Distributional Cues to Grammatical Categorization:  
Acquiring Categories in a  
Miniature Artificial Grammar  

by  
Patricia A. Reeder  

Submitted in Partial Fulfillment  
of the  
Requirements for the Degree  
Doctor of Philosophy  

Supervised by  
Professor Elissa L. Newport  

Department of Brain and Cognitive Sciences  
Arts, Sciences and Engineering  
School of Arts and Sciences  

University of Rochester  
Rochester, New York  

2010
Dedicated to Anne, Bill, Mom & Dad,

who have provided unending love and support.
Curriculum Vitae

The author was born in Buffalo, New York on December 14, 1982. She attended Cornell University in Ithaca, New York from 2000 to 2004, and graduated with a Bachelor of Arts in Computer Science, and a Bachelor of Arts with Summa Cum Laude honors in Psychology. She came to the University of Rochester in the Fall of 2005 and began graduate studies in the department of Brain and Cognitive Sciences on a departmental training fellowship. She pursued her research in language acquisition under the advisement of Elissa L. Newport, and received a Master of Arts degree in Brain and Cognitive Sciences from the University of Rochester in 2009.
Acknowledgements

This research was supported by NIH Grants HD037082 to Richard Aslin and DC00167 to Elissa Newport, and by an ONR Grant to the University of Rochester.

Many people played an invaluable role in the completion of this dissertation. First, I must thank my dissertation advisor, Elissa Newport. Her advice and constructive criticism throughout my graduate career have been greatly appreciated. Without her patience and guidance, this dissertation could not have been written. I am also indebted to the other members of my committee, Richard Aslin and Jeff Runner, for their generous assistance and encouragement throughout the past five years.

I am indebted to the members of the Aslin-Newport lab, past and present. Special thanks go to Ben Faber, Neil Bardhan, Sara Finley, Liz Wonnacott, David Ruskin, Celeste Kidd, Katie Schuler, Sarah Davis, Lizz Karuza and Masha Fedzechkina for many valuable discussions. I extend my gratitude to Amanda Robinson, Carrie Miller, Anna States, Katie Schuler and Lily Schieber, who were the fantastic lab managers and RAs that made this work possible. Many thanks go to Kathy Corser for her kindness, moral support, guidance and assistance.

I am very fortunate to have a close network of friends who have managed to keep me sane during this process. All of my BCS peers and colleagues have been fantastic in an uncountable number of ways over the past five years. I must give very special thanks to Katie Carbary and Sarah Davis for being two of the best officemate-friends a girl could ever ask for, and to Neil Bardhan for his constant friendship and fantastic sense of humor. Outside of academia, Joel G., Christine M.A., Hiren S., Jon R., Josh S., Kelly C., Amy C., Sammy LJ R., and many other Buffalo, Ithaca, and Rochester friends have helped me stay grounded in what really matters. Much
gratitude and love goes to Karen Riedesel, for her continuous support and camaraderie over the past 9 years; and Meghan Riley, for her unshakable belief in me (and unending patience with me). Last but not least, I am grateful to Davis Glasser, who has been the calm during the storm, a formatting wizard, and a constant source of encouragement. I think everyone owes him a pat on the back for putting up with my type-A personality.

My most loving words are for my family. My grandparents, aunts, uncles, and cousins have all aided and encouraged me throughout this endeavor, and I would like to express my heart-felt gratitude to them all. In particular, this dissertation is dedicated to four people: the memory of my uncle, William B. Forte, who made sure every morning started with a song and a smile; my aunt (aka, mom-away-from-mom), (Saint) Anne M. Forte, who taught me the value of persistence and a good sense of humor, and has been an inspiring presence throughout my entire life; the memory of my amazing mother, Susanne M. Reeder, who taught me how to set high goals, and instilled in me the confidence and passion to achieve them; and my father, Charles W. Reeder, who is always calm when I’m stressed, compassionate when I’m down, goofy when I need a good laugh, and a fantastic “Gears of War” partner. All of these individuals have been role models for how hard work can reap infinite success, and I could not have accomplished any of this without them.
Abstract

A crucial component of language acquisition involves organizing words into grammatical categories and discovering relations between them. The organization of words into categories, and the generalization of patterns from some seen word combinations to novel ones, account for important aspects of the expansion of linguistic knowledge in the early stages of language acquisition. One hypothesis of how learners handle the problem of categorization is that they exploit distributional information in the input to discover the category structure of natural languages (e.g., Braine, 1987). However, given the information processing limitations of young children and the complexity of the computational processes that would be entailed, this hypothesis has often been thought to be an unlikely contributor to categorization. Indeed, many studies have argued that in order for a learner to successfully utilize distributional information for category learning, there must be multiple correlated cues to category structure in the input (e.g., Gomez & Gerken, 2000). This, however, has been somewhat of a puzzle: grammatical categories and subcategories in natural languages do not always have reliable phonological, morphological, or semantic cues (Maratsos & Chalkley, 1980). Rather, learners must utilize distributional cues about the linguistic contexts in which words occur to acquire such categories.

In this thesis we hypothesize that the patterning of tokens in a corpus of linguistic input is sufficient, along with a small set of learning biases, to extract the underlying structural categories in a language. We present a series of artificial
grammar learning studies that examine how distributional variables will shift learners from forming a category of lexical items to maintaining lexical specificity. We begin with a series of experiments testing whether learners can acquire a single category, generalizing from some instances of the distributional contexts of individual words in the exposure set (but some withheld) to the full range of contexts for all the individual words in the set. To do this, we vary a number of distributional variables to category structure and test how adult learners use this information to inform their hypotheses about categorization. Our results show that learners are sensitive to the contexts across words, the non-overlap of contexts (or systematic gaps in information), and the size of the exposure. These variables taken together determine whether learners fully generalize or preserve lexical specificity.

We also explore whether subcategories are learnable from distributional information, if the learner is given adequate overlap inside each subcategory and adequate non-overlap between subcategories. Contrary to much of the existing literature, our results demonstrate that learners can use distributional information alone to find subcategory boundaries. Furthermore, we demonstrate that learners are able to learn lexical exceptions in their input while maintaining subcategory boundaries. Lastly, we explore the role of frequency variation for individual items, and we show that learners can overcome variations in input frequencies in order to maintain a rational generalization strategy based on their exposure to the language.

Overall, our results show that learners are able to take account of a rich set of variables to aid them in grammatical categorization, including degree of overlap
among category members, amount of input, consistency of gaps and overlaps in the
input, and conflicts or consistency among distributional cues.
# Table of Contents

Curriculum Vitae ........................................................................................................ iii
Acknowledgments ......................................................................................................... iv
Abstract ......................................................................................................................... vi
Table of Contents .......................................................................................................... ix
List of Tables ................................................................................................................ xii
List of Figures ................................................................................................................ xiii
Foreword ........................................................................................................................ 1
Chapter 1. Introduction ................................................................................................. 2

## Chapter 2. Categorization ......................................................................................... 13

2.1 Experiment 1: Dense Sampling With Complete Overlap ...................... 13
    2.1.1 Method ........................................................................................................... 14
    2.1.2 Results .......................................................................................................... 19
    2.1.3 Discussion .................................................................................................... 20

2.2 Experiment 2: Sparse Sampling With Complete Overlap ...................... 21
    2.2.1 Method ........................................................................................................... 21
    2.2.2 Results .......................................................................................................... 22
    2.2.3 Discussion .................................................................................................... 23

2.3 Experiment 3: Sparse Sampling With Incomplete Overlap ................... 24
    2.3.1 Method ........................................................................................................... 24
    2.3.2 Results .......................................................................................................... 25
    2.3.3 Discussion .................................................................................................... 27
2.4 Experiment 4: Incomplete Overlap With Extended Exposure ........ 27
   2.4.1 Method ................................................................. 28
   2.4.2 Results .............................................................. 28
   2.4.3 Discussion ......................................................... 29

2.5 General Discussion .................................................... 30

Chapter 3. Subcategorization ............................................. 34

3.1 Experiment 5: Subcategorization With Dense Sampling,
   Complete Overlap .......................................................... 35
   3.1.1 Method ................................................................. 35
   3.1.2 Results .............................................................. 38
   3.1.3 Discussion ......................................................... 40

3.2 General Discussion .................................................... 41

Chapter 4. Crossing Subcategory Boundaries ....................... 46

4.1 Experiment 6: Subcategorization Without Consistent
   Subcategory Boundaries .................................................. 50
   4.1.1 Method ................................................................. 51
   4.1.2 Results .............................................................. 54
   4.1.3 Discussion ......................................................... 60

4.2 Experiment 7: Subcategorization Without Consistent
   Subcategory Boundaries: Sparse Sampling, Incomplete Overlap .... 62
   4.2.1 Method ................................................................. 63
   4.2.2 Results .............................................................. 64
   4.2.3 Discussion ......................................................... 67

4.3 Experiment 8: Subcategorization Without Consistent
   Subcategory Boundaries: Increased Subcategory Exceptions .......... 69
   4.3.1 Method ................................................................. 69
   4.3.2 Results .............................................................. 71
   4.3.3 Discussion ......................................................... 74
List of Tables

Table 2.1:  Word-to-category assignments for language 1 and language 2 of Experiments 1-4 .......................................................... 15

Table 2.2:  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 & 4 are denoted ○ .......................................................... 16

Table 3.1:  Word-to-category assignments for the two languages used in Experiment 5. Light-grey AXB cells denote subcategory 1, and white AXB cells denote subcategory 2. Q and R words varied evenly across subcategories and were not a reliable cue for subcategorization ....................................................... 36

Table 5.1:  Mean raw ratings from Experiment 9 dividing up test strings by whether they contain $X_1$, $X_2$, $X_3$ or $X_4$. (Mean rating in boldface, standard errors in parentheses.) ........................................ 90
List of Figures

Figure 2.1: Pictorial depiction of the learning task in Experiments 1 and 2. Learners see 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 ................................................................. 17

Figure 2.2: Rating score results from Experiments 1-4, comparing familiar, grammatical novel, and ungrammatical test strings ... 23

Figure 2.3: Pictorial description of full overlap in the grammar space for the X-words in Experiment 2 (Fig 2.3A), compared to the partial overlap in Experiment 3 (Fig 2.3B) .............................. 25

Figure 2.4: Difference scores of raw ratings from Experiments 1-4 .......... 32

Figure 3.1: Pictorial depiction of subcategorization structure for Experiment 5 ................................................................. 37

Figure 3.2: Rating score results from Experiment 5 .............................. 39

Figure 3.3: Difference scores of raw ratings from Experiment 5 .......... 40

Figure 4.1: Adapted from Corbett (1994): the Romanian gender paradigm, where a subcategory of nouns (sometimes known as “mixed nouns” or “neuter gender”) cross a subcategory boundary. Neuter nouns behave as masculine nouns when they are singular, and they behave like feminine nouns when they are plural ........................................................................................................ 49

Figure 4.2: Pictorial description of the subcategory structure for Experiment 6. There is a clear boundary between the AXB contexts for Subcategory 1 (X_{1,2}) and Subcategory 2 (X_{3,4}), except for the single boundary-crossing context A_4 X_1 B_4 .......... 53

Figure 4.3: Rating score results for non-boundary crossing strings from Experiment 6 ................................................................. 56
Figure 4.4: Z-score results for non-boundary crossing strings from Experiment 6 ................................................................. 57

Figure 4.5: Raw score results for boundary crossing strings from Experiment 6. Ratings of ungrammatical (non-boundary crossing) strings are included for comparison. (BC: boundary crossing) ................................................................. 58

Figure 4.6: Z-score results for boundary crossing strings from Experiment 6. Z-scores of ungrammatical (non-boundary crossing) strings are included for comparison (BC: boundary crossing) ............ 59

Figure 4.7: Difference scores for boundary crossing test strings. Note that both novel boundary crossing strings ($X_1$ with $A_B$ contexts from Subcategory 2) and subcategory violation type 2 strings ($X_{2,3,4}$ with $A_B$ contexts from the opposite subcategory) are rated significantly different from the familiar boundary crossing string ($A_4X_1B_4$) (BC: boundary crossing) ................... 61

Figure 4.8: Pictorial depiction of the language in Experiment 7 where there is incomplete overlap within subcategories and one boundary crossing context for $X_1$ ................................................. 64

Figure 4.9: Raw rating scores of the non-boundary crossing strings in Experiment 7 ................................................................. 65

Figure 4.10: Raw rating scores of the boundary crossing strings in Experiment 7, with non-boundary crossing ungrammatical strings as a comparison. (BC: boundary crossing) ................... 67

Figure 4.11: Comparison of differences of raw rating scores of non-boundary crossing strings in Experiments 6 & 7 ............... 68

Figure 4.12: Pictorial description of the language in Experiment 8 where there is complete overlap and dense sampling within subcategories, but four boundary crossing contexts: one for each $X$ ................................................................. 70

Figure 4.13: Raw score results for Experiment 8 for non-boundary crossing test items ................................................................. 72
Figure 4.14: Raw score results for boundary crossing strings from Experiment 8. Ratings of ungrammatical (non-boundary crossing) strings are included for comparison. (BC: boundary crossing) ................................................................. 73

Figure 4.15: Difference scores of raw ratings for Experiment 8 for non-boundary crossing test strings ................................................................. 75

Figure 4.16: Difference scores of raw ratings for Experiment 8 for non-boundary crossing familiar and grammatical novel strings, and boundary crossing familiar and novel strings. (BC: boundary crossing) .................................................................................... 77

Figure 5.1: Raw rating score results from Experiment 9 ......................... 88

Figure 5.2: Difference scores of raw ratings from Experiment 9 .............. 88

Figure 5.3: Z-scores from Experiment 9 .................................................... 89

Figure 5.4: Raw scores from Experiment 9 comparing test strings containing $X_1$, $X_2$, and the average of strings containing $X_3$ & $X_4$. Note the different trends for $X_1$ versus $X_2$, while $X_3$ and $X_4$ follow the typical trend of familiar strings equal to grammatical novel strings ................................................................. 91

Figure 5.5: Z-scores from Experiment 9 comparing test strings containing $X_1$, $X_2$, and the average of strings containing $X_3$ & $X_4$ ............... 92

Figure 5.6: Mean raw scores of test items containing $X_2$ by participants who rated grammatical novel items no differently than familiar strings. Dotted lines are mean ratings from individual participants, and the bold solid line is the mean of these participants (error bars are standard errors) ................................. 94

Figure 5.7: Mean raw scores of test items containing $X_2$ by participants who rated grammatical novel items significantly lower than familiar strings. Dotted lines are mean ratings from individual participants, and the bold solid line is the mean of these participants (error bars are standard errors) ................................. 94
Foreword

The work described in Chapters 2, 3, 4, and 5 was carried out in collaboration with Elissa Newport and Richard Aslin. The work described in Chapters 2 and 3 (Experiments 1-5) were previously presented at the 31st Annual Meeting of the Cognitive Science Society in 2009 (Amsterdam, The Netherlands) and appeared as a proceedings paper (Reeder, Newport & Aslin, 2009).
Chapter 1: Introduction

Language acquisition crucially involves finding the grammatical categories of words in the input. This process involves determining what categories exist and also correctly classifying words into syntactic categories. The organization of elements into categories, and the generalization of patterns from previously experienced element combinations to novel ones, account for important aspects of the expansion of linguistic knowledge in early stages of language acquisition. This task is a challenge for both nativist and empiricist theories to explain.

One hypothesis regarding how learners solve the problem of categorization is that the categories (but not their contents) are innately specified prior to experiencing any linguistic input, with the assignment of tokens to categories accomplished with minimal exposure (McNeill, 1966). Assuming a universal innate set of syntactic categories is problematic, though, especially considering that some languages of the world lack certain categories or have multiple subclasses within a particular category. Furthermore, there is some evidence that full verb categories do not develop until noun categories have been established, based on how children generalize novel nouns and verbs (Olguin & Tomasello, 1993), and it is not obvious how innately defined category structures would lead to this pattern of evidence.

A second possibility is that the categories are formed around a semantic definition via semantic bootstrapping (e.g., Grimshaw, 1981). This hypothesis suggests that children associate semantic properties with syntactic classes, either by
coming equipped with innate knowledge about this mapping (e.g., Grimshaw, 1981), or through discovery of this mapping via the input (e.g., Brown, 1957). Bowerman (1973) argued that the patterns in two-word speech ordering could be defined in terms of semantic categories (“See baby” = action + object), and that after children learn the meanings of words, they are able to use this information to obtain the syntactic constraints of their language. At some point during learning, children then abandon the semantic criteria and allow syntactic categories to act as the main classifying devices.

This hypothesis is not without problems. Similar to the strong nativist categories view, an innatist approach to semantic bootstrapping assumes that there is a universal set of part-of-speech categories, which cannot be true for all of the syntactic categories. Another issue is that some classes of words are almost entirely semantically arbitrary, like gender classes (Maratsos & Chalkley, 1980), yet children are still able to figure out 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 (Maratsos & Chalkley, 1980). Thus it seems unlikely that semantic categories form the initial groundwork for grammatical categorization.

A third hypothesis, explored in the present research, is that the distributional information in the linguistic environment is sufficient (along with a set of learning biases) to extract the categorical structure of natural language. While it is likely that
innate learning biases and semantic sources of evidence make important contributions to language acquisition, this third hypothesis regarding distributional learning has often been thought to be an unlikely contributor, given the information processing limitations of young children and the complexity of the computational processes that would be entailed. Given infinite time and memory resources, a learner could use brute force methods to compute whether missing information in the input is an accidental gap (because that utterance hasn’t been heard yet) or a meaningful gap (because it is part of the category structure). However, children and adult learners never see an entire input corpus, and they must compute statistics over noisy, serially presented input.

Despite these complications, distributional information must play an important role in solving the problem of grammatical category formation. Maratsos and Chalkley (1980) pointed out that the existence of semantically arbitrary classes 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). There are a number of different types of distributional information correlated with syntactic categories. Maratsos & Chalkley (1980) noted that in English, words that take /-ed/ as a suffix also usually take /-s/ as a possible suffix (verbs). Discovering such patterns between properties of word roots (e.g., /-ed/ and /-s/ suffixing) might be one way to infer initial word classes that can be refined with more data.

The details of how a distributional learning mechanism would operate over
natural language stimuli has been difficult to ascertain, as many of the distributional cues to category structure in natural languages are highly correlated with other sources of information (e.g., semantic: Pinker, 1984; or phonological: Kelly, 1992; Farmer, Christiansen & Monaghan, 2006). However, artificial language learning paradigms offer the ability to test experimentally how learners utilize distributional information by permitting precise experimental control over the statistical properties of the input. A first step in using miniature languages to study categorization was by Smith (1966), who showed that learners were quite capable of learning the simple language:

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

where there are two categories of letters (\(\alpha\) and \(\beta\)) and one rule that requires \(\alpha\) words to be followed by \(\beta\) words. Subjects 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 subjects 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 categories are defined by positional information only (since the co-occurrence statistics between the two categories are uninformative in a distributional sense).

But in a similar paradigm by Smith (1969), participants needed to learn class dependencies in the language:
Pair $\rightarrow \alpha + \beta$
\[ \begin{align*}
\alpha & \rightarrow M, P \\
\beta & \rightarrow N, Q \\
M & \rightarrow m_1, m_2, m_3 \\
N & \rightarrow n_1, n_2, n_3 \\
P & \rightarrow p_1, p_2, p_3 \\
Q & \rightarrow q_1, q_2, q_3
\end{align*} \]

where strings of the language followed the basic pattern of $MN$ or $PQ$, and $M$, $N$, $P$, and $Q$ were categories of 3 items (letters) each. Exposure consisted of seeing $6/9$ of the possible $MN$ pairings and $6/9$ of the $PQ$ pairings. Subjects in this experiment learned that $M$- and $P$-words were supposed to occur first and $N$- and $Q$-words were supposed to occur last in the 2-word strings of the language, but they did not learn the co-occurrence dependencies that $N$-words were only allowed to follow $M$-words, and $Q$-words were only allowed to follow $P$-words. That is, they produced the illegal $MQ$ strings along with the legal $PQ$ strings, and showed no differentiation between the two. This “MN/PQ problem” (Braine, 1987) is the classic case of failure to categorize from distributional information alone.

Braine (1987) acknowledged how easily and quickly learners acquired positional cues to categories, such as “$M$-words come first” and “$N$-words come last.” But it seemed as if learners were only capable of acquiring serial dependencies, since they were unable to learn the rule, “$M$ words are obligatorily followed by $N$-words.” He concluded that learners required a “similarity relation” to cue the (slightly less salient) distributional structure of the subclasses in the MN/PQ problem. With the
addition of partially correlated semantic cues to the subcategory structure, subjects were able to restrict generalization in the MN/PQ experiment: they had fewer ungrammatical overgeneralizations when a “similarity relation” cued them into the co-occurrence structure of the subclasses.

A number of investigators followed up on this hypothesis, exploring the role of shared cues to category structure (e.g., Braine, 1966; semantic cues: Braine, Brody, Brooks, Sudhalter, Ross, Catalano & Fisch, 1990; morphological cues: Brooks, Braine, Catalano, Brody & Sudhalter, 1993; phonological cues: Frigo & McDonald, 1998; Gerken, Gomez & Nurmsoo, 1999; Gerken, Wilson & Lewis, 2005; Monaghan, Chater & Chrsitiansen, 2005; Morgan, Shi & Allopenna, 1996; Wilson, 2002; shared features: Gomez & Lakusta, 2004). The results from many of these artificial language studies seemed to show that the formation of linguistic categories (e.g., noun, verb) depended crucially on a perceptual property linking items within the category (as in Braine’s (1987) “similarity relations”). This correlated perceptual cue might arise from identity or repetition of elements in grammatical sequences, or a phonological or semantic cue identifying words across different sentences as similar to one another (for example, words ending in –a are feminine, or words referring to concrete objects are nouns). Many concluded that correlated perceptual cues were not only necessary to discovery the categorical structure in artificial languages, but they were also necessary in the acquisition of natural grammatical categories (Gomez & Gerken, 2000).

However, this has been somewhat of a puzzle: Maratsos & Chalkley (1980)
argued that in natural languages, grammatical categories do not have reliable phonological or semantic cues; rather, learners must utilize distributional cues about the linguistic contexts in which words occur to acquire such categories. More recently, a number of investigators have looked at whether computational models utilizing clustering algorithms over co-occurrence statistics could successfully acquire elementary form-class categories in natural language corpora (e.g., Cartwright & Brent, 1997; Finch & Chater, 1992; Mintz, Newport, & Bever, 1995, 2002; Mintz, 2003; Redington, Chater & Finch, 1998). All showed that models could exploit purely statistical distributional information in the input to lead to successful grammatical categorization, thus highlighting the potential importance of such a strategy during child language acquisition. Additionally, Mintz (2002) and Gerken et al. (2005) have shown that both adults and infants can learn a simple version of a grammatical category paradigm cued by only distributional information (at least when there are many correlated distributional cues).

How can the supposed necessity of perceptual cues to categories be reconciled with a lack of such cues in certain learning situations? One key distinction that might elucidate the connection between these two issues comes down to the way many investigators have framed the “categorization” problem. The work in this dissertation will utilize a framework for describing these cues that crucially focuses on the structure of the distributional information available to the learner. Importantly, in studying categorization, we will examine the distributional information required for learners to behave as if a set of words is a single category. 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.

Another problem is that many studies in this field have examined the formation of subcategories, rather than the acquisition of a single category. Subcategory learning has an important difference from single-category learning: the subcategorization task inherently involves a conflict of 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 (context words) signal that there are subcategories within the larger category. In the subcategorization case, then, not only must the learner figure out that there are categories, but the learner must also decide which gaps are the systematic omissions that create the subcategory structure and which are accidental gaps that arise from legal but withheld contexts. The MN/PQ problem is a case of subcategorization, where M and P are subcategories of the $\alpha$ category.

The hypothesis explored in this dissertation is that, under some circumstances, distributional cues alone are sufficient to form linguistic categories and subcategories. Given the incomplete and noisy input that any language learner receives, the main question facing the learner is whether missing information in the input is gapped accidentally (because the learner has simply not experienced an occurrence of that context yet), or purposefully omitted from the input (because that context is ungrammatical). In the case of the MN/PQ problem, the gapped strings MQ and PN are absent from the input because they are ungrammatical; yet learners still generalize
over them and consider them to be legal. This behavior demonstrates a failure to acquire the subcategory separating the M and P subcategories, but rather the acquisition of the \( \alpha \) and \( \beta \) categories, leading to generalizing to the contexts MQ and PN. An overarching goal of the work presented here will be to explore the distributional variables that lead learners to generalize to unseen contexts or restrict such generalization.

The series of experiments presented in this dissertation begins by asking what distributional cues are in the statistics of the input that learners use to form categories. The first step involves demonstrating that there are distributional properties that lead to successful learning of linguistic categories in artificial language paradigms. Importantly, however, in order to understand how this mechanism works in human learners and why many previous experiments have not found such learning, we present a series of experiments that manipulate various aspects of these distributional variables, in order to understand the computational requirements for successful category learning. Three main distributional variables of interest will be systematically varied across all experiments:

- **Density** (or, **sparseness**): how many of a word’s possible contexts a learner is exposed to
- **Overlap**: how much sharing of contexts occurs across words in the input
- **Number**: the number of contexts in the input (the size of the input set)

While a number of variables are likely to be influential in category formation from distributional information, little is known about the influence of each of these
factors (individually or combined) on category acquisition. If learners operate in an optimal way when using the statistics of their input corpus, then infrequent and non-systematic omissions in the input should be generalized over; the presence of low frequency and non-systematic omissions provides some information that those contexts are accidentally omitted from the exposure. On the other hand, very high frequencies of distributional information with systematic and recurring gaps should lead learners to increased certainty that the gaps are meaningful. In such a situation, with sparse input and recurring non-overlap among contexts, the plausibility of generalization declines. Thus, across the experiments in this dissertation, we will vary the density of contexts in the input, the overlap of contexts across words, and the number of contexts in the input set in order to assess the plausibility of this distributional hypothesis for category formation.

The remainder of this dissertation will be organized as follows:

Chapter 2 will present a series of experiments that tests whether learners can acquire a single category, generalizing from some instances of the distributional contexts of individual words in the exposure set (but some withheld) to the full range of contexts for all the individual words in the set.

Chapter 3 will explore whether subcategories are also learnable from distributional information, if the learner is given adequate overlap inside each subcategory and adequate non-overlap between subcategories.

Chapter 4 presents three experiments that look at how learners cope with “exceptions” in their input, where subcategory boundaries are occasionally crossed by
lexical exceptions.

Chapter 5 describes two experiments that look at the role of frequency variation for individual items, to see whether unbalanced lexical frequency leads learners to change their interpretation of sparse information and non-overlap in contexts across lexical items.
Chapter 2: Categorization experiments

In this chapter, we present four artificial language experiments with adult participants that investigate how learners behave in situations of varying distributional information to category boundaries. In the present series of experiments, we begin by demonstrating that there are distributional properties that lead to successful learning of linguistic categories in an artificial language paradigm. Additionally, by varying the density, overlap, and number of contexts for the words of a target category across the four experiments, we can explore how learners use these kinds of distributional information to license or restrict category generalizations. The following experiments describe the basic paradigm used throughout this dissertation and show how we can manipulate various aspects of the language landscape via certain distributional variables in order to gain insight into the computational requirements for successful category learning.

2.1 Experiment 1: Dense Sampling with Complete Overlap

In Experiment 1, learners were exposed to a very dense sampling of the language space of an artificial grammar. The structure of the strings in the language was \((Q)AXB(R)\), where each letter represented a category of 2 or 3 words: the Q and R categories had 2 words each, and the A, X, and B categories had 3 words each.
2.1.1 Method

Participants: 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, leaving 17 participants included in this experiment. Participants were randomly assigned to one of two languages, eight subjects exposed to language 1 and nine subjects exposed to language 2.

Stimulus Materials: The words of the language were: spad (/spæd/), klidum (/klaɪdʌm/), flairb (/fleɪrb/), daffin (/dæfɪn/), glim (/ɡlim/), tomber (/tɒmɜːr/), zub (/zʌb/), lapal (/lɛpɔl/), fluggit (/flʌɡɪt/), mawg (/mɔɡ/), bleggin (/bleɪɡɪn/), gentif (/dʒɛntɪf/) and frag (/fræɡ/). These words were selected such that they maintained the phonotactics of English words (and thus were easy to parse) while still being maximally different from real lexical items in English1. Words were read in isolation by a native English-speaking female and recorded with both non-terminal and terminal list intonation (so that any word could be used in a non-terminal or terminal position in a string; see below for how words were combined to form strings). Words were adjusted in Praat (Boersma, 2001) so the pitch, volume, and duration of words were fairly consistent. Languages 1 and 2 were created by assigning these words to the categories such that each category had a

---

1 An informal survey of 5 naïve native English-speaking University of Rochester students revealed that none of these words had common pre-existing meanings in English, and many of these words are successfully used in other experiments carried out in our lab. (“Frag” was reported as part of “defrag”, but not as a word on its own.) All subjects who participated in any of the studies described in this dissertation were required to have not participated in any other study that used any of these lexical items so that they were completely naïve to the nature of the items and the language.
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 syllable length) (see Table 2.1.) The words were not mapped to any sort of referential world, so they had no meaning associated with them.

Sentences were constructed by splicing together words into sequences in Sound Studio, with 50ms silence between each word and using the word token with a terminal intonation contour as the final word in the sentence. Exposure strings were recorded to mini-disc in a random order with 1.5s of silence between sentences.

Table 2.1 Word-to-category assignments for language 1 and language 2 of Experiments 1-4.

<table>
<thead>
<tr>
<th>Language 1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>A</td>
</tr>
<tr>
<td>spad</td>
<td>flairb</td>
</tr>
<tr>
<td>klidum</td>
<td>daffin</td>
</tr>
<tr>
<td></td>
<td>glim</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>A</td>
</tr>
<tr>
<td>frag</td>
<td>gentif</td>
</tr>
<tr>
<td>daffin</td>
<td>mawg</td>
</tr>
<tr>
<td></td>
<td>klidum</td>
</tr>
</tbody>
</table>

All sentences were constructed from the (Q)AXB(R) grammar. X was the target category under study, while A and B were the contexts elements that formed

---

2 Two languages were created to insure that our initial mapping of words to categories was not inadvertently biased to aid the learner with the categorization task.
the distributional cues to the 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 Q and R categories made words of the language observe patterning in terms of relative order but not fixed position in the sentences. (This was so that fixed position information could not be easily used as an informative cue to category membership.)

Focusing on just the AXB portion of the grammar, there were 3x3x3=27 possible word strings in the language. Of the 27 basic AXB sentence types, 18 were presented and 9 were withheld (see Table 2.2).

Table 2.2 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 & 4 are denoted ○.

<table>
<thead>
<tr>
<th>A1 X1 B1 ★</th>
<th>A1 X2 B1</th>
<th>A1 X3 B1 ★★ ○</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 X1 B3 ★★ ○</td>
<td>A1 X2 B3 ★</td>
<td>A1 X3 B3</td>
</tr>
<tr>
<td>A2 X1 B1</td>
<td>A2 X2 B1 ★★ ○</td>
<td>A2 X3 B1 ★</td>
</tr>
<tr>
<td>A2 X1 B2 ★★ ○</td>
<td>A2 X2 B2 ★</td>
<td>A2 X3 B2</td>
</tr>
<tr>
<td>A2 X1 B3 ★ ○</td>
<td>A2 X2 B3</td>
<td>A2 X3 B3 ★★</td>
</tr>
<tr>
<td>A3 X1 B1 ★★</td>
<td>A3 X2 B1 ★ ○</td>
<td>A3 X3 B1</td>
</tr>
<tr>
<td>A3 X1 B2 ★</td>
<td>A3 X2 B2</td>
<td>A3 X3 B2 ★★ ○</td>
</tr>
<tr>
<td>A3 X1 B3</td>
<td>A3 X2 B3 ★★ ○</td>
<td>A3 X3 B3 ★</td>
</tr>
</tbody>
</table>

Within these 18 AXB types, 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 4 sentence types AXB, QAXB, AXBR, and QAXBR).

**Figure 2.1** Pictorial depiction of the learning task in Experiments 1 and 2. Learners see 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.

**Procedure:** Participants were seated in a sound-attenuated both 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 try 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 4 times via mini-disc, forming 20 minutes of exposure to the language that the participant listened to over headphones. Importantly, the 18 AXB sentence types used during exposure included each X word in the presence of every A word and every B word.
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_B contexts among the various X words within the target category. (Figure 2.1 is a pictorial depiction of the categorization task facing the learner.)

After exposure, participants were presented with a pseudo-random ordering of individual test strings on mini-disc and were asked to rate each test string on a scale of 1 to 5 based on whether or not they thought the test sentence came from the language they heard during training. Subjects responded by circling their judgment on a scale on a sheet of paper: 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 following the rules of the language’s grammar. Test strings were all 3-word sentences 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). The order of presentation of test strings was such that the 9 familiar and 9 grammatical novel strings were randomized with 9 ungrammatical strings.

---

3 The experiments in this chapter were later piloted with additional types of ungrammatical items (AAB and ABB) to make sure that people were not memorizing the A and B item locations in order to rate the ungrammatical test items. These additional ungrammatical test items were not rated significantly differently than the AXA and BXB items and confirm that performance during test is not solely based on learning the positional information contained the A_B frame.
during the first half of the test, and then the same 9 familiar and 9 grammatical novel strings were presented again in random order along with 9 different ungrammatical strings during the second half of the test.

2.1.2 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 Figure 2.2). There was no significant difference between ratings of grammatical novel items and grammatical familiar items ($F(1,15)=1.845, p=0.19$) However, these items were rated significantly higher than ungrammatical items ($F(1,15)=45.651, p<0.001$).

Because individual subjects may utilize the rating scale in different ways (some may have used only 1s and 5s while others stuck to 2s, 3s and 4s), raw ratings scores were converted into z-scores in order to standardize ratings across subjects, using the formula $z_{ij} = \frac{\text{rating}_{ij} - \mu_j}{SE_j}$, where $z_{ij}$ is the z-score for the $i^{th}$ test item rated by subject $j$ based on the raw rating of item $i$ by subject $j$. Thus, a score below zero means that an item was rated lower than a subject’s average rating, and a score above zero means 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 language 1 and 2: $F<1$; no significant difference between grammatical novel and grammatical familiar: $F(1,15)=1.792$, $p=.2$; significant difference between novel and ungrammatical items: $F(1,15)=70.630$, $p<0.001$).

2.1.3 Discussion

In Experiment 1, the learner was exposed to a very dense sampling of the language space, with all of the words in the target category appearing in many highly overlapping A_B contexts. Under these conditions, learners 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 AXA or BXB. These findings show, therefore, that when the input densely samples the language space and words within a category appear in highly overlapping contexts, learners will fully generalize within the category to novel contexts and novel strings, even without any perceptual or semantic cues to indicate that the words form a single category.

In the subsequent experiments in this chapter, we investigate the degree to which category generalization is affected by manipulating these distributional variables, in learning a single category.
2.2 Experiment 2: Sparse Sampling with Complete Overlap

In Experiment 2, we explored what happens if we keep the number and overlap among X-word contexts in the language the same, but during learning we present learners with substantially fewer of the contexts that are possible in the language (see Table 2.2). We refer to this as reducing the density (or increasing the sparseness) of the contexts for X-words that are presented during learning.

2.2.1 Method

Participants: 19 monolingual native English-speaking students at the University of Rochester participated in Experiment 2, but three were excluded for not complying with experiment instructions (2) and equipment failure (1). This left 16 total subjects, with eight participants in each of languages 1 and 2. Subjects had not participated in any other categorization experiment and were paid for their participation.

Stimulus Materials: Strings were created in the same manner as in Experiment 1. Here, however, out of the 27 possible AXB combinations, only 9 were presented during exposure (see Table 2.2). Crucially, every X-word was still heard in combination with every A and every B word so the exposure set had complete overlap of contexts across X-words (see Figure 2.1 for a pictorial depiction). As in Experiment 1, each of the 9 sentence types was presented with category flanker elements Q and R present or absent, producing 36 sentences in the exposure set.

Procedure: The procedure was the same as in Experiment 1. The exposure set was
presented 4 times via mini-disc for a total exposure time of about 10 minutes. (Each input sentence type was thus presented with the same frequency in this experiment as in Experiment 1; the overall exposure was reduced in time and number of strings by reducing the size of the exposure set.) The test was the same as in Experiment 1, except that the grammatical novel test items were counterbalanced such that half of the participants in each language were tested on one subset of nine of the withheld (grammatical novel) items, and the other participants were tested on the other nine grammatical novel items.

2.2.2 Results

A repeated measures ANOVA with condition (familiar, novel, ungrammatical) as the within-subjects factors and language (1 or 2) and subtest (which counterbalanced set of novel items the subject received during test) as the between-subjects effects. 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 2.73 \( (SE=0.14) \). Grammatical novel strings were rated just as highly as grammatical familiar strings, and there was no significant difference between these two types of items \( F(1,12)=0.520, p>0.4 \). The analysis further revealed that the ungrammatical items were rated significantly lower than the grammatical novel items \( F(1,12)=19.805, p<0.001 \) (see Figure 2.2).
Figure 2.2 Rating score results from Experiments 1-4, comparing familiar, grammatical novel, and ungrammatical test strings.

Raw scores were transformed into z-scores, and another repeated measures ANOVA was conducted with condition as the within-subjects factors and language and subtest as the between-subjects factors. Once again, there was no difference between familiar and novel grammatical ratings \( (F(1,12)=.546, p>0.4) \), but ratings of grammatical novel strings were significantly higher than ungrammatical \( (F(1,12)=21.723, p<0.001) \). None of the interactions were significant.

2.2.3 Discussion

These results show that learners’ performance is unchanged from Experiment 1 when density/sparseness is reduced but other properties of the distributional
information are maintained, despite the fact that the exposure is half as rich and half as long. This permits us to ask what happens, in contrast, when the amount of overlap in the contexts of the X-words is reduced.

2.3 Experiment 3: Sparse Sampling with Incomplete Overlap

In Experiment 3, as in Experiment 2, we presented only 9 of the possible 27 AXB combinations. Here, however, we presented particular AXB combinations that reduced the degree of overlap among members of X in the contexts in which they were heard, in order to assess the importance of the overlap in distributional information for category formation and generalization. In the present experiment, the set of X-words, taken together, occurred in all of the A and B contexts, and the different X-words each overlapped in part with all the other X-words. However, individual X-words did not fully share all their contexts with one another (see Figure 2.3). The question addressed now, then, is the degree to which learners will restrict their generalization across the category as a function of this reduction in overlap.

2.3.1 Method

Participants: 24 monolingual native English-speaking students at the University of Rochester participated in Experiment 3, 12 in each of language 1 and 2. Subjects had not participated in any other categorization experiment and were paid for their participation.
Figure 2.3 Pictorial description of full overlap in the grammar space for the X-words in Experiment 2 (Fig 2.3A), compared to the partial overlap in Experiment 3 (Fig 2.3B).

**Stimulus Materials:**  Strings were composed in the same way as Experiment 1, and only 9 of the 27 possible AXB combinations were heard. X₁ was heard in the context of A₁, A₂, B₁ and B₂, but not in the context of A₃ or B₃. X₂ was heard in the context of A₂, A₃, B₂ and B₃, but not A₃ or B₃. X₃ was heard in the context of A₁, A₃, B₁ and B₃, but not in the context of A₂ or B₂. 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 do in Experiments 1 and 2, where each X word occurs with each A and each B.

**Procedure:**  The training and test procedures were the same as Experiment 1.

2.3.2 Results

A repeated measures ANOVA was conducted and revealed no differences
between languages 1 and 2 ($F(2,44)=1.581, p>0.1$). The mean rating of grammatical familiar items was 3.79 ($SE=0.1$), the mean rating of grammatical novel items was 3.48 ($SE=0.16$), and the mean rating of ungrammatical items was 2.85 ($SE=0.15$). In contrast to experiments 1 and 2, the ANOVA revealed significant differences between grammatical familiar and grammatical novel items ($F(1,22)=19.191, p<0.001$), as well as between grammatical and ungrammatical items ($F(1,22)=70271, p<0.001$).

Raw scores were transformed into z-scores, and another repeated measures ANOVA was conducted, with condition as the within-subjects factor and language as the between-subjects factor. There was no effect of language ($F<1$), but, again, grammatical novel items were rated significantly lower than familiar ($F(1,22)=23.812, p<0.001$) and significantly higher than ungrammatical ($F(1,22)=54.614, p<0.001$).

Because of the incomplete overlap imposed by this experiment, the novel test items form two types: “heard 2” items, where the subject heard both the $AX$ and $XB$ bigrams during exposure (but not the entire $AXB$ trigram); and “heard 1” items 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 mean = 3.62, $SE = 0.12$; heard 1 mean = 3.41, $SE = 0.12$; $t=2.54, p=0.018$). However, this is unsurprising given that heard 2 items have two times the amount of familiar bigram information. This difference is subtle and relies on too few items to make any strong claims, but this result might point to
the type of statistics that subjects might attending to (namely, bigram information) in order to successfully figure out the categories of the language. This possibility is returned to later in the last chapter of this dissertation.

2.3.3 Discussion

Whereas in Experiment 2 we tested how subjects would respond to fewer contexts but full overlap of the context environment, Experiment 3 greatly reduced the overlap in the exposure while keeping number the same as in Experiment 2 (see Figure 2.3A as compared to Figure 2.3B). It is important to note that, 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 items than for ungrammatical items. 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.

2.4 Experiment 4: Incomplete Overlap with Extended Exposure

One more variable that may impact generalization versus lexical distinctness
is the frequency or consistency with which each type of context is presented (and therefore the frequency with which contextual gaps recur). If learners operate in an optimal 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 and should 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.

2.4.1 Method

Participants: 17 monolingual native English-speaking students at the University of Rochester participated in Experiment 4. One was removed due to not understanding testing directions, which left eight subjects in each of languages 1 and 2. Subjects had not participated in any other categorization experiment and were paid for their participation.

Stimulus Materials: The corpus was the same as in Experiment 3; however, exposure was tripled, by presenting the exposure corpus 12 times rather than 4. The exposure thus lasted for approximately 30 minutes, which was somewhat longer exposure length than Experiment 1, but contained only 9 contexts, as in Experiments 2 and 3.
Procedure: The training and test procedures were the same as in Experiment 3.

2.4.2 Results

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 items were rated significantly lower than familiar items ($F(1,14)=19.40,\ p<0.005$), and were also rated significantly higher than ungrammatical items ($F(1,14)=31.744,\ p<0.001$) (see Figure 2.2).

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 within-subject conditions. Novel items were rated significantly lower than familiar items ($F(1,14)=21.01,\ p<0.001$) and significantly higher than ungrammatical items ($F(1,14)=39.963,\ p<0.001$).

2.4.3 Discussion

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, they are more likely to assume that gaps in the input are intentional. Nevertheless, the novel grammatical test strings are judged to be more grammatical than the ungrammatical strings. Previous work by Reeder, Newport and Aslin (2009) exposed subjects to two times through the Experiment 3 corpus (rather than three times through), and the results were in the same direction as those presented here but, as would be expected, showed a smaller difference between familiar and novel ratings. In contrast, we would expect that even more exposure to highly consistent gaps than given in Experiment 4 would increase even further the difference between familiar and novel strings and might bring the ratings of the novel strings down to the level of ungrammatical strings, a result that would suggest no generalization to those contexts at all. (On the other hand, extended exposure to less systematic and consistent gaps in the corpus should lead learners to generalize more. This question is explored further in later chapters of this dissertation.) Taken together, these manipulations and results suggest that learners are quite sensitive to a number of variables that signal whether to generalize across lexical items or restrict generalization to lexically specific contexts.

### 2.5 General Discussion

The four experiments described in this chapter tested whether learners can acquire a single category, generalizing from exposure to some instances of the distributional 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 evidence 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. Looking just at difference scores between familiar and novel test items and between familiar and ungrammatical test items (see Figure 2.4), it is clear that in Experiments 1 and 2, familiar and novel grammatical test items are rated no differently from each other (the ratings difference between the two types is no 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 begin to reduce their likelihood of generalizing. When the overlap in contexts is reduced, learners increase the difference in their ratings for familiar versus unfamiliar grammatical sentences. They restrict generalization even more sharply in Experiment 4, when the same exposure corpus (and its gaps) was repeated three times. These results show that adult learners can skillfully use the data in the input to determine whether to ignore gaps in the input (that is, whether to generalize over the gaps), or rather whether to restrict generalization over such gaps. Participants in these experiments were able to take account of 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 also 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 could 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 are carefully controlled. The results suggest that learners are keeping track of word co-occurrences at a mid-sized grain, such as bigram frequencies or probabilities. Alternatively, they could be keeping track of the network of occurring contexts for individual words (as illustrated in Figures 2.1 and 2.3) and collapsing the individual words into a category when these networks bear enough quantitative (as well as qualitative) similarities to one another.

**Figure 2.4** Difference scores of raw ratings from Experiments 1-4.
As noted in Chapter 1, a large body of previous research has claimed 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) are required for successful learning. Experiments 1-4 showed that additional cues are not necessary for adults to induce a category from distributional contexts alone. The following chapters explore how adult learners utilize distributional information when tackling a similar learning situation, that of learning *subcategories* – subsets of words with distinct privileges of occurrence (such as nouns of different genders).
Chapter 3: Subcategorization

Experiments 1-4 have investigated whether learners can acquire a single category, generalizing from hearing some instances of the distributional contexts of individual words to the full range of contexts for all the individual words in the set. By observing this generalization, we conclude that learners have formed a category rather than storing each word individually in terms of its specific experienced contexts. As previously noted, however, several investigators have claimed 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, are required for successful learning (e.g., Braine, 1987). While we have found in our QAXBR experiments that we can induce category learning from distributional contexts alone, many of the early experiments on linguistic category acquisition have often been studying a subcategory, rather than a single category, learning problem. A subcategory problem has an important distributional property that we hypothesize may differentiate it from a single category problem: in the subcategory case, some of the distributional cues signal that there is only one category, while other distributional cues signal that there are subcategories within this large category. It is therefore possible that these conflicting cues, rather than the difficulty of distributional learning itself, has created the problems observed by other investigators.

Experiment 5 explores whether subcategories are also learnable from
distributional information, if the learner is given adequate overlap inside each subcategory and adequate non-overlap between subcategories. The question of interest is whether learners can acquire subcategories, or whether they will tend to treat the X-words as a single category, ignoring the distributional information for the multiple and separate subcategories.

3.1 Experiment 5: Subcategorization With Dense Sampling, Complete Overlap

In Experiment 5, we expand on the (Q)AXB(R) language from Experiments 1-4 in order to test whether (contrary to much of the literature) learners are capable of forming functional subclasses of a grammatical category where the only cues to subcategorization are distributional cues. As this is a first step towards exploring subcategorization using distributional cues, Experiment 5 is similar to Experiment 1 in that learners will get very strong distributional cues to subcategories (dense sampling of the subcategory-obeying strings, and complete overlap within subcategories but complete non-overlap across subcategories).

3.1.1 Method

Participants: 24 monolingual native English-speaking students at the University of Rochester participated in Experiment 5 (12 each in languages 1 and 2). Subjects had not participated in any other categorization experiment and were paid for their participation.
**Stimulus Materials:** Experiment 5 utilized the same grammar as in Experiments 1-4, but more words were added to the language in order to allow for a subcategory structure: *mib* (/mɪb/), *bliffin* (/blɪfɛn/), *zemper* (/zɛmpər/), *roy* (/rɔɪ/), *nerk* (/nɛrk/), *prog* (/prɔg/), and *dilba* (/dɪlbə/). The updated language assignments are shown in Table 3.1.

**Table 3.1** Word-to-category assignments for the two languages used in Experiment 5. Light-grey AXB cells denote subcategory 1, and white AXB cells denote subcategory 2. Q and R words varied evenly across subcategories and were not a reliable cue for subcategorization.
Categories Q and R had 2 words each, as in Experiments 1-4, but categories A and B had 6 words each, while category X had 4 words. A subcategory structure was devised such that $A_{1,2,3}$ and $B_{1,2,3}$ were only heard with $X_{1,2}$. $A_{4,5,6}$ and $B_{4,5,6}$ were only heard with $X_{3,4}$ (see Figure 3.1).

**Figure 3.1** Pictorial depiction of subcategorization structure for Experiment 5.

```
Subcategory 1
A_1
A_2
A_3
X_{1,2}
B_1
B_2
B_3

Subcategory 2
A_4
A_5
A_6
X_{3,4}
B_4
B_5
B_6
```

**Procedure:** In this language, there are $6 \times 4 \times 6 = 144$ possible combinations of A, X, and B, but only 36 of those strings are legal according to the subcategory structure. Of those legal strings, 24 AXB combinations were presented during exposure and 12 AXB combinations were withheld. Optional Q and R elements were applied as in previous experiments, to create a training set of 96 strings. The sparseness and overlap within each category were proportional to the sparseness and overlap of Experiment 1. Pilot testing revealed that keeping exposure to 20 minutes (similar to Experiment 1) did not lead to systematic learning of the language (this is unsurprising given that the language was much larger). Therefore, exposure was increased to about 45 minutes (5 times through the training set).

The test stimuli were comprised of 12 grammatical familiar items, 12 grammatical novel items, 12 ungrammatical AXA or BXB items, and 12
ungrammatical subcategory violation items. The subcategory violation items had either the A word or the B word from the opposite subcategory as the X item. Crucially, the subcategory violation items 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 items and subcategory violation items therefore indicates that participants have discovered the subcategory structure in the language, and are not generalizing across the gaps created by this boundary.

The experiment was conducted via a custom software package for experimentation. The subject sat in a sound attenuated booth wearing headphones, seated at a comfortable distance away from a Dell desktop PC. Instructions were both presented on the screen and read aloud to the subject (as in Experiment 1). After the experimenter finished saying the instructions, the subject indicated that s/he understood them and the experimenter left the booth. The presentation of exposure sentences was randomized for each subject, and the test sentence presentation was also randomized such that half of each of the familiar, grammatical novel, subcategory violation, and ungrammatical test strings appeared during each half of the test. All other aspects of training and test were the same as in Experiment 1.

3.1.2 Results

A repeated measures ANOVA with language as the between-subjects effect and condition (familiar, grammatical novel, subcategory violation, and ungrammatical) as the within-subjects effect revealed no differences between
languages 1 and 2 ($F<1$). The mean rating of grammatical familiar items was 3.61 ($SE=0.1$), the mean rating of grammatical novel items was 3.7 ($SE=0.11$), the mean rating of subcategory violation items was 3.31 ($SE=0.12$), and the mean rating of ungrammatical items was 2.55 ($SE=0.12$) (see Figure 3.2 and Figure 3.3). Grammatical familiar and grammatical novel items were not significantly different from each other ($F(1,22)=1.559, p>0.1$). However, subcategory violation items were rated significantly lower than grammatical items ($F(1,22)=11.598, p<0.01$). Ungrammatical items were rated the lowest, significantly lower than subcategory violation items ($F(1,22)=19.648, p<0.001$).

**Figure 3.2** Rating score results from Experiment 5.

Raw scores were transformed into z-scores and another repeated measures ANOVA was conducted. There was no effect of language ($F=1.085, p=.309$). Once again, familiar and grammatical novel items were rated the same ($F(1,22)=0.96$,
$p=0.338$), but subcategory violation items were rated significantly lower than the grammatical items ($F(1,22)=13.623, \ p<0.005$) and significantly higher than ungrammatical items ($F(1,22)=6.643, \ p<0.001$).

**Figure 3.3** Difference scores of raw ratings from Experiment 5.

3.1.3 Discussion

As in Experiments 1-4, learning effects were observed based solely on distributional cues to subcategory structure. While the subcategorization results are weaker than the categorization results (as shown by the significant difference between subcategory violation items and ungrammatical items), it is important to keep in mind that this task involves a conflict of cues. The subcategory problem has an important distributional property that differentiates it from a single category problem: in the subcategory case, some of the distributional cues (e.g., word order) signal that there is only one category, while other distributional cues (A and B context words) signal that
there is subcategorization within this larger category. Not only must the learner figure out that there are categories, as in Experiments 1-4, but now the learner must also decide which gaps are systematic (the gaps that create the subcategory structure) and which are accidental (the gaps that are withheld-but-legal items). Furthermore, the higher scores for subcategory violation test items indicate that learners are attending to the structure of the language (in that they rate AXA and BXB strings differently than AXB strings). Figure 3.3 depicts difference scores of raw ratings for each type of test string, and shows that subcategory violation strings are significantly different from familiar and grammatical novel strings, while familiar and grammatical novel strings are not qualitatively different from each other.

3.2 General Discussion

Given the results of numerous studies by other investigators demonstrating that subcategories are unlearnable without multiple correlated cues (e.g., Braine, 1987), it is important to explore some of the reasons why we have found success where others have not. One important distinction that affects the interpretation of results from previous studies is our framework for defining subcategorization versus categorization. As described in the introduction, subcategorization crucially involves a conflict of cues. In a category that contains two or more subcategories, a learner must note that the distributional information signals a main X category (as in Experiments 1-4 in Chapter 2), but the gaps formed by the subcategory boundary signal multiple subcategories within X. It seems clear from past literature that
subcategorization is quicker and easier for the learner when there are multiple cues present to denote the subcategory structure. Furthermore, there is some evidence that natural languages are patterned such that certain subcategories have partially correlated cues to subcategorical structure (Monaghan et al., 2005). However, our work is the first to show that multiple subcategories can be acquired on the basis of distributional information alone.

First, consider the MN/PQ problem – the classic case of failure to acquire subcategories given distributional information alone. As described in Chapter 1 (Braine, 1987; Smith, 1969), subjects 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 were only allowed to follow M-words and Q-words were only allowed to follow P-words. That is, participants produced both illegal MQ strings and legal PQ strings during test, indicating that they were equally as grammatical to participants. Looking at the distributional information available to the learners as compared to our own hypotheses, this result is puzzling, as the input consisted of a dense sampling of the language with complete non-overlap across subcategories. However, one possibility is that the MN/PQ task utilizes some misleading distributional cues that are so salient and quickly learned that other, more relevant distributional cues might be ignored in favor of easier acquisition. Indeed, much of the early work on category learning showed that positional information is just such a cue (e.g., Braine, 1965; Smith, 1966; Smith, 1969): if subjects indeed attend to the positions of the words (which words occur first and last) and not the relationships among the word co-occurrences, they
will overgeneralize. Positional information is so salient in the very short and length-invariant strings within the MN/PQ problem that a rational subject would likely need to be exposed to the exposure set many times in order to ignore the positional cues to categories and pick up on the dependencies among the specific words forming the grammatical subcategories. At that point, however, subjects would have memorized the letter pairs, potentially failing to generalize at all. An important question that arises from the salience and use of positional cues for learning categories, is whether or not they are useful in natural languages. The clear answer to that question is that they are not: no natural linguistic category is defined by its absolute position in a sentence. Rather, natural categories and subcategories are defined in terms of the contexts in which words are licensed to occur. Indeed, the field of linguistics defines syntactic categories and subcategories via the use of distributional tests which are based on the assumption that two words with similar distributions (that is, similar surrounding contexts) might belong to the same grammatical category if they are syntactically interchangeable (Radford, 1988).

Braine (1987) acknowledged how easily and quickly learners acquired positional cues to categories, and concluded that learners required a “similarity relation” in order to cue the (slightly less salient) distributional structure of the subclasses in the MN/PQ problem. Braine supported this idea with the addition of partially correlated semantic cues to the subcategory structure in the MN/PQ experiment. With these correlated cues, subjects were able to restrict generalization across the subcategory boundary. In other words, there were fewer ungrammatical
overgeneralizations when a “similarity relation” cued them into the co-occurrence structure of the subclasses.

An important observation to make regarding Braine’s (1987) correlated cues version of the MN/PQ experiment is that the addition of the semantic cues did not remove the positional cue to the categories. The purpose of the semantic cue was to signal the existence of the subcategories and overcome the positional cues to the main categories. However, in our work, no positional information is available to the learner, due to the optional Q and R category flankers, which make the A, X, and B words have relative but not fixed positions of occurrence. Despite using a larger and more complex language, then, Experiment 5 demonstrated successful subcategory acquisition while using the same dense sampling and no overlap across the subcategory boundary as in the MN/PQ experiment.

Additionally, Gerken, Wilson, Gomez and Nurmsoo (2009) posited that many of the early failures to find subcategory learning in artificial language studies were due to the use of a referential field that was irrelevant to category formation. Much like the strength and saliency of positional cues, Gerken and colleagues 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. Experiment 5 utilized an implicit learning paradigm free of semantic cues to show that learners can, in fact, learn categorical structures without the need for semantic cues.

Other investigators (e.g., Gerken et al., 1999; Gerken et al., 2005; Gomez &
Lakusta, 2004; Wilson, 2002) have argued for the necessity of correlated phonological or morphological cues. Many of the test cases of subcategorization in these studies involve paradigm-completion methods for acquiring gender systems (e.g., gender in Russian). 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 it is quite possible that the full context involved in expressing such gender systems (both linguistic and extra-linguistic) provides enough distributional evidence to create initial subcategories, even without additional phonological cues. There is evidence (e.g., Gvozdev, 1969; Polinsky, 2008) that even though children acquiring the Russian gender paradigm may not consistently mark correct gender at an early age, they do have the correct number of subcategories of gender despite occasionally using them in the wrong contexts. This suggests that it is important to explore further the role of distributional information alone and in combination with morphophonological information as cues to gender subcategories.

Given the success of this paradigm for examining the effects of varying distributional information on subcategorization, the remaining experiments in this dissertation use the paradigm outlined in this chapter to explore how learners behave in two subcategorization situations that are common in natural language: the case of lexical items that cross subcategory boundaries (similar to homonyms, Chapter 4), and the case of uneven frequency information within a subcategory (Chapter 5).
Chapter 4: Crossing subcategory boundaries

Chapter 3 demonstrated that adults are capable of learning subcategory boundaries using distributional cues alone. However, the miniature language used in Experiment 5 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 from Experiment 5 could be modified in order to further test the limits of statistical learning would be to make the subcategory boundaries imperfect in ways that would be typical of real languages. As an example of how we could do this, consider the following sets of English strings, in which one word form occurs in the contexts for more than one word category:

(1) The postman just delivered the /ˈmeɪtl/.  
(2) Jim is a /ˈmeɪtl/.  
(3) The paint can dry overnight.  
(4) Don’t knock over the paint can.  
(5) We can paint the room blue.  
(6) I ate a fish yesterday.  
(7) We fish on Sunday afternoons.  
(8) What’s black and white and /red/ all over?

All of these are examples of homophony, in which there are apparent violations of category distribution due to the fact that two different words have the
same phonetic form. They demonstrate one way in which language users will encounter ambiguities of distributional cues in which they must use contextual information to provide possible resolutions for the ambiguity. The homophones in sentences (1) and (2) have different meanings, but the purely syntactic contextual cues only tell you that a noun should appear in the underlined slot, and they do not provide evidence of how to separate the sets of contexts in which the same word form appears. One potential solution is that the listener would use the word senses implied by the rest of the context to disambiguate whether /'meɪl/ is male or mail. The rest of the sentences demonstrate how a single phonological form of a word might occur in two different grammatical categories (a type of homonymy). Ambiguity of this type is found in many of the world’s languages (if not all), despite the complications that it introduces to language acquisition and processing. In this case, although there may be two distinct lexical items (for example, fish (verb) and fish (noun), or read (verb) and red (adjective)), from the learner’s point of view the same phonological form is heard in two different form class contexts. For example, fish might be heard in the context of a pronoun and a preposition (as in (7)), or in the context of a determiner and an adverb (as in (6)). The way to determine whether fish is a verb or a noun in any one particular sentence is to examine this particular context in comparison with other common noun contexts or verb contexts.

This type of ambiguity is found at many different levels of language (e.g., lexical level, like fish; morphological level, see Pertsova (2008); syncretism, see Baerman, Brown & Corbett, 2005). Such ambiguity also exists at a subcategory
level. One example is found in the “crossed” gender system of Romanian (Corbett, 1991, 1994).

Grammatical gender is predominantly an arbitrary designation of a noun that, in combination with grammatical case and number, defines the determiner and/or agreement markers that must be paired with that noun. (While there are some weak correspondences between gender class and semantic or phonological cues, these are frequently violated.) A well-known example comes from French, where nouns are divided into masculine and feminine gender subcategories that are reflected in the determiners and adjectival forms that accompany the nouns (Corbett, 1994). A more complicated example is the Romanian gender system. Much like French, the Romanian gender system divides nouns into masculine and feminine gender such that these two genders each have separate agreement markers. Unlike French, however, Romanian has a sizeable set of nouns that take the masculine agreement markers when singular, but the feminine agreement markers when plural. This third class of nouns does not belong grouped with either masculine nouns or feminine nouns, and thus they are said to form a third gender (neuter). Yet, Romanian does not have a separate class of agreement markers to match up with neuter nouns. In Corbett’s (1994) terminology, Romanian has three “controller genders” (the number of genders of nouns) but only two “target genders” (the number of classes of agreement markers). This system is depicted in Figure 4.1 (adapted from Corbett, 1994).
Figure 4.1 Adapted from Corbett (1994): the Romanian gender paradigm showing case marking on nouns, where a subcategory of nouns (sometimes known as “mixed nouns” or “neuter gender”) crosses a subcategory boundary. Neuter nouns behave as masculine nouns when they are singular, and they behave like feminine nouns when they are plural.

<table>
<thead>
<tr>
<th></th>
<th>SINGULAR</th>
<th>PLURAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>masculine</td>
<td>ø</td>
<td>masculine i</td>
</tr>
<tr>
<td>neuter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>feminine</td>
<td>â</td>
<td>feminine e</td>
</tr>
</tbody>
</table>

The subcategory-crossing neuter nouns may have arisen from a few possible sources. It is possible that before the Romanian language underwent the changes that brought it to its modern state, it had a separate target gender for neuter nouns. In that case, the table shown in Figure 4.1 would have six cells instead of four, and neuter nouns would have separately defined agreement markers. That situation would be similar to Experiment 5, where subcategories are clearly defined without any items consistently crossing subcategory boundaries. However, real languages commonly have boundary crossing exceptions that are similar to the Romanian gender system. Just as in the homophony cases demonstrated by sentences (1)-(8), subcategory exceptions arise from ambiguous situations where the syntax makes a particular distinction, but other systems (morphology, phonology) do not make the same distinction.

What is remarkable about the existence of an arbitrary system such as this is that children are capable of acquiring it without explicitly being told that some contexts for words are highly specific, whereas others are quite general. The learner
must be capable of realizing the correct representational scheme for determining when to generalize and when to encode something exceptional about a class/subclass uses. The experiments described in this dissertation show that human learners are remarkably sensitive to distributional information present in word contexts in order to make decisions about what class or subclass a word belongs to. But what if you have just started being exposed to a particular language, and you have a tenuous grasp of the possible contexts that are permissible for a word and its form class? When do you decide that the contextual information surrounding a particular experienced item is a mistake versus a meaningful extension of that item’s possible environments? And what sort of distributional information might you need to attend to in order to generalize that extension to other items in its (sub)category? The present experiments exploit this idea of distributional information that crosses subcategory boundaries to further test the strength of the distributional variables that learners use when deciding whether and how to generalize to unheard contexts.

4.1 Experiment 6: Subcategorization Without Consistent Subcategory Boundaries

The main difference between the experiments described in this chapter and Experiment 5 is the presence of one context that crossed the subcategory boundary. \((A_4 X_1 B_4)\). Based on the results of Experiment 5, we expect that this \(A_4 B_4\) context should clearly signal that the \(X_1\)-word belongs to subcategory 2; but the remainder of the contexts for \(X_1\) in the exposure set contains information strongly signaling that \(X_1\) belongs to Subcategory 1. Experiment 6 will look at whether subcategory boundaries
entirely break down because of the evidence of this boundary crossing context, or whether learners will treat \( A_4X_1B_4 \) as an exception and otherwise maintain the subcategory structure.

4.1.1 Method

**Participants:** 18 monolingual native English speaking students at the University of Rochester participated in Experiment 6 for payment, but 3 were removed for failing to understand the test instructions. None had participated in any other categorization experiment. Seven participants were assigned to language 1, and eight were assigned to language 2 (see Table 3.1).

**Stimulus Materials:** The words and category memberships were the same as in Experiment 5 (see Table 3.1). The strings of the language were constructed in the same fashion as in Experiment 5. The only difference between the exposure set of Experiment 5 and the exposure of Experiment 6 was the presence of the one context sequence coming from the opposite subcategory \( A_4X_1B_4 \).

**Procedure:** As in Experiment 5, there are \( 6 \times 4 \times 6 = 144 \) possible combinations of \( A \), \( X \), and \( B \). That set is divided in half so that only 36 of those strings are legal according to the subcategory structure. 24 of these 36 \( AXB \) combinations were included in the training set, plus the 1 boundary-crossing \( AXB \) combination, totaling 25 \( AXB \) exposure strings. After adding optional \( Q \) and \( R \) flanker elements (as in earlier experiments), the training set consisted of 100 strings. As in Experiment 5,
this training set was presented to the participants 5 times, for a complete exposure of 500 strings.

Due to the length of training as well as pilot testing indicating that participants had difficulty paying attention to this long exposure with the current task set-up, participants were told that while listening to the strings they would have to complete a 1-back task. The purpose of this secondary task was to keep participants attentive to the materials and willing to complete the experiment. A pilot experiment replicating Experiment 5 but using this task showed that the 1-back task did not impair performance: planned comparisons on the results of 6 subjects (3 in each of languages 1 and 2) showed no difference between familiar and grammatical novel items ($p=0.806$), 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 5, it appeared justified to use the 1-back task as a way to increase attention to the exposure strings 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 produce attention to the sound sequences). At the beginning of exposure, subjects 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. (Presentation was done via the custom software package and Dell PC setup described for
Experiment 5 such that exposure strings were randomly presented in a different order to each participant. Thus, there was a non-zero chance that the subject would hear items repeated back-to-back.)

Figure 4.2 Pictorial description of the subcategory structure for Experiment 6. There is a clear boundary between the AXB contexts for Subcategory 1 ($X_{1,2}$) and Subcategory 2 ($X_{3,4}$), except for the single boundary-crossing context $A_4X_1B_4$.

Test strings were comprised of a number of types and subtypes:

- 26 Familiar AXB Test Strings:
  - 12 grammatical familiar AXB strings presented during exposure, each presented twice during test (same as Experiment 5);
  - 1 familiar boundary-crossing $A_4X_1B_4$ string, presented twice

- 44 Novel AXB Strings:
  - 12 grammatical novel strings, presented twice (same as Experiment 5);

---

4 Compared to other experiments, this experiment’s test had a different proportion of clearly grammatical (i.e., familiar) items to clearly ungrammatical items. However, pilot results from a replication of Experiment 1 (see Chapter 2) showed that a test with half ungrammatical items and half familiar/grammatical novel items did not lead to different results. Thus, subjects do not appear to adhere to an implicit need to rate approximately half of the items as ungrammatical (or “no votes”); it does not appear to matter in this paradigm whether familiar items are perfectly matched in number to ungrammatical items.
• 4 novel boundary-crossing strings, with X₁ presented with novel A_B contexts from subcategory 2;
• 12 subcategory violation strings in which either the A or the B comes from the opposite subcategory as X (called “type 1” strings);
• 4 subcategory violation strings in which both the A and the B come from the opposite subcategory as X (two X₂ strings, one X₃ and one X₄) (called “type 2” strings)

• 12 Ungrammatical Non-AXB Strings:
  • 12 ungrammatical strings of the types AXA, BXB, AAB, or ABB

Just as in Experiment 5, the “subcategory violation” strings would be judged as grammatical if learners have not separated the two subcategories or if this distinction weakens with exposure to the boundary-crossing strings. The different types of novel items included in this test will distinguish whether learners treat novel AXBs as an exception, whether they generalize all Subcategory 2 contexts to X₁, whether they generalize Subcategory 2 contexts to all Subcategory 1 members, and whether the existence of the boundary-crossing item leads to the loss of a subcategory boundary at all.

4.1.2 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. It therefore appears that the longer test length did not significantly affect participants and their performance on later test questions was consistent with their decisions on earlier test questions. Furthermore, no difference was found between language 1 and language 2, so the two languages have been collapsed for all remaining analyses.

The variety of test string types allows for a number of interesting planned comparisons. The first comparisons of interest are the comparisons used in Experiment 5, (Chapter 3) to explore the relationships among familiar and novel grammatical strings within each subcategory as compared to ungrammatical strings; and these compared with subcategory violations, for those $X$-words that consistently observed the subcategory structure in the exposure materials. Looking at only test strings that did not involve a category-boundary crossing, a repeated measures ANOVA was conducted with test type (familiar, grammatical novel, subcategory violation type 1, ungrammatical) as the within-subjects factor, collapsing across languages 1 and 2. Grammatical novel strings (mean=3.61, $SE=0.10$) were rated just as high as familiar strings (mean=3.64, $SE=0.09$) ($F(1,14)=0.402$, $p=0.536$). Grammatical novel strings were rated higher than subcategory violation type 1 strings (mean=3.42, $SE=0.105$) with marginal significance ($F(1,14)=3.621$, $p=0.078$), and
subcategory violation type 1 strings were rated significantly lower than grammatical familiar strings ($F(1,14)=10.293, p<0.01$). These results indicate that participants generalized fully from familiar strings to novel strings within the subcategory boundaries, but they did not extend this generalization to strings where either A or B was from the opposite subcategory as X. The mean rating of ungrammatical items was 2.8056 (SE = 0.18), which was significantly lower than subcategory violation strings ($F(1,14)=16.923, p<0.005$) (see Figure 4.3).

Figure 4.3 Rating score results for non-boundary crossing strings from Experiment 6.

Since one of the comparisons above was marginally significant, it was important to do the same comparisons on standardized rating scores. When the same repeated measures ANOVA was conducted on z-scores (see Figure 4.4), the same pattern of results was obtained. However, Mauchly’s test of sphericity indicated that
the assumption of sphericity had been violated (chi-square=59.834, \( p<0.001 \)); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (epsilon=0.433). The results showed that there was a significant main effect of test string type \( (F(1.298, 18.166)=21.891, \ p<0.001) \). Post-hoc tests revealed that familiar and grammatical novel strings were rated the same \( (p>0.1) \), but that both were rated significantly higher than subcategory violation type 1 strings \( (p<0.05 \ for both) \). Subcategory violation type 1 strings were in turn rated significantly higher than ungrammatical strings \( (p=0.007) \).

**Figure 4.4** Z-score results for non-boundary crossing strings from Experiment 6.
The next comparison of interest is among the boundary-crossing types of test strings, comparing familiar grammatical boundary crossing strings $A_4X_1B_4$, novel strings in which $X_1$ was presented with other $A_B$ contexts from subcategory 2, and subcategory violation type 2 strings in which other $X$-words were presented with $A$ and $B$ contexts coming from the opposite subcategory as $X$. A repeated measures ANOVA was conducted with test type as the within-subjects factor, collapsing across language 1 and 2. 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 3.17 ($SE=0.16$) (see Figure 4.5).

**Figure 4.5** Raw score results for boundary crossing strings from Experiment 6. Ratings of ungrammatical (non-boundary crossing) strings are included for comparison.

(BC: boundary crossing)
The results of the ANOVA revealed that familiar boundary crossing strings were rated significantly higher than novel boundary crossing strings ($F(1,14) = 4.713, p < 0.05$), and novel boundary crossing strings were rated the same as subcategory violation type 2 strings ($F(1,14) = 0.79, p = 0.388$). This indicates that learners treat $A_4 X_1 B_4$ (the familiar boundary crossing string) as an exception and not generalizing across subcategory boundaries to other possible $A_B$ contexts for that $X$-word or for other $X$-words. Performing the same repeated measures ANOVA on $z$-scores reveals the same pattern of results: the difference between familiar and novel boundary crossing ratings is marginally significant ($F(1,14) = 4.123, p = 0.062$), and novel boundary crossing strings are not different from subcategory violation type 2 strings ($F(1,14) = 1.228, p = 0.286$).

**Figure 4.6** Z-score results for boundary crossing strings from Experiment 6. Z-scores of ungrammatical (non-boundary crossing) strings are included for comparison. (BC: boundary crossing)
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 can 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, either for
raw ratings (\( p > 0.5 \)), or \( z \)-scores (\( p > 0.1 \)). 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 (\( p > 0.1 \) for using raw
scores and \( z \)-scores). These results indicate that learners treated the \( A_4X_1B_4 \) 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.

4.1.3 Discussion

In Experiment 6, as in Experiment 5, we see that a dense sampling of the
language with complete overlap within subcategories and almost no overlap between
subcategories (except for the boundary crossing context) leads learners to treat novel
strings within the subcategories as grammatical. Furthermore, the distributional
properties of the exposure also encourage learners to reject novel strings that cross
subcategory boundaries, either with one context element (subcategory violation type 1 strings) or both (subcategory violation type 2 strings). The results also show that learners did not extend the $A_4B_4$ context that they heard with $X_1$ to the other members of $X_1$'s subcategory (i.e., $X_2$). The low ratings of novel boundary crossing test strings (where $X_1$ was heard in other contexts from Subcategory 2) demonstrate that participants did not treat $X_1$ as belonging to both subcategories. Additionally, by giving low ratings to both types of subcategory violation strings, learners demonstrated that they did not simply become more flexible with their subcategory boundaries once they were exposed to the boundary crossing item (see Figure 4.7).

**Figure 4.7** Difference scores for boundary crossing test strings. Note that both novel boundary crossing strings ($X_1$ with $A_B$ contexts from Subcategory 2) and subcategory violation type 2 strings ($X_{2,3,4}$ with $A_B$ contexts from the opposite subcategory) are rated significantly different from the familiar boundary crossing string ($A_4X_1B_4$). (BC: boundary crossing)
These data are an interesting extension of Experiment 5, not only showing that learners can overcome slight variations in their input where clear subcategory boundaries are violated, but they can also 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 helps learners decide that \( A_4X_1B_4 \) is a special case, and the boundary crossing behavior should not be allowed for other \( X \)-words. Next, we will manipulate the exposure by giving the learner sparse input and incomplete overlap within subcategories, along with one boundary crossing context string (\( A_4X_1B_4 \)). This will allow us to test whether learners will restrict generalization in the same way as in Experiments 3 and 4 (Chapter 2).

4.2 Experiment 7: Subcategorization Without Consistent Subcategory Boundaries: Sparse Sampling and Incomplete Overlap

In Experiment 7, we present only 3 of the 9 possible \( AXB \) strings for each \( X \) (a sparse sample) plus the one boundary crossing string \( A_4X_1B_4 \). Furthermore, as in Experiment 3, each \( X \)-word is only heard with 2 of the 3 legal A’s or B’s for its subcategory, and only partially overlaps with the other member of its subcategory. The question now is whether learners will restrict their generalization within the subcategory but continue to maintain a subcategory boundary (similar to the result from Experiment 3), or whether the weakened subcategory cues in the presence of a repeating non-subcategory conforming string will be enough for learners to disregard subcategory boundaries.
4.2.1 Method

Participants: 20 monolingual native English speaking students at the University of Rochester participated in Experiment 7 for payment. 10 were assigned to each of languages 1 and 2 (see Table 3.1).

Stimulus Materials: Strings were constructed in the same manner as Experiment 6. However, the exposure did not have the dense sampling of the language or the complete overlap within subcategories that Experiment 6 had. To achieve this, \( X_1 \) was only heard in the context of \( A_1, A_3, B_1, \) and \( B_3 \) (and also the boundary crossing context, \( A_4 \_B_4 \)); \( X_2 \) was only heard in the context of \( A_2, A_3, B_2, \) and \( B_5 \); \( X_3 \) was only heard in the context of \( A_4, A_5, B_4, \) and \( B_5 \); and \( X_4 \) was only heard in the context of \( A_4, A_6, B_4, \) and \( B_6 \) (see Figure 4.7). This creates some differences from the language of Experiment 3: notably, the predictability of \( X \) given \( A \) or \( B \) is changed. For example, in Experiment 3, \( A_1 \) and \( B_1 \) were always heard with \( X_1 \) or \( X_3 \) but never with \( X_2 \). In this experiment, \( A_1 \) and \( B_1 \) are only ever heard with \( X_1 \), so they are 100% “predictive” of \( X_1 \). This can be contrasted with \( A_4 \) and \( B_4 \), which are each heard 33% of the time with \( X_3 \), 33% with \( X_4 \), and 33% with \( X_1 \). Thus, at an abstract level, the sparseness and incomplete overlap of this experiment is comparable to Experiment 3/4. However, the actual ratios of context element overlap are different. Lastly, the exposure consisted of two times through the training set (so 10 times through all of the possible strings, not 5).

Procedure: The procedure was the same as in Experiment 6. Due to the size of the
training set (even with the reduced sampling), this experiment doubled the amount of exposure to the training set (not tripling it as in Experiment 4: Chapter 2).

**Figure 4.8** Pictorial description of the language in Experiment 7 where there is incomplete overlap within subcategories and one boundary crossing context for $X_1$.

4.2.2 Results

Split-half reliability item analyses were conducted to see if order effects existed between the first and second half of the test. Cronbach’s alpha values were at acceptable levels for all test item types (all $\alpha$ values >0.70). A repeated measures ANOVA with language as the between subjects contrast and test type (non-boundary crossing familiar, grammatical novel, subcategory violation type 1, ungrammatical) as
the within-subjects contrast showed no differences between languages 1 and 2 \((F<1)\), so they have been collapsed for all following analyses.

Starting with the non-boundary crossing test strings, the mean rating of familiar strings was 4.00 \((SE=0.10)\), the mean rating of grammatical novel strings was 3.63 \((SE=0.11)\), the mean rating of subcategory violation type 1 strings was 3.52 \((SE=0.12)\), and the mean rating of ungrammatical strings was 2.53 \((SE=0.12)\) (see Figure 4.9).

**Figure 4.9** Raw rating scores of the non-boundary crossing strings in Experiment 7.

![Figure 4.9](image)

A repeated measures ANOVA with test type (non-boundary crossing familiar, grammatical novel, subcategory violation type 1, ungrammatical) as the within-subjects contrast showed that familiar strings were rated significantly higher than grammatical novel strings \((F(1,19)=13.556, p<0.005)\). However, grammatical novel strings were no different than subcategory violation type 1 strings \((F(1,19)=2.399, p>0.05)\).
Ungrammatical strings were significantly lower than subcategory violation type 1 strings \( (F(1,19)=62.319, p<0.001) \).

Z-scores patterned in the same way. Another repeated measures ANOVA on z-scores showed that familiar and grammatical novel strings were significantly different \( (F(1,19)=29.992, p<0.001) \), grammatical novel and subcategory violation type 1 strings were no different \( (F(1,19)=1.621, p>0.20) \), but ungrammatical strings were significantly lower than subcategory violation type 1 strings \( (F(1,19)=72.187, p<0.001) \).

Next, we explored the differences between the boundary crossing test strings. The mean rating of the familiar boundary crossing strings \( (A_dX_1B_4) \) was 4.05 \( (SE=0.17) \), the mean rating of the novel boundary crossing strings (novel strings in which \( X_1 \) was presented with other \( A_B \) contexts from Subcategory 2) was 3.12 \( (SE=0.19) \), and the mean rating of the subcategory violation type 2 strings was 3.145 \( (SE=0.15) \) (see Figure 4.10). A repeated measures ANOVA with test type (familiar boundary crossing, novel boundary crossing, subcategory violation type 2) as the within-subjects factor revealed significant differences between familiar and novel boundary crossing strings \( (F(1,19)=20.975, p<0.001) \) but no difference between novel boundary crossing strings and subcategory violation type 2 strings \( (F<1) \). Performing the same repeated measures ANOVA on z-scores gave the same pattern of results: familiar and novel boundary crossing strings were rated significantly different from each other \( (F(1,19)=16.36, p<0.005) \) but novel boundary crossing strings and subcategory violation type 2 strings were no different from each other \( (F<1) \).
4.2.3 Discussion

Interestingly, learners show the same pattern in this experiment as in Experiment 3 (Chapter 2) when we reduce overlap among contexts within a subcategory: they rate non-boundary crossing grammatical novel strings lower than strings that they have heard during exposure to the language. Additionally, these non-boundary crossing grammatical novel strings are rated just as low as strings that have one contextual element crossing a subcategory boundary. While the overall non-boundary crossing familiar, novel, and subcategory violation type 1 strings show a different pattern from Experiment 6, the results comparing the boundary crossing test strings are qualitatively equivalent to those in Experiment 6. Learners do not generalize the exception $A_4\_B_4$ context to other $X$-words, nor do they become more
flexible with generalizing new contexts to $X_1$ (or other $X$’s), despite the increased uncertainty about the language introduced by more systematic gaps presented more frequently. Furthermore, comparing the difference scores of ratings of non-boundary crossing strings from Experiments 6 and 7 shows the same trend we saw in Chapter 2 between Experiments 1 and 3/4 (see Figure 4.11). Despite the complications of having an exception inserted into exposure, learners continued to restrict generalization to novel boundary crossing strings and novel non-boundary crossing strings.

**Figure 4.11** Comparison of differences of raw rating scores of non-boundary crossing strings in Experiments 6 & 7.

The pattern of contexts that learners are encoding for learning the subcategory boundaries is apparently specific enough to record exceptions like $A_4 X_1 B_4$. At the
same time, it is broad enough in Experiment 6 to suggest the grammaticality of the non-boundary crossing novel grammatical strings, and more conservative in Experiment 7 to restrict generalization of this type. Next, we return to a dense sampling of the language with complete overlap within subcategories to see how learners behave when exposed to multiple boundary crossing contexts, where there are four subcategory exceptions (one for each \( X \)).

**4.3 Experiment 8: Subcategorization Without Consistent Subcategory Boundaries: Increased Subcategory Exceptions**

Experiment 8 examines how learners treat a systematic increase in the number of boundary crossing strings, when the exposure set includes boundary crossing exemplars from *both* subcategories (see Figure 4.12). All other aspects of the input remain the same as in Experiment 6. The question we ask is whether the increased uncertainty, in the form of more violations of subcategory boundaries in exposure, is enough to encourage learners to be more conservative with their generalizations to the non-boundary crossing grammatical novel strings or less conservative with their generalizations to the boundary crossing novel strings.

**4.3.1 Method**

*Participants:* 22 monolingual native English speaking adults from the University of Rochester participated in Experiment 8. They were paid for their participation and had not participated in any other categorization experiment. Eleven participants were assigned to each of the two languages.
**Stimulus Materials:** Exposure strings were the same strings as in Experiment 6, plus three additional boundary crossing strings (A6X2B5, A2X3B1, A3X4B2).

**Procedure:** All aspects of the procedure during training were the same as in Experiment 6 (dense sampling, complete overlap within subcategories) except for the addition of the three additional boundary crossing strings and their QAXB, AXBR, and QAXBR forms. Thus, there were 560 total strings in the training set. The same 1-back task as in Experiment 6 was used to help focus participants during the long training time.

The procedural aspects of the test were the same as in Experiment 6. Experiment 8, however, used fewer types of test strings:
• 28 Familiar AXB Strings:
  o 12 grammatical familiar AXB strings presented during exposure, presented twice (same as Experiment 6);
  o 4 familiar boundary crossing strings (A4X1B4, A6X2B5, A2X3B4, A3X4B2)

• 40 Novel AXB Strings:
  o 12 grammatical novel strings, presented twice (same as Experiment 6);
  o 4 novel boundary crossing strings (A5X1B5, A4X2B6, A1X3B2, A2X4B3) (note that these strings are subcategory violation “type 2” strings from Experiment 6/7);
  o 12 subcategory violation strings (same as subcategory violation “type 1” strings from Experiment 6), where either the A or the B comes from the opposite subcategory as X

• 12 Ungrammatical Non-AXB Strings:
  o 12 ungrammatical strings of the types AXA, BXB, AAB, or ABB

4.3.2 Results

Cronbach’s alpha was computed on the ratings for each type of test strings to check split-half reliability. All types of test items had acceptable alpha values (lowest α = 0.71), so the length of the test did not impact the split-half reliability of the ratings within test type. There was no difference between the two languages (F<1), so the two languages have been collapsed for all further analyses.

Turning to the non-boundary crossing test strings first, (non-boundary crossing familiar, non-boundary crossing grammatical novel, subcategory violation
“type 1”, and ungrammatical), the mean rating of familiar strings was 3.65 (SE=0.09), the mean rating of grammatical novel strings was 3.62 (SE=0.08), the mean rating of subcategory violation strings was 3.427 (SE=0.10), and the mean rating of ungrammatical strings was 2.518 (SE=0.11) (see Figure 4.13).

Figure 4.13 Raw score results for Experiment 8 for non-boundary crossing test items.

A repeated measures ANOVA revealed that there was no significant difference between familiar and grammatical novel strings \((F(1,21)=0.3, p>0.5)\), but that there was a significant difference between grammatical novel and subcategory violation strings \((F(1,21)=4.663, p<0.05)\) and a significant difference between subcategory violation strings and ungrammatical strings \((F(1,21)=60.15, p<0.001)\). Transforming the raw ratings into z-scores and conducting another repeated measures
ANOVA revealed the same pattern of results: familiar and grammatical novel strings were rated the same ($F(1,21)=0.8, p>0.7$), but grammatical novel strings were rated significantly higher than subcategory violation strings ($F(1,21)=5.77, p<0.05$) and ungrammatical strings were rated significantly lower than subcategory violation strings ($F(1,21)=40.22, p<0.005$).

Comparing the boundary crossing test strings only, the mean rating of familiar boundary crossing strings was $3.89$ ($SE=0.08$) and the mean rating of novel boundary crossing strings was $3.03$ ($SE=0.15$) (see Figure 4.14).

**Figure 4.14** Raw score results for boundary crossing strings from Experiment 8. Ratings of ungrammatical (non-boundary crossing) strings are included for comparison. (BC: boundary crossing)

A repeated measures ANOVA with test type (familiar boundary crossing, novel boundary crossing) as the within-subjects contrast revealed that familiar boundary crossing strings were rated significantly higher than novel ones.
Conducting the same repeated measures ANOVA on z-scores also showed that familiar boundary crossing strings were rated significantly different from novel boundary crossing strings ($F(1,21)=30.00, p<0.001$).

4.3.3 Discussion

The results of Experiment 8 once again show that learners are highly sensitive to the contexts in which each $X$ appears, and they are able to use the presence and absence of contextual cues to partition the language space into two subcategories. Participants were able to discover the subcategorical structure of the language despite being exposed to four “exception” contexts that crossed subcategory boundaries for each $X$. Furthermore, participants must have utilized the absence of cues to the novel boundary crossing strings in order to decide to restrict generalization across subcategory boundaries. In this experiment, we carefully chose the $A_B$ contexts for the boundary crossing strings in the exposure to be contexts that were already present in grammatical boundary-obeying items in the exposure. The $A_B$ contexts used in the novel boundary crossing test strings were also present in the exposure set, so learners experienced these exact $A-B$ pairings with the subcategory-appropriate $X$. Thus, the low rating of novel boundary crossing strings cannot be due to learning a nonadjacent dependency of $A_B$ without any regard to $X$. Instead, it seems clear that learners have marked the boundary crossing strings present in the exposure as grammatical exceptions in their input, but not meaningful extensions of an entire subcategory.
Comparing the results from Experiment 8 with those from Experiment 5 (Chapter 3), it becomes apparent that the difference between subcategory violation (type 1) strings and familiar strings is much smaller in Experiment 8. It should be mentioned that earlier pilot testing on this design without utilizing the 1-back task resulted in subcategory boundaries breaking down such that subcategory violation strings were rated the same as familiar strings. However, in this pilot testing, results were weak regarding the one comparison that should have the largest difference (the mean of familiar strings was 3.43, \( SE=0.22 \), and the mean of ungrammatical strings was 3.01, \( SE=0.10 \)), and a post-hoc analysis of test items showed that Cronbach’s alpha was low for some tests of grammatical novel strings and subcategory violation strings. This seems to be an indication that the experiment was too long without a
task to engage subjects and encourage them to listen to the exposure set. Though training was just as long in the final version of Experiment 8 as in this pilot version (approximately 45 minutes), the 1-back task engaged participants so as to ensure that they were listening to the sentences during exposure without imposing large demands on working memory. The results of Experiment 8 show a smaller difference between the subcategory violation strings and the grammatical strings than was seen in Experiment 5 (see the “Familiar – Subcategory Violation” bar in Figure 4.15 compared with Figure 3.3), but this might indicate that with less exposure (or less attention to the task), learners would lose the subcategory distinction. If the gaps were made more salient (by increasing the number of times learners were exposed to the training set), the boundary crossing strings might become even more salient as exceptions to the subcategory rule, and we would expect to see subcategory violation and novel boundary crossing strings rated lower.

However, even though subcategory violation strings are not rated as low as in earlier experiments, it’s quite clear that learners do not generalize to the novel boundary crossing strings where both A and B come from the opposite subcategory (see the comparison in Figure 4.16). This is striking evidence that the robustness of the distributional cues for subcategories cannot be undone by systematic (though rare and infrequent) “exception” cases of contexts crossing subcategory boundaries.
4.4 General Discussion

The series of experiments described in this chapter attempt to answer 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. In all of these experiments, the boundary crossing context could be viewed as a grammatical extension of the language if learners did not entertain the possibility of subcategory boundaries (as in
the simple categorization case described in Chapter 2)\textsuperscript{5}. Experiment 6, however, showed that learners considered the boundary crossing context to be an exception in their input, and that they showed the same overall effects as Experiment 5 by encoding the subcategory boundary consistent with the majority of the distributional information. Experiment 7 builds on this finding by showing a similar pattern to the effects described in Chapter 2, where sparser sampling and decreased overlap lead to restricted generalization. That is, learners in Experiment 7 showed reduced generalization to grammatical novel strings within subcategory boundaries. However, they also rejected strings that did not obey subcategory boundaries, an effect that demonstrates the ability to encode the boundary crossing string as an exception in their input even when there is increased uncertainty or more lexical idiosyncrasy for the language overall. Experiment 8 showed the same overall pattern as Experiment 5 (learners generalized to grammatical novel strings that obeyed subcategory boundaries but rejected strings that did not), despite an increase in exposure to strings that cross subcategory boundaries. Apparently, the distributional information that licensed crossing subcategory boundaries could not overcome the distributional information supporting the existence of the two subcategories.

These experiments strengthen the idea that the absence of a cue in the input can be just as informative as the presence of a cue (discussed further in Chapter 6). It is clear that learners were highly sensitive to the lack of information licensing novel

\textsuperscript{5} One question that this chapter cannot answer is whether the learner maintains two “lexical entries” for the X-words heard in boundary crossing contexts, or whether learners maintain an exception context for the boundary crossing X-words (such as \textit{paint} (noun) versus \textit{paint} (verb)). But the answer to this question does not change how the experimental results speak to our larger theoretical interests about how learners use distributional information in boundary crossing situations.
boundary crossing strings or supporting the dissolution of the subcategory boundaries. This behavior restricts what kind of a mechanism we might propose that learners use to perform subcategorization.

First, consider the comparison between these data and the data reported by Reeder, Newport and Aslin (2010). Using a similar paradigm to the categorization experiments described in Chapter 2, Reeder, Newport and Aslin showed that when learners are exposed to a novel $X$-word ($X_4$) that is only heard in just a single context (a previously-experienced $AXB$ context), they extend the same context generalization to $X_4$ that they do to the other $X$-words that they heard much more extensively and in many contexts during exposure ($X_1$-$X_3$). However, this was in a circumstance where the cues to category formation for $X_1$-$X_3$ were quite strong and clear. In other words, when there are strong enough distributional cues to license generalization to novel $A_B$ contexts for items that learners have grouped into one category, learners extend those generalizations to $X_4$. On the other hand, when there are weak distributional cues to categorization (sparse sampling, incomplete overlap and increased number of times through the training set), learners restrict generalization of $A_B$ contexts for $X_4$. But this is not the pattern we see across Experiments 6-8, where learners do not apply the same generalizations they make within subcategories to the novel $X$ that is seen in the boundary crossing context. The difference in results between these two sets of experiments could be due to a number of factors: Reeder et al. (2010) use a novel $X_4$ that is very rare in the exposure set, whereas Experiments 6-8 presented here use familiar $X$-words presented in novel contexts. It could also be that the many types of
distributional cues (distributional information signaling categories combined with information signaling subcategories, in addition to the information added by the boundary crossing contexts) make the learner much more careful in storing what they do and do not hear, thus making them less likely to extend the boundary crossing contexts to the other members of the subcategory. Despite these different findings, both Experiments 6-8 and Reeder et al. (2010) show that learners care a great deal about missing information, and over the course of differing amounts of exposure they update their beliefs regarding the validity of absent contexts based on the strength of the distributional information supporting the experienced contexts.

The next set of experiments continues our exploration of how learners handle variations in the structure of categories and subcategories that are solely defined by distributional information. Whereas this chapter dealt with exposing the learner to imperfect subcategory boundaries, Chapter 5 will explore the role of varying frequency information for different X-words in the categorization and subcategorization problems.
Chapter 5: Varying Frequency

Returning to the MN/PQ problem outlined in Chapter 1, Braine (1987) noted a major difference between Smith’s (1969) grammar and natural language. Let us liken the MN/PQ subcategorization problem to noun phrases in a natural gender system, where M and P refer to subclasses of determiners that are associated with different genders, and N and Q refer to masculine or feminine nouns (respectively). Gender paradigms in natural language have a systematic imbalance in class size and frequency between the M/P and N/Q classes. As determiners, M and P would be members of a small, closed-class category that are some of the most frequent words in natural speech. As nouns, N and Q are very large subcategories, but individual items within N and Q are vastly less frequent than the items in the determiner subcategories, and we might consider the exposure the learner gets to the items within N and Q as sparse.

The problem becomes more complicated when we consider the natural variation of individual item frequencies in the input with regards to individual item frequencies. Even within a category or subcategory, certain lexical items are heard more frequently than others. This might be particularly true in the case of child language acquisition, where vocabulary in the input is skewed towards certain lexical items (e.g., labels of toys) but not others (e.g., names of plants). Thus, being sensitive to frequency variations while not being mislead into rejecting infrequent exemplars is an important aspect of natural language acquisition process.
Given the problem of learning from a sparse and uneven input corpora, and we arrive at the main question of interest in this chapter: how does frequency variability in the input affect the robustness of learning, as well as the ability to generalize to novel contexts? If information for specific lexical items is particularly sparse, how will learners use consistency and frequency to make decisions regarding categorization and generalization? It is possible that learners might use information from more frequent items to form a category, and then apply the properties of this category to strongly overlapping items (regardless of their frequency). On the other hand, when items are less frequent, learners might inherently be less certain about those items (or the category as a whole).

Along these lines, having uneven frequency information within a category might encourage learners to be more cautious with generalizing over gaps. This raises the question of whether learners will show the same overall categorization behavior in our miniature language paradigm if we systematically manipulate the frequency of certain items. That is, when given a rich sampling of the language space with some items presented more frequently than others, will learners continue using the available distributional cues in the same way as we have seen in Experiments 1-8? Conversely, will learners show more conservatism and less generalization in situations where there is uneven frequency information about particular items? Furthermore, what will be the effect of uneven frequency information on individual items? We attempt to answer these questions in the domains of category learning (as in Chapter 2) and subcategory learning (as in Chapter 3).
5.1 Frequency Variation in Categorization (Reeder, Newport, Aslin & Schuler, in prep)

In Reeder, Newport, Aslin and Schuler (in prep), frequency variations were explored in a (Q)AXB(R) categorization experiment that was similar to Experiment 4 (Chapter 2). Exposure consisted of a sparse sampling of the language and incomplete overlap of contexts among the X words (each X was heard with only 2 of the three A’s and B’s). Frequency was manipulated by using a 3:2:1 ratio of X₁ to X₂ to X₃ strings. Learners were exposed to a corpus containing 288 X₁ strings, 192 X₂ strings, and 96 X₃ strings across all of the possible sentence types (AXB, QAXB, AXBR, QAXBR). Participants rated the grammaticality of the same types of 3-word test items as in Experiment 4 (Chapter 2): familiar (items presented during training), grammatical novel (items withheld from the training set), and ungrammatical (AXA, BXB, AAB, ABB).

The frequency manipulation employed in this experiment is a first step to exploring how learners interpret missing information. One possibility is that learners regard low frequency and high frequency words differently. Learners might have less certainty regarding gaps in the input for low frequency words, interpreting the input for these words as an imperfect and unbalanced sampling of the language space. In this case, learners’ responses might pattern differently than those in Experiment 4: we would expect to see an increase in generalization across the missing contexts and no difference in ratings of grammatical novel and familiar strings, despite the sparse sampling and incomplete overlap for the language as a whole.
Another possibility, however, is that learning is robust to small frequency manipulations in this paradigm, either because frequency is outweighed by the strength of the other distributional cues in the paradigm, or because the frequency difference in such a small language is too tiny to impact generalization. The frequency differences between grammatical categories in natural languages are on a much larger scale than 3:1, and as such, the results from this type of miniature language might reflect a kind of ceiling effect that does not allow for the frequency manipulation to affect a change at the level of item type analyses.

The results of this experiment were in line with the second hypothesis. Participant ratings of familiar test strings were significantly higher than their ratings of grammatical novel strings \((F(1,14)=11.625, p<0.01)\), and grammatical novel strings were rated significantly higher than ungrammatical strings \((F(1,14)=100.719, p<0.001)\). Despite the frequency manipulation, then, the results pattern precisely like those in Experiment 4. Individual item analyses showed no differences among ratings of test strings containing \(X_1\), \(X_2\), and \(X_3\), except that \(X_3\) familiar strings were rated significantly lower than \(X_2\) familiar strings.

Comparing these results with those of Experiment 4, we can conclude that learners are capable of overcoming variations in frequency information while continuing to use a rational learning mechanism for categorization. That the low frequency \(X_3\) items were rated significantly lower than the higher frequency \(X_2\) items indicates that learners are sensitive to the frequencies in their input, and their ratings reflect an attempt to match the frequency statistics of the exposure set. The lack of
any difference between X-word novel strings or ungrammatical strings indicates that learners might move away from frequency matching and move towards a more rule-governed approach when faced with a generalization task.

5.2 Experiment 9: Frequency Variation in Subcategorization

In Experiment 9, we extend frequency variation to the subcategorization problem. Here we systematically vary the frequency information for the items within one subcategory but keep the frequency information of the other subcategory the same as in our earlier subcategory experiments. The learner thus receives the same amount of exposure to each subcategory, but within one subcategory the possible contexts for $X_1$ are infrequent and sparsely sampled, while possible $X_2$ contexts are densely sampled and very frequent during exposure.

5.2.1 Method

Participants: 18 monolingual native English-speaking students at the University of Rochester participated for payment. They had not participated in any other categorization experiment. One subject was removed for inattentiveness during training, and 1 subject was removed for not understanding the test procedure. Of the remaining 16 participants, eight were in language 1 and eight were in language 2.

Stimulus Materials: Materials were created in the same manner as in Experiment 5 (Chapter 3). The assignment of words to categories was the same as in Table 3.1, and the subcategory structure of the language was the same as depicted in Figure 3.1.
Procedure: The possible strings of the language were almost the same as in Experiment 5 with one major difference: while the frequencies and presentation of items in Subcategory 2 remained unchanged (there was a dense sampling and complete overlap of contexts within the subcategory), frequencies and sampling density were changed within Subcategory 1 such that $X_1$ strings were sparsely sampled and infrequent while $X_2$ strings were densely sampled and frequent. By doing this, the frequency of exposure and number of Subcategory 1 strings in training was equal to the frequency of exposure and number of Subcategory 2 strings. Thus there was equal information regarding the subcategory boundaries of the language. For Subcategory 2 strings, the learner was exposed to a dense sampling (6 of the 9 possible basic $AXB$ strings for $X_3$ and $X_4$) and complete overlap, such that $X_3$ and $X_4$ were heard with every possible subcategory-conforming $A$ and $B$. It was only within Subcategory 1 that frequency and sampling density were varied. Of the 9 possible $AX_1B$ strings, the learner was exposed to 3 (a sparse sample); however, of the 9 possible basic $AX_2B$ strings, the learner heard 6 (a dense sample). Complete overlap within the subcategory was maintained such that both $X_1$ and $X_2$ were heard with every subcategory-conforming $A$ and $B$.

As in Experiment 5, the exposure set contained 5 repetitions of the training set. To achieve equal exposure to each subcategory, each training set included $\frac{1}{8} X_1$ strings, $\frac{3}{8} X_2$ strings, $\frac{1}{4} X_3$ strings, and $\frac{1}{4} X_4$ strings across the four possible types of string ($AXB$, $QAXB$, $AXBR$, and $QAXBR$). All procedural aspects of training were the same as in Chapter 4: subjects were seated in a sound attenuated
booth, and the experiment was presented via a custom software package on a Dell PC with the subject wearing headphones. Exposure strings were presented randomly, and training lasted for approximately 40 minutes. During training, participants used the 1-back task described in Chapter 4.

Test strings were constructed in the same fashion as those in Experiment 5: there were familiar AXB strings, grammatical novel AXB strings, subcategory violation AXB strings, and ungrammatical strings (AXA, BXB, AAB or ABB). Test items were divided evenly between X1, X2, X3, and X4 (except for AAB and ABB ungrammatical test strings) so that item analyses could be conducted to explore the effect of the frequency manipulation. Test strings were presented pseudo-randomly such that half of each of the familiar, grammatical novel, subcategory violation, and ungrammatical test strings appeared during each half of the test, and there were an approximately even number of each X test string for each half.

5.2.2 Results

A repeated measures ANOVA with language as the between-subjects effect and condition (familiar, grammatical novel, subcategory violation, and ungrammatical) as the within-subjects effect revealed no differences between languages 1 and 2 ($F<1$). The mean rating of grammatical familiar items was 3.63 ($SE=0.08$), the mean rating of grammatical novel items was 3.51 ($SE=0.09$), the mean rating of subcategory violation items was 3.19 ($SE=0.09$), and the mean rating of ungrammatical items was 2.59 ($SE=0.09$) (see Figures 5.1 and 5.2).
Grammatical familiar and grammatical novel items were marginally significantly different from each other ($F(1,14)=4.439, p=0.054$). However, subcategory violation items were rated significantly lower than grammatical items ($F(1,14)=19.00$, $p=0.001$).
Ungrammatical items were rated the lowest, significantly lower than subcategory violation items ($F(1,14)=24.075, p<0.001$).

Raw scores were transformed into z-scores (see Figure 5.3) and another repeated measures ANOVA was conducted. Familiar and grammatical novel items were rated the same ($F(1,14)=3.983, p=0.066$), but subcategory violation items were rated significantly lower than the grammatical items ($F(1,14)=18.580, p<0.005$) and significantly higher than ungrammatical items ($F(1,14)=20.861, p<0.001$) (see Figure 5.3).

**Figure 5.3** Z-scores from Experiment 9.

Individual item analyses were conducted to further explore the influence of the frequency variation. Mean raw ratings for test items containing $X_1-X_4$ are shown in Table 5.1. Because there were no differences between language 1 and 2, they have been collapsed for the item analyses.
Table 5.1 Mean raw ratings from Experiment 9 dividing up test strings by whether they contain $X_1$, $X_2$, $X_3$ or $X_4$. (Mean rating in boldface, standard errors in parentheses.)

<table>
<thead>
<tr>
<th></th>
<th>Raw Rating Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_1$</td>
</tr>
<tr>
<td>Familiar</td>
<td>3.74(0.11)</td>
</tr>
<tr>
<td>Grammatical Novel</td>
<td>3.76(0.09)</td>
</tr>
<tr>
<td>Subcategory Violation</td>
<td>2.94(0.13)</td>
</tr>
<tr>
<td>Ungrammatical</td>
<td>2.21(0.17)</td>
</tr>
</tbody>
</table>

A repeated measures 4 ($X_1$, $X_2$, $X_3$, $X_4$) x 4 (familiar, grammatical novel, subcategory violation, ungrammatical) ANOVA was conducted to explore the difference between test item types for each $X$-word. The ANOVA revealed that $X_3$ and $X_4$ each showed the same trends across test type as the overall results: familiar test strings were no different from grammatical novel strings, but these were rated significantly higher than subcategory violation strings with ungrammatical strings significantly lower than all test string types.

For $X_1$ (the sparsely sampled and infrequently presented word), familiar strings were rated no different from grammatical novel strings ($F<1$), but subcategory violation strings were rated significantly lower than the grammatical strings ($F(1,15)=25.34, p<0.001$) and ungrammatical strings were rated significantly below these ($F(1,15)=11.00, p<0.01$) (see Figure 5.4).
Figure 5.4 Raw scores from Experiment 9 comparing test strings containing $X_1$, $X_2$, and the average of strings containing $X_3$ & $X_4$. Note the different trends for $X_1$ versus $X_2$, while $X_3$ and $X_4$ follow the typical trend of familiar strings equal to grammatical novel strings.

When converted to z-scores, the results are the same: familiar strings containing $X_1$ are no different from the grammatical novel strings ($F<1$), but these were significantly higher than subcategory violation strings ($F(1,15)=22.18, p<0.001$) which were, in turn, higher than ungrammatical strings ($F(1,15)=10.76, p<0.01$) (see Figure 5.5).

For $X_2$ (the densely sampled and frequently presented word), however, familiar strings were rated significantly higher than grammatical novel strings ($F(1,15)=5.501, p<0.05$), and grammatical novel strings were no different than subcategory violation strings ($F<1$). Ungrammatical strings were rated significantly lower than these ($F(1,15)=8.67, p<0.01$). When using z-scores in the ANOVA, the results were similar: familiar strings were rated higher than grammatical novel strings ($F(1,15)=5.50, p<0.05$); grammatical novel and subcategory violation strings were no
different from each other \((F(1,15)<1)\); and ungrammatical strings were rated significantly lower than these \((F(1,15)=6.76, p<0.01)\). This is a very different pattern than that for the other X-words.

**Figure 5.5** Z-scores from Experiment 9 comparing test strings containing \(X_1, X_2, \) and the average of strings containing \(X_3 \& X_4\).

5.2.3 Discussion

As in the subcategorization experiments described in Chapters 3 and 4, the results of Experiment 9 once again show that subcategory structure can be learned based on distributional information alone. One major difference between Experiment 9 and the other experiments with dense sampling of the language and complete overlap of (sub)category-appropriate contexts is that this experiment resulted in a much larger difference between familiar and grammatical novel test strings. While overall there is not a *significant* difference between familiar and grammatical novel test strings, the familiar items are almost significantly higher than the novel strings, even when transformed into standardized z-scores. This marginal difference is
certainly noteworthy given how indistinguishable the two test types were to learners in Experiments 5, 6, and 8 (the other subcategory experiments with complete overlap and dense sampling of the language). This small overall difference between familiar and novel grammatical is undoubtedly due to the more strikingly reduced generalization to grammatical novel test strings containing X2. The lower rating of grammatical novel X2 test strings fits in with our hypothesis about how learners utilize distributional information within the subcategory to decide when to generalize. Since learners are exposed to the X2 training strings with increased frequency, the gaps made by the withheld X2 strings (i.e., the grammatical novel test strings) become all the more obvious. Given that the information regarding X1 is very sparse, infrequent and uncertain, it might not be rational to generalize to the withheld X2 strings because Subcategory 1 lacks the distributional strength of Subcategory 2 (or of the subcategories in any of the other experiments presented in this dissertation). Learners may be using their more complete knowledge of Subcategory 2 to license some generalization within Subcategory 1, but not as much as when there are equally strong distributional cues for all of the elements within the subcategory.

However, closer inspection of individual participants’ ratings of X2 test items reveals individual differences on this outcome. Some participants are generalizing to grammatical novel test strings and rejecting subcategory violation and ungrammatical strings just as they did for all of the items in Experiment 5 (see Figure 5.6). In contrast, other participants used lower ratings for both grammatical novel and subcategory violation test strings (see Figure 5.7).
Figure 5.6 Mean raw scores of test items containing $X_2$ by participants who rated grammatical novel items no differently than familiar strings. Dotted lines are mean ratings from individual participants, and the bold solid line is the mean of these participants (error bars are standard errors).

Figure 5.7 Mean raw scores of test items containing $X_2$ by participants who rated grammatical novel items significantly lower than familiar strings. Dotted lines are mean ratings from individual participants, and the bold solid line is the mean of these participants (error bars are standard errors).
These differences would likely become insignificant with more participants and it is likely that, overall, $X_2$ grammatical novel items would be rated no differently from familiar items. It is also true that the item analyses were conducted on a relatively small number of test strings (6 for each $X$ within each test type). However, it is good to note that the overall results for Subcategory 2 strings ($X_3$ and $X_4$) were qualitatively no different from those in Experiment 5. If these results had been different from Experiment 5, it might demonstrate that the frequency manipulation negatively impacted learners’ overall attainment of the language since there was no difference in exposure to Subcategory 2 strings between Experiments 5 and 9. Instead, learners show only minimal differences because of the frequency manipulation (as in Reeder et al. (in prep)), and those differences are restricted to the subcategory in which the manipulation was carried out.

5.3 General Discussion

In both Reeder et al. (in prep) and Experiment 9, we see learners overcoming variations in input frequencies in order to maintain a rational generalization strategy based on the learner’s exposure to the language. In the sparse and incomplete overlap exposure set used in Reeder et al. (in prep) learners replicated the broad results from Experiment 4 by restricting generalization to novel strings. Upon closer inspection, learners showed some indication of frequency matching for ratings of familiar strings, but resorted to the same mechanisms employed for the other experiments in this dissertation when confronted with novel instances of the language. In Experiment 9,
learners replicate the results from Experiment 5 by encoding two subcategories within the X-category and generalizing within subcategories but not across subcategory boundaries. When the individual test items are inspected, we see sensitivity to the frequency manipulation for the subcategory in which the frequency was varied, but the results remain inconclusive given the bimodal distribution of responses regarding the X-word that was frequently and densely sampled.

Overall, we can conclude that the use of density, number, and overlap to form categories is not greatly affected by making our paradigm more “language-like” through the introduction of frequency variations. However, the frequency differences that were used in these experiments are clearly not on the order of the differences in frequency among natural language categories. In future research it would be of interest to utilize larger frequency differences. That being said, showing that learners do not discard other kinds of distributional information when frequency is varied is a good first step towards arguing that our paradigm scales up to studying the distributional information that exists in natural language.
Chapter 6: General Discussion and Conclusions

6.1 Distributional Learning: A Reasonable (and Useable) Hypothesis

A major criticism of a learning mechanism that is mainly reliant on
distributional information is that such a learning scheme would result in
overgeneralization (Gleitman & Wanner, 1984). Consider the strings:

(1) I gave a pie to John.
(2) I gave John a pie.
(3) I donated a pie to John.
(4) *I donated John a pie.
(5) John told the secret to me.
(6) John told me the secret.
(7) John whispered the secret to me.
(8) *John whispered me the secret.
(9) Mary likes fish.
(10) Mary likes dogs.
(11) Mary might fish.
(12) *Mary might dogs.

Many arguments against distributional analyses stem from the possibility of deriving
strings (4), (8), and (12) from the preceding input sentences. However, the results
from the studies described in this dissertation provide strong evidence that, under
some circumstances, distributional cues alone are sufficient to form linguistic
categories and to generalize across an appropriate range of contexts. Moreover, 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. The general problem learners face when given sparse, ambiguous and incomplete input about the underlying structure of natural language is how to distinguish accidental omissions in the input (such as, “Mary will eat the fish she caught when fishing”) from systematic gaps that arise from rules or lexical idiosyncrasies (such as, “*Mary might dogs”). Across the experiments in this dissertation, we observed remarkable sensitivity to distributional information. Not only are learners sensitive to distributional information to categories; they use this information in very rational and sophisticated ways to categorize the input.

At the start of this dissertation, we proposed a systematic set of computational variables that could potentially explain the types of distributional information that are important for categorization. The results of the experiments demonstrate that deciding whether to generalize across words or preserve lexical specificity depends on (at least) 3 distributional variables: the number of linguistic contexts in which each word in the input set occurs, the density or proportion of these contexts that are present in the input, and the degree of overlap of contexts across words.

**Number and Density:** The first finding regarding these variables was that we observed no great differences in categorization when learners were exposed to more sparse input, as long as only the number of contexts, but not the overlap in contexts across words, was altered (Experiments 1 and 2). The same finding appeared in
studies of subcategorization: when the exposure contained complete overlap among contexts for words within subcategories but no overlap across subcategory boundaries, we saw the same result: learners generalized to withheld strings within the subcategory, but rejected novel strings that violated the subcategory boundary (Experiment 5). Learners continued to maintain a strict subcategory boundary in the face of exceptions or imperfections in their input, even when the input had consistent evidence of subcategory boundary violations (Experiments 6 and 8).

Overlap and Frequency: Learners began to reduce the likelihood of generalizing (that is, increased the difference in their ratings for familiar versus novel grammatical sentences) when the overlap in contexts for different words within the category/subcategory was reduced (Experiments 3 and 7). Learners also restricted their tendency to generalize quite sharply when the same exposure corpus (and its gaps) was repeated (Experiment 4). In contrast, learners were not affected by frequency differences among individual X-words, instead showing the same degree of generalization in this case as when frequency was matched for each of the X-words (Reeder, Newport, Aslin & Schuler, (in prep) and Experiment 9). Even though frequency is a very strong distributional cue, learners largely ignored unequal frequencies of words and contexts within a category/subcategory, in favor of the stronger context-based distributional cues, such as overlap, that more sharply signal category and subcategory boundaries. These results demonstrate that adult learners can skillfully use the data in the input to determine when to ignore gaps and when to generalize over them.
These results also highlight specific 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 exposure sentences as complete sequences, or the trigrams or quadrigrams (e.g., $AXB$, $AXBR$) and their frequencies of occurrence during exposure, they would have been able to discriminate between the familiar and novel grammatical sentences in all of the experiments. Since learners cannot always discriminate perfectly between familiar and novel, they must not have been encoding information in these forms. In contrast, if they were only keeping track of simple word frequencies, they would have failed in all experiments, since these are carefully controlled. While the experiments presented in this dissertation are compatible with the possibility of bigram encoding (e.g., $AX$, $XB$), this is incompatible with several of our other studies (Reeder, Newport and Aslin, 2010). Reeder et al. (2010) asked when, under the same varying circumstances of category strength and category learning as in Experiments 1-4 (Chapter 2), learners would extend their knowledge of the target category to a novel word ($X_4$), one for which learners had only minimal context information. $X_4$ was presented in the exposure set in only one context: $A_1X_4B_1$. The results demonstrated that learners would skillfully transfer their knowledge of category structure to a novel item that was only weakly represented in the input. Storing bigram frequencies could not account for generalization to novel $X_4$ strings. Only 2 bigrams for $X_4$ were part of exposure, but generalization to new $X_4$ contexts was very strong when, based on contexts of $X_1-X_3$, the overall category $X$ was strongly formed. Learners must therefore be keeping track
of word co-occurrences via a network of occurring contexts for individual words (as in Figures 2.1 & 2.3) and collapsing the individual words into a category when these networks bear enough quantitative (as well as qualitative) similarities to one another. This process can be idealized in terms of a Bayesian model estimating whether sample data are drawn from one hypothesis space or another. But there are potentially a number of models, in addition to a Bayesian model, that could simulate such results.

6.2 Future Directions

One question that is raised by our results is whether infants and young children can coordinate multiple variables as adults do. We are in the process of testing child learners to determine how they weigh the large number of variables involved in forming categories in these tasks. One possibility is that young children are as skillful as adults at weighing variables to decide how to generalize. Another possibility is that they are more likely to follow only one or a few of these variables (as found in related studies of child learners), or that they are more likely overall to generalize than adults are, regardless of the input.

We also know very little about how category and subcategory structures emerge over the course of learning. An interesting extension of this work would be to use an online measure of learning. Categorization is, after all, a two tier process: not only do learners need to figure out that there are categories and how many there are,

---

6 Pilot results running a modified version of Experiment 1 with 7-10 year olds point to this conclusion. However, there is research that demonstrates that children younger than 7 might only track a subset of the necessary variables (Austin, Newport & Wonnacott, 2006; Austin & Newport, 2009).
but they also need to figure out which words belong in which category. When faced with conflicting data, will learners maintain their internally represented category structure and shift their more flexible subcategory structure? Since subcategory structure might only emerge after learners determine basic category structure, when will a learning mechanism decide to switch representations, and what kind of information is required in order to do so? What constraints guide this learning mechanism over the course of learning?

The current paradigm allows us to probe the state of the learner after a set period of exposure, but we do not yet know how the state of the learner changes over the course of exposure to different types of distributional information. The differences between Experiments 3 and 4 give us a clue as to the effect of different amounts of exposure, but we do not yet know the exact nature of the threshold that determines generalization or conservatism. Another way to state this problem is that the results of our experiments strengthen the idea that gaps in the input can be just as informative as the presence of that information, but we do not know how much gaps get strengthened when we vary density, overlap, or number. In a cue-weighting paradigm where learners needed to determine how much a particular cue led to a certain reaction, Wasserman and Castro (2005) studied the information learners gleaned from the nonoccurrence of an event. They found that even when a cue was not present during a particular trial, learners continued to change the strength of the
absent cue. By recording some type of measure after every new piece of information, we could better understand how learners compile the vast amount of distributional information in their input.

A first step toward accomplishing this goal was made by Hunt & Aslin (2001, 2010). Hunt and Aslin created a non-linguistic category learning task that utilized serial reaction time as the dependent measure. Participants pressed sequences of buttons that corresponded to strings of visual symbols. The strings contained sequential statistics that organized the symbols into categories. Learners were then tested on novel sequences that either obeyed or violated the category structure of the grammar. Reaction time measures showed that subjects were consistently slower to respond to ungrammatical novel strings, ones that disobeyed the categorical structure of the grammar, but not to grammatical novel strings that observed the category structure. The methodology allowed the investigators to record reaction times during every single training and test trial, thus permitting observation of the time course of learning.

While the results of an online learning task would speak to the types of statistical learning mechanisms learners utilize during categorization of an artificial language, computational models would also provide some insight into the computational strategies that learners might use in a wide range of structured domains. The results of our experiments demonstrate that learners cannot be tracking

---

7 This finding mirrors much older research in the animal literature regarding associative learning. For example, Holland (1981) paired food with a tone. Later, the tone was paired with a lithium chloride injection, which caused illness. Even though the food was absent during the illness-causing injection, when presented with the food again, the animal had revalued its representation of the food, which became aversive.
certain types of statistics such as simple word frequencies or trigrams. Instead, learners must be forming a more abstract representation of the data in order to generalize their knowledge to novel strings. This process can be idealized within a Bayesian model estimating whether sample data are drawn from one hypothesis space or another. We are currently exploring a variety of Bayesian accounts of this behavior. One promising framework is the Infinite Relational Model, or IRM (Kemp, Tenenbaum, Niyogi & Griffiths, 2010). Such a model would allow us to make predictions about human behavior given varied distributional information, as we have manipulated across our experiments. Within such a framework, comparisons of different particular models (differing in the specific statistics from the input that are utilized and different ways in which this information is combined) can elucidate how human learners might be weighing different statistical variables in the input in order to decide whether a particular novel string is a grammatical extension of the grammar or an illegal one. While the results we have obtained in our experiments may point to a rational account of human behavior, there are potentially a number of models, in addition to a Bayesian model, that could simulate such results. Other accounts of our data also warrant exploration, particularly utilizing some of the clustering methods utilized by the concepts and categories literature (e.g, Anderson, 1991; Kruschke, 1992; Nosofsky, 1998), in order to address whether common mechanisms might be used across linguistic and non-linguistic categorization.

Another direction that warrants further investigation is how distributional variables combine with other types of information in natural language acquisition. As
learners face the problem of inferring category membership, there are a number of
cues, correlated with but in addition to statistical cues about linguistic context
distribution, that they could use to help them extract category information, such as
phonological, prosodic, or semantic cues to category membership. While
phonological, prosodic, and semantic cues are known to be imperfect cues to category
membership, they do have some correlation with distributional cues and
categorization, and therefore could be used as correlated cues to solve the problem.
In our own experiments we have eliminated these additional cues, in order to estimate
how much learners might be able to rely on distributional cues alone. But in natural
language acquisition we do expect, along with other researchers (cf. Monaghan,
Chater & Christiansen, 2005), that the integration of multiple imperfect and uncertain
cues – including the distributional ones we have studied here – is likely to be the way
real learners determine when to generalize an when to restrict generalization in a
complex problem space.

A final issue that needs to be addressed in any artificial language learning
study concerns the difference in scale between experimental languages and natural
language. Experimentally, we isolate individual computational variables (or use
carefully controlled cue combinations) so that we may better understand how a
human learner operates over those particular cues and what the outcome of learning
is. If we are able to demonstrate that our miniature languages produce learning
outcomes similar to those we see in natural language learning, we believe that such
findings create one part of an argument that the same mechanisms are used in natural
language acquisition. In addition to this, however, artificial languages need to be created that mirror the learnable statistics and the realistic statistical frequencies and patterns of cues that we see in natural language. Perhaps by using computational models over natural language corpora we can test the existence and usefulness of the cues we study on real language data, while also having control over the necessary constraints that guide learning mechanisms. Models of this sort can be constrained by the experimental evidence we get from artificial language learning studies that isolate the distributional cues of interest. At the same time, these models can confirm the validity of the experimentally tested cues by successfully using them over corpora of natural language input.

6.3 Conclusions

One major aspect of successful language acquisition is the ability to organize words into form class categories and generalize from properties of experienced items to novel items. The experimental results presented here suggest that the number of categories and their functional roles in a grammar are determined, at least in part, by a form of constrained statistical learning. Learners are able to take account of a rich set of variables to aid them in grammatical categorization, including degree of overlap among category members, amount of input, consistency of gaps and overlaps, and conflicts or consistency among cues. This work shows that the patterning of tokens in a substantial corpus of linguistic input appears to be sufficient, along with a small set of learning biases, to extract the underlying structural categories in a language.
References


