Journal of Memory and Language 97 (2017) 17–29

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Distributional learning of subcategories in an artificial grammar: Category generalization and subcategory restrictions

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

[a] Department of Psychological Science, Gustavus Adolphus College, St. Peter, MN, USA
[b] Center for Brain Plasticity and Recovery, Georgetown University, Washington, DC, USA
[c] Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY, USA

ARTICLE INFO

Article history:
Received 27 December 2015
Revision received 12 July 2017

Keywords:
Categorization
Linguistic subcategories
Language acquisition
Distributional learning

ABSTRACT

There has been significant recent interest in clarifying how learners use distributional information during language acquisition. Many researchers have suggested that distributional learning mechanisms play a major role during grammatical category acquisition, since linguistic form-classes (like *noun* and *verb*) and subclasses (like *masculine* and *feminine* grammatical gender) are primarily defined by the ways lexical items are distributed in syntactic contexts. Though recent experimental work has affirmed the importance of distributional information for category acquisition, there has been little evidence that learners can acquire linguistic *subclasses* based only on distributional cues. Across two artificial grammar-learning experiments, we demonstrate that subclasses can be acquired from distributional cues alone. These results add to a body of work demonstrating rational use of distributional information to acquire complex linguistic structures.

Introduction

Natural languages are highly structured systems, governed by particular organizational rules and representations. Language learners are tasked with acquiring these rules and representations in a primarily unsupervised environment, without initial access to the full set of sounds, word combinations, or structures that are necessary to produce and comprehend the infinite set of possible sentences in their language. One of the main linguistic structures that support a language's generativity are its syntactic categories. These form-class categories are primarily defined based on how groups of words are distributed with certain syntactic arguments. For example, certain words can occur as the subject of a *verb* or the object of a *preposition*. Words that have these syntactic properties (among others) are grouped together as *nouns*. Having the category *noun* allows a language user to use new nouns in syntactic contexts where they have previously heard other nouns occur; that is, the distributional properties of the category *noun* can be generalized across words in the category.

Languages not only have major form-class categories like *noun* and *verb*; some of these categories may be further divided into subcategories. Like major form-class categories, language subcategories are partly defined and differentiated based on the different types of linguistic contexts in which words in the subcategory may occur (e.g., Bloomfield, 1933; Chomsky, 1965; Harris, 1954). One well-studied example of noun subcategories is grammatical gender. In many languages, nouns differ in the form of the determiner that goes with them (e.g., in French, masculine nouns take the definite determiner *le*, whereas feminine nouns take the definite determiner *la*) or in the endings that must occur on the noun or on co-occurring adjectives. Importantly, linguistic gender is arbitrarily defined: grammatical gender does not clearly relate to natural biological/social gender, linguistic gender assignments are inconsistent across languages, and the number of grammatical genders in a language varies cross-linguistically. Though not all languages have grammatical gender, nouns in many languages contain other types of subcategories, such as the distinction between count nouns and mass nouns. In English, determiners serve as one type of distributional cue to these subcategories: whereas mass nouns may occur with the determiners *much* and *some*, count nouns occur with determiners such as *many* or *one*. Verbs can be subdivided based on whether or not the verb takes an object, forming transitive and intransitive subcategories, or in many languages are subdivided into conjugations, differing in the endings the verb takes for person and number. While the distinction between transitive and intransitive subcategories is related to verb semantics and argument structure, verb conjugations are distributionally defined.

* Corresponding author at: Department of Psychological Science, 800 West College Avenue, Gustavus Adolphus College, Saint Peter, MN 56082, USA.
E-mail addresses: preeder@gustavus.edu (P.A. Reeder), eln10@georgetown.edu (E.L. Newport), aslin@cvs.rochester.edu (R.N. Aslin).

http://dx.doi.org/10.1016/j.jml.2017.07.006
0749-596X/

Because linguistic categories and subcategories are crucial components of natural language structure, there has been sustained interest in studying the mechanisms underlying their acquisition. However, the exact process underlying their acquisition has been particularly difficult to define. Categories and subcategories lack consistent perceptual or semantic cues to their organization, and distributional cues are often ambiguous and overlapping (e.g., Braine, 1987). Despite the complexity of this system, however, even young children demonstrate early knowledge of the form-class organization of their native language (e.g., Maratsos & Chalkley, 1980). This knowledge allows them to use syntactic categories and subcategories to learn the meanings of new words (e.g., Scott & Fisher, 2009; Yuan & Fisher, 2009) and to produce grammatical utterances based on form-class category knowledge (e.g., Berko, 1958). Even though children may not have perfect subcategory representations by the time they demonstrate productive use of form-class categories, there is evidence that they at least have basic knowledge of relevant subcategories at a very early age. For example, children acquiring the Russian gender paradigm do not consistently mark the correct gender at an early age, but they do have the correct number of gender subcategories despite occasionally using them in the wrong contexts (e.g., Gvozdev, 1961; Polinsky, 2008). Thus, although there may be imperfect production of subcategory knowledge (perhaps due to performance limitations), grammatical subcategories are clearly being formed early in language development (e.g., Valian, 1986).

Given the potential complexity of category acquisition, a large body of work has explored the types of information that learners could in principle – and do in practice – exploit for discovering the categories and subcategories in their language. Though natural language categories are associated with many possible sources of cues, distributional information has proven to be a reliable cue to major form class category structure (e.g., Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998). Additionally, human learners have been shown to use the distributional cues that define categories – sometimes along with other types of cues – in order to acquire them (e.g., Braine et al., 1990; Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Mintz, 2002; Mintz, Wang, & Li, 2014; Reeder, Newport, & Aslin, 2013; Schuler, Reeder, Newport, & Aslin, in press; Scott & Fisher, 2009; St. Clair, Monaghan, & Christiansen, 2010).

A first step in studying the role of distributional information for categorization was provided by Smith (1966), who showed that learners were quite capable of learning a simple language consisting of two categories:

$\text{Pair} \rightarrow \alpha + \beta$
$\alpha \rightarrow$ D, V, H, R, X
$\beta \rightarrow$ M, F, G, K, L

where there are 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 recall as many strings as possible. The results showed that learners recalled both the presented strings and "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 the categories are defined by positional information (the co-occurrence statistics between the two categories were distributionally uninformative in this study).

However, in a similar paradigm by Smith (1969), participants had to learn dependencies between words within a pair of contingent categories:

$\text{Pair} \rightarrow \alpha + \beta$
$\alpha \rightarrow$ M, P
$\beta \rightarrow$ N, Q
$M \rightarrow m_1, m_2, m_3$
$N \rightarrow n_1, n_2, n_3$
$P \rightarrow p_1, p_2, p_3$
$Q \rightarrow q_1, q_2, q_3$

Importantly, strings of the language followed the basic pattern of MN or PQ; no MQ or PN strings were presented. M, N, P, and Q were categories of 3 items (letters) each. Exposure consisted of seeing 2/3rds of the possible MN pairings and 2/3rds of the PQ pairings. However, while participants learned that M- and P-words occurred first and that N- and Q-words occurred last in the 2-word strings of the language, they did not learn the co-occurrence dependencies that M-words were only followed by N-words and P-words were only followed by Q-words. They produced MQ strings as well as PQ strings, and showed no differentiation between the two. This "MN/PQ problem" (Braine, 1987) is a classic case, widely cited in the literature, of failure to acquire categories from distributional information alone.

Other problems have also plagued learning theories that primarily rely on distributional analyses for category formation. As Pinker (1984, 1987) noted, it is not always obvious which contexts a learner should learn from in any particular utterance, and overly simplistic distributional analyses could lead a learner astray. Likewise, Braine (1987) recognized how easily and quickly learners acquired positional cues to categories in the MN/PQ problem, such as "M-words come first" and "N-words come last." Though positional cues are a type of distributional information, they do not reveal the full set of rules governing the MN/PQ language. Unfortunately for proponents of distributional analyses, it seemed as if learners were *only* capable of acquiring these serial dependencies in Smith (1969), since they were unable to learn the rule "M words are obligatorily followed by N-words." Braine (1987) concluded that learners required an additional salient cue (called a "similarity relation") to overcome this positional information and highlight the distributional structure of the categories in the MN/PQ problem – for example, associating the M subclass with males and the P subclass with females, thus building in a semantic similarity relation. With the addition of partially correlated semantic cues, subjects were able to restrict generalization in the MN/PQ experiment: they made fewer ungrammatical overgeneralizations when a semantic similarity relation cued them into the co-occurrence structure of the MN/PQ subclasses.

A number of investigators have followed up on this hypothesis, exploring the role of shared cues to category structure (e.g., Braine, 1966; semantic cues: Braine et al., 1990; morphological cues: Brooks et al., 1993; phonological cues: Frigo & McDonald, 1998; Gerken, Gomez, & Nurmsoo, 1999; Gerken, Wilson, & Lewis, 2005; Monaghan, Chater, & Christiansen, 2005; Morgan, Shi, & Allopenna, 1996; Shi, Morgan, & Allopenna, 1998; Wilson, 2002; shared features: Gomez & Lakusta, 2004). The results from many of these artificial language studies suggest that the formation of linguistic classes crucially depends on overlapping perceptual properties that link the items together. These correlated perceptual cues might arise from identity or repetition of elements in grammatical sequences, or from a phonological or semantic cue identifying words across different sentences as similar to one another (for example, words ending in *–a* are feminine). On this view, correlated cues are necessary and sufficient to discover the categorical structure in artificial languages, and in the acquisition of natural grammatical classes (Gomez & Gerken, 2000).

However, most categories (and most subcategories) are arbitrary: though they may have partially correlated semantic,

phonological, or morphological cues, none of these sources of information perfectly define natural language categories (cf. Maratsos & Chalkley, 1980). Given this complexity, how can a child acquire such structures without presupposing that the foundations of these structures are innate (e.g., Chomsky, 1965; McNeill, 1966)?

There is some evidence that languages exhibit interactions between different types of correlated cues (e.g., phonological, semantic, distributional). In some situations, the redundancy of correlated cues might help young language learners hone in on the relevant linguistic components for bootstrapping other aspects of grammar. In other cases, some cues may be more reliable markers of linguistic structures when other cues are unreliable. Monaghan et al. (2005) and Monaghan, Christiansen, and Chater (2007) showed that some languages exhibit a trade-off between phonological information and distributional information, such that phonological cues are better predictors of linguistic structures when distributional cues are weak. It is likely that young language learners utilize these correlated patterns to refine their linguistic representations; however, it is difficult to ascertain how a language learning mechanism would use these patterns for the initial stages of category learning, since that would require some a priori knowledge of the category structure or basic knowledge of how the distributional and perceptual cues are correlated.

It is also clear that some types of correlated distributional cues can be useful during grammatical category learning. For example, Mintz et al. (2002) demonstrated that there are sets of linguistic contexts (words before and/or after a particular target category) that co-occur with words belonging to the same grammatical category. Mintz (2002) demonstrated that adult learners can use these co-occurring dependencies to induce categories in artificial grammars. Mintz (2003) further suggested that a particular subset of these – the pairs of non-adjacent contexts that occur together, before and after target items, called 'frequent frames' – provide particularly useful information, and that in some categorization tasks adult learners rely on frequent frames more heavily than uncorrelated lexical bigrams (Mintz et al., 2014). Mintz (2003) also showed that frequent frames are present in corpora of child-directed speech, supporting the idea that these non-adjacent context words might provide sufficient information for a young distributional learner to induce categories. However, while research has shown that infants as young as 7 months can learn a simple repetitive "a-B-a" non-adjacency (Gervain & Werker, 2013), it appears that infants younger than 12 months cannot learn non-repetitive non-adjacent dependencies, such as the frequent frames that are most useful during natural language categorization (Gomez & Maye, 2005).

Since many previous studies have failed to find successful categorization when learners are relying solely on distributional cues, it is worthwhile to re-examine the types of distributional information available to the learner in those studies. Though one might suggest that learners in Smith (1969) failed to acquire categories and their dependencies, another interpretation is that the MN/PQ problem is a special case of *subcategory* learning. M and P are subcategories of the α category, and N and Q are subcategories of the β category.[1] Subcategory learning has an important difference from single-category learning: the subcategorization task inherently involves conflicting cues. For subcategories in natural languages, some of the distributional information (e.g., word order) signals that there is one category, whereas other distributional cues (e.g., contextual function words and grammatical morphemes) signal that there are distinct subcategories within the larger category. In the subcategorization case, then, the learner must figure out that there is a major form class category that encompasses all the words in terms of their word order and argument structure, and also that there are subsets of items in this category that each have their own specific linguistic contexts. The latter problem is complicated by the fact that some contextual gaps are systematic omissions that arise from the subcategory structure, while others may be accidental gaps of legal contexts that did not occur in a particular linguistic sample. Given the incomplete and noisy input that any language learner receives, the main question facing the learner is whether particular word combinations are absent from the input accidentally (because the learner just has not heard that context yet), or are absent from the input because that context is ungrammatical. To successfully acquire the MN/PQ grammar, learners must simultaneously restrict generalization (i.e., M's shouldn't be seen with Q's because "MQ" is ungrammatical), while also generalizing appropriately (i.e., even if I haven't heard this particular M with this particular N, it is still an acceptable string because "MN" is a legal structure). The "gaps" in the input created by the missing MQ and PN strings occur because those combinations are ungrammatical; the "gaps" created by missing MN and PQ strings, in contrast, are accidental. Unfortunately, Smith's (1969) participants were unable to make this distinction and overgeneralized. Apparently the participants failed to acquire the boundary separating the M and P *subcategories*; instead, learners only acquired α and β as major form-class categories, generalizing to contexts MQ and PN. The goal of the experiments presented here is to explore the distributional variables that can lead learners to appropriately limit generalizations and reject MQ- and PN-like strings.

Recent behavioral work on form-class category learning has highlighted a number of distributional cues that lead to successful category learning from distributional cues alone (e.g., Mintz, 2002; Mintz et al., 2002, 2014; Reeder et al., 2013). In particular, Reeder et al. (2013) investigated whether learners of an artificial language could learn that a set of nonsense words formed a linguistic category based solely on distributional information: that is, could learners generalize from hearing some of the distributional contexts for individual words to the full range of contexts for all the words in the category. In order to examine distributional learning, all other potential cues to the category were removed; only the distributional contexts for words were available as cues to categorization. The results showed that, given specifiable distributional information, word co-occurrence statistics were enough to cause learners to induce a representation of a category – rather than storing each word individually in terms of its specific experienced contexts – and to generalize or withhold generalization to new contexts and new words based on distributional cues.

Our goal in the present paper is to examine how this distributional learning mechanism might operate during *subcategory acquisition*, and to explore why many previous experiments like Smith (1969) have failed to see successful subcategory learning from distributional cues alone. Across two artificial grammar-learning experiments, we systematically manipulate the distributional information that signals subcategory structure and a subcategory boundary. Given that word categories and subcategories are defined based on how they are distributed with respect to other lexical items in a sentence, the main distributional variable of interest is the amount of *contextual overlap* among words. Complete contextual overlap means that words within a category share all of their possible linguistic contexts with each other. In the idealized case of a perfect subcategory boundary, learners should expect complete overlap of contexts across words *within* a

[1] The representations that learners have acquired in the experiments we describe here and in the previous literature are compatible with two interpretations: acquisition of subcategories, or acquisition of multiple form-class categories. Given the distributional information framework we lay out in this paper, we argue that subcategories are the cognitive representation that is most compatible with our paradigm and previous work. However, our conclusions and framework would remain the same if the outcome of learning is interpreted to be multiple form-class categories.

subcategory, but no overlap for words *across* a subcategory boundary. This circumstance is investigated in Experiment 1. In Experiment 2 we investigate what happens to learning when there is imperfect overlap and imperfect distinctiveness of contexts. In order to demonstrate subcategory acquisition, learners in our experiments will need to achieve two outcomes: strong generalization from words in their experienced contexts to the larger range of contexts allowable for other words in the same subcategory, but restricted (or no) generalization to the contexts of words in a different subcategory.

Experiment 1

In Experiment 1, we adapt the artificial grammar from Reeder et al. (2013) in order to test whether (contrary to much of the literature, particularly Smith (1969) and Braine (1987)) subcategories are learnable from distributional information alone. Building upon Reeder et al.'s findings on major category learning, we hypothesize that learners can acquire subcategories if they are given adequate overlap in distributional cues for the items inside each subcategory and adequate non-overlap in distributional cues for the items in different subcategories. As a first step towards exploring subcategorization using distributional cues, adult learners in this experiment receive very strong distributional cues to the subcategory structure in our artificial grammar. These cues involve a dense sampling of the subcategory-obeying strings generated by the grammar, with complete overlap of the presented contexts across the words within each subcategory. Importantly, there is no overlap of immediate distributional contexts for the words in different subcategories. By using adult participants, we can extend Reeder et al.'s (2013) findings and set a baseline for what to expect from a mature distributional learner. This will set the stage for future experiments with children and infants.

Our basic artificial grammar paradigm is similar to that of Smith (1969): learners receive no evidence that words from different subcategories share contexts, just as Smith's participants received no MQ or PN strings. However, in contrast to Smith (1969), our grammar provides additional distributional richness: it is larger and more complex, and it removes absolute position as a possible distributional cue ("M words come first; Q words come last"), since the absolute position of words (e.g., first, second, last in the string) is rarely useful in defining linguistic categories in natural languages. The question of interest is whether learners can acquire the subcategories embedded in our language from the cues of relative word order and immediate linguistic contexts, or whether they will tend to treat all words as belonging to a single category (as they did in Smith, 1969), despite the strong distributional information for separate subcategories. By asking this question, we can examine whether the types of distributional cues that are useful for single category acquisition (Reeder et al., 2013) are also used to acquire subcategories.

Method

Participants

Twenty-four monolingual native English-speaking students at the University of Rochester were paid to participate in Experiment 1. Participants were randomly assigned to one of two exposure languages such that twelve participants were exposed to Language 1 and twelve were exposed to Language 2. All participants were naïve to the materials and goals of this experiment.

Stimulus materials

Experiment 1 exposed learners to strings generated from an artificial grammar of the form **(Q)AXB(R)** (adapted from Reeder et al. (2013)). Each letter corresponds to a category of nonsense words: **X** was the target category of interest and contained 4 words; **A** and **B** were contexts for X, and each contained 6 words; **Q** and **R** were optional flanker categories, and each contained 2 words. Based on the structure of the language, words were concatenated to build strings of the form **AXB**, **QAXB**, **AXBR**, or **QAXBR**. The optional presentation of the Q and R categories prevented learners from relying on absolute position as a cue for categorization.

Because there were 6 A-words, 6 B-words, and 4 X-words, there could be 144 unique AXB combinations generated by our grammar. Importantly, in contrast to Reeder et al.'s (2013) experiments, here we divided the X category into two subcategories by restricting which A- and -B contexts occurred with each X-word. Subcategory 1 contained words X_1 and X_2, which only occurred with A_1, A_2, A_3, B_1, B_2, and B_3 as allowable contexts. Subcategory 2 consisted of words X_3 and X_4, which had only A_4, A_5, A_6, B_4, B_5, and B_6 as allowable contexts (see Fig. 1).

This subcategory structure reduced the number of legal AXB combinations that could be presented during exposure to 36 AXB strings (3 A-words $*$ 2 X-words $*$ 3 B-words $*$ 2 subcategories). Of these 36 AXB sentence types, 24 were presented during exposure and 12 were withheld for testing (see Table 1). The exposure strings were selected such that every X-word was seen with every subcategory-appropriate A-word and B-word, but with *no* A-words or B-words from the other subcategory. In the terminology of Reeder et al. (2013), this created a *dense* sampling of the possible contexts for X, because learners were exposed to 2/3rds of the possible subcategory-appropriate contexts for each X-word. Additionally, there was *complete overlap* of contexts across X-words within each subcategory. This is because each X-word is seen with all of the same context words as the other X-word in its subcategory. There was complete *non-overlap* of contexts between the X-words from different subcategories. The sparseness and overlap

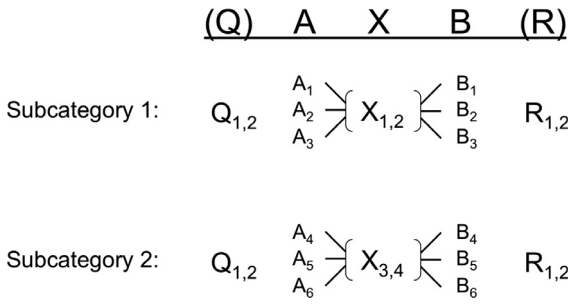

Fig. 1. Pictorial depiction of subcategorization structure for Experiment 1.

Table 1
Possible legal AXB combinations for Experiment 1. Combinations that appeared in the exposure for Experiment 1 are bolded. Subcategory 1 consisted of X_1 and X_2, with A_1, A_2, A_3, B_1, B_2, and B_3 as legal contexts for X_1 and X_2. Subcategory 2 consisted of X_3 and X_4, with A_4, A_5, A_6, B_4, B_5, and B_6 as legal contexts for X_3 and X_4.

Subcategory 1		Subcategory 2	
$A_1X_1B_1$	$A_1X_2B_1$	**$A_4X_3B_4$**	**$A_4X_4B_4$**
$A_1X_1B_2$	**$A_1X_2B_2$**	**$A_4X_3B_5$**	$A_4X_4B_5$
$A_1X_1B_3$	**$A_1X_2B_3$**	$A_4X_3B_6$	**$A_4X_4B_6$**
$A_2X_1B_1$	**$A_2X_2B_1$**	**$A_5X_3B_4$**	$A_5X_4B_4$
$A_2X_1B_2$	**$A_2X_2B_2$**	$A_5X_3B_5$	**$A_5X_4B_5$**
$A_2X_1B_3$	$A_2X_2B_3$	**$A_5X_3B_6$**	**$A_5X_4B_6$**
$A_3X_1B_1$	**$A_3X_2B_1$**	$A_6X_3B_4$	**$A_6X_4B_4$**
$A_3X_1B_2$	$A_3X_2B_2$	**$A_6X_3B_5$**	**$A_6X_4B_5$**
$A_3X_1B_3$	**$A_3X_2B_3$**	**A_6X3B_6**	$A_6X_4B_6$

Table 2
Assignment of nonsense words to categories for Language 1 and 2.

Language 1				
Q	A	X	B	R
mib bliffin	flairb daffin glim gentif spad klidum	tomber zub lapal roy	fluggit bleggin mawg frag zemper nerk	prog dilba
Language 2				
Q	A	X	B	R
zub klidum	fluggit tomber mawg glim dilba frag	nerk bleggin zemper prog	daffin roy spad lapal mib bliffin	flairb gentif

of the exposure set were thus comparable to Experiment 1 of Reeder et al. (2013).

The optional 2 Q- and 2 R-words were added to these 24 AXB combinations to create AXB, QAXB, AXBR, and QAXBR strings. Each Q and R flanker word was equally frequent across each subcategory, so their presence was not an informative cue to the identity of each string's subcategory structure. Overall, this led to 96 unique (Q)AXB(R) strings in the input.

The 20 words in the language were recorded in isolation by a native English-speaking female with a terminal and a non-terminal list intonation for each. They were adjusted in Praat (Boersma, 2001) so that the pitch, volume, and duration of syllables were qualitatively consistent. Words were assigned to each of the categories in order to achieve a relative balance of phonological and syllabic properties across the categories. To insure that our word-to-category mapping was not biased in any way, we created two separate languages (Languages 1 and 2) that differed only in different assignments of words to categories (see Table 2). Strings were created by concatenating the recordings of individual words, with 50 ms of silence between each word and using a terminal intonation token at the end of each string.

Crucially, the only systematic cues to the category and subcategory structure of this language were relative order of the words and the distribution of A- and -B context words. There were no semantic or systematic phonological cues to the categories, and there was no referential world attached to this language.

Procedure

Participants were informed that they would be listening to sentences from a new language that they had never heard before. Their task was to try and pay attention to the sentences, because they would be tested on their knowledge of the language later. Participants were seated at a Dell desktop PC in a sound-attenuated booth and passively listened to the sentences via headphones. The exposure phase consisted of five repetitions of the 96 (Q)AXB(R) training strings presented in pseudo-random order (a total of 480 strings). There were 1.5 s of silence between each string, leading to an exposure phase of approximately 30 min.

Once the exposure phase was complete, participants began the testing phase of the experiment. As in Reeder et al. (2013), each test trial began by hearing a 3-word sentence and then rating that sentence on a scale from 1 to 5, based on whether or not the sentence came from the language heard during the exposure phase. 1 meant that the sentence definitely did not come from the language; 2 meant the string may not have come from the language; 3 meant that the string may or may not have come from the language; 4 meant the string may have come from the language; 5 meant the string definitely came from the language. Participants were instructed to rate test strings based on their "gut reaction" to each sentence and whether they thought a native speaker of the language would have said that particular sentence when following the rules of the language's grammar. There were four types of test strings: *grammatical familiar* (12 of the 24 AXB strings that were presented during exposure), *grammatical novel* (the 12 AXB strings from 36 possible strings that obeyed the grammar's subcategory restrictions but were withheld from exposure), *subcategory violation* (12 of the 72 AXB strings that violated the subcategory structure of the grammar, but still obeyed the overall AXB word order of the grammar), and *ungrammatical* (6 AXA and 6 BXB strings that violated the word order of the grammar without repeating word tokens). Subcategory violation strings contained *either* the A-word *or* the B-word from the opposite subcategory for that X-word, in the correct word order position. In other words, either the A- or the B-words came from the opposite subcategory as the X-words, but not both. Importantly, these strings would be grammatical if learners ignored the subcategory structure of the language and generalized to form a single X category. A difference in ratings between grammatical and subcategory violation strings therefore indicates that participants have learned the subcategory structure in the language and are not generalizing across the gaps created by this boundary. Each test string was presented twice during the test phase, in pseudo-random order.

Results

Because there were no significant differences between the ratings of learners exposed to language 1 versus language 2, we collapsed these two languages for all subsequent analyses. The mean rating of grammatical familiar strings was 3.61 ($SE = 0.10$), the mean rating of grammatical novel strings was 3.70 ($SE = 0.11$), the mean rating of subcategory violation strings was 3.31 ($SE = 0.12$), and the mean rating of ungrammatical strings was 2.55 ($SE = 0.12$) (see Fig. 2). A repeated measures ANOVA with test item type (grammatical familiar, grammatical novel, subcategory violation, and ungrammatical) as the within-subjects effect was conducted. Mauchly's test revealed that the assumption of sphericity was not met ($\chi^2 = 20.50$, $p < 0.05$), so the Greenhouse-Geisser correction was used ($\varepsilon = 0.656$). There was a significant main effect of test item type ($F(1.97, 45.23) = 34.238$, $p < 0.001$). Planned comparisons revealed no difference between grammatical familiar and grammatical novel strings ($t(23) = 1.24$, $p = 0.23$), but subcategory violation strings were rated significantly lower than familiar grammatical strings ($t(23) = 3.14$, $p < 0.01$) and novel grammatical strings ($t(23) = 3.47$, $p < 0.005$). Ungrammatical strings were rated the lowest, significantly lower than subcategory violation strings ($t(23) = 4.42$, $p < 0.001$).

Because different participants may exhibit biases in how they use our rating scale, we converted all raw ratings scores into z-scores and conducted another repeated measures ANOVA. Results were all qualitatively the same.

Discussion

As in Reeder et al. (2013), learning effects were observed based solely on the distribution of words and their surrounding contexts. However, it is important to note that the distributional cues in the present experiment are balanced quite differently than in Reeder et al. (2013) and Smith (1969). To instantiate a subcategory structure, while all of the words in category X occurred in the same relative word order positions, there were two subsets of X-words that differed in their immediate distributional contexts. X_1 and X_2 had strong overlap in the particular A- or -B context words with which

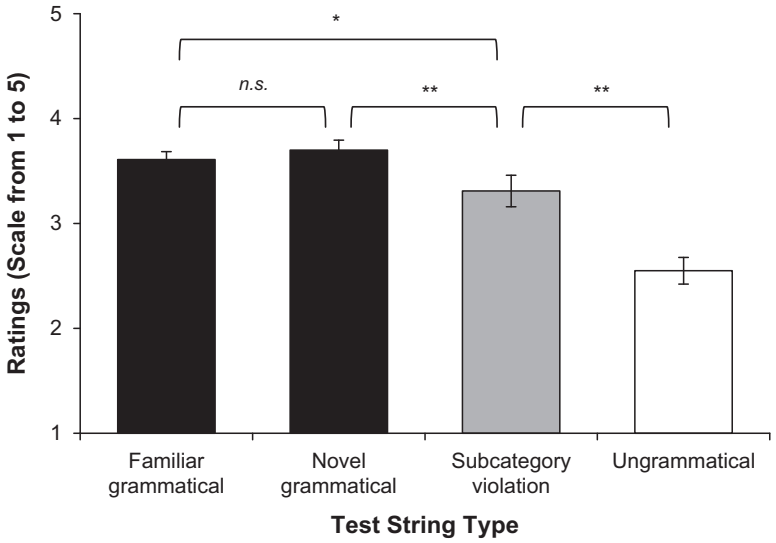

Fig. 2. Mean ratings from Experiment 1, comparing familiar, novel grammatical, subcategorical violation, ungrammatical test strings. * Significant at $p < 0.05$, ** significant at $p < 0.01$, error bars are standard error.

they occurred, but these X-words had no A- or -B contexts in common with X_3 or X_4. The complementary situation also applied to X_3 and X_4. Additionally, the optional Q and R flankers removed the absolute positional cues that were present in Smith's (1969) MN/PQ problem, which may have suggested to learners that all X-words belonged to the same large X category. The important result is that, under these circumstances, learners succeeded in acquiring the subcategory structure.

Firstly, Experiment 1 showed that learners *generalized* within subcategories when the co-occurrence statistics showed high overlap among X-words. Despite not hearing every legal AXB combination, learners rated grammatical novel AXB strings just as highly as familiar strings, indicating that they were willing to extend the appropriate withheld contexts to X-words within the same subcategory. Second, learners *restricted generalization* to contexts of the opposite subcategory when the distributional information suggested that these were systematic gaps created by subcategory boundaries. This was evidenced by significantly lower ratings of subcategory violation strings than either grammatical familiar or grammatical novel strings (see Fig. 2), suggesting that learners were not willing to share distributional properties across the subcategory boundary. In addition, the results of Experiment 1 demonstrate that learners are highly sensitive to the type and systematicity of the *missing* information in their input. Our subcategory paradigm created a perfect boundary, with no overlap of contexts across subcategories but perfect overlap of contexts within subcategories. Learners interpreted the consistent absence of subcategory-crossing strings from their input as purposeful (not accidental) omissions, signaling ungrammatical contexts for certain X-words. On the other hand, the more sparse and less systematic gaps within subcategories were not interpreted in this way.

Interestingly, however, subcategory violation strings were rated significantly higher than ungrammatical strings. As noted above, subcategories present two types of cues for learners: overall word order cues indicate that the words in different subcategories fall into the same major category, while immediate contextual cues indicate that they are in distinct subcategories. In our artificial grammar, the presence of stable flanker words (Q and R) and relative word order across all strings regardless of their subcategory membership suggests that all X words belong to the same category. However, the immediate A- and -B context cues suggest that X_1 and X_2 are fundamentally distinct from X_3 and X_4.[2] We interpret the finding that ratings for subcategory violations are lower than novel grammatical strings but higher than ungrammatical strings as further evidence that learners have acquired subcategories in an appropriate way. Though native adult speakers of a language would rarely (if ever) intentionally produce a subcategory violation, these types of errors do sound relatively more acceptable than a word order violation like those used in our ungrammatical test strings (e.g. *John disappeared the rabbit* vs. *John rabbitted the disappear.*). This difference is reflected in our ratings results, and we believe this reflects acquisition of both category-level (X) representations and subcategory-level ($X_{1,2}$ vs $X_{3,4}$) representations.

The results of Experiment 1 thus show successful subcategory acquisition using a distributional learning paradigm with no correlated phonological or semantic cues. Experiment 2 further explores the process of subcategory acquisition in a more complex learning environment.

Experiment 2

The miniature language used in Experiment 1 was a highly idealized case of the types of distributional cues one might encounter in a natural language learning situation. One way in which our miniature language could be modified in order to further test the limits of this type of statistical learning would be to make the subcategory boundaries imperfect in ways that are typical of real languages. One example in natural language is the case of homophony or homonomy of words that fall into different categories. For example, in the question "*What's black and white and* /rɛd/ *all over*," it is unclear whether the word /rɛd/ should be the past tense of the verb *read* or the color *red*. The same phonological form appears in two separate grammatical categories, as it does in other words like

[2] As mentioned in the Introduction, due to the small artificial grammar used here, an argument could be made that learners acquired multiple form-class categories rather than two subcategories. Given the distributional cues described here, we believe the most consistent interpretation of our grammar is that X is a form-class category with two subcategories. However, regardless of one's interpretation of the learner's abstract representation of the grammar, the demonstration of successful (sub)category acquisition in Experiment 1 informs us about the useful and usable distributional evidence for this type of learning problem.

fish. With words like this, two different sets of distributional contexts will appear to belong to a single word. This type of problem is likely overcome by a statistically oriented distributional analysis, noting that most words distinguish these two sets of contexts.

This type of ambiguity is found at many different levels of language (e.g., at the lexical level, like read/red; at the morphological level, see Pertsova (2008); syncretism, see Baerman, Brown, & Corbett, 2005). Such ambiguity also exists in linguistic subcategories. One example is the "crossed" gender system of a language like Romanian (Corbett, 1991, 1994). As in many other languages, Romanian divides nouns into masculine and feminine linguistic gender subcategories that are reflected in distinct sets of agreement markers. However, there is a sizeable set of nouns that take masculine agreement markers when singular, but feminine agreement markers when plural. This third class of nouns does not belong with either masculine or feminine nouns, and thus they are said to form a third gender (neuter). However, Romanian does not have a separate class of agreement markers that occur with neuter nouns. In Corbett's (1994) terminology, Romanian has three controller gender subcategories (the number of genders of nouns) but only two target gender subcategories (the number of classes of agreement markers).

Real languages commonly have category-crossing exceptions that are similar to the cases of homophony or the Romanian gender system. These exceptions arise from situations where the syntax makes one set of distinctions but other systems (morphology, phonology) do not always make the same distinction. What is remarkable about arbitrary systems like these is that children are capable of acquiring them, without explicitly being told that some contexts for words are highly specific while others are quite general. Learners must be capable of using the detailed patterning of these cues to determine when to generalize and when to encode something exceptional about the uses for a class, a subclass, or a token. Experiment 2 explores this process and expands on the results from Experiment 1 by introducing an "exception": a single string that crosses the subcategory boundary. We then compare the results from Experiments 1 and 2 to see whether subcategory boundaries entirely break down because of the distributional complications of this one exceptional boundary-crossing context, or whether learners maintain the subcategory structure and treat the boundary crossing string as a special exception.

Method

Participants

Eighteen monolingual native English-speaking students at the University of Rochester were paid to participate in Experiment 2. None of these individuals had participated in Experiment 1. Three subjects were removed for failing to complete the experiment according to the instructions (reporting that they did not pay any attention during training due to its length, failing to perform above chance at a 1-back task during training and subsequently rating all test items the same). Seven participants were assigned to language 1, and eight were assigned to language 2.

Stimulus materials

The stimuli were the same as in Experiment 1 (see Table 1) except for the addition of one "boundary-crossing" string to the exposure set. This string paired X_1 (from Subcategory 1) with the context $A_4\_B_4$ (a context from Subcategory 2) to create an exception to the subcategory boundary: $A_4X_1B_4$ (see Fig. 3). As in Experiment 1, the exposure consisted of a dense sampling of A- and -B contexts, with complete overlap of contexts within each subcategory. Of the 36 legal subcategory-obeying AXB combinations, 24 were presented and 12 were withheld for testing generalization. After the addition of the boundary-crossing string and optional Q and R flanker words (as in Experiment 1), the training set consisted of 100 strings. As in Experiment 1, this set of 100 strings was presented 5 times during exposure in uniquely randomized orders.

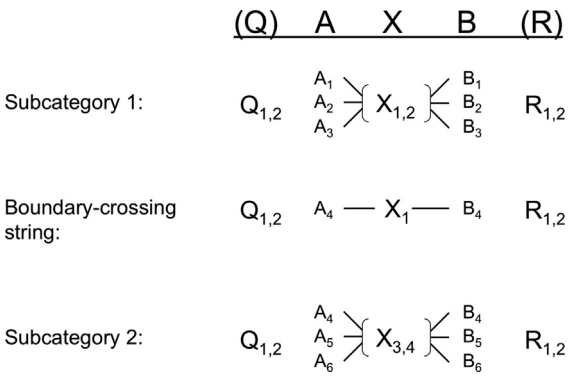

Fig. 3. Pictorial depiction of the subcategory structure for Experiment 2. There is a clear boundary between the A- and -B contexts for Subcategory 1 (X_1 and X_2) versus Subcategory 2 (X_3 and X_4), except for the single boundary-crossing string $A_4X_1B_4$.

Procedure

As in Experiment 1, participants were instructed to pay attention as they listened to strings from a new language, as they would be tested on their knowledge of the language later. Preliminary evidence suggested that the somewhat longer training phase (500 vs. 480 strings) created more difficulty for participants in maintaining their attention. Thus participants were told that while listening to the strings, they would have to complete a 1-back task.[3] The purpose of this secondary task was to keep participants attentive to the materials. At the beginning of the exposure phase, participants were instructed to make a hatch mark on a sheet of paper every time they thought they heard a repeated sentence. They were told that the number of repeated sentences was randomized for each subject, so there may be many repeated sentences or none at all.

After exposure, participants entered the test phase of the experiment. The instructions were the same as in Experiment 1. In order to fully investigate how category representations were affected by the presence of the single boundary-crossing string, a number of additional test strings were constructed.[4] There were 13 familiar AXB test strings: 12 *grammatical familiar*, subcategory-obeying AXB strings that were presented during exposure (same as Experiment 1), plus the 1 *familiar boundary-crossing* $A_4X_1B_4$ string. There were 12 *grammatical novel*, subcategory-obeying AXB strings (same as Experiment 1), and 4 *novel boundary-crossing* strings. These novel boundary-crossing strings combined X_1 with novel A- and -B con-

[3] A pilot experiment replicating Experiment 1 but using this 1-back task showed no change in performance relative to the original results: planned comparisons on the results of 6 participants (3 in each of languages 1 and 2) showed no difference between grammatical familiar and novel strings ($p = 0.81$), but a significant difference between grammatical novel and subcategory violation strings ($p < 0.05$) and also between subcategory violation strings and ungrammatical strings ($p < 0.01$). As these results are qualitatively the same as in Experiment 1, we felt justified in using the 1-back task as a means to increase attention during the exposure phase while allowing implicit learning of the rules of the language. Note that the 1-back task does not provide any explicit information about the grammatical structure of the exposure strings, but it does require attention to the sound sequences. Failure to notice at least half of the repeated strings was directly related to participants reporting that they were not paying attention during training ($N = 3$), and these participants were removed from our final sample.

[4] Compared to Reeder et al. (2013) and Experiment 1, this experiment had a different proportion of clearly grammatical (i.e., familiar) test items to clearly ungrammatical test items. However, pilot results from a replication of Reeder et al. (2013) Experiment 1 showed that a test with half ungrammatical strings and half familiar plus novel grammatical strings did not lead to qualitatively different results. Participants did not appear to be biased towards rating approximately half of the items as ungrammatical.

texts from Subcategory 2 (such as $A_5X_1B_6$). Importantly, all of the A- and -B contexts in these test types appeared (with other items) during the exposure phase, so a low rating of these strings with X_1 would confirm that learners are not simply responding on the basis of the A- and -B contexts themselves. Together these boundary-crossing test items assess whether X_1 is represented as allowing all contexts from both subcategories or only allowing the specific boundary-crossing context in which it was presented during exposure. There were 12 *subcategory violation "type 1"* strings, in which either the A- *or* the B-word came from the opposite subcategory as X (same as the subcategory violation strings in Experiment 1). There were also 4 *subcategory violation "type 2"* strings in which *both* the A- and the B-word came from the opposite subcategory as X (two X_2 strings, one X_3 string, and one X_4 string). Lastly, there were 12 ungrammatical word-order violation strings of the form AXA, BXB, AAB or ABB. Therefore, the full test contained 82 strings presented in pseudorandom order: 26 familiar AXB test strings (12 strings heard during exposure, repeated twice; 1 familiar boundary-crossing string, repeated twice), 44 novel AXB test strings (12 subcategory-obeying novel strings repeated twice, as in Experiment 1; 4 novel boundary-crossing strings, with X_1 in novel A_B contexts from subcategory 2; 12 subcategory "type 1" violation strings; 4 subcategory "type 2" violation strings), and 12 ungrammatical (word order violation) test strings.

Just as in Experiment 1, the subcategory violation strings would be judged as grammatical if learners did not acquire a subcategory boundary or if this distinction was weakened by exposure to the boundary-crossing string. The remaining types of novel strings distinguish between whether learners treat all novel AXB's as exceptions, whether they generalize all Subcategory 2 contexts to X_1, whether they generalize Subcategory 2 contexts to all Subcategory 1 members, and whether the existence of the boundary-crossing string leads to the loss of a subcategory boundary.

Results

Due to the length of the test, split-half reliability item analyses were conducted to see if order effects were present between the first and second half of the test. Cronbach's alpha (Cronbach, 1951) was computed for the ratings of each type of test string (familiar, familiar boundary crossing, grammatical novel, novel boundary-crossing, subcategory violation 1, subcategory violation 2, ungrammatical), and all Cronbach's alpha values were "acceptable" values for internal test reliability (lowest $\alpha = 0.691$, for subcategory violation type 1). This indicates that the ratings within a test type for the second half of the test were consistently correlated with ratings of that test type from the first half. We conclude that the longer test length did not significantly affect participants. Furthermore, no difference was found between language 1 and language 2, so the two languages have been collapsed for all remaining analyses.

The first comparisons of interest are those used in Experiment 1 to explore the relationships among words within a subcategory and across the subcategory boundary. Looking at only test strings that did not involve a category-boundary crossing, a repeated measures ANOVA was conducted with four test types (familiar, grammatical novel, subcategory violation type 1, ungrammatical) as the within-subjects factor, collapsing across languages 1 and 2. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2 = 25.71$, $p < 0.01$), so Greenhouse-Geisser correction was applied ($\varepsilon = 0.478$). The ANOVA revealed a significant main effect of test type ($F(1.435, 20.087) = 16.92$, $p < 0.001$). Planned comparisons showed that grammatical novel strings (mean = 3.61, $SE = 0.10$) were rated just as high as familiar strings (mean = 3.64, $SE = 0.09$), $t(14) = 0.634$, $p = 0.54$. Subcategory violation type 1 strings (mean = 3.42, $SE = 0.105$) were rated significantly lower than grammatical familiar strings ($t(14) = 3.57$, $p < 0.005$). These results indicate that participants generalized fully from familiar strings to novel strings within the subcategory boundaries, but they did not fully extend this generalization to strings where either A or B was from the opposite subcategory as X (the subcategory violation type 1 strings). The mean rating of ungrammatical items was 2.81 ($SE = 0.18$), which was significantly lower than subcategory violation strings ($t(14) = 3.90$, $p < 0.005$) (see Fig. 4). The same repeated measures ANOVA was conducted on z-scores, and the same qualitative pattern of results was obtained.

The next comparison of interest is among the boundary-crossing test strings: the familiar boundary-crossing $A_4X_1B_4$ string, the novel boundary-crossing strings in which X_1 was presented

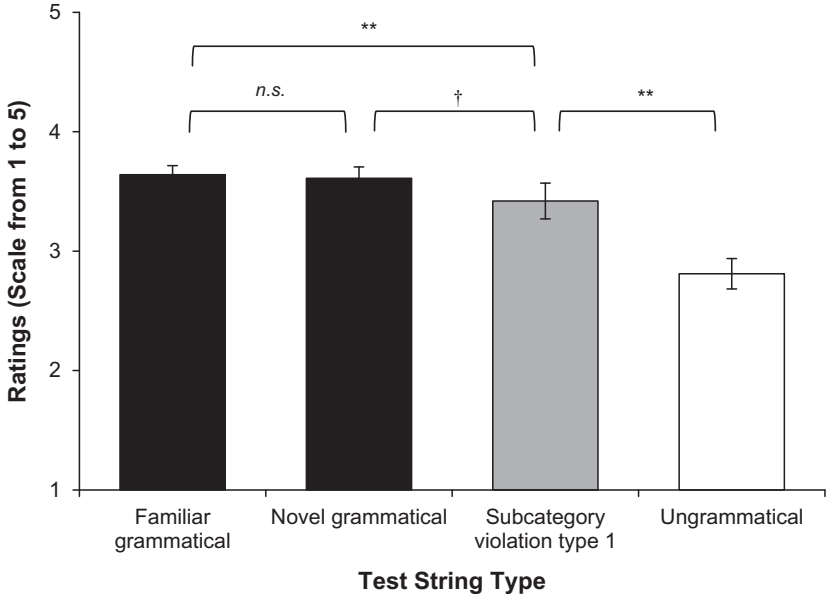

Fig. 4. Mean ratings for non-boundary-crossing test strings in Experiment 2. † Marginally significant at $p = 0.05$, * significant at $p < 0.05$, ** significant at $p < 0.01$, error bars are standard error.

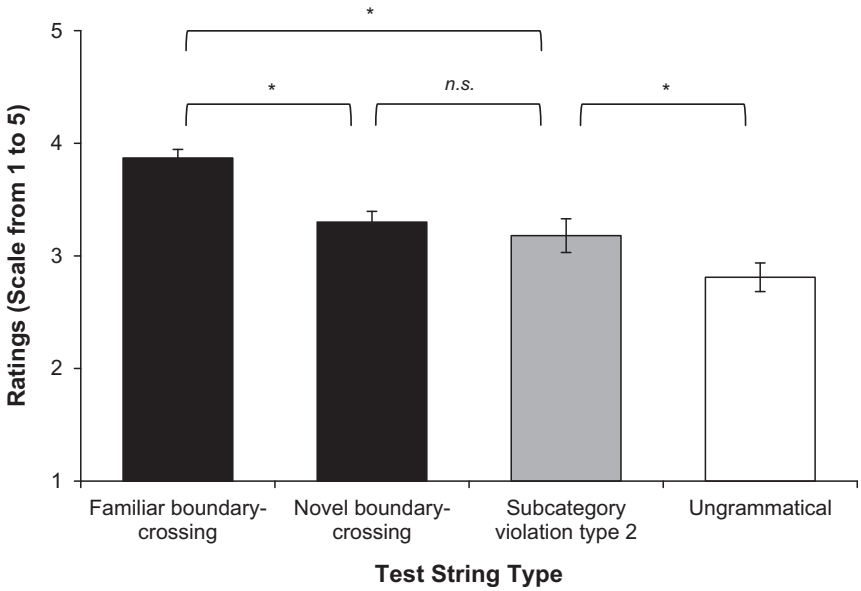

Fig. 5. Mean ratings for boundary-crossing strings from Experiment 2, comparing the familiar boundary-crossing string ($A_4X_1B_4$) to novel boundary-crossing strings, subcategory violation type 2 strings, and ungrammatical (AXA, BXB, AAB, ABB) strings. * Significant at $p < 0.05$, ** significant at $p < 0.01$, error bars are standard error.

with other A_B contexts from Subcategory 2 (e.g., $A_4X_1B_6$), and the subcategory violation type 2 strings where X_2, X_3 and X_4 were presented with A and B contexts from the opposite subcategory as X (e.g., $A_5X_2B_5$). The mean rating of familiar boundary-crossing strings was 3.87 ($SE = 0.16$), the mean rating of novel boundary crossing strings for X_1 was 3.30 ($SE = 0.16$), and the mean rating of subcategory violation type 2 strings was 3.18 ($SE = 0.16$) (see Fig. 5).

A repeated measures ANOVA was conducted with test type as the within-subjects factor. Mauchly's test revealed a violation of the assumption of sphericity ($\chi^2 = 6.54$, $p < 0.05$), so Greenhouse-Geisser corrections were used ($\varepsilon = 0.717$). The results of the ANOVA revealed a significant main effect of test type ($F(1.433, 20.068) = 5.34$, $p < 0.05$). Planned comparisons showed that familiar boundary crossing strings were rated significantly higher than novel boundary crossing strings ($t(14) = 2.17$, $p < 0.05$) and subcategory violation type 2 strings ($t(14) = 2.71$, $p < 0.05$), but novel boundary crossing strings were rated the same as subcategory violation type 2 strings ($t(14) = 0.89$, $p = 0.39$). This indicates that learners treated $A_4X_1B_4$ (the familiar boundary crossing string) as an exception, and they did *not* generalize across the subcategory boundary to other possible A_B contexts for that X-word or for other X-words.

Finally, the subcategory violation type 2 test strings can be divided into two groups: X_2 strings (belonging to Subcategory 1) and X_3 and X_4 strings (belonging to Subcategory 2). It was possible that learners would interpret hearing the $A_4\_B_4$ context with X_1 to mean that all the X-words in Subcategory 1 should have this extension. It was also possible that, due to this boundary-crossing context, learners would be more flexible overall in generalizing all of the contexts for Subcategory 1 members. However, a planned comparison of the X_2 subcategory violation type 2 test strings versus the X_3 and X_4 strings showed no difference between them ($p > 0.5$). Furthermore, comparing the X_2 subcategory violation type 2 test strings to the novel boundary crossing strings (X_1 strings with novel A_B contexts from Subcategory 2) showed no difference ($p > 0.1$). These results indicate that learners treated the unique $A_4X_1B_4$ string as an exception; they did not extend the $A_4\_B_4$ context to the other member of Subcategory 1, nor did they generalize the members of Subcategory 1 more broadly to encompass other contexts of Subcategory 2).

Discussion

As in Experiment 1, we found that a dense sampling of a language with complete overlap among contexts within subcategories – and almost no overlap across subcategories – leads to generalization within but not across subcategories. Learners generalized to novel combinations within subcategories, but rejected novel combinations that crossed the subcategory boundary, either with one inappropriate context element (subcategory violation type 1 strings) or two inappropriate context elements (subcategory violation type 2 strings). The results also show that learners did not extend the single $A_4\_B_4$ context that appeared with X_1 to the other member of X_1's subcategory (i.e., X_2). The low ratings of novel boundary crossing test strings demonstrate that participants did not treat X_1 as belonging to both subcategories. By rejecting both types of subcategory violations strings, learners demonstrated that they did not become more flexible with their subcategory boundaries from being exposed to the single boundary-crossing string. In essence, learners behaved the same as they did in Experiment 1, but maintaining a specific exception to the subcategory boundary (the unique $A_4X_1B_4$ boundary crossing string). These data show that learners can acquire subcategories even in the face of some input strings in which subcategory boundaries are violated, and also that they can learn an exception, such as X_1 being legal in a context typically restricted to Subcategory 2. The absence of information licensing other subcategory boundary crossings apparently is adequate for learners to determine that $A_4X_1B_4$ is a special case and that boundary crossing is not allowed for other X-words.

Experiment 2 answers the question of how learners use distributional information to categorize their input if their exposure to the language contains imperfect distributional cues in the form of specific exceptions to subcategory boundaries. The boundary-crossing strings they heard could have been viewed as evidence of a single major X-category, with no subcategory structure. However, the results from Experiment 2 show that learners considered the boundary-crossing string to be an exception; they showed the same overall effects as in Experiment 1, encoding the subcategory boundary consistent with the majority of the distributional information. These experiments strengthen the idea – suggested in Reeder et al. (2013) – that the absence of a cue in the input can

be just as informative as the presence of a cue. It is clear that learners were highly sensitive to the lack of information licensing novel boundary crossing strings or supporting the formation of a single category without subcategory boundaries.

General discussion

The present experiments provide robust evidence that learners can use distributional information in a sophisticated way in order to form linguistic subcategories. We showed that learners are highly rational in their interpretation of particular distributional cues, especially the amount and type of overlap across lexical contexts. Learners used this information to determine whether to restrict generalization across gaps in their input, generalize across gaps, or maintain a particular context as an exception to the overall rules of the grammar. Our results support the idea that language learners take advantage of several different co-occurrence statistics in parallel in order to learn word order, form categories, and discover the subcategories of their language. As described in Reeder et al. (2013), these cues include the overlap of contexts across words and the density of sampling of a particular category.

Given the results of earlier studies arguing that subcategories are not learnable without multiple correlated cues (e.g., Braine, 1987, which suggests that perceptual or semantic cues are required for successful subcategory learning), it is important to consider why we found evidence of subcategory acquisition here when others have not. Successful distributional learning depends on the types of distributional cues and their patterning in the input, especially (a) whether this information is readily available and provides reliable cues to the learnable structures in the environment (i.e., is it *useful*), and (b) whether the learner is able to access this information and perform the necessary computations over it (i.e., is it *usable*). One important distinction between our results and previous studies is our framework for defining subcategorization. As described in the introduction, linguistic subcategorization crucially involves a conflict of distributional cues. When a major category contains multiple subcategories, a learner must recognize that the overall word order information signals a major X-category, but their co-occurrence with some context elements and the absence of their co-occurrence with other context elements – that is, contextual gaps formed by the subcategory boundary – signal multiple sub-categories inside X.

First, in contrast to these characteristics, consider the MN/PQ problem – the classic case of failure to acquire subcategories given distributional information alone. Participants in Smith's (1969) task learned which words could occur first in a string (M, P) and last in a string (N, Q), but not the dependency that N-words could only follow M-words and Q-words could only follow P-words. Participants produced both legal PQ strings *and* illegal MQ overgeneralizations during test, indicating that they were equally grammatical to participants. Using our terminology, the distributional information available during exposure for Smith (1969) did consist of a dense sampling of the language, with complete overlap of contexts within subcategories and complete non-overlap across subcategories. This is equivalent to the sampling density and contextual overlap available during the present Experiment 1, where we saw generalization within (but not across) subcategories. However, there are a number of differences between our task and previous artificial language investigations of subcategorization that can explain the dissimilarity in results.

One difference is that we used a rating scale, whereas other investigators have used recall, 2-alternative forced choice, binary choice (yes/no) responses, or other production measures to obtain information about the learner's category knowledge. The greater sensitivity of a rating scale may have allowed us to demonstrate distinctions between category representations that are not reflected in other measures.

On the other hand, the sensitivity of our rating scale may make it susceptible to revealing low-level surface differences in our stimuli rather than true subcategory representations. For example, in the current design, our test of subcategory knowledge is the difference in ratings between familiar strings and strings that contain one novel bigram (the AX or XB bigram that violates the subcategory structure) and a novel non-adjacency (such as $A_{subcategory1}$ – $B_{subcategory2}$).[5] Learners may rate subcategory violation strings lower because of these differences in surface statistics, which are not present in familiar versus novel grammatical strings. However, for a number of reasons we think it is unlikely that learners are relying solely on bigram statistics to guide their ratings. Reeder et al. (2013, Experiment 5A) demonstrated that bigram statistics do not provide a good account of category learning and generalization: learners in that study generalized to completely novel bigrams when there was strong distributional evidence that those word combinations should be licensed based on the category structure of the language. Additionally, in the present Experiment 2 we saw that adjacent and non-adjacent bigram familiarity did not boost ratings of certain types of ungrammatical, subcategory-violating test strings. Grammatical novel strings (where either the AX or XB bigram and the A_B frame were familiar) were rated higher than subcategory violation "type 1" strings (where either the AX or XB bigram was familiar, but the A_B frame was unfamiliar). Learners also rated novel boundary-crossing strings and subcategory violation "type 2" strings (where the AX and XB bigrams were both unfamiliar but the A_B frame was familiar) lower than either grammatical novel or subcategory violation "type 1" strings. These results suggest that participant ratings are not driven by adjacent bigram familiarity or non-adjacent bigram frame familiarity (though clearly bigram statistics must play a role in the initial stages of category and subcategory formation). Other studies of category learning have also found that learners generalize to novel bigrams if the category structure of the language licenses it. In a serial reaction time task, Hunt and Aslin (2010) found that learners were just as fast to produce novel transitions that were legal category extensions as they were to produce item transitions that they had practiced during training; but they were significantly slower to produce novel transitions that violated the category structure of the grammar. Lastly, work by Schuler et al. (in press) demonstrates that even when there is large variation in bigram and lexical frequencies, learners rely on the patterns of number, density, and overlap of linguistic contexts across words to determine category representations. Schuler et al. introduced Zipfian frequency differences in the X-words of the (Q)AXB(R) languages from Reeder et al. (2013). The words within the X category thus differed in lexical frequency by as much as a 7:1 ratio; this likewise created drastic differences in the bigram frequencies of familiar and novel grammatical test strings. Despite such bigram frequency differences, however, learners relied on the same category-level statistics as in Reeder et al. (2013) to determine whether to form a single X category and generalize to unseen legal combinations. Taken together, these results suggest that bigram frequency variations did not control how learners interpreted the distributional cues to categories in these previous studies. We therefore believe it is unlikely that learners in the present experiments relied on surface-level cues like bigram frequencies when rating grammatical novel test items. However, future work should incorporate lexical frequency variations into subcategory-learning paradigms such as ours to

[5] We thank Michael C. Frank and an anonymous reviewer for suggesting this possibility.

determine how a distributional learning mechanism handles these cues.

In related work, Mintz et al. (2014) have also shown that distributional information can be used as the sole cue for category acquisition. However, they argue that frequent frames (nonadjacent dependencies in the form of fixed A_B frames) are necessary in order to acquire categories and subcategories, and that a high degree of overlap of contexts across category members is not sufficient for category learning (as we show here). The language in Mintz et al.'s artificial grammar learning experiments is significantly larger than the artificial language used here, making it more naturalistic but also leading to proportionally larger gaps in the learner's input. From this perspective our results are consistent with theirs: in a sparse sampling of a large language without frequent frames (Mintz et al. (2014), Experiment 2), a learner is exposed to highly inconsistent overlap of contexts across words. Our results suggest that learners will not readily group these words into categories based on such weak evidence from the distributional information and will restrict generalization based on their input. This conclusion is supported by the results from Mintz et al.'s Experiment 2. However, if a language does contain frequent frames, as in Mintz et al.'s Experiment 1, learners are exposed to words with a significant degree of overlap across contexts, which licenses them to collapse the words into a category. Our present results extend those discussed in Mintz et al. (2014) by showing that learners can form category and subcategory representations if exposed to a sufficiently dense sampling of a language with consistent overlap of contexts across words, even when frequent frames are not present. In our view, this overlap can come in terms of immediate contexts (A_ or _B), or in surrounding nonadjacent dependencies (i.e., A_B or frequent frames). We have also demonstrated elsewhere that nonadjacent dependency learning cannot fully explain our results. In Reeder et al. (2013), learners encountered all possible A_B frames during exposure. The results showed that category learning could not be predicted by amount of exposure to A_B frames. Instead, it was predicted by the amount of contextual overlap with each X-word, sampling density, and the overall size of the exposure set (how often learners encountered "gaps" in their input).

Returning to the MN/PQ problem, Smith (1969) and similar artificial grammar learning tasks offer particular types of distributional cues that are so salient and quickly learned – sometimes misleading ones, such as absolute word position – that other more relevant distributional cues might be ignored. Indeed, much of the early work on category learning showed that positional information is just such a cue (e.g., Braine, 1965): if subjects attend to the salient absolute positions of words (which words occur first and last) and not the relationships between words or their co-occurrences, they will overgeneralize in the way that Smith and others have found. Positional information is so salient in the very short and length-invariant strings of the MN/PQ problem that a rational subject might never acquire the dependencies among the words forming the grammatical subcategories. (Indeed, participants in that experiment eventually memorized the specific letter pairs in the exposure before they acquired any dependencies for generalization.) However, absolute position information is not useful in natural language acquisition: natural linguistic categories are not defined by their absolute position in a sentence. Rather, the distributional information that defines natural language categories and subcategories are their linguistic contexts – the surrounding words and morphemes with which particular words occur. Indeed, syntactic categories and subcategories are *defined* via the use of distributions: two words with similar distributions (similar surrounding contexts) belong to the same grammatical category if they are syntactically interchangeable (Radford, 1988).

It is important to note that Braine's (1987) correlated cues version of the MN/PQ experiment, which added semantic cues, did not remove the salient positional cue to categories (being first or last in the two-word sentences). The purpose of the semantic cue was to signal the subcategories and hence overcome the positional cues to the main categories. In our own experiments, the optional Q and R category flankers meant that relative but not absolute positional information was available to learners: since Q and R could optionally begin and end a sentence, the A, X, and B words have relative but not fixed positions of occurrence. This is more like natural languages, which do not generally have fixed positions for words or word categories. Despite using a larger and more complex language, then, Experiment 1 demonstrated successful subcategory acquisition while using similar levels of sampling and context overlap as in the MN/PQ experiment.

We do not dispute that subcategorization is quicker and easier when there are multiple cues to the subcategory structure. Indeed, some subcategories in natural languages have partially correlated cues to subcategory structure (e.g., Monaghan et al., 2005), and as described earlier, many investigators have found successful subcategorization learning when distributional cues are correlated with phonological, semantic, or morphological cues (e.g., Gerken et al., 1999, 2005; Gomez & Lakusta, 2004). Some have also suggested that artificial grammar learning in a semantically-empty world significantly impairs syntax learning (e.g., Moeser & Bregman, 1972; though see Arnon & Ramscar, 2012, for evidence that distributional analyses may be impaired if the learner relies on some types of semantic cues). All of this evidence clearly shows that correlated perceptual cues are relevant for categorization when they are present and that learners can utilize correlated cues to induce categories in experimental settings.

Our goal is not to suggest that correlated cues have no value during linguistic categorization, or that only distributional information is useful for learning. Rather, our results add to a body of work that systematically explores how distributional information can be used in higher-order language acquisition (e.g., Reeder et al., 2013; Schuler et al., in press). By removing all other cues in our artificial grammar, we were able to investigate whether distributional information alone can allow a learner to acquire the category and subcategory structure of a language. This may explain how learners acquire arbitrary subclass systems in natural languages, where non-distributional information is not available, is very sparse, or is inconsistent with syntactic distribution. Gerken, Wilson, Gomez, and Nurmsoo (2009) have also suggested that early failures to find subcategory learning in artificial language studies may be due to the presence of a referential field that was irrelevant to category formation. Much like our arguments above regarding the saliency of positional cues, Gerken and colleagues have suggested that learners might be overly engaged in learning the associations between the semantic cues and lexical items rather than learning the structure of the language. Our experiments utilized an artificial grammar that was free of semantic and phonological cues in order to show that learners can, in fact, learn categorical structures without the need for perceptual cues. Several other examinations of subcategory acquisition involve morphological paradigm-completion methods for acquiring grammatical gender systems (e.g., Cyr & Shi, 2013; Gerken et al., 2005). While these experiments have the nice feature of utilizing natural language input, they do so by exposing the learner to isolated words (stems or their inflected forms), without giving the learner access to the full linguistic system that surrounds the paradigm. Our results indicate that the full contexts involved in expressing such gender systems (both linguistic and extra-linguistic) may provide enough distributional evidence to create initial subcategories, even without additional phonological or semantic cues. Future

work must clarify how different cues work together with a distributional learning mechanism to result in mature category representations.

Much of the grammatical categorization literature has focused on identifying the cues that learners rely on, rather than identifying precisely how these cues shape a learner's underlying category representation. We believe that the results described here and elsewhere (e.g., Reeder et al., 2013; Schuler et al., in press) suggest that learners can and do engage in a rational distributional analysis of their input in order to arrive at a reasonable representation of the underlying grammar. This conclusion is supported by modeling work described in Qian, Reeder, Aslin, Tenenbaum, and Newport (2012) and Qian (2014). Models that rely primarily on surface statistics (lexical bigrams) succeed in some types of categorization and subcategorization problems, particularly when the input has carefully balanced lexical frequencies and includes a dense sampling of the language (see also St. Clair & Monaghan, 2005; St. Clair et al., 2010). In Experiment 1, this type of model arrives at the same grammatical structure as our human participants in Experiment 1: when given a dense sampling of a language with complete context overlap *within* subcategories and no overlap *across* subcategories, lexical bigrams provide a significant amount of information about the subcategory structure, since the AX and XB bigrams of different X-subcategories never overlap. However, when given a very sparse sampling of a language (e.g., Reeder et al., 2013, Experiment 4), when a word only minimally overlaps with other words in the category (e.g., Reeder et al., 2013, Experiment 5), or when lexical bigram frequencies are not equated (e.g., Schuler et al., in press), the lexical bigram model breaks down. These versions of our artificial grammar are much more like the complex input that child language learners face. In these circumstances, models that hypothesize *categories* based on *category*-level bigrams (category-to-category transitions) better mirror the sophistication that human learners bring to their distributional analyses. By adding the ability to represent higher-order category structures, these models can adjust the granularity of their category and subcategory representations in a rational way based on the evidence (or lack thereof) that generalization is warranted, much the same way that adult learners behave across the Reeder et al. (2013) and present experiments. Though in the present paper we are not proposing a full model of category and subcategory acquisition, we believe that our results add to the body of literature suggesting that humans have a rational distributional learning mechanism which operates over multiple levels of analysis (surface-level and category-level) and can use the gaps in the input (or lack thereof) to determine when to generalize and when to restrict generalization.

Overall, our results add to a body of work highlighting the relevant distributional cues for forming initial categories and subcategories. However, there remain many unanswered questions about the details of the relevant distributional learning mechanism. First, our experiments do not clarify whether learners start with a "coarse grain" of representation and refine their category representations as they gain more evidence from the input, or whether they start with narrow category representations and eventually collapse words into broader categories once the evidence warrants. It is possible that a series of between-subjects experiments, each testing learners' representations after a particular amount of exposure, could address this question. Another approach would be to use a paradigm (as in Hunt & Aslin, 2010), which allows an online measure of the timecourse of learning, like serial reaction time, mouse-tracking, or anticipatory eye-tracking.

Additionally, the present experiments were conducted with adult participants; it is important to ask how young learners utilize distributional information, given their more limited cognitive resources. Schuler, Reeder, Lukens, Aslin, and Newport (in preparation) found that 5–7 year old children behave much the same as adults when acquiring the major categories of the Reeder et al. (2013) (Q)AXB(R) grammar. However, experiments using the more complex category and subcategory representations of the present paradigm, or experiments using highly unbalanced lexical frequencies as in Schuler et al. (in press), have not been done with children. Behavioral research with infants does suggest that these very young learners can utilize distributional information to create higher-order linguistic representations (e.g., Zhang, Shi, & Li, 2014). Future work is needed to determine whether young learners utilize the full array of distributional cues that we have explored in the present research.

Acknowledgements

We thank Amanda Robinson, Carrie Miller, and Anna States for assistance with stimulus creation and data collection. We also thank the Aslin-Newport lab at the University of Rochester for helpful comments on this work. This research was supported by NIH Grants HD037082 to RNA and ELN, DC00167 to ELN, by an ONR Grant to the University of Rochester, and by the Center for Brain Plasticity and Recovery at Georgetown University.

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292–305.

Baerman, M., Brown, D., & Corbett, G. G. (2005). *The syntax-morphology interface: A study of syncretism*. Cambridge: Cambridge University Press.

Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.

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

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

Braine, M. D. S. (1965). The insufficiency of a finite state model for verbal reconstructive memory. *Psychonomic Science, 2*, 291–292.

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., & Fisch, S. M. (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.

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.

Corbett, G. C. (1991). *Gender*. Cambridge: Cambridge University Press.

Corbett, G. C. (1994). Gender and gender systems. In R. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 1347–1353). Oxford: Pergamon Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

Cyr, M., & Shi, R. (2013). Development of abstract grammatical categorization in infants. *Child Development, 84*, 617–629.

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

Gerken, L. A., Wilson, R., Gomez, R., & Nurmsoo, E. (2009). The relation between linguistic analogies and lexical categories. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition*. Oxford: Oxford University Press.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249–268.

Gervain, J., & Werker, J. F. (2013). Learning non-adjacent regularities at age 0;7. *Journal of Child Language, 40*, 860–872.

Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178–186.

Gomez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*, 567–580.

Gomez, R. L., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*, 183–206.
Gvozdev, A. N. (1961). *Voprosy izučenija detskoj reči*. Moscow: Izd. Akadem. Nauk RSFSR.
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.
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*, 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. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393–424.
Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (bigram) parts. *Cognitive Psychology, 75*, 1–27.
Moeser, S. D., & Bregman, A. (1972). The role of reference in children's acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behavior, 11*, 759–767.
Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorization. *Cognition, 96*, 143–182.
Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259–305.
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.
Pertsova, K. (2008). *Learning form-meaning mappings in presence of homonymy: A linguistically motivated model of learning inflection*. Unpublished doctoral dissertation. The University of California, Los Angeles.
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.
Polinsky, M. (2008). Gender under incomplete acquisition: Heritage speakers' knowledge of noun categorization. *Heritage Language Journal, 6*, 40–71.
Qian, T. (2014). *Rational perspectives on the role of stimulus order in human cognition*. Unpublished doctoral dissertation. The University of Rochester.
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.
Radford, A. (1988). *Transformational grammar: An introduction*. Cambridge: University Press.
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. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology, 66*, 30–54.
Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R.N. (in press). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development*.
Schuler, K. D., Reeder, P. A., Lukens, K. R., Aslin, R. N., & Newport, E. L. (in preparation). Children can use distributional contexts to acquire grammatical categories in an artificial language.
Scott, R. M., & Fisher, C. (2009). 2-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and Cognitive Processes, 24*, 777–803.
Shi, R., Morgan, J. L., & Allopenna, P. (1998). Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*, 169–201.
Smith, K. H. (1966). Grammatical intrusions in the free recall of structured letter pairs. *Journal of Verbal Learning and Verbal Behavior, 5*, 447–454.
Smith, K. H. (1969). Learning co-occurrence restrictions: Rule learning or rote learning? *Journal of Verbal Learning and Verbal Behavior, 8*, 319–321.
St. Clair, M. C., & Monaghan, P. (2005). Categorizing grammar: Differential effects of preceding and succeeding contextual cues. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 1913–1918). Austin, TX: Cognitive Science Society.
St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition, 116*, 341–360.
Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology, 22*, 562–579.
Wilson, R. (2002). *Syntactic category learning in a second language*. Unpublished doctoral dissertation. The University of Arizona.
Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?": Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science, 20*, 619–626.
Zhang, Z., Shi, R., & Li, A. (2014). Grammatical categorization in Mandarin-Chinese-learning infants. *Language Acquisition, 22*, 104–115.