of a category, it inherits the defining properties of category membership. Often these defining properties are based on perceptual similarity (things that are green), semantic relations (things that float), or functional roles (things that can be sat upon). In the domain of language, however, there is a very loose relationship between perceptual, semantic, or functional properties and grammatical categories. A noun that serves as the subject of a sentence does not always sound like other subjects, express uniform semantics, or even play the same role in sentences that convey the same meaning (e.g., The *frog* ate the bug vs. The *bug* was eaten by the frog).

How, then, do naïve learners master the assignment of exemplars to grammatical categories in natural language? This is a crucial first step in language acquisition, since sentences of languages are organized in terms of grammatical form-classes (such as noun, verb, and adjective). Language learners must determine when they should treat words as a category (thus generalizing from properties of experienced words to novel words) and when they should treat words separately, as lexically idiosyncratic (with no generalization from properties of experienced words to novel words). Importantly, words of both types do in fact occur in natural languages. This process of organizing words into categories, and the selective generalization of patterns from experienced word combinations to novel ones, account for important aspects of the expansion of linguistic knowledge in the early stages of language acquisition. As highlighted above, linguistic categories are rarely defined on the basis of perceptual similarity; assignment of words to most grammatical categories is independent of the surface features of its members.

There are a number of additional complicating factors that make the acquisition of grammatical categories different from non-linguistic categorization. We hear individual words in a limited number of specific contexts. However, the rules that languages are built on involve patterns defined over *categories* of words, not the individual words themselves. Additionally, language input is serially presented – we hear words in their various sentence contexts spread out over hours or days – so learners continually need to predict the proper contexts for words they have not yet heard in their full range of possible contexts. Learners never see the entire input corpus, so they must figure out the proper contexts for new words, keeping in mind that sometimes there are lexically specific restrictions on words (such as *give* versus *donate*: despite having similar meanings, Joe can *give David a book*, but Joe cannot **donate David a book*). In acquiring grammatical categories, then, the learner must tease apart lexically specific restrictions and small-sample omissions from the corpus, asking whether contexts are absent by accident or because they are ungrammatical.

This question is particularly difficult to resolve when a new item is encountered in a single context and therefore only minimally overlaps with previously encountered words. For example, consider hearing the sentence: *I remembered to nerk yesterday*. Should one generalize from this context to other contexts that are grammatical for the category 'verb', such as *She will make him nerk tomorrow*, or *I saw the cat nerk earlier*?

Despite the difficulty of this problem, learners are able to determine how to use a new word even when there is very sparse information regarding its acceptable contexts. A number of hypotheses have been considered to explain this (Gleitman & Wanner, 1982). One hypothesis regarding how learners solve the problem of categorization is that linguistic categories (though not their lexical instantiations) are innately specified prior to experiencing any linguistic input, with the assignment of tokens to categories accomplished with minimal exposure (e.g., Chomsky, 1965; McNeill, 1966). However,

assuming a universal innate set of syntactic categories does not resolve the problem: some languages lack certain categories (such as the distinction between adjectives and verbs) or have multiple subclasses particular to that language (such as noun gender or verb subcategories). In addition, having innate abstract categories does not solve the massive problem of how the lexical items are assigned to these categories.

A second possibility is that the categories are formed around a semantic definition and extended via semantic bootstrapping (e.g., Grimshaw, 1981; Pinker, 1984, 1987). This hypothesis suggests that children associate semantic properties with syntactic classes, either via innate knowledge about this mapping (e.g., Brown, 1957; Grimshaw, 1981; Pinker, 1984, 1987), or through discovery of this mapping via the input (e.g., Bates & MacWhinney, 1979, 1982; Bowerman, 1973; Macnamara, 1972; Schlesinger, 1974). This hypothesis also has some significant problems. Similar to the strong nativist view of syntactic categories, an innatist approach to semantic bootstrapping assumes that there is a universal set of part-of-speech categories, which is not true for all syntactic categories. In addition, some classes of words are almost entirely semantically arbitrary, like gender classes (Maratsos & Chalkley, 1980), yet children are still readily able to acquire these categories. Furthermore, semantic bootstrapping requires a referential completeness assumption that is not upheld in natural language. Semantic features do not neatly match syntactic categories; yet, despite this lack of fit, there is little evidence that children miscategorize words based on their semantic properties (Gordon, 1985; Maratsos & Chalkley, 1980).

While it is likely that innate learning biases and semantic or phonological sources of evidence make important contributions to the task of linguistic categorization, it is also clear that grammatical categories must eventually be represented in terms of the syntactic contexts that are allowable for the words in a category. This important role for context suggests a third hypothesis for how learners might solve the problem of categorization: they exploit *distributional* information in the input to discover the category structure of natural languages (e.g., Bloomfield, 1933; Braine, 1987; Cartwright & Brent, 1997; Finch & Chater, 1992, 1994; Fries, 1952; Harris, 1951, 1954; Maratsos & Chalkley, 1980; Mintz, 2002, 2003; Mintz, Newport, & Bever, 1995, 2002; Redington, Chater, & Finch, 1998). The "distributional learning" hypothesis stems from the idea that learners could group words together into categories when those words occur in the same linguistic environments (e.g., Bloomfield, 1933), thus utilizing the same type of information that linguists use to find grammatical categories in a language (Harris, 1951, 1954). Given infinite time, input, and memory resources, a language learner could use such methods to compute the similarities among words in their linguistic contexts and determine whether missing contexts in the input signal an accidental gap (because that utterance has not yet been heard) or a meaningful gap (because it is part of the category structure). There are a number of different types of distributional information correlated with syntactic categories that could be exploited in this manner: for example, in English, words that take /-ed/ as a suffix also usually take /-s/ as a possible suffix and are in the category *verb* (Maratsos & Chalkley, 1980). Discovering these patterns between properties of word roots (e.g., /-ed/ and /-s/ suffixing) might be an important part of the learning process. Indeed, computational analyses using very large linguistic corpora show some success in solving the categorization problem via distributional analyses alone (e.g., Cartwright & Brent, 1997; Finch & Chater, 1992, 1994; Mintz, 2003; Mintz et al., 1995, 2002; Redington et al., 1998).

However, distributional learning has often been thought to be insufficiently powerful to play a major role in the category acquisition process. Human learners never see an entire input corpus, and to perform a distributional analysis they must compute statistics over noisy, highly variable and serially presented input. Given the information processing limitations of young children and the complexity of the computational processes that would be entailed, this hypothesis has often been viewed as implausible. However, there is a wealth of recent evidence that human infants and adults *can* learn other aspects of language based on distributional evidence. But could a distributional learning mechanism lead learners to know which distributional contexts are the relevant ones for grammatical categorization? Pinker (1984, 1987), for example, suggested that a distributional learning mechanism must work in tandem with semantic information; otherwise children would be unable to resolve ambiguous input such as in the sentences: *Jim could fish; Jim likes fish; Jim eats fish; Jim eats beef; Jim eats quietly.* If a learner were to have access to this input and only tallied word co-occurrences, the learner would be

likely to generalize to the erroneous and ungrammatical *Jim could quietly, *Jim likes quietly and *Jim could beef.

Similarly, many have argued that, in order for a learner to successfully utilize distributional information for category acquisition, there must be multiple correlated cues to category structure in the input (e.g., Braine, 1966). In accord with this suggestion, a large number of artificial language learning studies have explored the utility of correlated non-distributional cues to enable category learning (for example, semantic cues: Braine et al., 1990; morphological cues: Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; phonological cues: Frigo & McDonald, 1998; Gerken, Gomez, & Nurmsoo, 1999; Gerken, Wilson, & Lewis, 2005; Monaghan, Chater, & Christiansen, 2005; Morgan, Shi, & Allopenna, 1996; Wilson, 2002; shared features: Gomez & Lakusta, 2004). The consensus interpretation of their results is that the formation of linguistic categories depends crucially on the presence of some perceptual property that links items within the category, such as Braine's (1987) "similarity relations" (see also Gomez & Gerken, 2000). Examples of correlated perceptual cues are the identity or repetition of elements in grammatical sequences (Gomez & Gerken, 1999), or – more commonly proposed – a phonological or semantic cue identifying words across different sentences as similar to one another (e.g., words ending in *–a* are feminine, or words referring to concrete objects are nouns).

These correlated-cues hypotheses, however, suffer from the following puzzle: grammatical categories in natural languages do not always have reliable phonological, morphological, or semantic cues (Gleitman, 1990; Maratsos & Chalkley, 1980). The absence of reliable correlated cues suggests that learners must acquire such categories at least in part by utilizing the distributional cues to the linguistic contexts in which words occur. (For example, Maratsos and Chalkley (1980) pointed out that the existence of semantically arbitrary grammatical categories necessitates some form of distributional analysis.) Furthermore, when the semantic and distributional properties of a word conflict, it is usually the distributional information that determines the syntactic class of the word (Braine, 1987; Gordon, 1985).

As mentioned above, a number of investigators have demonstrated that computational models utilizing clustering algorithms over co-occurrence statistics can successfully acquire elementary form-class categories in natural language corpora (e.g., Cartwright & Brent, 1997; Finch & Chater, 1992, 1994; Mintz, 2003; Mintz et al., 1995, 2002; Redington et al., 1998). These models exploit purely distributional information in the input, highlighting the potential importance of such a strategy during child language acquisition. However, the details of how a distributional learning mechanism actually operates in natural language acquisition has been difficult to ascertain; many distributional cues to category structure in natural languages are correlated with other sources of information (e.g., semantic: Pinker, 1984; or phonological: Farmer, Christiansen, & Monaghan, 2006; Kelly, 1992). This makes it difficult to isolate distributional cues in studies of natural language to determine their contribution to linguistic category learning, unconfounded by these other cues.

Fortunately, artificial language learning paradigms offer the ability to test how learners utilize distributional information, by permitting precise experimental control over the various properties of the input and then testing to find the circumstances under which learners acquire categories. A first step in using miniature languages to study categorization was by Smith (1966), who showed that learners were quite capable of learning a simple language where there are only two categories of letters (α and β) and one rule that requires α words to be followed by β words. Participants saw some of the possible strings of the language and were then asked to do written recall of as many strings as possible. The results showed that participants recalled both the presented strings as well as "intrusions" (legal strings according to the pairing rule of the language that were not presented during exposure). The recall of grammatical intrusions is evidence of category-level generalizations, where categories are defined by positional information only (since the co-occurrence statistics between the two categories are distributionally uninformative). More recently, both Mintz (2002) and Gerken et al. (2005) have used artificial grammar learning paradigms with many correlated distributional cues to show that adults and infants can learn a simple version of a grammatical category cued only by distributional information.

However, none of these earlier studies articulated the principles governing the use of distributional variables that enable learners to solve the problem of category learning. In the present series of

experiments, we introduce a framework for describing the *structure* of the distributional information available to the learner. We focus on one part of the categorization problem: namely, how does the learner cope with incomplete evidence about the allowable contexts for particular words?[1] We then ask what type of distributional information will lead learners to behave as if a set of words is a single category, and what type will lead learners to restrict generalization and treat words as lexical exceptions. If learners demonstrate generalization from experienced words and their contexts to the full range of contexts for all words in the target set, they will have demonstrated formation of a category. If they restrict generalization to specific words, they will have demonstrated that they have stored particular contexts as being lexically specific. Overall, we view such behavior as probabilistic, rather than as sharply divided between category-general versus lexically specific representations. As we will see, human learners appear to weigh distributional information carefully and probabilistically, tending to generalize or restrict generalization as they learn and obtain more evidence, depending on the precise structure of the information provided.

Our series of experiments begins by outlining the distributional cues in the input that we hypothesize learners could use to form categories, without correlated perceptual or semantic cues. We then demonstrate that these cues alone can lead to successful learning of linguistic categories in an artificial language learning paradigm. In a series of four experiments (Experiments 1–4), we manipulate these distributional variables, showing that modulating these variables does indeed shift learners' tendency to generalize. The main distributional variables of interest are: the *number* of linguistic contexts in which each word in the input set occurs, the *density* or proportion of these contexts present in the input, and the degree of *overlap* of contexts across words. In addition, we investigate the importance of the *frequency* of these cues (or size of the input corpus). If learners operate in a principled way when using the statistics of their exposure corpus, then infrequent and non-systematic omissions in the input should still result in generalization to the appropriate category; the low frequency and non-systematic character of such omissions suggest that those contexts are accidentally omitted from the exposure corpus. On the other hand, systematic and recurring gaps should lead learners to increase their certainty that the gaps are meaningful. In this situation, where there is frequently recurring non-overlap among contexts, generalization should decline.

To ask whether human learners can exploit distributional information in such systematic ways, Experiments 1–4 vary the density of contexts in the input, the overlap of contexts across words, and the number of contexts in the input in order to assess the effects of these variables on learners' willingness to generalize novel words to a potential category. In Experiment 5, we ask how, under these same circumstances of varying category strength, learners extend the target category to the special case of a novel word for which they have only minimal context information. This last experiment thus asks if there is a point in category learning where hearing only one context for a novel word is enough to obtain full category privileges for that word, or whether every novel word must be heard in a number of overlapping contexts in order to be treated as a member of the category.

All of the experiments reported here employed adults as participants. Although there may be differences in the principles that guide learning by adults, who already possess a rich linguistic system, and infants who are just acquiring their native language, we believe that the relative ease of exposure and test among adults justifies an exploration of their ability to acquire a simple artificial language in the lab. Moreover, adult performance on many learning tasks is sufficiently reliable that subtle comparisons are possible as the distributional properties of the language are manipulated. Of course, in future work, it will be important to extend these studies, if possible, to young children and infants who are acquiring their first language.

[1] There are, of course, other important components of the categorization problem that we and other researchers have explored elsewhere. In this work, we focus on acquisition of a single category surrounded by context words, and we ask which contexts should be assigned to particular category members based on distributional analyses alone. But we could also ask how learners categorize the context cues themselves (e.g., Mintz, 2002), or how learners figure out that the *noun* category belongs with one set of contexts, but the *verb* category does *not* belong with those contexts. The results reported here have implications for these other aspects of categorization, but our work on subcategorization (e.g., Reeder, Newport, & Aslin, 2009, Experiment 5) more directly speaks to these other questions.

2. Experiment 1

In Experiment 1, learners were exposed to a fairly dense sampling of a language generated by an artificial grammar, with a small number of exemplars withheld from the total set of possible grammatical strings. We did this by presenting two-thirds of the possible sentence types in the exposure set and withholding one-third for later test. The goal of this experiment was to give learners a rich input set (while still allowing for tests of novel strings), in order to establish a baseline level of performance on category formation in our artificial language learning paradigm.

2.1. Method

2.1.1. Participants

A total of 19 monolingual native English-speaking students at the University of Rochester participated in Experiment 1 and were paid for their participation. Two subjects were excluded from the analysis for not complying with experimental instructions. Participants were randomly assigned to one of two languages: eight subjects were assigned to language 1, and nine subjects were assigned to language 2. All of the participants, in this and all of the remaining experiments, were naïve to each experiment and were not allowed to participate in any other categorization study.

2.1.2. Stimulus materials

All sentences in the language were constructed from a grammar of the form (Q)AXB(R), where Q, A, X, B, and R were categories of nonsense words. X was the target category under study, while A and B were the context categories that formed the distributional cues to the X category. Q and R served as optional categories that made sentences of the language vary in length from 3 to 5 words (thus, sentences could be of the form AXB, QAXB, AXBR, or QAXBR). The optional status of Q and R categories ensured that the words of the language observed regular patterning in terms of relative order and co-occurrence but did not have fixed positions in the sentences. As in natural languages, then – but in contrast with several other artificial language experiments on this topic – fixed or absolute position information (such as 'initial position' or 'second word in the string') could not be used as an informative cue to category membership.

Two versions of the language (languages 1 and 2) were created to insure that the mapping of words to categories was not inadvertently biased to aid the learner with the categorization task. The same 13 words were used in each of the two languages (see Table 1). These words were read in isolation by a native English-speaking female and were each recorded with both a non-terminal and terminal list intonation. Words were adjusted in Praat (Boersma, 2001) so that the pitch, volume, and duration

Table 1
Word-to-category assignments for languages 1 and 2.

Q	A	X	B	R
Language 1				
spad (/spæd/)	flairb (/fleɪrb/)	tomber (/tɑmbɚ/)	fluggit (/flʊgɪt/)	gentif (/ʤɛntɪf/)
klidum (/klaɪdʌm/)	daffin (/dæfɪn/)	zub (/zʌb/)	mawg (/mɔg/)	frag (/fræg/)
	glim (/glɪm/)	lapal (/lʌpɔl/)	bleggin (/blɛgɪn/)	
Language 2				
frag (/fræg/)	gentif (/ʤɛntɪf/)	spad (/spæd/)	zub (/zʌb/)	lapal (/lʌpɔl/)
daffin (/dæfɪn/)	mawg (/mɔg/)	fluggit (/flʊgɪt/)	tomber (/tɑmbɚ/)	flairb (/fleɪrb/)
	klidum (/klaɪdʌm/)	bleggin (/blɛgɪn/)	glim (/glɪm/)	

of syllables were qualitatively consistent. Languages 1 and 2 differed only on the assignment of words to categories. In both cases, we ensured that each category had a relatively balanced number of one- and two-syllable words, and no category was strongly imbalanced in terms of phonological properties of the category members (onset, offset, and number of syllables) (see Table 1). The words were not mapped to any referential world, and they had no meanings associated with them. Sentences were constructed by splicing together words into sequences using Sound Studio, with 50 ms of silence between each word and selecting the word token with a terminal intonation contour as the final word in the sentence.

Focusing on just the AXB portion of the grammar, there were 27 possible word strings in the language (3 A-words × 3 X-words × 3 B-words). Of the 27 basic AXB sentence types, 18 were presented and 9 were withheld (see Table 2). Within these 18 AXB types, AXB, QAXB, AXBR and QAXBR strings were created by varying whether the 2 Q- and 2 R-words were present or absent. Q- and R-words were added such that each X-word was seen with all Q- and R-words. Bigram frequencies of Q–A and B–R pairs were controlled such that the flanker words could not be an informative cue to the sentence type. With the use of the optional flanker Q- and R-words, the 18 AXB sentence types used for exposure generated a total of 72 different (Q)AXB(R) sentences (18 of each of the four sentence types AXB, QAXB, AXBR, and QAXBR). Each possible A_B frame was also heard equally often during the exposure phase; learners heard each of the nine different frames 32 times during exposure. (See Supplementary materials for exact frequencies of all adjacent and nonadjacent bigrams for each experiment.)

2.1.3. Procedure

Participants were seated in a sound-attenuated booth and were informed that they would be exposed to some sentences from a new language that they had never heard before. They were told to just listen to the sentences and to pay attention to them because they would be tested on their memory of them in the second portion of the experiment. The exposure set of 72 sentences was presented four times (288 sentences) via headphones, forming 20 min of exposure to the language. Exposure strings were presented in pseudo-random order with 1.5 s of silence between sentences. Importantly, the 18 AXB sentence types used during exposure included each X-word in the presence of every A-word and every B-word and two-thirds of the possible AXB sentences. Thus, the exposure set for this language is *dense* (covering a high proportion of the overall language space) and has complete *overlap* of the possible A_ and _B contexts among the various X-words within the target category (see Fig. 1).

After exposure, participants were presented with a pseudo-random ordering of individual test strings and were asked to rate each test string on a scale of 1–5 based on whether or not they thought the test sentence came from the language they heard during training: 1 meant that the string sounded like it definitely did *not* come from the language; 2 meant the string *might not* have come from the language; 3 meant the string *may or may not* have come from the language; 4 meant the string *might* have come from the language; 5 meant the string *definitely* came from the language. If subjects asked what it meant to "come from the language," they were instructed to go with their gut reaction as to whether the string might have been something a native speaker of the language would have said when

Table 2
Possible AXB strings in Experiments 1–4. Strings presented in Experiment 1 are denoted ★; strings presented in Experiment 2 are denoted ✠; strings presented in Experiments 3 and 4 are denoted ○.

$A_1\ X_1\ B_1$ ★	$A_1\ X_2\ B_1$	$A_1\ X_3\ B_1$ ★ ✠ ○
$A_1\ X_1\ B_2$	$A_1\ X_2\ B_2$ ★ ✠	$A_1\ X_3\ B_2$ ★ ○
$A_1\ X_1\ B_3$ ★ ✠ ○	$A_1\ X_2\ B_3$ ★	$A_1\ X_3\ B_3$
$A_2\ X_1\ B_1$	$A_2\ X_2\ B_1$ ★ ✠ ○	$A_2\ X_3\ B_1$ ★
$A_2\ X_1\ B_2$ ★ ✠ ○	$A_2\ X_2\ B_2$ ★	$A_2\ X_3\ B_2$
$A_2\ X_1\ B_3$ ★ ○	$A_2\ X_2\ B_3$	$A_2\ X_3\ B_3$ ★ ✠
$A_3\ X_1\ B_1$ ★ ✠	$A_3\ X_2\ B_1$ ★ ○	$A_3\ X_3\ B_1$
$A_3\ X_1\ B_2$ ★	$A_3\ X_2\ B_2$	$A_3\ X_3\ B_2$ ★ ✠ ○
$A_3\ X_1\ B_3$	$A_3\ X_2\ B_3$ ★ ✠ ○	$A_3\ X_3\ B_3$ ★

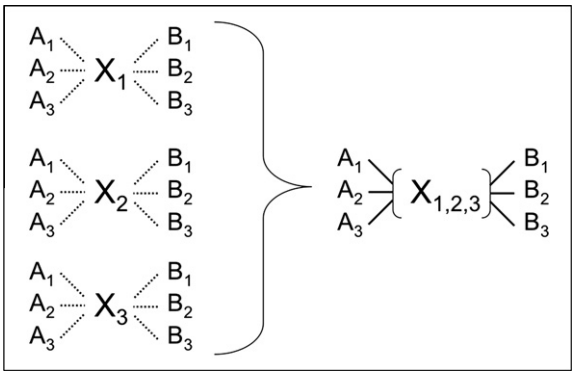

Fig. 1. Pictorial depiction of the learning task in Experiments 1 and 2. Learners hear each X-word with every A-word and with every B-word (though not every A_B context), such that there is completely overlapping contextual information across the X-words.

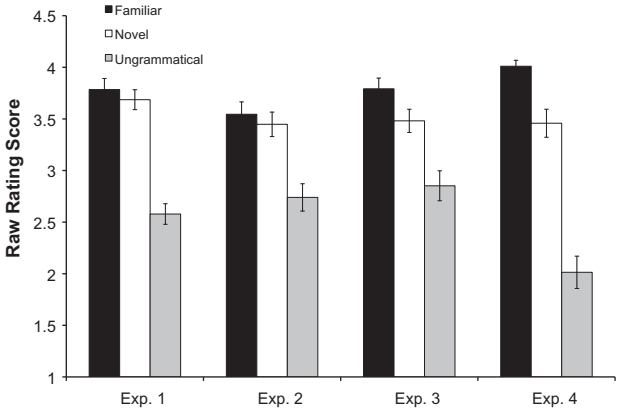

Fig. 2. Rating score results from Experiments 1–4, comparing familiar, grammatical novel, and ungrammatical test strings.

following the rules of the language's grammar. All test strings were 3-word sentences and consisted of three types: *grammatical familiar* (9 AXB strings presented during training), *grammatical novel* (9 AXB strings withheld during training), and *ungrammatical* (strings of the form AXA or BXB).[2] Although ungrammatical test strings contained repeated categories (such as **AXA**), no test string had repeated word tokens. The nine familiar and nine grammatical novel strings were randomized with nine ungrammatical strings during the first half of the test, and then the same nine familiar and nine grammatical novel strings were presented again in random order along with nine different ungrammatical strings during the second half of the test. (See the Supplementary materials for all test items used in the reported experiments.)

[2] Experiments 1–4 were later each piloted with two additional types of ungrammatical items (AAB and ABB) to make sure that participants were not simply using the A_B frame in order to identify the ungrammatical test items. AAB and ABB strings, like the AXA and BXB strings, had no repeated word tokens. These additional ungrammatical test items were not rated significantly differently than the AXA and BXB items ($p > 0.05$ for each experiment), and confirm that performance during test is not solely based on learning the positional information contained in the A_B frame.

2.2. Results

A repeated measures ANOVA was conducted with condition (familiar, novel, and ungrammatical) as the within-subjects factor and language (1 or 2) as the between-subjects factor. There were no significant effects of language ($F < 1$). The mean rating of grammatical familiar strings was 3.78 ($SE = 0.11$), the mean rating of grammatical novel strings was 3.69 ($SE = 0.10$), and the mean rating of ungrammatical strings was 2.58 ($SE = 0.10$) (see Fig. 2). There was no significant difference between ratings of grammatical novel strings and grammatical familiar strings ($F(1,15) = 1.85$, $p = 0.19$). However, these items were rated significantly higher than ungrammatical strings ($F(1,15) = 51.992$, $p < 0.001$).

Because individual subjects may utilize the rating scale in different ways, raw ratings scores were converted into z-scores in order to standardize ratings across subjects, using the formula $z_{ij} = \frac{rating_{ij} - \mu_j}{SE_j}$, where z_{ij} is the z-score for the ith test item rated by subject j based on the raw rating of item i by subject j. Thus, a score below zero indicates that an item was rated lower than a subject's average rating, and a score above zero indicates that an item was rated higher than a subject's average rating. Using the z-scores, another repeated measures ANOVA was conducted with condition (familiar, novel, and ungrammatical) as the within-subjects factor and language (1 or 2) as the between-subjects factor. Overall effects were the same as when computed over raw ratings (no significant difference between languages 1 and 2: $F < 1$; no significant difference between grammatical novel and grammatical familiar strings: $F(1,15) = 1.792$, $p = 0.2$; significant difference between grammatical novel and ungrammatical strings: $F(1,15) = 70.630$, $p < 0.001$).

2.3. Discussion

In Experiment 1, learners were exposed to a dense sampling of the language space, with two thirds of the possible AXB contexts presented and with all of the words in the target category appearing in many highly overlapping A_ and _B contexts. Under these conditions, learners fully generalized, treating the X-words as belonging to a category of words that all had the same set of permissible linguistic contexts. They did not discriminate between the presented and the withheld AXB's, both of which were rated as highly grammatical and strongly preferred to ungrammatical sentences in which one word in the string occurred in an ungrammatical position. These findings show, that when the exposure set densely samples the language space and words within a category appear in highly overlapping contexts, learners will successfully form a linguistic category. This occurs without any perceptual or semantic cues to indicate that the words form a single category, and with no negative evidence about which strings are illegal.

In Experiments 2–4, we investigate the degree to which category generalization is affected by manipulating the distributional variables of density and overlap in learning a single X category. Furthermore, we explore whether learners use distributional information to avoid overgeneralization when deciding if particular contexts are lexically specific.

3. Experiment 2

In Experiment 2, we kept the number of contexts presented for each X-word and the overlap among X-word contexts the same as in Experiment 1, but we substantially reduced the number of different A_B contexts that were presented during the exposure phase (see Table 2). We refer to this as *reducing the density* (or *increasing the sparseness*) of the contexts for X-words that are presented during learning.

3.1. Method

3.1.1. Participants

A total of 19 monolingual native English-speaking students at the University of Rochester were paid to participate in Experiment 2; three were excluded for not complying with experiment

instructions (2) or for equipment failure (1). This left 16 total subjects, with eight participants assigned to each of languages 1 and 2.

3.1.2. Stimulus materials and procedure

Strings were created in the same manner as in Experiment 1. However, out of the 27 possible AXB combinations, only nine were presented during exposure (see Table 2). Crucially, every X-word was still heard in combination with every A- and every B-word; therefore, as in Experiment 1, the exposure set had complete *overlap* of contexts across X-words (see Fig. 1). As in Experiment 1, each of the 9 AXB sentence types was presented with category flanker elements Q and R present or absent, producing 36 sentences in the exposure set (rather than the $18 \times 4 = 72$ sentences presented in Experiment 1). Each of the nine possible A_B frames was heard 16 times during the exposure phase.

The procedure was the same as in Experiment 1. The input corpus consisted of presenting the exposure set four times in pseudo-random order. Since the exposure set included fewer sentences than Experiment 1, total exposure time was about 10 min. The test was the same as in Experiment 1, except that the 18 grammatical novel test strings were counterbalanced such that half of the participants in each language were tested on one subset of nine of the withheld (grammatical novel) strings, and the rest of the participants were tested on the other nine grammatical novel strings.

3.2. Results and discussion

A repeated measures ANOVA was used to analyze the ratings, with condition (familiar, novel, ungrammatical) as the within-subjects factor, and language (1 or 2) and subtest (which counterbalanced set of novel items the subject received during test) as the between-subjects factors. As in Experiment 1, there was no difference between languages 1 and 2, ($F < 1$), nor was there any effect of subtest ($F < 1$) or interactions ($F < 1$ for all). The mean rating of grammatical familiar strings was 3.54 ($SE = 0.12$), the mean rating of grammatical novel strings was 3.47 ($SE = 0.12$), and the mean rating of ungrammatical strings was 2.74 ($SE = 0.14$). Grammatical novel strings were rated just as highly as grammatical familiar strings, and there was no significant difference between these two types of strings ($F(1,12) = 0.810$, $p > 0.3$). Ungrammatical strings were rated significantly lower than the grammatical strings ($F(1,12) = 19.022$, $p < 0.001$) (see Fig. 2).

As in Experiment 1, raw scores were transformed into *z*-scores, and another repeated measures ANOVA was conducted with condition as the within-subjects factor and language and subtest as the between-subjects factors. Once again, there was no difference between familiar and novel grammatical ratings ($F(1,12) = 0.693$, $p > 0.4$), but ratings of ungrammatical strings were significantly lower than grammatical strings ($F(1,12) = 21.842$, $p < 0.001$). None of the interactions were significant.

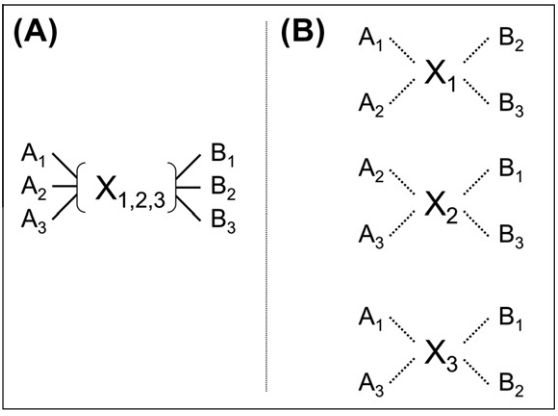

Fig. 3. Pictorial description of full overlap in the grammar space for the X-words in Experiment 2 (Fig 3A), compared to the partial overlap in Experiment 3 (Fig 3B).

These results show that learners' performance is qualitatively unchanged from Experiment 1, despite the change in density/sparseness while other properties of the distributional information were maintained – that is, despite the fact that the exposure was half as rich and also half as long. This outcome suggests that a change in the density/sparseness of the input does not alter generalization in category learning, so long as the input contains a systematic pattern of *overlap* in contexts among the members of the category. In Experiment 3, we ask what happens when the amount of overlap in the contexts of the X-words is reduced. In Experiment 4, we then ask what happens when the reduced or missing overlap is a consistent property of the input.

4. Experiment 3

In Experiment 3, as in Experiment 2, we presented only 9 of the possible 27 AXB combinations. Here, however, we presented a slightly different set of AXB combinations that reduced the overlap of contexts among members of X. This allows us to assess the importance of overlap in distributional information for category formation and generalization. In the present experiment, each of the three X-words occurred in all of the A and B contexts. However, individual X-words did not fully share all their contexts with one another (see Fig. 3): each X-word occurred with only 2 A-words and 2 B-words, out of the possible 3. The question addressed, then, is the degree to which learners will restrict their generalization across the category as a function of this reduction in overlap.

4.1. Method

4.1.1. Participants

A total of 24 monolingual native English-speaking students at the University of Rochester were paid to participate in Experiment 3; 12 were assigned to language 1, and 12 were assigned to language 2.

4.1.2. Stimulus materials and procedure

Strings were composed in the same way as Experiment 2, with only 9 of the 27 possible AXB combinations presented during the exposure phase. X_1 was heard in the context of A_1, A_2, B_1 and B_2, but not in the context of A_3 or B_3. X_2 was heard in the context of A_2, A_3, B_2 and B_3, but not A_1 or B_1. X_3 was heard in the context of A_1, A_3, B_1 and B_3, but not in the context of A_2 or B_2. Thus, the overlap among contexts is maintained over the X category as a whole, but individual words in X do not have the degree and type of overlap in distributional contexts that they did in Experiments 1 and 2 (where every X-word occurred with each A- and each B-word). As in Experiment 2, each of the 9 AXB sentence types was presented with category flanker elements Q and R present or absent, producing 36 sentences in the exposure set. Also, as in Experiment 2, the exposure set was presented 4 times and had a total exposure time of about 10 min. Each of the 9 A_B frames was heard 16 times during the course of exposure. Thus, the only difference between Experiments 2 and 3 was in the co-occurrence of individual X-words with the individual context A- and B-words, not in the A_B frame frequencies. The training and test procedures were otherwise the same as in Experiments 1 and 2.

4.2. Results and discussion

The mean rating of grammatical familiar strings was 3.79 ($SE = 0.1$), the mean rating of grammatical novel strings was 3.48 ($SE = 0.16$), and the mean rating of ungrammatical strings was 2.85 ($SE = 0.15$). A repeated measures ANOVA, with condition as the within-subjects factor and language as the between-subjects factor, revealed no difference between languages 1 and 2 ($F < 1$), but a significantly higher rating for grammatical strings than for ungrammatical strings ($F(1,22) = 40.691$, $p < 0.001$). In contrast to Experiments 1 and 2, however, the ANOVA revealed significant differences between grammatical familiar and grammatical novel strings ($F(1,22) = 18.981$, $p < 0.001$).

Raw scores were transformed into *z*-scores, and another repeated measures ANOVA was conducted. There was no effect of language ($F < 1$), but, again, grammatical novel strings were rated

significantly lower than grammatical familiar strings ($F(1,22) = 23.852$, $p < 0.001$) and significantly higher than ungrammatical strings ($F(1,22) = 56.230$, $p < 0.001$).

Because of the incomplete overlap imposed by this experimental design, the grammatical novel test items can be divided into multiple types according to bigram information: "heard 2 bigram" test strings, where the subject heard both the AX and XB bigrams during exposure (but not the entire AXB trigram); and "heard 1 bigram" test strings, where the subject heard only one of the AX or XB bigrams during exposure. There was no effect of language on these ratings, so the two languages were collapsed for this analysis. Paired samples *t*-tests revealed that the two types of grammatical novel test items were rated differently (heard 2 bigrams mean = 3.62, $SE = 0.12$; heard 1 bigram mean = 3.41, $SE = 0.12$; $t = 2.54$, $p = 0.018$). This string difference is subtle, and so is the rating difference it produces; but it suggests that learners are extremely sensitive to the details of the exposure corpus. In line with the overall result for this experiment, learners apparently utilize the pattern of specific contexts in which words do and do not occur to determine their likelihood of generalizing to novel contexts. This result might also point to the type of statistic that subjects are storing (bigram information) in order to acquire the categories of the language. We return to this possibility in Section 8.

Whereas Experiment 2 tested how subjects would respond to fewer contexts but full overlap of the context environment, Experiment 3 tested the effect of reducing the overlap in the exposure while keeping the amount of exposure the same as in Experiment 2 (see Fig. 3A as compared to Fig. 3B). Of course, as the size of an input corpus is reduced, some of the contexts that are possible for a particular word are likely not to occur, simply by chance. A naïve learner would not be sure whether such absences were chance omissions, or were reflections of the unacceptability of the non-occurring contexts. However, at some point along the sparseness and non-overlap dimensions, learners must stop concluding that X is a category and must acquire lexical restrictions or shift to word-by-word learning. The results of Experiment 3 give insight into the computational details of how this occurs by showing that, despite full coverage over lexical items, the incomplete *overlap* between words led to a slight decrease in generalization. At the same time, however, learners did continue by and large to generalize, showing a much higher rating for grammatical novel strings than for ungrammatical strings. These results suggest that learners take into account both the overlap and the non-overlap among items, modestly reducing their willingness to generalize when the data supporting generalization are less strong. In Experiment 4, we investigate how repeated exposure to these partially overlapping items influences the decrease in generalization that we witnessed in Experiment 3.

5. Experiment 4

One more variable that may impact generalization versus lexical distinctness is how often each type of context is presented (and therefore the frequency with which contextual gaps recur). If learners operate in a principled way when using the statistics of their input corpus, the prediction is that very high frequencies of sparse distributional information, with systematic and recurring gaps, should lead learners to increased certainty that the gaps are meaningful. This should lead learners to restrict generalization. Indeed, this is the result obtained in work by Wonnacott, Newport, and Tanenhaus (2008) in a miniature verb-argument structure-learning paradigm, as well as in work on concept acquisition by Xu and Tenenbaum (2007). In Experiment 4, we explored how an increase in the amount of exposure to the very same corpus used in Experiment 3 would affect categorization.

5.1. Method

5.1.1. Participants

A total of 17 monolingual native English-speaking students at the University of Rochester were paid to participate in Experiment 4. One was removed for failing to understand the testing directions, which left eight subjects assigned to each of languages 1 and 2.

5.1.2. Stimulus materials and procedure

The corpus was the same as in Experiment 3; however, exposure was tripled, by presenting the exposure set 12 times rather than 4. The exposure therefore lasted for approximately 30 min, which was a somewhat longer exposure than Experiment 1, but contained only 9 contexts, as in Experiments 2 and 3. The training and test procedures were the same as in Experiment 3.

5.2. Results and discussion

A repeated measures ANOVA, with condition (familiar, novel, ungrammatical) as the within-subjects factor and language (1 or 2) as the between-subjects factor, revealed no differences of language ($F < 1$). The mean grammatical familiar rating was 4.01 ($SE = 0.06$), the mean grammatical novel rating was 3.458 ($SE = 0.136$), and the mean ungrammatical rating was 2.014 ($SE = 0.157$). There were highly significant differences between all conditions. Novel grammatical strings were rated significantly lower than familiar strings ($F(1,14) = 19.40$, $p < 0.005$), and were also rated significantly higher than ungrammatical strings ($F(1,14) = 31.747$, $p < 0.001$) (see Fig. 2).

Raw scores were transformed into z-scores and another repeated measures ANOVA was conducted on the transformed familiar, novel, and ungrammatical mean ratings. Once again, there was no significant effect of language, but there were highly significant differences between all three within-subject conditions. Novel grammatical strings were rated significantly lower than familiar grammatical strings ($F(1,14) = 21.473$, $p < 0.001$) and significantly higher than ungrammatical strings ($F(1,14) = 38.919$, $p < 0.001$).

The results from Experiment 4 reveal that increased exposure to a corpus containing incomplete overlap reduces the likelihood that learners will generalize based on this input. Instead, learners are more likely to assume that these systematic gaps in the input are not accidental omissions, but instead they signal potential idiosyncratic behavior of individual lexical items. The increase in the difference between grammatical familiar and grammatical novel strings that occurs between Experiments 3 and 4 highlights the learner's sensitivity to these frequent and consistent gaps. This conservatism may be a component of the learner's strategy to avoid overgeneralization. Despite this reduced generalization in Experiment 4 (compared to Experiments 1–3), participants still judged novel grammatical strings as more familiar than ungrammatical strings, thereby documenting an initial generalization bias. We return to this conflict between over- and under-generalization in Section 8.

6. Discussion of Experiments 1–4

The first four experiments tested whether learners can acquire a category that is defined solely by distributional information, generalizing from exposure to some instances of the contexts of individual words (with some withheld) to the full range of contexts for all the individual words in the set. The results lend strong support to the hypothesis that learners can extract the category structure of an artificial language based solely on the distributional patterning of the words and their surrounding contexts. These results run counter to a large body of previous research claiming that linguistic categories in artificial language experiments cannot be formed on the basis of distributional contexts alone, and that additional information (such as phonological or semantic cues) is required for successful learning (e.g., Braine, 1987; Gomez & Gerken, 2000). The results of Experiments 1–4, however, show that such additional cues are not necessary for adults to induce a category from distributional contexts alone. We return to the question of why our experiments may be different from previous experiments in Section 8.

Looking just at difference scores between familiar and novel test strings and between familiar and ungrammatical test strings (see Fig. 4), it is clear that in Experiments 1 and 2, familiar and novel grammatical test strings are rated no differently from each other (the ratings difference between the two types is not different from zero). Experiments 1 and 2 were cases in which only the *number* of contexts differed (the sampling of the language space became sparser, but the overlap in contexts across words was not changed). But in Experiment 3, learners start to reduce their likelihood of generalizing when the overlap in contexts is reduced. This reduction in overlap leads learners to increase the difference in

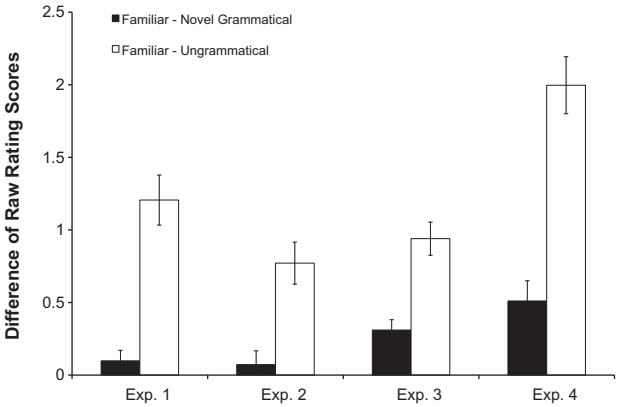

Fig. 4. Difference scores of raw ratings from Experiments 1–4 for familiar and novel grammatical AXB test strings (dark bars), and familiar and ungrammatical test strings (white bars). The difference between familiar and novel grammatical strings is negligible in Experiments 1 and 2, indicating that learners generalized to the novel contexts. In Experiment 3, and then especially in Experiment 4, this difference increases as a function of decreased overlap in contexts and increased exposure, respectively.

their ratings for familiar versus unfamiliar grammatical sentences. They restrict generalization even more sharply in Experiment 4, when the same reduced-overlap exposure corpus was repeated three times. In the limit, with much more extensive exposure to gaps created by reduced overlap, we should see an even larger increase in the difference between ratings of familiar and novel grammatical strings, such that learners would eventually rate the novel strings at the level of ungrammatical strings (a result that would suggest no generalization to those contexts whatsoever). Taken together, the manipulations implemented in Experiments 1–4 suggest that adult learners are quite sensitive to the distributional information in the input that signals whether to generalize across lexical items (indicating that gaps are likely to be accidental), or restrict generalization to lexically specific contexts. Participants in these experiments were able to skillfully balance a rich set of variables to aid them in this task – degree of overlap among category members, amount of input, consistency or systematicity of gaps and overlaps, and conflicts or consistency among cues.

These results highlight some types of information that learners might be encoding or computing during learning, and other types that they do not appear to be relying on. If learners were encoding the full set of exposure sentences, or the trigrams or quadrigrams (e.g., AXB, AXBR) and their frequencies of occurrence during exposure, they should discriminate between the familiar and novel grammatical sentences in all of the above experiments. In contrast, if they were only keeping track of simple word frequencies, they would fail in all experiments, since these frequencies were carefully controlled. It is clear that neither of these explanations can fully explain behavior across all four experiments.

Another possibility is that learners were simply responding on the basis of perceived surface similarity between familiar and novel items, and not based on forming a higher-level category structure.[3] One such explanation for our results might argue that learners were matching training and test items based on the similarity of adjacent or nonadjacent bigrams, without constructing an X category. A similar explanation might hypothesize that learners responded on the basis of familiar and unfamiliar A_B frames, without taking X into consideration. To rule out such possibilities, we carefully controlled multiple aspects of the input and test strings across all experiments. Moreover, though word order violations may be the reason why ungrammatical strings are always rated lower than grammatical strings, simple surface strategies such as these cannot explain our results with regards to our comparison of central interest: familiar vs. novel grammatical strings. First, we re-ran Experiments 1–4 each on naïve

[3] We thank Toby Mintz for helpful discussions of this alternative hypothesis.

participants and included two additional types of ungrammatical test strings: AAB and ABB (see footnote 2; none of the ungrammatical strings included repetitions of individual words). These ungrammatical items allowed us to examine whether learners were assessing A_B frames without attending to the X-word. We found that participants did not rate these new ungrammatical test strings higher than our old types of ungrammatical test strings, AXA and BXB ($p > 0.05$), and overall, the results of these experiments were qualitatively the same as those of Experiments 1–4 reported above.

Moreover, it is important to consider how A_B frames are distributed across our experiments. The crucial pattern of results from Experiments 1–4 is that the difference between familiar and novel grammatical items systematically increases across experiments, as the overlap among the contexts of X-words declines (Experiment 3) and the consistency of these gaps increases with extended exposure (Experiment 4). However, all possible A_B frames were heard equally often during the exposure phase of each experiment, so the pattern of generalization across experiments cannot be due to A_B frame learning. If learners relied solely on the A_B frames without encoding their relationships to the X-words, we would expect to see stable ratings of novel grammatical strings across the experiments. Instead, we see no relation between frequency of A_B frame exposure and generalization. For example, in Experiments 2 and 3, each A_B frame was heard exactly 16 times during exposure. However, in Experiment 2 there is no difference in ratings of familiar and novel grammatical strings, whereas in Experiment 3 there is a significant difference. What *did* change across these experiments was the overlap of contexts for each X-word: Experiment 2 had complete overlap of contexts, whereas Experiment 3 had only partial overlap. This shift in the overlap of shared contexts among individual X-words, not the familiarity of A_B frames, drives our effect across the four experiments. (See Supplementary materials for more details on frequencies of adjacent and nonadjacent bigrams.)

However, it is still unclear what type of information extracted by the learner best accounts for the difference in results across the four experiments. Our hypothesis is that participants are highly sensitive to the statistics in the input and conduct distributional analyses over multiple levels of input based on their current representation of the language's category structure, but we have not yet asked about the specific type of statistical information that learners are acquiring. For example, learners might extract local (adjacent) pair-wise (bigram) statistics over words to form categories, and then extract pair-wise statistics for new words with respect to these categories. Also unknown is how learners exploit these statistics, in conjunction with their current knowledge about the language, in order to decide whether to incorporate new words into existing category representations. One might imagine, for example, that when a category is strongly formed, new words that share some of the category's linguistic contexts will then inherit all of that category's other linguistic contexts, indicating that the category has crystallized and acts as a unit of representation. On the other hand, given the graded and probabilistic way in which generalization to novel strings operated in Experiments 1–4, we might find that extension to a new word will also show lexical specificity or a graded degree of generalization. The goal of our final experiment is to investigate these questions within the framework of our artificial language learning paradigm.

Experiments 1–4 showed how we can manipulate various aspects of the language landscape via certain distributional variables, all of which are based on *shared* contexts across words, in order to gain insight into the computational requirements for successful category learning. We now turn to exploring the variables involved in extending category knowledge to a new word presented in a *single* context. We investigate whether learners always maintain lexical specificity when they have very limited distributional information for a new word, or whether they show varying degrees of generalization to a new word, depending on the type and strength of the distributional information available for other words. In these experiments, learners face the same question for the new word as they did in Experiments 1–4 for more familiar words: are the unattested contexts for this new word absent by accident, or because they are ungrammatical?

7. Experiment 5

The goal of the remaining series of experiments is to assess whether learners will generalize the distributional properties of a learned category to a word that shares just one context with the other

Table 3
Possible AXB strings in Experiments 5. Strings presented in Experiment 5A are denoted ★; strings presented in Experiment 5B are denoted ✠; strings presented in Experiments 5C and 5D are denoted ○.

A_1 X_1 B_1 ★	A_1 X_2 B_1	A_1 X_3 B_1 ★ ✠ ○	A_1 X_4 B_1 ★ ✠ ○
A_1 X_1 B_2	A_1 X_2 B_2 ★ ✠	A_1 X_3 B_2 ★ ○	A_1 X_4 B_2
A_1 X_1 B_3 ★ ✠ ○	A_1 X_2 B_3 ★	A_1 X_3 B_3	A_1 X_4 B_3
A_2 X_1 B_1	A_2 X_2 B_1 ★ ✠ ○	A_2 X_3 B_1 ★	A_2 X_4 B_1
A_2 X_1 B_2 ★ ✠ ○	A_2 X_2 B_2 ★	A_2 X_3 B_2	A_2 X_4 B_2
A_2 X_1 B_3 ★ ○	A_2 X_2 B_3	A_2 X_3 B_3 ★ ✠	A_2 X_4 B_3
A_3 X_1 B_1 ★ ✠	A_3 X_2 B_1 ★ ○	A_3 X_3 B_1	A_3 X_4 B_1
A_3 X_1 B_2 ★	A_3 X_2 B_2	A_3 X_3 B_2 ★ ✠ ○	A_3 X_4 B_2
A_3 X_1 B_3	A_3 X_2 B_3 ★ ✠ ○	A_3 X_3 B_3 ★	A_3 X_4 B_3

category members. Recall that in the full overlap design of Experiments 1 and 2, every one of the three X-words appeared with every A-word and every B-word (a total of 9 contexts), and in the partial overlap design of Experiments 3 and 4, every X appeared with 2/3 A's and 2/3 B's (a total of 4 contexts). Here we introduce a new X-word that occurs in only a *single* A and B context. Over a series of four different experiments mirroring the distributional manipulations in Experiments 1–4, we now explore how the amount and type of exposure to the input corpus influences whether learners extend full category privileges to the minimally overlapping X_4 word. As in Experiment 1, we first explore how the learner behaves in situations where there are strong distributional cues (high density) to X being a category (Experiment 5A). Then we test the outcome of weakening one distributional cue (moderate density) while maintaining others (overlap of contexts; Experiment 5B). Lastly, we explore the effect of further weakening the distributional cues to the X category, by first reducing overlap in contexts across X-words (Experiment 5C), and then by increasing exposure to systematic gaps in the input (Experiment 5D). By manipulating the contexts across X-words, we can assess the degree to which learners restrict generalization within X_1–X_3 as we did in Experiments 1–4, and we can also explore how this affects extension of category membership to the novel X_4 word.

7.1. Method

7.1.1. Participants

Separate groups of 16 monolingual native English-speaking students at the University of Rochester were paid to participate in each subcomponent of Experiment 5 (eight in each of the two languages created by different assignments of words to categories for each subcomponent). None of these participants took part in any other categorization study; a total of 64 naïve participants were involved in Experiments 5A–D.

7.1.2. Stimulus materials

The languages used in these experiments were identical to the languages in Experiments 1–4 except that there were 4 X-words instead of 3. Thus, the full language had $3 \times 4 \times 3 = 36$ grammatical AXB strings. However, one of the 4 X-words was presented in only *one* AXB context, rendering its density as minimal as possible. As with Experiments 1–4, the presence of the 2 Q- and 2 R-words was varied evenly in order to remove stable position cues to the A, X, and B categories. Each of the 9 possible A_B frames was heard equally often during the exposure phase of each experiment (see Supplementary materials for adjacent and nonadjacent bigram frequencies).

We started by exposing participants to a dense sampling of the language by presenting a high proportion of the possible X_1–X_3 strings, mirroring the distribution of Experiment 1. Thus, in Experiment 5A, 19 of the possible 36 AXB sentence types in the language were presented to participants, and the remainder were withheld for testing generalization (see Table 3). Including the optional addition of Q and R words, the exposure set was expanded to 76 possible (Q)AXB(R) sentences. However, the input to the learner was very sparse for X_4. The exposure set contained only four X_4 strings: $A_1X_4B_1$, $Q_1A_1X_4B_1$, $A_1X_4B_1R_1$, and $Q_2A_1X_4B_1R_2$, which presented the X_4 word in only one A_B context (A_1_B_1); the remaining 72 sentences included equal numbers of sentences containing X_1, X_2, and X_3

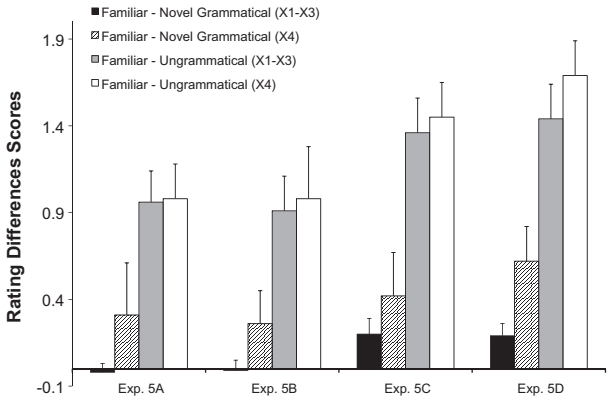

Fig. 5. Difference scores of raw ratings from Experiment 5. The dark bars show the difference in ratings between familiar $AX_{1-3}B$ test strings and novel grammatical $AX_{1-3}B$ test strings; the striped bars show the difference between familiar AX_4B test strings and novel grammatical AX_4B test strings. The gray and white bars show the difference in ratings between familiar and ungrammatical strings containing X_{1-3} and X_4.

such that every one of these three X's appeared with every A-word and every B-word. This meant that there was complete overlap of contexts among X_1, X_2, and X_3, but X_4 shared only one context with X_1–X_3. Training consisted of four times through this exposure set, forming 22 min of exposure.

In Experiment 5B, we explored whether an increase in sparseness for X_1–X_3 affected learners' generalizations to the novel X_4 item. We decreased the density of the contexts for X_1–X_3 words such that the exposure set contained only 10 (versus 19 in Experiment 5A) of the 36 possible AXB combinations (see Table 3), but we kept the number and overlap among X_1–X_3 contexts the same. As in Experiment 5A, every X_1–X_3 word was heard in combination with every A-word and every B-word, but X_4 was only heard in a single context (A_1_B_1). With the addition of AXB strings with optional Q and R flanker words, there were 40 sentences in the exposure set. The exposure set was repeated four times for a total duration of about 12 min.

Exposure for Experiment 5C consisted of only 10 of the 36 possible AXB combinations, as in Experiment 5B. However, in order to test how overlap in contexts influences generalization of category knowledge to new X-words, this experiment reduced the overlap of contexts among members of X_1–X_3. X_1 only occurred with A_1, A_2, B_1, and B_2, but not A_3 or B_3; X_2 was heard with A_2, A_3, B_2, and B_3, but not A_1 or B_1; X_3 was heard with A_1, A_3, B_1, and B_3, but not A_2 or B_2. Thus, the overlap among contexts is maintained over the X_1–X_3 category as a whole, but individual X-words do not have the degree and type of overlap in distributional contexts that they do in Experiments 5A and 5B, where each X-word occurs with every A-word and every B-word. This partial-overlap situation is analogous to the design of Experiment 3.

The language for Experiment 5D was the same as in Experiment 5C, except that training was tripled by presenting the exposure set 12 times rather than 4 (which is qualitatively equivalent to the exposure for Experiment 4). Training lasted for approximately 22 min.

7.1.3. Procedure

As in the earlier experiments, a female native English speaker recorded the words from Experiment 1 plus two new X_4-words (*nerk* /nɚk/ and *sep* /sɛp/). The words were adjusted in Praat (Boersma, 2001) such that pitch, volume, and duration were roughly consistent. Sentences were constructed in the same manner as for Experiments 1–4. The order of sentences in the exposure set was randomized for each subject and presented via a custom software package on a Dell PC. Each sentence was separated by 1.5 s of silence. Participants wore headphones and passively listened to the exposure sentences during training.

Training and test instructions were the same as in the earlier experiments. All test strings were 3-word sentences of the following forms: grammatical familiar strings (10 AXB strings presented during training), grammatical novel strings (13 AXB strings withheld during training), or ungrammatical strings (of the form AXA or BXB).[4] Although ungrammatical test strings had one category repeated, no word tokens were repeated in any string. Of the grammatical novel test strings, 4 of the 13 were strings testing generalization of X_4: $A_2X_4B_2$, $A_2X_4B_3$, $A_3X_4B_2$, and $A_3X_4B_3$. (See Supplementary materials for the list of test items.) With these strings we can ask whether learners have generalized X_4 to the full range of grammatical contexts for X-words even though they have only seen X_4 in one of these contexts. These strings can then be compared to the 6 ungrammatical strings that contain X_4 (three of the form AX_4A, and three of the form BX_4B).

7.2. Results

For each manipulation in Experiment 5, we ran a repeated measures ANOVA with condition (familiar, novel, ungrammatical) as the within subjects factor and language as the between subjects factor. For each of the four variations, there were no significant effects of language ($F < 1$), leading us to collapse across the two languages for all subsequent analyses.

Our analyses examine generalization separately within test strings that contain X_4 and test strings that contain X_1, X_2, and X_3. We do not compare ratings of the X_1–X_3 test items directly with those for the X_4 items, because of the lower statistical power of the X_4 test (4 trials versus 9 trials) and the large difference in frequency of exposure to X_4 vs. X_1–X_3 (up to 18 times more). For all experiments, we take the pattern of learning for familiar and novel grammatical items of the same type to be more informative than the size of the differences between X_1–X_3 and X_4 (see Fig. 5).

7.2.1. Results for Experiment 5A: Dense and complete overlap

For test items without X_4, the mean rating of grammatical novel strings was 3.87 ($SE = 0.14$), the mean rating of grammatical familiar strings was 3.86 ($SE = 0.13$), and the mean rating of ungrammatical strings was 2.90 ($SE = 0.15$). We found no significant difference between ratings of grammatical novel strings and grammatical familiar strings ($F < 1$). These strings were rated significantly higher than ungrammatical test strings ($F(1,15) = 26.40$, $p < 0.001$).

For the test items that contained X_4, the mean rating of grammatical novel strings was 3.28 ($SE = 0.18$), the mean rating of grammatical familiar strings was 3.59 ($SE = 0.24$), and the mean rating of ungrammatical strings was 2.61 ($SE = 0.21$). These items showed the same pattern as the without-X_4 items: there was no significant difference between ratings of grammatical novel X_4 strings and familiar X_4 strings ($F(1,15) = 1.71$, $p = 0.21$), however there was a significant difference between these strings and ungrammatical X_4 strings ($F(1,15) = 13.10$, $p < 0.005$).

7.2.2. Results for Experiment 5B: Sparse and complete overlap

For test items without X_4, the mean rating of grammatical novel strings was 3.55 ($SE = 0.09$), the mean rating of grammatical familiar strings was 3.54 ($SE = 0.10$), and the mean rating of ungrammatical strings was 2.61 ($SE = 0.15$). Just as in Experiment 5A, as well as Experiments 1 and 2, there were no significant differences between ratings of grammatical novel items and grammatical familiar items ($F(1,15) = 0.008$, $p = 0.93$), but grammatical test strings were rated significantly higher than ungrammatical test strings ($F(1,15) = 23.12$, $p < 0.001$).

For the test strings that contained X_4, the mean rating of grammatical novel strings was 3.27 ($SE = 0.15$), the mean rating of grammatical familiar strings was 3.59 ($SE = 0.22$), and the mean rating of ungrammatical strings was 2.45 ($SE = 0.16$). This is the same trend as demonstrated for the X_{1-3} items and the comparable analyses in Experiment 5A. While there was a significant difference

[4] As with Experiments 1–4, a separate experiment confirmed that AAB and ABB ungrammatical strings were rated similarly to AXA and BXB ungrammatical strings, indicating that subjects learned more than just the A_B frame. This argues against the possibility that learners are only responding to the surface similarity between training and test items when making grammaticality judgments of the novel X_4 strings.

between grammatical X_4 strings and ungrammatical X_4 strings ($F(1,15) = 13.42$, $p < 0.005$), there was no significant difference between ratings of grammatical novel X_4 strings and familiar X_4 strings ($F(1,15) = 2.343$, $p = 0.147$).

7.2.3. Results for Experiment 5C: Sparse and incomplete overlap

For test items without X_4, the mean rating of grammatical novel strings was 3.71 ($SE = 0.12$), the mean rating of grammatical familiar strings was 3.91 ($SE = 0.09$), and the mean rating of ungrammatical strings was 2.55 ($SE = 0.15$). We found significant differences between ratings of grammatical novel strings and grammatical familiar strings ($F(1,15) = 9.12$, $p < 0.01$). Additionally, grammatical strings were rated significantly higher than ungrammatical test strings ($F(1,15) = 26.82$, $p < 0.001$).

For the test items that contained X_4, the mean rating of grammatical novel strings was 3.25 ($SE = 0.16$), the mean rating of grammatical familiar strings was 3.66 ($SE = 0.24$), and the mean rating of ungrammatical strings was 2.21 ($SE = 0.16$). Although the mean ratings of the X_4 strings and the X_{1-3} strings showed similar differences between familiar and novel grammatical items, we did not see any significant difference between novel grammatical X_4 strings and familiar X_4 strings ($F(1,15) = 2.98$, $p = 0.11$), perhaps due to the lower statistical power for these test strings. There was, nevertheless, as in all prior experiments, a significant difference between ratings of grammatical and ungrammatical X_4 strings ($F(1,15) = 26.21$, $p < 0.001$).

7.2.4. Results for Experiment 5D: Increased exposure to sparse and incomplete overlap

For test items without X_4, the mean rating of grammatical novel strings was 3.86 ($SE = 0.12$), the mean rating of grammatical familiar strings was 4.05 ($SE = 0.10$), and the mean rating of ungrammatical strings was 2.61 ($SE = 0.21$). These results show a significant difference between ratings of grammatical novel strings and grammatical familiar strings ($F(1,15) = 8.60$, $p = 0.01$). Additionally, grammatical novel strings were rated significantly higher than ungrammatical test strings ($F(1,15) = 26.37$, $p < 0.001$).

For the test items that contained X_4, the mean rating of grammatical novel strings was 3.44 ($SE = 0.19$), the mean rating of grammatical familiar strings was 4.06 ($SE = 0.21$), and the mean rating of ungrammatical strings was 2.37 ($SE = 0.21$). Similar to the X_{1-3} strings, we now find a significant difference between novel grammatical X_4 strings and familiar X_4 strings ($F(1,15) = 8.33$, $p = 0.011$), along with a significant difference between these and ungrammatical X_4 strings ($F(1,15) = 14.261$, $p < 0.005$).

7.3. Discussion

As in Experiments 1–4, learners strongly preferred familiar and novel grammatical sentences to ungrammatical sentences. In Experiment 5A, learners showed generalization to the novel grammatical X_4 strings, but not to the ungrammatical X_4 strings. Thus, subjects generalized the novel X_4-word to the full range of grammatical contexts for X-words, even though they had heard X_4 in only one of these contexts. These results show that, when learners are exposed to a dense sampling of the language space for words in the target category (X_1–X_3) and presented with many overlapping contexts, they generalize their knowledge within the X_1–X_3 category and also extend the category to X_4. Importantly, the generalized contexts are novel contexts for X_4, but are well represented in the learner's exposure to the permissible contexts for X_1–X_3. Neither the X-words nor their contexts were cued by any semantic or perceptual cues, indicating that learners were able to use distributional information alone to extend their knowledge of the X category to a novel X-word.

Also in accord with the results of Experiments 1–4, the degree to which learners extend their category generalization to X_4 in other conditions depends on the strength of the category formed for X_1–X_3. In Experiment 5A, with high density and overlap among the X_1–X_3 words, novel and familiar contexts for the X_4 word were judged to be equivalently acceptable, mirroring the results of Experiment 1. In Experiment 5B, when we decreased the density of the contexts for X_1–X_3 words, (but kept the number and overlap among X_1–X_3 contexts the same), the results mirrored those in Experiment 2. Reduced density did not greatly affect learners' performance, as long as there was full overlap of contexts

among X_1–X_3 words. The generalization to X_4 was maintained despite greatly reduced exposure, due to a sparser sampling of the language space.

In Experiment 5C, we reduced the overlap among contexts in the exposure set by a third, while keeping the number of contexts in the input the same as in Experiment 5B. The results show that the incomplete overlap between X_1–X_3 words led to decreased generalization within X_1–X_3 and also led to decreased generalization to X_4. However, learners still showed a much higher rating for X_4 grammatical novel strings than ungrammatical strings, indicating that they were still willing to generalize, though somewhat more conservatively than when there was complete overlap of contexts.

As we saw in Experiment 4, the decision to generalize over a gap in the input or to maintain lexical distinctness is also influenced by the frequency of contexts (and gaps) in the input. If a context is consistently and repeatedly absent, learners show even more conservatism in their generalizations and more certainty that gaps in the input are systematic and not accidental (e.g., Wonnacott et al., 2008; Xu & Tenenbaum, 2007). This manipulation is particularly important with regards to X_4, where we can observe how an increase in the exposure to the one context for X_4 (and an increase in the gaps formed by the non-occurring contexts for X_4) affects how learners generalize their knowledge of the category defined by X_1–X_3. In Experiment 5D, when we increased exposure to the sparse data of Experiment 5C (with incomplete overlap among the X-words and recurring gaps that presumably become more prominent with repetition), learners were even less likely to generalize over such gaps. Not only did this lead to reduced ratings of novel X_1–X_3 strings, but the increase in exposure to one context for X_4 led to reduced ratings of novel X_4 strings as well. While novel grammatical test strings continued to be rated as more acceptable than the ungrammatical strings, further exposure to the sparse input set might push learners to judge all novel strings as ungrammatical.

Overall, as we move along the dimensions of sparseness, overlap, and frequency explored in Experiments 5A–D, we see that learners use the same variables investigated in Experiments 1–4 to weigh the likelihood that X_4 shares the same contexts as X_1–X_3. The more strongly learners generalize within X_1–X_3, the more strongly they also generalize to X_4. Looking at these results in another way, we can use the degree of generalization to a novel word that is observed in a single context as a diagnostic for how strongly the X-category has been formed. Below we consider what these results suggest regarding the type of information learners are extracting from their input and the type of category representation they may be constructing of the linguistic category and of specific lexical strings.

8. General discussion

The present experiments provide compelling evidence that adult learners can use their sensitivity to systematic patterns of distributional information to acquire a grammatical category. Moreover, the pattern of data across the experiments demonstrates that learners generalize across gaps in the input by weighting distributional information in a principled manner. They accomplish this task in the absence of correlated cues such as phonological similarity or semantic relatedness, which is important given that natural language does not always contain consistent and reliable correlated cues (Gleitman, 1990; Maratsos & Chalkley, 1980). The first finding regarding these variables was that we observed strong generalization and category formation with fairly dense input (when learners were exposed to 18 of the 27 possible basic AXB sequences in the language) and a high degree of overlap in contexts among words (Experiment 1). We also found no decline in categorization when learners were exposed to sparser input of the same type (reducing exposure to 9 of the 27 AXB sequences in the language), as long as only the number of contexts was reduced, but not the overlap in contexts across words (Experiment 2). Learners began to decrease their generalization (that is, increased the difference in their ratings of familiar versus novel grammatical strings) when the overlap in contexts for different words within the category was reduced (Experiment 3). Learners restricted their tendency to generalize even more sharply when the same exposure corpus (and its gaps) was repeated three times (Experiment 4). Taken together, these findings indicate that learners are highly sensitive to the details as well as the overall patterning of distributional information in their linguistic input, and they use this information in sophisticated ways to determine when it is appropriate to generalize words to new contexts or to be cautious about generalization.

For many decades, the literature on syntax acquisition has focused on concerns about generalization – especially about the danger of overgeneralization and the impossibility of recovering from overgeneralization without negative evidence. These concerns have arisen, in part, from the assumption that learners must be working from individual input sentences and the misleading information they potentially provide. (Consider, for example, the uncertainty about whether to generalize dative movement to the verb '*donate*' after hearing sentences in which '*give*' undergoes dative movement; cf. Baker, 1979; Pinker, 1984.) Our results suggest, however, that adult learners decide whether to generalize based not on individual input sentences, but based on the statistical patterning of evidence in the input corpus. Even without negative evidence, they can compute the likelihood of certain types of gaps occurring by chance, depending on the size and structure of the corpus to which they have been exposed. Given a sizeable corpus, sophisticated statistical learners can determine the likelihood that gaps are recurring and systematic or are accidental, and can base their generalizations on such information. These results are consistent with other findings (see especially Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths, & Kemp, 2006; Xu & Tenenbaum, 2007) that show human learners are able to weigh observed evidence (the likelihood) with expectations about what evidence should be observed (the prior) in a Bayesian framework applied to many cognitive domains.

Importantly, the current experiments also show that learners can skillfully transfer their knowledge of category structure and category cues to a novel item that is only minimally represented in the input (Experiment 5). When given a dense sampling of the language space with almost complete overlap of contexts for many words in a target X-category, learners generalize a novel word (X_4) to the full range of grammatical contexts of the other X-words, even when they have only seen X_4 in one of those contexts. This willingness to add X_4 to the strongly established X_1–X_3 category is most robust when the X_1–X_3 contexts are dense and overlapping. When contexts are more sparse and less overlapping across different X words, we see more conservative generalization to a new X_4 word. The most extreme case is when we increase the number of times the learner hears the sparse exposure set, thus increasing the frequency of recurring gaps in the input for X_1–X_3 strings: learners in this situation rate the withheld X_4 contexts as more unfamiliar, while rating the one context in which X_4 was actually heard as highly familiar. These findings are in line with results from Wonnacott et al. (2008) on verb-argument learning. In their studies, if a language contains many verb-specific constructions, participants do not show generalization of a minimally exposed verb (like X_4) to other contexts. In contrast, if the language allows the same contexts for all verbs, then participants show strong generalization of a minimally exposed verb to contexts in which it has not been heard. The results of Experiment 5 again show that learners use the pattern of distributional information across the language to tell them when to generalize and when to be lexically conservative. They also suggest a new finding: with sufficiently well-structured input (as in Experiments 5A and B), linguistic categories can become 'crystallized', allowing all of their properties to be passed on as a whole to novel items that are only seen a few times, but are nevertheless consistent with category membership. We return to this point below in considering what representational hypotheses are consistent with this behavior.

Why have so many other artificial language learning studies failed to show category formation with only distributional information, thus necessitating correlated perceptual or semantic cues in order to attain successful category learning? One contrast between our experiments and earlier studies is the way many investigators have framed the categorization problem. Most prior studies have looked at the formation of multiple categories in a single artificial language, whereas the present work looks at the formation of a single category. The ability to acquire multiple categories – in essence, knowing that X's allowable contexts are A_ and _B, while Y's allowable contexts are C_ and _D, and that these are not interchangeable – is obviously an important aspect of natural category acquisition. This component of the category acquisition problem is one that we have not directly addressed in the present set of experiments. However, one way to view the multiple form-class categories previously studied in small artificial grammars is as *subcategories* (such as subcategories of grammatical gender of nouns), rather than major form-class categories (such as noun or verb). Categorization and subcategorization involve similar processes, since in both, the learner must distinguish between the gaps that are accidental omissions from the input and the systematic gaps that signal structural aspects of the category or subcategory. Subcategory learning has an important difference from single-category learning, though: the subcategorization task inherently involves a conflict of cues. For subcategories of a larger

form-class category, some distributional information (namely, word order) signals that there is one category, while other distributional cues (such as the patterning of context words) signal that there are distinct categories within the larger category. In the present experiments we have been careful to study only basic category learning. However, in more recent work we have applied the same distributional variables to the problem of subcategory acquisition. We have found the same type of outcomes as in the present studies, though (as expected from conflicting cues) with somewhat reduced sharpness of subcategory formation (see Reeder et al., 2009: Experiment 5).

A second difference between our present findings and those of prior studies is that we have systematically manipulated a number of distributional variables in order to understand not only *whether* distributional information can support category learning, but also *how* and *when* it can do so. As we have seen from this exploration, category learning shows graded effects, depending on the nature of the distributional patterns contained in the linguistic input. Such results may explain why earlier experiments have shown either chance or weak performance in category learning from distributional cues alone – often the learner must contend with very small languages that have weak distributional evidence for categories, and conflicting distributional cues that are inherent to a subcategory structure.

8.1. Formulating the precise mechanisms underlying categorization

A comprehensive approach to formulating the mechanisms underlying linguistic categorization will require computational modeling work, which is already in progress. However, the results from the present experiments allow us to discern something about the types of information that participants are extracting from the input and utilizing for generalization during learning. First, learners could not have been relying on a simple encoding of the exposure sentences as complete sequences or in terms of their trigrams or quadrigrams. If they were storing any of these types of information, they would have discriminated sharply in every experiment between familiar and novel grammatical strings. This is because these strings always differed in the specific AXB trigrams they contained, as well as in the quadrigrams or complete sequences that included these AXB trigrams. At the opposite extreme, learners could not have relied solely on storing the individual word frequencies in the exposure, as this information was carefully balanced between test items in all of the experiments. What types of information storage, then, are compatible with the results we have obtained?

The results of Experiments 1–4 are compatible with the possibility that learners are encoding bigram frequencies (e.g., AX, XB) and using them to rate test strings. However, this is incompatible with the results from Experiments 5, as storing bigram frequencies could not account for generalization to novel X_4 strings. Because the X_4 word was presented in only one A_B context, only 2 bigrams for X_4 were part of the exposure set (A_1X_4 and X_4B_1). Nonetheless, generalization to new X_4 contexts was very strong when, based on the contexts in which X_1–X_3 had appeared, the overall X category was robustly learned. These results suggest that, as modeled in corpus work on category acquisition (cf. Mintz et al., 1995, 2002), learners might be keeping track of word co-occurrences by storing a network of occurring contexts for each individual word, and then collapsing the individual words into a category when these networks bear enough quantitative (and qualitative) similarities to one another.[5] By this mechanism, the category network could be applied as a whole to a new word that shared any of the contexts. As already mentioned, ongoing work is testing this and other models against the details of the experimental data (Qian et al., 2012).

An additional question of interest concerns whether this type of distributional analysis is specific to language, or can be performed by a more domain-general statistical learning mechanism. In order to study category learning in a non-linguistic domain, Hunt and Aslin (2010) used a non-linguistic serial reaction time categorization task in which sets of buttons formed the possible serial order patterns to which subjects were exposed. Their results suggest that the type of mechanism involved in learning

[5] One possible interpretation of our data is that the X-words have not formed a true category, but instead are merely just the by-product of a set of linked contexts. The work presented here, in combination with our modeling results (Qian, Reeder, Aslin, Tenenbaum, & Newport, 2012) and other behavioral results from our lab on the acquisition of multiple categories and subcategories, makes this interpretation unlikely.

categories of items based on shared contexts may well apply to non-linguistic as well as linguistic category tasks.

8.2. Extending the results to natural language learning

How do the results of these experiments apply to natural language learning? Of course, the present studies involve adult subjects, whereas the most important natural language learners are young children. Perhaps adults are better able to take the rich details of differing distributional environments into account (however, see Mintz et al. (2002) and Wang and Mintz (2008) for evidence that a resource-limited system can still perform successful categorization just by utilizing certain types of distributional information). We are in the process of performing these same experiments with children – a methodological challenge, since young children do not readily participate in listening experiments for the length of time required to learn these languages. We have found that they look very much like adults in their ability to generalize in designs like those in Experiments 1 and 2; whether they look precisely the same as adults across the input variations in Experiments 3–5 remains to be seen. In the meantime, though, we can ask which of the experimental conditions is most similar to the distributional environments in typical natural language input to children. From there, we can ask what the similarities and differences from natural language input can tell us about extending our findings to language acquisition.

In contrast to our experiments, where we have removed all phonological and semantic information, natural language categories do sometimes have partially correlated phonological or semantic cues that learners could use in acquiring categories, and many studies have shown that category learning is enhanced when category membership is correlated with such surface cues (e.g., Monaghan et al., 2005). It is also true that learners could utilize a distributional learning mechanism in tandem with performing semantic analyses on the input, as suggested by Pinker (1984, 1987). Like other investigators, we expect that learners will exploit correlated cues when they are available. But an important question in this literature is whether category learning can utilize distributional information as a way to break into the category learning problem. As we have noted, our results indicate that adult learners are able to skillfully employ a statistical learning mechanism as the primary tool with which to extract category information from the input, even in cases where other correlated cues are incomplete or absent. Furthermore, in response to arguments such as those made by Pinker (1984, 1987), our results suggest that a few shared lexical contexts are not sufficient to collapse grammatical categories. In particular, our findings indicate that a distributional learning mechanism can utilize the consistency or inconsistency of contextual cues, as well as the breadth and overlap of contextual cues across lexical items, to decide whether to collapse words into a category. This mirrors results from Xu and Tenenbaum (2007) and Gerken (2006), where infants and children generalize only when the contextual evidence suggests that tokens are interchangeable because of strong context overlap. These results suggest that some contextual ambiguities will not mislead learners into major category formation errors.

Which of our experiments best represents the type of distributional information that is likely to be present in real linguistic input? We expect that a large corpus of speech directed to young children would contain a mixture of the types of distributional patterns presented to learners in Experiments 1–5. It is likely that only a small number of words will have patterning similar to the X-words from Experiments 1 and 2, where the input includes full and overlapping coverage of the possible grammatical contexts for each word. Even so, our results from Experiment 5 suggest that encountering just a few words with this type of patterning is enough to allow learners to extend category properties to other minimally overlapping words.

Linguistic input to young language learners likely involves many words with partially overlapping contexts (as in Experiment 3). Although this might be seen as a problem, the results of Experiment 3 indicate that adult learners do show generalization to novel strings that follow the patterns of the overlapping familiar contexts, though with some uncertainty. (Recall that although grammatical novel strings are rated significantly lower than familiar strings, they are rated significantly and substantially higher than ungrammatical strings.) If the learner has encountered some words with stronger cues to category membership (as in Experiments 1 and 2), this could be enough to bootstrap category

membership to the words experienced in partially overlapping contexts. Indeed, a computational analysis of maternal speech corpora from CHILDES reported in Mintz et al. (2002) found that the 100 most frequent nouns and verbs in the corpora each occurred in a wide variety of overlapping linguistic contexts; this was the basis for the successful classification of nouns and verbs achieved in that work. The present findings indicate that human learners also utilize these patterns of overlapping distributional contexts. Furthermore, given a large set of words that vary in their contextual overlap (as exhibited across our experiments), human learners appear to be able to discern precisely where to generalize and where to withhold generalization based on these distributional patterns.

In conclusion, our findings suggest a new framework for thinking about the linguistic category-learning problem. According to this view, a critical question concerns the *structure* of the distributional information that the learner receives. Across our experiments we observed remarkable sensitivity to the character, patterns, and reliability of distributional information, indicating not only that learners are sensitive to this information, but also that they are capable of using this information in principled and sophisticated ways to induce form-class categories from language input.

Acknowledgments

We would like to thank Amanda Robinson, Carrie Miller, Anna States, Kathryn Schuler, Kathryn Lukens, and Caitlin Hilliard for assistance with stimulus creation and data collection. We also thank two anonymous reviewers, Chuck Clifton, Toby Mintz, Josh Tenenbaum, and the Aslin-Newport lab at the University of Rochester for helpful comments on this work. This research was supported by NIH Grants HD037082 to R.N.A. and DC00167 to E.L.N., and by an ONR Grant to the University of Rochester.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2012.09.001.

References

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry, 10*, 533–581.

Bates, E., & MacWhinney, B. (1979). The functionalist approach to the acquisition of grammar. In E. Ochs & B. Schieffelin (Eds.), *Developmental pragmatics*. New York, NY: Academic Press.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the state of the art*. Cambridge: Cambridge University Press.

Bloomfield, L. (1933). *Language*. New York: Holt, Rinehart and Winston.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bowerman, M. (1973). Structural relationships in children's utterances: Syntactic or semantic? In T. Moore (Ed.), *Cognitive development and the acquisition of language*. New York, NY: Academic Press.

Braine, M. D. S. (1966). Learning the position of words relative to a marker element. *Journal of Experimental Psychology, 72*, 532–540.

Braine, M. D. S., Brody, R. E., Brooks, P., Sudhalter, V., Ross, J. A., Catalano, L., et al (1990). Exploring language acquisition in children with a miniature artificial language: Effects of item and pattern frequency, arbitrary subclasses, and correction. *Journal of Memory and Language, 29*, 591–610.

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.

Brooks, P. B., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language, 32*, 79–95.

Brown, R. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology, 55*, 1–5.

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

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

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 12203–12208.

Finch, S., Chater, N., 1992. Bootstrapping syntactic categories using statistical methods. In: *Background and experiments in machine learning of natural language: Proceedings of the 1st SHOE workshop on statistical methods in natural language, ITK proceedings 92/1* (pp. 229–235). Berlin: Springer.

Finch, S., & Chater, N. (1994). Learning syntactic categories: A statistical approach. In M. Oaksford & G. D. A. Brown (Eds.), *Neurodynamics and psychology* (pp. 294–321). London: Academic Press.

Fries, C. C. (1952). *The structure of English: An introduction to the construction of English sentences*. New York: Harcourt, Brace and Company.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory & Language, 39*, 218–245.
Gerken, L. A. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*, B67–B74.
Gerken, L.A., Gomez, R., & Nurmsoo, E. (1999, April). The role of meaning and form in the formation of syntactic cateogries. In: *Paper presented at the Society for Research in Child Development*, Albuquerque, NM.
Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249–268.
Gleitman, L. R. (1990). The structural sources of verb meaning. *Language Acquisition, 1*, 3–55.
Gleitman, L. R., & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3–48). Cambridge, England: Cambridge University Press.
Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109–135.
Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*(5), 178–186.
Gomez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*(5), 567–580.
Gordon, P. (1985). Evaluating the semantic categories hypothesis: The case of the count/mass distinction. *Cognition, 20*, 209–242.
Grimshaw, J. (1981). Form, function, and the language acquisition device. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 165–187). Cambridge, MA: MIT Press.
Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.
Harris, Z. S. (1954). Distributional structure. *Word, 10*, 146–162.
Hunt, R. H., & Aslin, R. N. (2010). Category induction via distributional analysis: Evidence from a serial reaction time task. *Journal of Memory and Language, 62*, 98–112.
Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*(2), 349–364.
Macnamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79*, 1–14.
Maratsos, M., & Chalkley, M. A. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.). *Children's language* (Vol. 2, pp. 127–189). New York: Gardner Press.
McNeill, D. (1966). Developmental psycholinguistics. In F. Smith & G. Miller (Eds.), *The genesis of language* (pp. 15–84). Cambridge, MA: The MIT Press.
Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*(5), 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 North East Linguistics Society 25* (Vol. 2, pp. 43–54). Amherst: 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.
Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorization. *Cognition, 96*, 143–182.
Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of grammatical categories. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 263–283). Mahwah, NJ: Lawrence Erlbaum Associates.
Pinker, S. (1984). *Language learnability and language development*. Cambridge: 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.
Qian, T., Reeder, P. A., Aslin, R. N., Tenenbaum, J. B., & Newport, E. L. (2012). Exploring the role of representation in models of grammatical category acquisition. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 881–886). Austin, TX: Cognitive Science Society.
Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425–469.
Reeder, P. A., Newport, E. L., & Aslin, R. N. (2009). The role of distributional information in linguistic category formation. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
Schlesinger, I. M. (1974). Relational concepts underlying language acquisition. In R. L. Schiefelbusch & L. Lloyd (Eds.), *Language perspectives: Acquisition, retardation, and intervention*. Baltimore, MD: University Park Press.
Smith, K. H. (1966). Grammatical intrusions in the free recall of structured letter pairs. *Journal of Verbal Learning and Verbal Behavior, 5*, 447–454.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629–641.
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*, 309–318.
Wang, H., & Mintz, T. H. (2008). A dynamic learning model for categorizing words using frames. In H. Chan, E. Kapia, & H. Jacob (Eds.), *Proceedings of the 32nd annual Boston University conference on language development* (pp. 525–536). Somerville, MA: Cascadilla Press.
Wilson, R. (2002). *Syntactic category learning in a second language*. Unpublished doctoral dissertation. The University of Arizona.
Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 56*, 165–209.
Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272.