Learning Hierarchies:
Acquiring Hierarchical Structure from Miniature Artificial Languages

by
Benjamin M. Faber

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

Supervised by
Professor Elissa L. Newport
and
Jeffrey Runner

Department of Brain and Cognitive Sciences
Department of Linguistics
The College
Arts and Sciences

University of Rochester
Rochester, New York

2008
Curriculum Vitae

The author was born in Buffalo, New York on December 24, 1979. He attended the University of California at Santa Cruz from 1996 to 2000, and graduated with Honors with a Bachelor of Arts in Linguistics and Psychobiology. He came to the University of Rochester in the fall of 2000 and began graduate studies in Brain and Cognitive Sciences and Linguistics under the guidance of Elissa L. Newport. He received a University of Rochester Sproull Fellowship in 2000 and earned an M.A. in Brain and Cognitive Sciences in 2005.
Acknowledgments

This research was supported by National Institute of Health research grant DC-00167 awarded to Elissa Newport.

This dissertation would not have been possible without the support of my advisor, Elissa Newport. She has provided me with exemplary guidance and support in all my endeavors at the University of Rochester. Were it not for the freedom she granted me to pursue my research interests, combined with her thoughtful evaluation and advice, this dissertation would never have coalesced into its present form. I also owe her a great debt of gratitude for giving me numerous opportunities to teach during my time as a graduate student; it has made the past few years incredibly educational and rewarding.

Second, I must thank Jeff Runner for being a wonderful advisor to me in Linguistics. He has made it possible for me to think and work across disciplines throughout my graduate years. I have many fond memories of the hours we spent discussing linguistic theory in his office. His flexibility and willingness to let our discussion roam the linguistic landscape directly shaped the thinking that led to much of the research here. Thank you as well to the other members of my committee, Dick Aslin, Greg Carlson, and Loisa Bennetto, who helped me organize and analyze 5 years of experiments into a coherent thesis.

Special thanks are due to my colleagues and friends in the department who have provided insightful conversation and criticism over the years. It has been a pleasure working with Patty Reeder, Celeste Kidd, David Feder, Alison Austin, and
Natalie Klein both in research settings and in the classroom. I also owe Kathy Corser my deepest gratitude for making every aspect of my life in the department possible. It has been wonderful knowing that she is always there to help, answer any question, and provide a much needed ribbing now and then as well.

My entire family, the Fabers and Kunzes, have been an unending source of support during this entire process. While I came to the University of Rochester for its academic environment, it has also been a blessing to live so close to so much of my family. Finally, and most importantly, I want to thank my wife, Kiera, who is my life, my inspiration, my joy, and my happiness.
Abstract

One of the central features of natural language syntax is that it is organized hierarchically but expressed linearly. While it has been suggested that much of this hierarchical structure must be innate, it is plausible that what is innate is not specific hierarchical structures, but biases to learn hierarchical patterns. If this is true, then there must be information in the organization of natural languages from which hierarchical structures can be learned. Three possible types of information that are known to occur in natural language patterns, and known to be salient for language learning, are statistical regularities between units of language, prosodic groupings of words into acoustically marked phrases, and the semantically meaningful relationships between words. Past research has shown that some types of statistical regularities between syllables or phonemes can result in segmentation of a continuous speech stream into word like units. Furthermore, statistical regularities between words can also result in the segmentation of word strings into simple phrase-like units. In the present research, we extend these findings to ask what types of phrases can be formed as a result of statistical regularity cues between words, and how additional cues might combine with these regularities to produce more complex hierarchical structure learning.

One hundred sixty college students were exposed to sentences from one of sixteen miniature languages which varied the statistical regularities between categories of words and the prosodic and semantic relationships between words. In the first experiment, meaning and prosody were held constant, but statistical regularity was
systematically varied between groups of subjects. In the second experiment, statistical
regularities, prosody, and semantic relationships were systematically varied to explore
their interaction in the learning of a small set of hierarchical structures. In the final
section, learners were presented with more complex languages with multi-tiered
hierarchical structures to be acquired from a both adjacent and non-adjacent statistical
regularities and two types of semantic representations. All subjects were tested on a
variety of measures of hierarchical phrase groupings as well as measures of their
learning of the linear order of words in sentences.

Results showed that participants are incredibly flexible in their ability to learn a
wide variety of hierarchical patterns from information in the language environment.
This flexibility is tempered by the presence of some consistent biases in how they learn
from the input, but a combination of several linguistic cues can overcome these biases.
Finally, when faced with more complex miniature languages, the contributions of both
statistical regularities and semantic relationships appear to work in an additive fashion
for the learning of hierarchical phrase structures.

These results, taken together, suggest that statistical regularities play a
significant role in the acquisition of hierarchical phrase structure, but more
importantly work in concert with prosodic information and semantic relationships to
produce the complex learning that is required for natural language acquisition.
# Table of Contents

Curriculum Vitae ii
Acknowledgments iii
Abstract v
Table of Contents vii
List of Tables ix
List of Figures xi

**Chapter 1. Introduction** 1

1.1 Hierarchical Representation in Syntax 5
1.2 Hierarchical Representations for Non-Language Sequences 7
1.3 Learning Hierarchical Structures from Miniature Languages 8
1.4 The Current Work 11
  1.4.1 Statistical Regularities 11
  1.4.2 Prosodic Grouping 14
  1.4.3 Semantic Reference Worlds 15
  1.4.4 Multi-Tiered Hierarchical Structure 17
  1.4.5 Overview of Experiments 19

**Chapter 2. Experiment 1** 22

2.1 Method 23
  2.1.1 Participants 25
  2.1.2 Description of the linguistic systems 25
  2.1.3 Procedure 28

2.2 Results 32
  2.2.1 Sentential Word Order test 32
  2.2.2 Two-Word Phrase Grouping Preference test 34
  2.2.3 Three-Word Phrase Grouping Preference test 42

2.3 Discussion 45
# Chapter 3. Experiment 2

3.1 Method

3.1.1 Participants

3.1.2 Description of the linguistic systems

3.2 Results

3.2.1 Sentential Word Order test

3.2.2 Two-Word Phrase Grouping Preference test

3.2.3 Three-Word Phrase Grouping Preference test

3.3 Discussion

# Chapter 4. Experiment 3

4.1 Methods

4.1.1 Participants

4.1.2 Description of the linguistic systems

4.1.3 Procedure

4.2 Results

4.2.1 Sentential Word Order test

4.2.2 Two-Word Phrase Grouping Preference test

4.2.3 Three-word Phrase Grouping Preference test

4.3 Discussion

# Chapter 5. General Discussion and Conclusions

5.1 Plato's Problem

5.2 Implications for Natural Language

5.3 Future Directions

5.4 Conclusions

References
**List of Tables**

Table 1.1: Calculation of $\phi$ for a sample language. 13

Table 1.2: Calculation of $\phi$ for a second sample language. 14

Table 2.1: Lexicon and pronunciation of words for Experiment 1. 24

Table 2.2: Sentences of Language a1, b1, c1, and d1. 25

Table 2.3: $\phi$ statistics for all neighboring category bigrams. 26

Table 2.4: "Phrase Grouping Preference test types. Each row represents one possible 2 Alternative Forced Choice test. Participants received three trials of each test type per testing block." 29

Table 2.5: "Phrase Grouping Preference test types. Each row represents one possible 2 Alternative Forced Choice test. Participants received three trials of each test type per testing block." 29

Table 2.6: Sentential Word Order test types. Each row represents one possible 2-AFC test. Participants received one trial of each test type per block. Violations of correct word order are bolded. 31

Table 2.7: Sentential Word Order test scores for Experiment 1 languages. 33

Table 2.8: Trial types for Two-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. phrases not supported by statistical regularities. Each test is listed as “Supported - Unsupported”. 34

Table 2.9: Trial types for Two-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. pairs that cross the phrase boundary. 36

Table 2.10: Trial types for Three-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. phrases not supported by statistical regularities. Each test is listed as “Supported - Unsupported”. 43

Table 3.1: Sentences of Languages a2, b2, c2, and d2 for each cue type. 53

Table 3.2: $\phi$ statistics for all neighboring category bigrams. 53
Table 3.3: Semantic Reference world groups in the ‘Semantics and Prosody’ languages. Categories separated by spaces are visually distant, those in parentheses are close.

Table 3.4: Semantic Reference world groups in the ‘Cue Combination’ languages. Categories separated by spaces are visually distant, those in parentheses are close.

Table 3.5: Trial types for Two-word Phrase Grouping Preference tests: Phrases supported by statistical and/or semantic/prosodic cues vs. phrases not supported by cues. Each test is listed as “Supported - Unsupported”.

Table 3.6: Trial types for Narrow Two-word Phrase Grouping Preference tests: Phrases supported by cues vs. phrases not supported by cues which cross the three-word phrase boundary. Each test is listed as “Supported - Unsupported”.

Table 3.7: Trial types for Edge and Non-Edge Two-word Phrase Grouping Preference tests: Phrases supported by cues vs. phrases not supported by cues which cross the three-word phrase boundary. Each test is listed as “Supported - Unsupported”.

Table 4.1: Lexicon of words for Experiment 3.

Table 4.2: \(\phi\) statistics and adjacency percentage for all possible category bigrams. Bigrams involving category A can not have a \(\phi\) calculated because it would involve dividing by 0.

Table 4.3: Semantic reference word-to-world mappings for Dependency Semantics.

Table 4.4: Sentential Word Order test types. Violations of correct word order are bolded.

Table 4.5: Two-word Phrase Grouping Preference test types.

Table 4.6: Three-word Phrase Grouping Preference test types.
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Different instantiation of Semantics in Miniature Languages, Reproduced from Morgan and Newport (1981), Figure 4.</td>
<td>17</td>
</tr>
<tr>
<td>1.2</td>
<td>Hierarchical phrase structure without, and with, crossed dependencies.</td>
<td>19</td>
</tr>
<tr>
<td>2.1</td>
<td>Phrase structures of the languages used in each experiment.</td>
<td>22</td>
</tr>
<tr>
<td>2.2</td>
<td>Sample sentences with visual referents (samples drawn from language a1).</td>
<td>27</td>
</tr>
<tr>
<td>2.3</td>
<td>Sentential Word Order test scores for Experiment 1 languages. Dashed line represents chance. Error bars represent 95% CI.</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>Two-word Phrase Grouping Preference test results: Phrases supported by statistical regularities vs. phrases not supported.</td>
<td>35</td>
</tr>
<tr>
<td>2.5</td>
<td>Two-word Phrase Grouping Preference test results by language: Phrases supported by statistical regularities vs. pairs that cross the phrase boundary.</td>
<td>37</td>
</tr>
<tr>
<td>2.6</td>
<td>Two-word Phrase Grouping Preference test results: Bars depict preference for statistically regular pair for Languages (ABC)-D-E and A-B-(CDE) and depict preference for BC over CD for Languages A-(BCD)-E and (AB)-C-(DE).</td>
<td>39</td>
</tr>
<tr>
<td>2.7</td>
<td>Two-word Phrase Grouping Preference test results: Antisymmetric X-bar Phrase structure preference within Statistically cued Ternary Group.</td>
<td>41</td>
</tr>
<tr>
<td>2.8</td>
<td>Three-word Phrase Grouping Preference test results: Phrases supported by statistical regularities vs. phrases not supported by statistical regularities.</td>
<td>44</td>
</tr>
<tr>
<td>3.1</td>
<td>Phrase structures supported by cues in languages of Experiment 2.</td>
<td>50</td>
</tr>
<tr>
<td>3.2</td>
<td>Sample sentences with semantic referents (samples from SemPro-a2).</td>
<td>54</td>
</tr>
<tr>
<td>3.3</td>
<td>Sentential Word Order test results for Experiment 2. Error bars represent 95% CI.</td>
<td>57</td>
</tr>
</tbody>
</table>
Figure 3.4: Two-word Phrase Grouping Preference test results. Phrases supported by cues vs. phrases not supported by cues.

Figure 3.5: Narrow Two-word Phrase Grouping Preference test results. Phrases supported by cues vs. unsupported phrases that cross the three-word phrase boundary.

Figure 3.6: Edge Two-word Phrase Grouping Preference test results. Phrases supported by cues vs. unsupported phrases that cross the phrase boundary.

Figure 3.7: Non-Edge Two-word Phrase Grouping Preference test results. Phrases supported by cues vs. unsupported phrases that cross the phrase boundary.

Figure 3.8: Three-word Phrase Grouping Preference test results by language: Phrases supported by cues vs. phrases not supported by cues.

Figure 3.9: Three-word Phrase Grouping Preference test results by language: Preference for ABC phrase.

Figure 3.10: Three-word Phrase Grouping Preference test results by language: Preference for CDE phrase.

Figure 4.1: Languages for Experiment 3.

Figure 4.2: $\phi$ statistics and adjacency percentage for all possible category bigrams.

Figure 4.3: Semantic Reference World examples for Constituent Grouping Semantics. Sentences represented from left to right are ABCDE, ADEFG, and AFGBA.

Figure 4.4: Four frames of a visual reference animation for Dependency Semantics. Sentence represented is beta zami gesor vigo nebu, or “Bounce two squares left-of tree”. Semi-transparent images represent previous frames and are only shown for reference.

Figure 4.5: Sentential Word Order test results for Experiment 3 languages.

Figure 4.6: Two-Word Phrase Grouping Preference test: “A test”.

Figure 4.7: Two-Word Phrase Grouping Preference test: “Phi test”.

Figure 4.8: Two-word Hierarchy without Statistics preference test.
Figure 4.9: Three-Word Phrase Grouping Preference tests: (a) Phi final, (b) Phi initial, (c) Non-adjacent Linear.

Figure 4.10: Three-Word Phrase Grouping Preference tests: (a) Phi final > Phi initial, (b) Phi final > Non-Adjacent Linear, (c) Phi initial > Non-adjacent Linear, (d) Tests for Adjacent/Dependency language.

Figure 4.11: Alternative multi-tiered hierarchical phrase structures for Non-Adjacent Statistical Regularities with Dependency Semantics language.
Chapter 1. Introduction

One of the great and long-standing problems of psychological inquiry is how we come to acquire vast amounts of knowledge from a chaotic world with little effort. This has been called “Plato’s problem” by Chomsky (1986) and others, in reference to Plato’s *Meno* dialogue in which Socrates discusses how we come to have knowledge of complex ideas such as virtue. Plato’s Socratic answer to his own problem is that knowledge is not learned, but simply recalled memories from past lives, an early attempt to explain innateness. Darwin (1859) offered a biological account of inborn knowledge, or instincts, arguing that they, like corporeal organs, evolve via natural selection and are shared across related species. Nearly all theories of knowledge and learning since Plato (perhaps excluding strongly behaviorist ideologies) rely upon some inborn component, but there is strong disagreement on the nature of the innate knowledge and how much can be drawn from the environment.

Plato’s problem is especially central to domain of Language. As Chomsky says:

> “Much of the interest of the study of language, in my opinion, lies in the fact that it offers an approach to Plato’s problem in a domain that is relatively well circumscribed and open to inspection and inquiry, and at the same time deeply integrated in human life and thought. If we can discover something about the principles that enter into the construction of this particular cognitive system, the principles of the language faculty, we can progress toward a solution for at least one special and quite important case of Plato’s problem.” (Chomsky, 1986, xxvi)

Chomsky’s answer to Plato’s problem for language is that the requisite knowledge cannot be learned from the environment, but instead relies upon an innate Universal Grammar.
A second key question for psychology is why humans alone seem to possess, and rely on, a number of complex cognitive abilities that are not closely shared by other species. These faculties include numbers and mathematics, tool making and use, and moral thought. Recent comparative research has suggested analogues for each of these domains among varied non-human species (Gelman and Cordes, 2001; Flombaum, et al., 2005; Hunt and Gray, 2003; Flack and de Waal, 2000; Herman and Uyekama, 1999), and even components of language, once thought to be unique to humans, seem to be present in animal communication systems. These include the use of arbitrary symbols in chickens (Evans, Evans, and Marler, 1995), intentionality in Rhesus Macaques (Hauser, 1992), voluntariness in Antshrikes and Tanagers (Munn, 1986), referentiality in Honeybees (von Frisch, 1967), and discreet call types in Vervets (Seyfarth, Cheney, and Marler, 1980). While these basic similarities suggest a difference of degree, rather than of kind, a number of features of human language do not seem to have close relatives amongst the communications of other species, including complex Phonology, Morphology, Syntax, and Semantics. There has been much recent debate over what specifically it is that sets human language and its cognitive underpinnings apart (Hauser, Chomsky, and Fitch, 2002; Pinker and Jackendoff, 2005), but there continues to be widespread agreement that human language has a crucial innate component that sets it apart from all other animal communication systems.

The problem for cognition is therefore two fold: to attempt to answer Plato’s problem of how a complex domain can be acquired from a chaotic environment, and simultaneously to discover the nature of the innate and species specific cognitive
systems that permit only humans to solve this problem. Two potential research paths are obvious to approach this dual problem: studying the genetic basis of innate knowledge, or studying learning in a controlled environment. The first option, to date, has not proven particularly successful, as current models of genetics and genomics do not provide a direct path from genes to behavior (McGuffin, et. al., 2001), but the second has repeatedly proven quite tenable. As early as Braine (1963), research conducted using miniature artificial languages has demonstrated the learning capacities of subjects exposed to carefully controlled language environments. These studies have been used to reveal the computations underlying the acquisition of phonology (Maye, Werker, and Gerken, 2002), word segmentation (Saffran, Aslin, and Newport, 1996; Saffran, Newport, and Aslin, 1996), syntactic categories (Mintz, Newport & Bever, 2002), syntactic structures (Gomez & Gerken, 1999; Moeser and Bregman, 1972; Morgan, Meier, and Newport, 1987; Saffran, 2001, 2002; Thompson and Newport, 2007), word order universals (Christiansen, 2000), and algebraic rule learning (Marcus, et. al., 1999).

One of the domains of language that has proven quite accessible to miniature language research is syntax, including one of its key properties that seems to go unshared by non-human communication: hierarchical structure\(^1\). The acquisition of hierarchical structure is of particular interest in the debate over Plato’s problem, because it involves learning a mental representation that does not have a direct superficial representation in the external world. This is an instance of a mapping problem: The input to a language learner is simply a linear, or one dimensional, 

---

\(^1\) While many analyses of birdsong and whale-song suggest a hierarchical structure, there is no evidence that these structures convey varying or significant semantic relationships.
sequence of words unfolding in time, but the hierarchical structure which correlates only partially with the linear string is multidimensional. Just as learning about virtue must either be inborn or derive from experiencing sporadic and varying instances where virtue is indirectly present, learning hierarchical structures in language must too be inborn in some way, or learned via extraction from chaotic input. The fact that non-human animals do not seem to acquire or use hierarchical structures for communication indicates that there must be a species specific innate component to hierarchical representations in language; however, the exact nature of this inborn component is entirely unsettled and seems inaccessible to a purely genetic-genomic approach.

Hierarchical representations also pose an interesting dilemma because of their distribution in human and non-human cognition. Far from being an exclusive attribute of language, hierarchies are actually quite widespread in mental representations. Perhaps the earliest significant study of hierarchical organization comes from Lashley’s (1951) work on motor planning. Further work on short-term memory chunking (Miller, 1956; Ericcson, Chase, and Faloon, 1980) and concepts and categories (Rosch, 1978) suggests that hierarchical organization may be a universal pattern rather than a domain specific property of language. The universality of hierarchical organization seems to extend beyond our own species as well, showing up in the imitation learning of motor sequences in great apes including gorillas (Byrne, 2005), tracking of dominance relationships in baboons (Seyfarth, Cheney, and Bergman, 2005), and stacking of seriated cups in Chimpanzees, Bonobos, and Capuchins (Johnson-Pynn, et. al., 1999). If hierarchical representations are shared among primates, but only arise
in humans for communicative purposes, then it raises the question of how it is that people extract hierarchical representations from a complex linguistic environment. The present experiments seek to explore sensitivities to potential cues in the linguistic environment to hierarchical structure during language acquisition and how they might combine to help solve Plato’s problem in this domain.

1.1 Hierarchical Representation in Syntax

Hierarchical structures have been a tradition in syntactic theory for nearly its entire existence. Some of the earliest, and perhaps most influential, work dates back to Paṅini in the Fourth Century BC (approximately) whose analysis of Sanskrit laid much of the groundwork for modern descriptive and generative grammars. In current theories of syntax, hierarchical representations are nearly universal, encompassing almost all major approaches\(^2\). The role that hierarchical structures play in syntactic theories is one of mapping and interfacing. Although specific approaches vary in the exact role of structures, all share a focus on how hierarchical relationships are expressed in the linear order of words and morphemes. Thus, linear ordering, and its associated distributional phenomena (formulated in modern approaches as statistical relations), have long been associated with models of linguistic form and hierarchical structures (Bloomfield, 1933; Harris, 1951), although an explicit focus on this aspect fell out of favor with Chomsky’s 1957 review of Skinner’s *Verbal Behavior*. The second major role of hierarchical structure is to capture semantic relationships between morphemes, words, and phrases. Theories vary greatly in the way they deal with this

\(^2\) Some schema models and connectionist approaches do not rely on explicit structures or hierarchies, although even in these theories, there may still be a place for hierarchical relationships if not hierarchical representations.
role including Kernel Sentences (Chomsky, 1957), Deep Structures (Katz and Postal, 1964; Chomsky, 1965), Logical Forms (Chomsky, 1981), Semantic Structures (Jackendoff, 1976, 1990, 2002), S-Structures (Bresnan, 2001), SYNSEM (Pollard and Sag, 1994), and numerous others.

While nearly all theories share the basic notion and functions of hierarchical structures, theories differ significantly in the exact nature of the hierarchical forms which are appropriate for a given sentence. Within mainstream generative grammar, two basic types of theories have been used to limit the types of hierarchies involved. Beginning with Chomsky (1970) and Jackendoff (1977), X-bar syntax became a major component of grammar. X-bar theory is a representational and structural hypothesis that all phrases share a common arrangement of projections ($X^0$, X-bar, and X-Phrase) and relationships (Head-Complement and Head-Specifier). Further work has attempted to further motivate and limit this analysis, arguing that all languages share not only the basic X-bar phrase structure, but the same ordering of Specifier-Head-Complement as is standardly found in analyses of English (Kayne, 1994). A more recent and quite opposite trend in hierarchical analyses is that of Bare Phrase Structure (BPS) as proposed by Chomsky (1995), which builds hierarchical structure from the bottom up using the processes of Merge and Move: a derivational rather than representational approach to phrase structure. Taking yet a different approach, Culicover and Jackendoff (2005) have argued for a “Simpler Syntax Hypothesis” which rejects both of these theories in favor of a less hierarchically complex set of structures. What is lacking in this debate, however, is psycholinguistic evidence
regarding the types of hierarchical structures that learners actually posit during acquisition and the processes involved in building these forms.

1.2 Hierarchical Representations for Non-Language Sequences

Although current research on learning does not address the specific issues brought up by the linguistic debates described above, there is a long history of work on learning hierarchical structures in varied domains. Beginning with Lashley's (1951) classic paper on “The Problem of Serial Order in Behavior”, it became evident that motor activity did not just involve a series of local associations, but was instead based on hierarchically organized plans. Lashley’s evidence for this included the fact that many sequences of action occur without sensory feedback or before sensory feedback from a previous step is available and that people make errors in typing and speech that involve upcoming actions, as in the spoonerism "It is kisstomary to cuss the bride". He also noted that interpretation of words or stimuli was highly context dependent, as in the auditory presentation of “The mill-wright on my right thinks it right that some conventional rite should symbolize the right of every man to write as he pleases” [p. 116], only possible if word sequences were treated as hierarchical combinations and not simply interpreted piecemeal. Further work also demonstrated that planning larger hierarchical components of action tended to require longer preparation time (Henry and Rogers, 1960).

Following on Lashley’s heels, George Miller (1956) demonstrated hierarchical organization in short term memory, or chunking. This work has been expanded to reveal that with training, subjects can memorize extremely long arbitrary strings by forming hierarchical chunks of chunks (Ericcson, Chase, and Faloon, 1980). Another
line of work on manipulating objects to reflect underlying hierarchical structures and combinations, such as stacking seriated cups, has shown that children develop complex object combining abilities at approximately the same time as they develop combinations of words in language productions (Greenfield, et.al., 1972; Greenfield and Schneider, 1977). These same abilities also correlate with language learning abilities in non-human apes (Johnson-Pynn, et.al. 1999), suggesting a domain general ability between forming hierarchical structures for early language and motor combinations. Taken together, these, and numerous other studies, have consistently revealed a central role for hierarchical mental organizations whenever individuals are faced with complex and highly varied environmental sequences.

1.3 Learning Hierarchical Structures from Miniature Languages

While there is clear evidence for hierarchical representations in language and other domains, the acquisition of these mental representations is still controversial. Some of the earliest research on acquisition of grammatical structure comes from Braine (1963), who revealed that children as young as four are sensitive to linear position and phrasal constituent grouping in miniature languages. Braine’s languages, however, were quite simple and did not truly involve hierarchical structures. Following on the heels of the earliest miniature language studies, Moeser and Bregman (1972) expanded on the complexities of the languages involved, the variety of measurements taken, and the variables controlled. Their study varied both the complexity of the language subjects were exposed to as well as the visual reference world associated with the sentences of their languages. The most central of their findings is that learners only succeeded at a high level when the visual reference world
included information that correlated with the syntactic rules of the language they were attempting to learn. They suggested that, when semantic information was available, learners utilized it in structuring their analyses of the syntax of the language. When this information was not available, they suggested, learners shifted to a more position based approach, as was found in Braine’s work. This explanation is not without problems, however, as explored by Morgan and Newport (1981), who showed that it was not necessary to have semantic referents, but rather to have some information beyond the order of words in the language to help reveal the phrasal groupings of the language. Learners exposed to a visual reference world where phrasally grouped referents appeared spatially grouped was in fact just as potent of a cue to phrase structure as the fully incorporated visual reference world. This work was again extended and clarified in Morgan, et.al., (1987, 1989) which revealed that not only were visual reference worlds effective in inducing phrase structure learning, but so were auditory cues such as prosody and grammatical markers such as function words and morphological endings. Altogether, the results of these studies indicate that as long as there are concordant cues to phrase structure, these structures are generally learnable.

While all of these studies dealt with cues that supplemented the basic distributional information to phrase structure, Saffran (2001,2002) has shown that there are potential cues present within the order of words in the sentences of the learning set that can serve this function. When specific words occur consistently within a phrase, there is a statistically predictive relationship between these words and their syntactic categories. By controlling these predictive relationships, Saffran
revealed that even without additional concordant cues, the phrase structures of languages like those used in prior studies was still learnable, although not as thoroughly as in the previous work. Thompson and Newport (2007) showed, however, that the statistical regularities between phrase internal categories in the Saffran studies were not consistently stronger than the regularities that crossed phrase boundaries. By more carefully controlling the predictive dependencies in a series of experiments, Thompson and Newport were able to show that the cue which allowed learners to succeed in learning phrase structure was the same cue that other miniature language learners had used to segment words from running speech (Saffran, Aslin, and Newport, 1996; Aslin, Saffran, and Newport 1998): tracking local Transitional Probabilities between words and categories.

All of these studies share one feature, however: the phrase structures involved and tested for are almost entirely at a single level or tier. In other words, items (words or categories) group together to form a phrase, but there is little necessity in these languages to then group these phrases together with other items (words, categories, or other phrases) to form hierarchically nested, or multi-tiered, phrase structures. The studies also leave open the question of what sorts of limits or biases there might be in forming phrasal grouping, as has been suggested by X-bar grammars and Bare Phrase Structure theories, since they all share very similar phrase structures. Finally, there is still much work to be done exploring the integration of varied cues to phrase structure and how they might allow for complex phrase structure learning and offer a potential solution to Plato’s problem in this domain.
1.4 The Current Work

The experiments in this dissertation probe how a number of cues - statistical regularities between words and categories, prosodic groupings, and semantic reference worlds – together and separately contribute to the learning of multi-tiered hierarchical phrase structure grammars from miniature languages. It is important at this point to make clear what is meant by the above terms.

1.4.1 Statistical Regularities

Distributional patterns of words and categories was the foundation of many early grammatical analyses ranging from Phonology and Morphology to Syntax and Semantics, with constituents constructed out of highly correlated neighboring elements. More recently, much miniature language research has explored the influence on learners of statistical regularities in the distribution patterns of linguistic units. Saffran, Aslin, and Newport, 1996 suggested that an important statistical regularity for learning to segment words in a speech stream is transitional probability, or more specifically, forward transitional probability ($fTP$):

\[ P(Y | X) = \frac{\text{frequency}(XY)}{\text{frequency}(X)} \]

Simply put, $fTP$ represents the ability to predict an upcoming item given knowledge of the item that immediately precedes it. This is just one of a number of different conditionalized statistics including mutual information, backward transitional probability, conditional entropy, and correlation, all of which show a great deal of overlap in both natural languages and most miniature languages that have been studied. While $fTP$ seems to be a logical choice for representing units and structure in
linguistic stimuli, since sentences unfold linearly in time, it does not quantify information about the frequency of the predicted element occurring without its preceding predictor. There is evidence that this information might be useful, however, from the work of Rescorla (1968, 1988) on non-contiguity learning in conditioning and from a long-standing and well replicated finding in memory research, the recency effect, in which the last item encountered in a list is especially salient and better remembered later (Murdock, 1962). If final items are particularly salient, then it would be quite surprising if their frequency of occurrence was unimportant to a learner.\(^5\) In light of this concern, and some unpublished prior results indicating a high sensitivity of learners to sentence final elements, we will not utilize TP but rather will focus on a related but symmetric statistic for quantifying the predictive relationships among words and word categories that might be used by learners to acquire phrase structure. While this still leaves many options, the one most similar to a combination of forward and backward transitional probability is a derivative of $\chi^2$, Cramer's $\phi$ coefficient. In the first two experiments of the present work, we will quantify predictive relationships among words only in terms of immediately neighboring items – though it is possible that learners might, under certain circumstances, acquire nonadjacent relationships as well (see Newport & Aslin, 2004, for discussion and evidence), and this issue will be addressed in the third experiment. Limiting statistical calculation to only immediately neighboring items, we can calculate the strength of the association between a pair of items as follows:

\(^5\) There is some slight evidence of a trend in the results of Braine (1963) to suggest that learners in his miniature language experiments did perform better on sentence final items than sentence initial items.
\[ \phi(XY) = \frac{\left( f(XY) \times f(\neg X \neg Y) \right) - \left( f(X \neg Y) \times f(\neg X Y) \right)}{\sqrt{f(X) \times f(Y) \times f(\neg X) \times f(\neg Y)}} \]

This formula can be clarified in a concrete example. If we take a miniature language consisting of the seven sentences \{X, Y, Z, XY, YZ, XZ, XYZ\} in equal proportions, we can calculate the frequency at which any pair of elements do or do not occur and co-occur locally as seen in Table 1.1.

**Table 1.1 Calculation of \( \phi \) for a sample language.**

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>( \neg X )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>2a</td>
<td>2b</td>
<td>4c</td>
</tr>
<tr>
<td>( \neg Y )</td>
<td>2c</td>
<td>1d</td>
<td>3f</td>
</tr>
<tr>
<td>Total</td>
<td>4g</td>
<td>3b</td>
<td>7n</td>
</tr>
</tbody>
</table>

And we can calculate \( \phi(XY) \) as:

\[ \phi(XY) = \frac{ad - bc}{\sqrt{efgh}} = \frac{(2 \times 1) - (2 \times 2)}{\sqrt{4 \times 3 \times 4 \times 3}} = -0.333 \]

In other words, in this example there is a small negative relationship between X and Y as a result of the higher frequency at which they occur separately \{in X, Y, YZ, and XZ\} relative to the frequency at which they occur together \{in XY, XYZ\}. The advantage of this statistic over \( iTP \), which is \( f(XY)/f(X) = 0.5 \), is that it is sensitive to changes involving either element of a pair. For instance, if we removed the sentences Y and YZ, making the occurrence of Y independent of X, this would be reflected as a change in \( \phi \) to 0, though this would leave \( iTP \) from X to Y unchanged at 0.5.

Alternatively, if the language was one where X and Y always co-occurred, as in \{XY, Z, XYZ\}, we would then calculate:
It is worth noting here that $\phi$ is not an obtuse statistic, but in fact is exactly equivalent to Pearson’s Correlation coefficient $r$ for pairwise binomial distributions like these.\footnote{The logic behind using $\phi$ rather than $r$, is that $r$ is not specific to pairs of binomial distributions, and therefore involves more complicated calculations that are far less transparent.}

\begin{table}[ht]
\centering
\begin{tabular}{|c|c|c|}
\hline
 & $X$ & $\neg X$ & \textbf{Total} \\
\hline
$Y$ & $2_a$ & $0_b$ & $2_c$ \\
$\neg Y$ & $0_c$ & $1_d$ & $1_f$ \\
\textbf{Total} & $2_g$ & $1_b$ & $3_n$ \\
\hline
\end{tabular}
\caption{Calculation of $\phi$ for a second sample language.}
\end{table}

\[
\phi(XY) = \frac{ad - bc}{\sqrt{efgh}} = \frac{(2*1) - (0*0)}{\sqrt{2*1*2*1}} = 1.0
\]

1.4.2 Prosodic Grouping

Phrasal prosody in natural languages relies on a set of overlapping cues, including pitch shifts near phrasal boundaries, speed and length difference within phrases, and brief pauses between phrases. In the sentence:

\textit{[The bear and the sheep]NP [ran to the park]VP}

all of these cues would be present near or at the clause boundary. These cues have been used in previous studies (Morgan, Meier, and Newport, 1987) to successfully aid in phrase structure learning. Infants are also sensitive to these cues (Hirsh-Pasek, et. al., 1987); although at first this ability seems to be a general acoustic capacity and only

\footnote{Throughout the remainder of this paper, all statistical regularities will be discussed in terms of $\phi$, but this is not necessarily the exact statistical relationship that learners are computing. As in many miniature language experiments, conditionalized statistics overlap strongly with each other. A bigram or trigram that has a high $\phi$ regularity is also likely to have a strong forward transitional probability, conditional entropy, etc. as well. One statistical factor that is not controlled for in these studies is bigram and trigram frequency; however, previous studies on word segmentation (Aslin, Saffran, and Newport, 1998) and phrase structure learning (Thompson and Newport, 2007) have controlled for frequency and shown that it had little effect on the learners. In these studies learners acquired words and phrases from predictive relationships rather than frequency, so it is likely that any learning here will be due to the conditionalized $\phi$ cue rather than frequency as well. In order to control for frequency in the present work, a further set of studies dissociating frequency and $\phi$ would need to be performed.}
sensitive to infant directed speech (Kemler Nelson, 1989), it develops into a powerful
sensitivity by 9 months of age, at which time infants are able to detect not only clause
boundaries, but phrase boundaries within a clause as well (Jusczyk, et al., 1992). One
potential pitfall of relying on prosody, however, is that due to its interaction with
phonology, it is not always consistent with syntactic phrases. Furthermore, prosody is
unable to represent the many levels of embedding in the hierarchical structure of a
complex sentence, especially in the simplified forms utilized in miniature language
experiments. Nonetheless, prosodic cues have been offered as a potential early cue to
syntactic structure in Prosodic Bootstrapping theories (Morgan and Demuth, 1996).
Thus prosody may be a useful cue to some parts of hierarchical phrase structure,
especially early in learning, but it is unlikely to be sufficient on its own to support
complete multi-tiered syntactic hierarchy acquisition.

1.4.3 Semantic Reference Worlds

Since language is most often used to describe and refer to the external world,
it may also be partially learned by a word to world mapping procedure. This is clearly
evident in word learning at and after the word spurt (Smith, 1926; Benedict, 1979),
which involves a variety of cognitive processes including fast-mapping (Carey, 1977),
categorization (Rosch and Mervis, 1975), and schematic scripts (Nelson, 1986).
Lexical development also correlates with syntactic development (Bates and Goodman,
1997), which may be due to early learning of individual verb based constructions such
as *Eat* requiring an *Eater* and possibly an *Eaten* (Tomasello, 2000). Early semantic
development has also been proposed as a central component of syntax learning,
through a Semantic Bootstrapping approach (Grimshaw, 1981) which relies on
patterns such as verb meaning correlating with permissible syntactic structures (Fisher, Gleitman, & Gleitman, 1991). The process may also work in the opposite direction, with syntactic knowledge aiding semantic learning (Gleitman, 1990). Semantic structures, as discussed above, also play a central role in many linguistic theories of syntax, especially those of Bresnan (1982, 2001), Harris (1951, 1968, 1982), Jackendoff (1972, 1990, 1997, 2002), Croft (1991, 2001), Lakoff (1987), Langacker (1987), and Talmy (2000). While few theories go so far as to argue for synonymy of syntax and semantics, a common trend is to attempt to explain as much of syntactic structure as possible as deriving from semantics.

Within the miniature language learning literature, semantics is instantiated as a visual reference world associated with the language. While a reference world may not be necessary to induce learning of word order and basic phrase structure, its presence has a number of benefits beyond providing the learner with potential information about the syntactic structure. Since miniature language learning tasks in adults can last anywhere from 15 minutes to over 3 hours (split over multiple days), they often require a task that is interesting enough to hold the attention of subjects, and the presence of a visual stimulus often fulfills this role. Second, and perhaps more importantly, by providing a visual reference world, learners are presented with an obvious task of trying to determine the meanings of words. Even if this is unimportant to syntactic learning, it is a sufficient distractor to relegate structure learning to implicit and unconscious processes, which may actually improve learning by preventing conscious hypothesis generating and testing, thus better matching natural language acquisition.
Semantics has traditionally taken one of two forms in miniature language learning studies: as a simple word and category based associate, or as a more incorporated dependency structure marker, as depicted in Figure 1.1, taken from Morgan and Newport (1981). In the first approach (Figure 1.1.3), each word picks out a visual object, but there is no relationship or incorporation between different objects. This is simpler but does not provide information about hierarchical structure. This can be partially included by the addition of spatial grouping of word referents that are structurally related (usually members of the same phrase). This method, however, is not closely related to natural language semantics. The second option (Figure 1.1.4) involves different word classes identifying different kinds of visual features, and phrasally related items interacting visually through an incorporation. This method seeks to better match natural language semantics, although its success depends greatly on the nature of the visual world and the types of semantic relationships chosen by the experimenters.\(^6\)

1.4.4 Multi-Tiered Hierarchical Structure

While many of the miniature language studies of hierarchy learning have been based on grammars in which words group into phrases and phrases group together to

\(^6\) For further discussion, see Chapter 4.
form sentences, they do not test for groupings larger than a single phrase. For instance, in the grammar used by Morgan and Newport (1981) and others:

\[
\begin{align*}
S &\rightarrow AP + BP + (CP) \\
AP &\rightarrow A + (D) \\
BP &\rightarrow \{E, CP + F\} \\
CP &\rightarrow C + (D)
\end{align*}
\]

the fragment tests measured whether subjects had formed A phrase, B phrase, and C phrase groupings, but not whether they had formed a multi-tiered hierarchical phrase structure (as in the BP, which within it contains a CP). This means that subjects might not have learned a hierarchically organized phrase structure, but only a set of phrases at a single level of syntactic analysis, a mono-stratal\(^7\) structure. Within syntactic theories of natural language, hierarchical structures are often extremely complicated, involving the nesting of smaller phrases within larger phrases, and the organization of these structures are quite important for the syntactic system. Notions such as C-command (Reinhart, 1976), for instance, depend on a particular arrangement of a multi-tiered hierarchical structure. Nested hierarchical structures also follow a few simple but critical rules regarding linear order and constituency. First, traditional hierarchical structures must be built from immediately neighboring elements, as in Figure 1.2a, and cannot involve crossed dependencies as in Figure 1.2b. Second, once an item is a daughter (or member) of a phrase, it can only become a daughter (or perhaps “granddaughter”) of another phrase if its entire existing constituent becomes a daughter, a property that can be referred to as Constituency. Thus phrasal membership is inherited, and constituents cannot be split. In Figure 1.2, this means that item \(D\)

\(^7\) I use the term “mono-stratal” not in the linguistic sense of non-derivational, but with its original meaning of “one-layered”.

cannot become part of EP without all of the DP and its component parts, C and D, becoming a daughter of EP. These basic rules must be followed by the hierarchical structure of miniature languages, and can provide useful test cases of what learners have acquired after exposure to sentences of these languages.

1.4.5 Overview of Experiments

This dissertation reports the results from three experiments which explored the learning of hierarchical phrase structure from sentences of a variety of miniature languages (4 in Experiment 1, 8 in Experiment 2, and 4 in Experiment 3). In Experiment 1 (Chapter 2), we explore whether learners can acquire a hierarchical structure from local statistical regularities (φ) alone and whether learners impose additional structure due to pre-existing biases. Three of the languages in this experiment involve sentence structures in which three neighboring lexical categories have a strong statistical relationship between them, while the other two categories have weak relationships to each other and/or to the group of three categories. In the fourth language, statistical regularities are strong within two separate pairs of categories. By testing for two- and three-category grouping preferences in each of these languages, we can determine whether learners are attentive to the statistical information and
whether they form any constituent groupings that would be unexpected from the statistical information, indicating a pre-existing learning bias.

In Experiment 2 (Chapter 3), we consider how other cues, such as prosodic or visual world constituent grouping, similar to those used in Morgan and Newport (1981) and subsequent work, contribute to hierarchy learning, and the interaction of these cues with local statistical regularities. Two of these languages are similar to the final language in Experiment 1, consisting of statistical regularities within a pair of two category phrases. Another pair of languages does not have any statistical cues to phrase structure, but has semantic and prosodic information indicating the same hierarchical structures as the previous two languages in this experiment. The final four languages contain both statistical cues and semantic and prosodic grouping information, in languages that are structurally like those used in this experiment and like the last language in Experiment 1. By comparing the results of these languages with each other and with the results from Experiment 1, it will be possible to determine what cues are important to multi-tiered hierarchical phrase structure learning and how they can combine.

In Experiment 3 (Chapter 4), we examine how these cues might work in a more complex syntactic hierarchy learning task, more akin to natural language, and ask whether the type of semantic relationships expressed by the visual world might actually matter more than was found by Morgan and Newport (1981). Two different types of visual reference worlds are used here, one identical to Experiment 2 and one that attempts to mimic natural language semantics and its dependency and modifier relationships among categories. Two different word orders are also used, one in which
the statistical regularities between categories are strong between adjacent categories (we will refer to this as Local $\phi$), and one in which the statistical regularities are not consistent between adjacent categories but are strong between hierarchically related categories that are nonadjacent (we will refer to this as Non-Local $\phi$). This experiment allows us to determine the contribution of different sorts of semantic and statistical relationships, as well as their interaction. Because the languages used in this experiment are more complicated and therefore harder to learn, success here ought to better suggest the processes occurring during natural language acquisition.
Chapter 2. Experiment 1

In the first experiment, learners were exposed to one of four miniature languages containing statistical regularities between some lexical categories. There were no prosodic or semantic differences between the languages, and all relied on identical lexicons, thus making the statistical regularities the only variable on which languages differed. Any differences in hierarchical phrase structure learning between languages must therefore be attributed to the influence of the statistical information. The statistical regularities in each language did not provide enough information to determine a full phrase structure parse, but instead were ambiguous as to the complete phrase structure of the language, allowing the testing of learners’ biases in disambiguating hierarchical structures. In three of the languages (Figure 2.1 a1, b1, c1)

Figure 2.1: Phrase structures of the languages used in each experiment.
the statistical regularities indicated a three-category phrase, allowing for testing of binary grouping preferences within the phrase where statistical relationships were identical. In the fourth language (Figure 2.1 d₁), statistical regularities indicated two two-category phrases with a single category in between them, allowing for testing whether any three-category phrase group preference was formed by combining one of these phrases with the individual category. We predict that learners will attend to the statistical information in forming phrase grouping preferences, and that this information will interact with any biases the learners bring with them to the task to disambiguate the hierarchical structure of the languages.

2.1 Method

2.1.1 Participants

Forty adult monolingual English-speaking undergraduate students participated in the experiment. Each participant was randomly assigned to one of the four language conditions. All participants gave informed consent prior to participating and were paid for their participation upon completion of the experiment.

2.1.2 Description of the linguistic systems

All of the languages shared a number of properties, including an identical set of nonsense words as the lexicon, a visual reference world, auditory pronunciation and prosody of stimuli, number and order of exposure and testing trials, and instructions as described below.
Vocabulary:

The vocabulary of each language consisted of fifteen nonsense words (henceforth ‘words’) organized into five lexical categories (called A, B, C, D, and E). Each category contained a CVC, CVCV, and CVCVC word to avoid any potential rhythmic consistency in the stimuli, which might effect phrase structure learning. The words used in these languages can be found in Table 2.1. Two versions of the lexicon were utilized, one in which the categories were assigned to the word groups left to right along Table 2.1, and the other where the categories were assigned to word groups in a right to left order on Table 2.1. Half the learners of each language learned the first version of the lexicon, while the other half learned the reversed version. This alternation provided counterbalancing of specific lexical items with particular sentence positions, but in all experiments, no difference was found between these two versions of the lexicon, so results will combine both versions unless otherwise noted.

### Table 2.1: Lexicon and pronunciation of words for Experiment 1

<table>
<thead>
<tr>
<th>Version 1</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>siz /sɪz/</td>
<td>mun /mʌn/</td>
<td>tav /tæv/</td>
<td>lod /lɒd/</td>
<td>rup /rʌp/</td>
</tr>
<tr>
<td></td>
<td>heta /hɛtə/</td>
<td>zami /zæmi/</td>
<td>vigo /vɪgəʊ/</td>
<td>nehu /nɛhu/</td>
<td>boke /boukeɪ/</td>
</tr>
<tr>
<td></td>
<td>folub /fʊlʌb/</td>
<td>gesor /gɛsɔr/</td>
<td>kufes /kʊfəs/</td>
<td>pabif /pæbɪf/</td>
<td>diral /dɪrəl/</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Version 2</th>
<th>E</th>
<th>D</th>
<th>C</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
</table>

Grammar and Statistics:

Each language consisted of equal numbers of sentences that contained all possible linear combinations of phrases and individual categories that followed the

---

8 The syllables in the language were chosen so that each scored between 60 and 70 on an index of meaningfulness to English speakers of all possible CVC trigrams (Archer, 1960). This made each word reasonably distant from real words of English to prevent easy mnemonic strategies.
phrase structure grammars shown in Figure 2.1. The categories, A, B, C, D, and E must occur in that order, whenever they occur together, and each category can only be followed or preceded by its immediately neighboring category. In other words, sequences such as ABC are permissible, but ABD is not as it contains a “gap” between B and D⁹. Any member of a category can fill its location, and categories can only occur once per sentence, making the longest possible sentence 5 words long (1 each of categories A-E), while the shortest possible sentence is a single word: the member of a category that is not part of a larger phrase. Table 2.2 lists the sentences, in terms of their category symbols, for languages a₁, b₁, c₁, and d₁.

Since all of the languages rely on a maximal sentence of ABCDE, it is only through the optional appearance of certain categories that they differ. This design is very similar to the Optional Condition used in Thompson and Newport (2007), but modifies it in a few important ways. First, all of the lexical items in Thompson and Newport were CVC syllables and all of the phrases were two-word groups, creating the potential for a strong rhythmic cue to phrase structure. Second and more important is that the languages here vary in which sets of categories optionally occur.

Table 2.2: Sentences of Language a₁, b₁, c₁, and d₁.

<table>
<thead>
<tr>
<th></th>
<th>a₁</th>
<th>b₁</th>
<th>c₁</th>
<th>d₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>B</td>
<td>E</td>
<td>AB</td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td>AB</td>
<td>BCD</td>
<td>DE</td>
<td></td>
</tr>
<tr>
<td>ABC</td>
<td>CDE</td>
<td>ABCD</td>
<td>ABC</td>
<td></td>
</tr>
<tr>
<td>ABCD</td>
<td>BCDE</td>
<td>BCDE</td>
<td>CDE</td>
<td></td>
</tr>
<tr>
<td>ABCDE</td>
<td>ABCDE</td>
<td>ABCDE</td>
<td>ABCDE</td>
<td></td>
</tr>
</tbody>
</table>

⁹ These restrictions make it nearly impossible to create simple sets of phrase structure rules for these languages, but still allows for simple tree-like figures to be drawn.
In languages $a_1$, $b_1$, and $c_1$, there are two one-word optional “phrases” and a three-word optional phrase; in language $d_1$, there are two two-word optional phrases and a single one-word optional phrase. This experiment tests the ability of learners to use this cue to form varied hierarchical structures.

Using these distributions of sentences, we can calculate the $\phi$ for each pair of adjacent categories in the languages. As is evident in Table 2.3, the phrase internal category pairs (AB and BC in $a_1$, CD and DE in $b_1$, BC and CD in $c_1$, and AB and DE in $d_1$) have perfect predictive relationships between them because they always co-occur. Other category pairs have $\phi = 0$ because there is no predictiveness between them.

Table 2.3: $\phi$ statistics for all neighboring category bigrams.

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$b_1$</th>
<th>$c_1$</th>
<th>$d_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$(AB)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\phi$(BC)</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\phi$(CD)</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\phi$(DE)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Each grammar generates over 300 possible unique sentences ($a_1$: 366, $b_1$: 366, $c_1$: 438, $d_1$: 318) which are randomly sampled with replacement during exposure trials.

Because the different sentence types occur in equal proportions, there is greater repetition of shorter sentences.

**Semantic Reference World**

Each word in the language is matched to a single solid color filled shape which appears on the screen during the presentation of any sentence containing that word.
All the words within the same lexical category occur in a single consistent screen position; each category has its own position. Category members also have family resemblance in their shapes and colors, so one category might contain words matching a red isosceles triangle, a pink equilateral triangle, and a maroon scalene triangle. Figure 2.2 shows examples of sentences with their semantic reference worlds. All of the languages share the same set of word-to-world mapping between lexical items and visual positions and shapes, so the visual reference world only varies in accordance with the set of possible sentences. Visual reference scenes are not used during testing, however, so participants must rely on their knowledge of the words, categories, and phrases when they are tested on their knowledge of the language.

Figure 2.2: Sample sentences with visual referents (samples drawn from language a₁)

Prosody and Auditory materials

Auditory presentation of each word was synthesized using Apple’ speech synthesis voice “Victoria High Quality” and edited in SoundEdit 16 and Sound Studio for length and volume. Words were recorded individually and concatenated into sequences with minimum time between each word (approximately 20 msec) during the experiment, creating a list like pronunciation of the sentences. This permitted the randomization of sentences individually for each participant. The same auditory concatenation was used for test items.
2.1.3 Procedure

The experiment was conducted in 6' by 8' sound attenuated booths using an iMac computer running PsyScope software and Sennheiser 570 headphones for language presentation. Participants were given the following instruction at the beginning of the experiment:

“In this experiment you will be attempting to learn an artificial language. You will see and hear sentences from the language and see pictures they describe. Try to learn as much of the language as you can.

After approximately 10 minutes of watching sentences, you will do a series of tests. For each test click on the answer you think comes from the language, or feel is better. You will then watch 3 more sets of sentences, and do 3 more sets of tests. The entire experiment should take approximately 50 minutes to 1 hour.”

In each of the four exposure blocks, participants heard a series of 150 sentences matched with their appropriate semantic reference world pictures, separated by a 2 second inter-sentence interval. After each exposure block, participants completed a set of two tests, one set for Phrase Grouping preference (27 trials) and the other for correct Sentential Word Order (6 trials); the order of these tests was alternated by blocks. All participants received identical test items, appropriate to the version of the lexicon they experienced during the exposure phase. Participants were permitted to take a brief break after each Exposure + Testing block to avoid fatigue. The entire procedure lasted approximately 45-60 minutes.

Phrase Grouping Preference test

In the Phase Grouping Preference test, participants were asked to choose between a pair of two- or three-word groups that were permissible in the language. Before beginning, all participants were given the following instructions:
"In the following section you will be answering a set of tests. In each test you will see two or three words at the top of the screen and bottom of the screen. You will also hear both sets of words.

Both groups of words come from the language, your job is to choose which group is "better" or which group you think is more likely to occur in the language. For each test, click on the group of words you think is better."

The test block included 18 trials assessing preferences for two-word phrases and 9 trials assessing preferences for three-word phrases. These were organized as shown in Tables 2.4 and 2.5. Three trials of each type displayed in the table were performed in every test block. Participants were allowed as much time as needed to choose between fragment groups, and the computer only moved on to the next trial after a group was selected with the mouse. Specific lexical items were varied over the trials to include every possible lexical item in each test block and every possible lexical combination.

<table>
<thead>
<tr>
<th>Two Word Phrase Grouping</th>
<th>Three Word Phrase Grouping</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choice 1</strong></td>
<td><strong>Choice 2</strong></td>
</tr>
<tr>
<td>AB</td>
<td>BC</td>
</tr>
<tr>
<td>AB</td>
<td>CD</td>
</tr>
<tr>
<td>AB</td>
<td>DE</td>
</tr>
<tr>
<td>BC</td>
<td>CD</td>
</tr>
<tr>
<td>BC</td>
<td>DE</td>
</tr>
<tr>
<td>CD</td>
<td>DE</td>
</tr>
</tbody>
</table>

Table 2.4, 2.5: Phrase Grouping Preference test types. Each row represents one possible 2 Alternative Forced Choice test. Participants received three trials of each test type per testing block.
during the experiment. Lexical items were selected to avoid any repetition of a word within a given trial.\textsuperscript{10} Position of the two answers was randomized by trial and subject.

This test was designed to assess the preferences participants had formed for particular groupings of words and categories into phrases, reflecting their acquisition of phrase structure. The design of these trials permits preferences for all possible two- and three-category combinations to be calculated and allows the same testing procedures to be used on all subjects regardless of the specific language they were exposed to. While a more focused set of planned comparisons might have been possible with fewer trials, this would have varied by the particular grammatical system of each language. The present organization ensures that the results of these tests can be directly compared across different languages.

Sentential Word Order test

In the Sentential Word Order test, participants were asked to choose between a pair of five word sentences, only one of which was permissible in the language. Before beginning, all participants were given the following instructions:

“In the following section you will be answering a set of tests. In each test you will see a sentence at the top of the screen and bottom of the screen and hear both sentences.

Only one of these sentences comes from the language, your job is to choose which sentence could occur in the language. For each test, click on the sentence you think is possible.”

The test block included six trials assessing preferences for possible over impossible five-word sentences. Each impossible sentence involves a B, C, or D category word in

\textsuperscript{10} This means that if subjects were in an AB vs. BC trial, the B word was guaranteed to be different between the two options.
an impossible non-neighboring position and was paired with a possible sentence. The impossible sentence types are shown in Table 2.6. One trial of each type was used in each testing block. Participants were allowed as much time as needed to choose between sentences, and the computer only moved on to the next trial after a sentence was selected with the mouse. Three of the words in the possible and impossible items of each pairs were identical; only two words varied between the choices: the misplaced word and one within-category distractor. This simplified the trials and helped to insure that the choices made were based on word order and not other lexical factors.

Table 2.6: Sentential Word Order test types.
Each row represents one possible 2-AFC test. Participants received one trial of each test type per block. Violations of correct word order are bolded.

<table>
<thead>
<tr>
<th>Correct Word Order</th>
<th>Incorrect Word Order</th>
<th>Example Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABCDE</td>
<td>ABCBE</td>
<td>siz zami kufes lod boke</td>
</tr>
<tr>
<td></td>
<td></td>
<td>siz zami vigo <strong>mun</strong> boke</td>
</tr>
<tr>
<td>ABCDE</td>
<td>ABCDB</td>
<td>heta mun tav nehu diral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>heta mun kufes nehu <strong>gesor</strong></td>
</tr>
<tr>
<td>ABCDE</td>
<td>CBCDE</td>
<td>folub gesor vigo pabif rup</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>tav</strong> mun vigo pabif rup</td>
</tr>
<tr>
<td>ABCDE</td>
<td>ABCDC</td>
<td>siz mun vigo nehu boke</td>
</tr>
<tr>
<td></td>
<td></td>
<td>siz mun vigo lod <strong>kufes</strong></td>
</tr>
<tr>
<td>ABCDE</td>
<td>DBCDE</td>
<td>heta zami tav pabif diral</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>nehu</strong> zami kufes pabif diral</td>
</tr>
<tr>
<td>ABCDE</td>
<td>ADCDE</td>
<td>folub gesor kufes lod rup</td>
</tr>
<tr>
<td></td>
<td></td>
<td>folub <strong>pabif</strong> vigo lod rup</td>
</tr>
</tbody>
</table>
without making the purpose of the test obvious to participants. Position of the two answers was randomized by trial and subject.

This test was designed to assess the ability of participants to learn the basic word order of the language. Because all of the languages share the same possible five word sentences, the same testing procedures were used on all subjects. Participants generally do quite well on the Sentential Word Order tests, performing at 90% or better in many cases, regardless of the grammatical regularities within the language being learned; the data of focus will therefore be on phrasal learning. Thus we can also use the Sentential Word Order tests as a metric of whether subjects were attending to the experiment, or whether they were just choosing answers at random.\footnote{A previous experiment, not reported here, was designed to measure whether a semantic reference world was necessary at all, but was abandoned because participants became bored and failed the word order test entirely.}

\section*{2.2 Results}

\subsection*{2.2.1 Sentential Word Order test}

In order to determine whether participants had learned the basic structure of the language, regardless of the grammar, we began by measuring how accurate they were at distinguishing a possible five-word sentence from an impossible five-word sentence. Overall, participants averaged 90\% correct (SD = 8\%) on tests of Sentential Word Order, suggesting that they had learned the possible word orders extremely well. There was little variation between language groups on the word order test, as can be seen in Table 2.7 and Figure 2.3. A one-way ANOVA across the four language groups, $F(3, 36) = 1.25$, $p > .05$, confirmed that there was no difference across the
Table 2.7: Sentential Word Order test scores for Experiment 1 languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Word Order Score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₁: (ABC)-D-E</td>
<td>91.8% (6.6%)</td>
</tr>
<tr>
<td>b₁: A-B-(CDE)</td>
<td>91.5% (8.6%)</td>
</tr>
<tr>
<td>c₁: A-(BCD)-E</td>
<td>85.7% (8.7%)</td>
</tr>
<tr>
<td>d₁: (AB)-C-(DE)</td>
<td>89.8% (8.0%)</td>
</tr>
</tbody>
</table>

Figure 2.3: Sentential Word Order test scores for Experiment 1 languages. Dashed line represents chance. Error bars represent 95%CI.

conditions, and a one sample T-test for each language confirmed that all scores were significantly greater than chance, $T_{a₁}(9) = 19.99$, $T_{b₁}(9) = 15.50$, $T_{c₁}(9) = 15.524$, $T_{d₁}(9) = 15.82$, all $p<.001$. Regardless of which grammar, and corresponding set of shorter sentences in the language, participants all learned the correct word order of the longest possible sentences equally well and significantly above chance.
2.2.2 Two-Word Phrase Grouping Preference test

The most local statistical regularities occur between neighboring words and categories, as reflected by the calculations of $\phi$ in Table 2.3. If learners base their phrase structure hierarchies on these regularities, then we would expect the highly regular pairs ($\phi = 1$) to be preferred as groups over those which lack a predictive relationship ($\phi = 0$). In each language there are four specific combinations of two-word phrase grouping preference comparisons that contrast the phrases supported by local statistical regularities against those that are not supported by local statistical regularities, as listed in Table 2.8. Across the languages, the average preference for two-word phrases supported by statistical regularities was only 57.4% (SD = 17.8%), which, while significantly different from chance, $t(39) = 2.624$, $p < .05$, is not very impressive performance given the perfectly consistent organization of the statistical information in the exposure sentences. Separating out the performance on the two-word phrase grouping preference test by language reveals a heterogeneous pattern across languages, as seen in Figure 2.4.

The languages are in fact quite different from each other, $F(3,36) = 24.565$, $p < .001$, with the effect almost entirely driven by the very high scores in language d1.

Table 2.8: Trial types for Two-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. phrases not supported by statistical regularities. Each test is listed as “Supported - Unsupported”

<table>
<thead>
<tr>
<th>$a_1$: (ABC)-D-E</th>
<th>$b_1$: A-B-(CDE)</th>
<th>$c_1$: A-(BCD)-E</th>
<th>$d_1$: (AB)-C-(DE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB - CD</td>
<td>CD - AB</td>
<td>BC - AB</td>
<td>AB - BC</td>
</tr>
<tr>
<td>AB - DE</td>
<td>DE - AB</td>
<td>CD - AB</td>
<td>AB - CD</td>
</tr>
<tr>
<td>BC - CD</td>
<td>CD - BC</td>
<td>BC - DE</td>
<td>DE - BC</td>
</tr>
<tr>
<td>BC - DE</td>
<td>DE - BC</td>
<td>CD - DE</td>
<td>DE - CD</td>
</tr>
</tbody>
</table>
(AB)-C-(DE). Post-hoc tests with a Bonferroni correction revealed that only this language was significantly different from the other three, $t_{a_1-d_1}(18) = 7.50$, $t_{b_1-d_1}(18) = 5.44$, $t_{c_1-d_1}(18) = 7.33$, all $p < .001$. Furthermore, the (AB)-C-(DE) language was also the only language significantly better than chance on this test, $t(9) = 11.35$, $p < .001$.

There are at least two possible explanations for this heterogeneous pattern in the two-word phrase grouping results. First, the (AB)-C-(DE) language is statistically quite different from the other languages, because the two phrases that have $\phi = 1$ do not share a common category in (AB)-C-(DE) as they do in the other languages. This means that the statistical regularities in the first three languages may point the learner to a 3-word, or ternary phrase, while the statistical regularities in language d clearly points learners to a pair of 2-word, or binary phrases. Since the test here is for binary phrase preference, the lack of preference in the (ABC)-D-E, A-B-(CDE), and A-(BCD)-E languages may reflect this ternary phrase structure. An alternative
explanation is that participants might be doing two kinds of learning: building phrase structure from statistical regularities -what might be considered ‘learning’ from cues - and then forming a complete hierarchical parse by combining the phrases already learned with the other words and categories, a ‘secondary learning’ process. One way this process might function in language (ABC)-D-E and A-B-(CDE) is by forming a two-word phrase from the categories that remain outside of the ternary phrase which was built upon the statistical regularities. In a₁, this would produce a complete parse of (ABC)-(DE), and in b₁, (AB)-(CDE). No such complete parse is possible in language c₁, A-(BCD)-E, because any other binary combination would necessarily cross a phrase boundary.

To distinguish these two alternative hypotheses, the two-word phrase grouping preference analysis was performed again considering only the statistically regular phrases compared to those that would cross the phrase boundary, crucially leaving out the potential remaining pairs from the calculations. This produces the contrasts in Table 2.9. The results of this analysis are depicted in Figure 2.5. What is strikingly clear from this, especially compared to Figure 2.4, is the vast and significant improvement in the (ABC)-D-E and A-B-(CDE) scores, \( t_{a₁}(9) = 3.91, p < .01, t_{b₁}(9) = \)

<table>
<thead>
<tr>
<th>a₁: (ABC)-D-E</th>
<th>b₁: A-B-(CDE)</th>
<th>c₁: A-(BCD)-E</th>
<th>d₁: (AB)-(CDE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB - CD</td>
<td>CD - BC</td>
<td>BC - AB</td>
<td>AB - BC</td>
</tr>
<tr>
<td>BC - CD</td>
<td>DE - BC</td>
<td>CD - AB</td>
<td>AB - CD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BC - DE</td>
<td>DE - BC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CD - DE</td>
<td>DE - CD</td>
</tr>
</tbody>
</table>

Table 2.9: Trial types for Two-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. pairs that cross the phrase boundary
6.259, p < .001. This indicates that DE and AB, respectively, were also being treated like phrases and the problem was not due to the statistics only indicating a ternary phrase in these two languages. This is confirmed by comparing DE against CD in the (ABC)-D-E language, Preference = 79.2%, t(9) = 4.72, p < .001, and AB against BC in the A-B-(CDE) language, Preference = 78.3%, t(9) = 6.218, p < .001.

The improvement in scores within this analysis explains the earlier apparent failure of the (ABC)-D-E and A-B-(CDE) languages as clearly resulting from an additional phrasal preference that was not provided by the statistical regularities. It also clearly rejects the hypothesis that the problem for learners was in trying to acquire a ternary phrase. However, this does not explain the failure of the A-(BCD)-E
language as it is unaffected by the change in this analysis. A discussion of this issue will be taken up below in section 2.3.

One issue of concern in these languages, which might contribute to the failure of learners in the A-(BCD)-E condition, is that the other three languages all involve two-word statistical regularities that occur in sentence initial or sentence final position (AB and DE), whereas the A-(BCD)-E condition exhibits its statistical regularities for phrases only in the middle of the sentence. Given the large body of literature on Primacy and Recency effects and the U-shaped serial position curve for short term memory tasks, as well as extensive evidence in the language acquisition literature that beginnings and ends of sentences and words are learned early (Shipley Smith & Gleitman 1969; Slobin, 1973), it is possible that the results here are not being primarily driven by the statistical regularities, but instead by simple serial position or edge effects. If this were the case, then the phrase preference effects in these languages would be expected to only hold for tests involving category pairs on the edge of the sentence (AB and DE) pitted against category pairs in the middle of the sentence (BC and CD). There would also be a predicted lack of difference between sentence internal category pairs, as in comparing BC and CD phrases, across all of the languages, not just in those conditions which do not distinguish these pairs based on statistical regularities. This concern turns out to be incorrect in its extreme form, as shown by the comparison of BC and CD in Figure 2.6. Language a_1, (ABC)-D-E, shows a strong preference for BC over CD, t(9) = 5.49, p < .001, while language b_1, A-B-(CDE), shows a strong preference for CD over BC, t(9) = 5.873, p < .01. These preferences match the statistical regularities in these languages. Language c_1, A-
(BCD)-E, and d₁, (AB)-C-(DE), on the other hand, do not show a preference between these two-word phrase groupings, \( t_{c1}(9) = 1.64, \ p > .05, \) and \( t_{d1}(9) = .64, \ p > .05, \) thus further confirming that it is statistical regularities that are driving these non-edge preference effects.

While these non-edge phrases are learned properly, however, there is also an overarching edge effect that is likely contributing to the failure of participants to learn the A-BCD-E language and is a strong component of the learning in the other three languages. This is especially apparent for the sentence final DE phrase, which across all the languages has an average preference of 70.5%, \( t(39) = 8.88, \ p < .001. \) Even in the A-BCD-E language, the preference for DE, which violates the statistical regularities in the language, is significant at 63%, \( t(9) = 3.057, \ p < .05. \) It is likely that
this recency effect, or edge bias, is interfering with the learning of the A-BCD-E language while supporting learning in the other languages.

Finally, the ternary structure of the statistically regular phrases in language a₁, b₁, and c₁ allow the question of biases in hierarchical structure learning to be addressed. Most instantiations of X-bar theory (Chomsky, 1970; Jackendoff, 1977) as well as Bare Phrase Structure (Chomsky, 1995) predict that all hierarchical structures are inherently binary, and that in any phrase that appears to be ternary, there is an underlying multi-tiered hierarchical structure. Some theories (Kayne, 1994) go even further to suggest that the underlying structure of apparently ternary phrases all have exactly the same binary hierarchical structure, (Specifier-(Head-Complement)). On the other hand, some recent alternatives, such as the Simpler Syntax Hypothesis, accept the possibility of truly ternary (or perhaps larger) phrase structures without any organized substructure (Culicover and Jackendoff, 2005). Given the ambiguous nature of the ternary phrases in languages a₁, b₁, and c₁, an X-bar analysis would predict a consistent preference for either the first binary phrase or second binary phrase across all three languages¹², while a Simpler Syntax Hypothesis would predict no preference whatsoever, or at the very least, no consistent preference across the languages.

As Figure 2.7 shows, no consistent sub-structural preferences are present across these languages. While the A-B-(CDE) language does seem to lead to a significant preference for DE as a binary phrase over CD, t(9) = 4.659, p < .01, the other two languages do not show a significant distinction between their ternary-

¹² Because Kayne (1994) suggests that all ternary phrases should have a binary phrase preference for the middle and final word pair, the preference graphs will use this as their metric.
internal binary groups, $t_{a1}(9) = .352, p > .05$, $t_{c1}(9) = 1.64, p > .05$. These results can be interpreted two ways: The data could be indeterminate between the conflicting theories, with the lack of preference in the (ABC)-D-E and A-(BCD)-E languages confirming a flat ternary structure and the A-B-(CDE) preference for DE supporting an X-bar internal binary structure. Alternatively, the data can be seen as suggesting the interaction of a flexible learning system, willing to accept a ternary phrase with no internal structure, with a bias to group sentence final pairs into phrases as is expected from the Recency effect discussed earlier. Given the fact that a recency effect shows up even in the A-BCD-E language, the later alternative seems more likely.

---

Language d1, (AB)-(DE), is not shown in Figure 2.7 because these theories make no prediction about the relationship between AB and DE, but a statistical analysis has revealed no significant difference in the preference for these two binary phrases, $t(9) = 1.748, p > .05$. 

---

Figure 2.7: Two-word Phrase Grouping Preference test results: Antisymmetric X-bar Phrase structure preference within Statistically cued Ternary Group.
These analyses of Two-word Phrase Grouping preferences have revealed that learners are sensitive to statistical regularities between words and categories in forming phrase structure groups. This ability, however, seems to have some limits, as revealed in the failure to form consistent preferences in the A-(BCD)-E language. The phrase group learning applies to pairs that lie at the edges of sentences as well as those in the middle of sentences. It also seems to lead to secondary learning to form a complete parse of sentences. Finally, the data here suggest that not all large phrases need to have a hierarchically organized binary substructure, but that there is flexibility in the types of phrase structures that learners can acquire, although these options may be limited or effected by learning or structural biases.

2.2.3 Three-Word Phrase Grouping Preference test

If learners go beyond the purely local statistical regularities between immediate neighbors and consider patterns holding across larger sets of categories, then they might show three-word phrase groupings along with the two-word phrases already analyzed in section 2.2.2. In languages a1, b1, and c1, these would be based on a combination of primary statistical regularities, leading learners to form ABC, CDE, and BCD three-word phrase preferences, respectively. In language d1, (AB)-C-(DE), however, a three-word phrase would have to be formed via a secondary structure building relationship of combining one of the two-word phrases, built upon statistical regularities, with the remaining single C word in the sentence. There are two ways in which this combination could occur, by combining the single word with the sentence initial AB phrase as in [(AB)C]-(DE), or by combining the single word with the sentence final DE phrase as in (AB)-[C(DE)]. Since both of these are possible
combinations, they can be analyzed together to measure the average preference for these secondary possible phrase structures for this language. This creates two types of test trials in each language to measure learner’s preferences for supported three-word phrases over unsupported phrases, as seen in Table 2.10. While learning was significant for these three-word phrases overall (70.7% average), learners were not equally good at forming three-word groups, as can be observed in Figure 2.8, F(3,36) = 8.05, p < .001. Bonferonni adjusted post-hoc tests revealed significant differences between each of the three well learned languages, a₁ (71.6%), b₁ (81.8%), and d₁ (77.3%) as compared with language c₁ (52.2%), all p < .05, but no significant differences among the three well learned languages. T-tests also confirmed that learners of the three successful languages performed above chance, tₐ₁(9) = 4.58, p < .001, t₉₁(9) = 9.29, p < .001, t₉₁(9) = 4.48, p < .01, while learners of the A-(BCD)-E language performed at chance, t₉₁(9) = .602, p > .05.

In the (AB)-C-(DE) language, it is possible that success on this test was driven by only one of the two plausible three-word groups, ABC or CDE, being treated as a preferred phrase. Under closer scrutiny, however, this does not appear to be the case, as ABC is preferred over BCD 72.5% of the time, t(9) = 2.82, p < .05, and CDE is also preferred over BCD 81.7% of the time, t(9) = 5.30, p < .001. While there does seem to

<table>
<thead>
<tr>
<th>a₁: (ABC)-D-E</th>
<th>b₁: A-B-(CDE)</th>
<th>c₁: A-(BCD)-E</th>
<th>d₁: (AB)-C-(DE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC - BCD</td>
<td>CDE - ABC</td>
<td>BCD - ABC</td>
<td>ABC - BCD</td>
</tr>
<tr>
<td>ABC - CDE</td>
<td>CDE - BCD</td>
<td>BCD - CDE</td>
<td>CDE - BCD</td>
</tr>
</tbody>
</table>

Table 2.10: Trial types for Three-word Phrase Grouping Preference tests: Phrases supported by statistical regularities vs. phrases not supported by statistical regularities. Each test is listed as “Supported - Unsupported”.
be a trend toward stronger preferences for statistically consistent sentence final three-word phrases in A-B-(CDE) and (AB)-[C(DE)] (81.75%), compared to statistically consistent sentence initial three word phrases in (ABC)-D-E and [(AB)C]-(DE) (72.05%), this difference is only marginally significant, t(38) = 1.73, p = .093. It is surprising, however, that even in the A-BCD-E language, there is a small but consistent preference (58%) for CDE as a three-word phrase, t(9) = 2.59, p < .05. Thus, there is no cross-language consistent three-word phrase bias that overwhelms the statistical regularities, but there is an underlying bias towards sentence final groups being preferred over other possible combinations.

The results of the three-word phrase preference tests are quite consistent with the two-word phrase preference tests. The A-(BCD)-E language learners failed to show a hierarchical phrase structure analysis in both tests, while learners of the other
three languages all acquired phrase structure preferences that are in accordance with the primary statistical regularities for the A-B-(CDE) and (ABC)-D-E languages, and secondary derivative relationships for the [(AB)C]-(DE) and (AB)-[C(DE)] languages. Throughout, there was a small but sometimes significant bias towards treating sentence final groups (DE and CDE) as better phrases.

2.3 Discussion

The most important result of this experiment was to demonstrate that learners are capable of acquiring a phrase structure from statistical regularities between words and categories alone. This ability is flexible enough to acquire hierarchies involving flat ternary structures, as well as more traditional binary phrase groups. With the exception of the A-BCD-E language, learners showed preferences that were always consistent with the statistical regularities, preferring two-word and three-word groups that had high statistical regularity between them, and dispreferring pairs and triplets that crossed over the boundaries of statistically based phrases. This indicates that statistical cues across a set of sentences are sufficient to induce the learning of simple phrase structural grouping.

Learners also demonstrated a great deal of flexibility in the nature of the hierarchical structure they were willing to posit and upon which they based their decisions on the phrase grouping tests. When the statistical cues indicated a ternary structure, learners seemed to form a strong preference for a ternary structure, with additional preferences for any binary combinations within the larger structure, but no consistent preferences between these smaller combinations. This represents a ‘primary learning’ ability which builds structure out of statistical regularities. Similarly, when
the statistical cues were only consistent with two-word phrases, the learners showed strong preferences for these two word phrases, another instance of primary cue learning.

Going beyond the statistical information, learners also consistently performed in accordance with logical ‘secondary’ phrase structure learning. In the AB-C-DE language, they did this by positing three-word phrases that were possible combination of two-word phrases and an individual category. Importantly, not any ternary structure was possible, but only those that did not break up the constituency of the two-word phrases they had formed from their ‘primary’ cue learning. In the ABC-D-E and A-B-CDE languages, learners formed consistent preferences for binary combinations of the two individual categories. These preferences are not based on statistical cues, as $\phi = 0$ between these pairs, but instead suggest a learning system that seeks to find a logical and complete phrase structure wherever it is possible and builds it upon a foundation of phrases that were formed by attention to statistical information. In combination, these two abilities should allow learners to acquire a great deal of hierarchical structure from a statistically reliable environment, thus potentially helping to solve Plato’s problem for the domain of syntax.

As impressive as the learning here is, it is tempered by the failure of subjects exposed to the A-BCD-E language to learn any phrase structure preferences consistent with the statistical information that was available. There are a number of possible explanations for the failure of learners to acquire this particular language, which seems surprising given its similarity to the other languages in this experiment. First, there does seem to be an underlying predilection across all the languages to learn
phrase groupings that occur at the beginning or end of a sentence, that is, an *Edge Bias*. Since the two-word and three-word phrases in this particular language are exactly the opposite of those that coincide with the edge bias, this could be considered a conflict situations for the learners, pitting an underlying learning bias against the statistical information. All of the other languages in this experiment have primary statistical information and secondary learning that corresponds with the edge bias, leading to strong preferences for AB and DE pairs as well as ABC or CDE ternary phrases. This bias to form stronger groups near a natural edge is very reminiscent of the results of Shukla, Nespor and Mehler (2007) who showed that learners are more likely to treat statistically regular sequences of syllables as word like units when they occur at prosodic boundaries, especially at the end of a prosodic phrase. These languages could be so easily learned because of the compatibility between learning biases and the statistical information. Another alternative is that learners are unable to acquire the A-BCD-E language because there is not a complete parse that is possible for it. If learners are attentive to statistical information, as they seem to be, and begin to treat BCD as a phrase, it becomes hard to coordinate this with the remaining A and E categories to form a complete phrase structure.14 This inability to completely parse the sentence would then have to cause learners to throw out, or at least temper, their attention to the statistical information, however, to explain why the three-word phrase test shows no preference for BCD. The other languages do not face this problem, as

14 There are three possible parses of this language built upon this three word phrase, one in which BCD combines with E first, and then this four-word phrase combines with A; another in which BCD combines with A first, then this four-word phrase combines with E; and a final one in which BCD combines with A and E simultaneously in a ternary combination. All of these phrase structures, however, are far more complex and multi-tiered than any of the apparent phrase structures of the other languages in this experiment.
ABC-D-E can form a phrase out of the neighboring DE, A-B-CDE can do the same for AB, and AB-C-DE can form one of two different three word phrases combining C with either of the statistically based binary phrases. Clearly, it is difficult to determine why learners are unable to acquire this specific language, and while this is an important question for theories of hierarchical learning, it is beyond the scope of this dissertation to address in full.

While this experiment does demonstrate phrase structure learning from statistical regularities alone, it does not address an important issue: The phrase structures learned here are not necessarily multi-tiered. The lack of a preference for binary phrases within a ternary phrase may not be due purely to flexibility in the learning system, but could be a result of not building multi-tiered hierarchical phrase structures from statistical cues alone. While the formation of three-word phrase preferences in learners exposed to the AB-C-DE language does suggest a possible multi-tiered hierarchical structure, the participants did not distinguish between the two possible three-word phrases, indicating that they had not fully disambiguated the phrase structure of this language. The reason for the selection of potentially ambiguous hierarchical phrase structures in these languages was to measure underlying learning biases in the absence of overwhelming statistical cues. An unfortunate consequence of this choice, however, is a lack of clarity about the specific hierarchical structures learners had acquired. This problem will be addressed in Experiment 2, which involves modified, multi-tiered, versions of languages \(a_1\) and \(b_1\), as well as languages which attempt to disambiguate the potential ternary phrases resulting from a language like AB-C-DE.
Chapter 3. Experiment 2

In Experiment 1, we demonstrated that miniature language learners can acquire a number of different simple phrase grouping preferences from statistical information alone. This ability seems to function as long as the information does not contradict a learner’s bias towards grouping words into phrases at the edges of sentences. Experiment 1 does not address a number of important issues, however, which are critical to a more thorough examination of the acquisition of multi-tiered hierarchical structures. First, the previous results do not conclusively explain why participants failed to learn the A-BCD-E language, nor do they explore the nature of the edge preference or how it fully interacts with statistical information. Second, statistical regularities between words and categories is only one of numerous possible cues to phrase structure, and the prior experiment does not address how these other cues might function or how statistical information might combine with other cues to aid in hierarchical structure learning. Finally, the previous work does not necessarily implicate the acquisition of a multi-tiered hierarchical structure when the languages were learned, but instead could just involve the acquisition of a single level of phrase grouping preferences. The languages of Experiment 2 are designed to address all three of these questions.

Experiment 2 involves eight miniature languages that vary in both their statistical regularities and the presence of semantic and prosodic cues to their hierarchical structure. In Figure 2.1, there are four hierarchical structures shown, but the figure does not indicate which cues are present in each language, only the appropriate hierarchical structure which the cues are designed to reflect. Figure 3.1
above details all eight of the languages in this experiment. These languages will be referred to by the cues which they contain, as well as the hierarchical structure that the cues correlate to. The possible cue types are Statistical Regularities (Stat), Semantics and Prosody (SemPro), and a Combination of both types of Cues (CueCombine). Thus, a language with Semantic and Prosodic cues to an AB-(CD-E) phrase structure, would be called “SemPro-b2”.

Languages a2, and b2 are closely related to languages a1, and b1 from Experiment 1. In the ‘a’ languages, a three-word phrase is followed by a two-word phrase; in the ‘b’ languages, a two-word phrase is followed by a three-word phrase. One obvious difference, however, is that in Experiment 1, the relationships between
the words in the three-word phrase are all equivalent, while here there is sub-structure and asymmetries within the three-word phrase that are signaled by differences in cue expression. The remaining pairs of words, DE in $a_2$, and AB in $b_2$, are also now signaled as phrases by cues. These two hierarchical structures are represented by statistical regularities alone, semantic and prosodic information alone, or a combination of both cue types. These variations allow us to ask whether learning for both languages is the same and whether the type of cues to hierarchical structure present during exposure causes changes in learning. These languages also crucially involve multi-tiered hierarchical structures and cannot be entirely learned through an edge bias guided process. The final two languages, $c_2$ and $d_2$, both involve a combination of statistical regularities and semantic/prosodic cues to their hierarchical structure. These are the two alternative complete parses of language $d_1$, and the statistical regularities in both of these languages are identical to language $d_1$; only the semantic and prosodic cues vary between these languages. While learners of $d_1$ did not distinguish between ABC and CDE as possible three-word groups, these two languages allow us to ask whether the addition of semantic and prosodic information is enough to disambiguate the hierarchical structure of these languages and whether both three-word groups are equally learnable.

3.1 Method

The vocabulary of these languages and procedure for all participants was identical to Experiment 1. Unless specifically noted below, other components of the method are also identical to Experiment 1.
3.1.1 Participants

Eighty adult monolingual English-speaking undergraduate students participated in the experiment. Each participant was randomly assigned to one of the eight language conditions. All participants gave informed consent prior to participating and were paid for their participation upon completion of the experiment.

3.1.2 Description of the linguistic systems

Grammar and Statistics:

The sentences used in these languages include statistical regularities between categories in the statistically cued languages (Stat and CueCombine), and exhibit non-grouped, equivalent statistical relations among all neighboring word categories in the semantics and prosody only languages (SemPro). In order to achieve this, equal numbers of the possible sentence types are used for the ‘Statistical Regularities’ and ‘Cue Combination’ languages, while the ‘Semantics and Prosody’ languages involve controlled numbers of each sentence type as listed in Table 3.1. This creates the $\phi$ relationships between categories listed in Table 3.2. Once again, as in Experiment 1, the optionality of certain categories and sets of categories in each language creates the statistical regularities represented by $\phi$.

The sentences used here create perfect predictive relationships for the bigrams BC and DE in the statistically cued A-BC-DE languages ($a_2$), the AB and CD bigrams in the statistically cued AB-CD-E languages ($b_2$), and the AB and DE bigrams in the AB-C-DE languages ($c_2$ and $d_2$). When statistical cues are not present, there are no differences between the various bigram predictive relationships; however, these
languages have semantic and prosodic cues that point the learner to the same hierarchical structure as the equivalent statistical information.

**Table 3.1: Sentences of Languages $a_2$, $b_2$, $c_2$, and $d_2$ for each cue type.**

<table>
<thead>
<tr>
<th>Statistical Regularities</th>
<th>Semantics and Prosody</th>
<th>Cue Combination: Statistics + Semantics and Prosody</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_2$</td>
<td>$b_2$</td>
<td>$a_2$ and $b_2$</td>
</tr>
<tr>
<td>A</td>
<td>E</td>
<td>50 × {A, E}</td>
</tr>
<tr>
<td>BC</td>
<td>AB</td>
<td>28 × {AB, DE}</td>
</tr>
<tr>
<td>DE</td>
<td>CD</td>
<td>60 × {BC, CD}</td>
</tr>
<tr>
<td>ABC</td>
<td>CDE</td>
<td>28 × {ABC, CDE}</td>
</tr>
<tr>
<td>BCDE</td>
<td>ABCD</td>
<td>60 × BCD</td>
</tr>
<tr>
<td>ABCDE</td>
<td>ABCDE</td>
<td>248 × ABCDE</td>
</tr>
</tbody>
</table>

**Table 3.2: $\phi$ statistics for all neighboring category bigrams.**

<table>
<thead>
<tr>
<th>Statistical Regularities</th>
<th>Semantics and Prosody</th>
<th>Cue Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_2$</td>
<td>$b_2$</td>
</tr>
<tr>
<td>$\phi(AB)$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\phi(BC)$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$\phi(CD)$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$\phi(DE)$</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Semantic Reference Worlds**

The ‘Statistical Regularities’ languages rely on the same semantic reference world as in Experiment 1. The added semantic cue to hierarchical structure in the ‘Semantics and Prosody’ and ‘Cue Combination’ languages involves visual grouping of within-phrase category locations as can be seen in Figure 3.2. In the ‘Semantics and
Prosody’ languages, because a greater variety of sentence types was required to ensure there were no statistical cues to the hierarchical structure, there was also, by necessity, a greater variety of forms of the semantic reference world. Semantic grouping is a very different sort of cue than statistical regularities, in that it holds within individual sentences and not just across a corpus of different sentence types. This makes semantic grouping for a given bigram an all-or-none process: either word referent shapes appear close together or separated on the screen. This issue creates an interesting dilemma of how to make the semantic cue neutrally informative, like the statistical cues with a $\phi$ of 0, without making it negatively informative, which would be the equivalent of categories never co-occurring, or in statistical terms, a $\phi$ of -1. If a pair of shapes always occurred separated on the screen, while others occurred close together, this might be a cue that the separate shapes were definitely not in the same phrase. In order to avoid this potential issue, category bigrams were carefully matched with semantic groupings to either always appear close together, indicating a phrasal constituent, or sometimes appear together, indicating a neutral cue to phrase structure. Each sentence type was consistent in its relationship to the visual world, but the set of sentence type to visual world groupings was manipulated for neutrality. For example, potential ternary phrases, such as ABC in language $a_2$, appear clustered together when

![Sample sentences with semantic referents](image)
all three categories are present, but pairs such as AB appear separately visually. The
semantic groupings for all of the sentence types in the ‘Semantics and Prosody’ cue
languages are noted in Table 3.3, with parentheses marking visual neighbors. What
can be seen in Table 3.3 is that the semantic groupings offer a potential cue to both the
binary phrase combination as well as the larger ternary phrase groupings, in a very
similar way that the statistical regularities create primary and secondary cues to
hierarchical structure.

Table 3.3: Semantic Reference world groups in the ‘Semantics and Prosody’ languages.
Categories separated by spaces are visually distant, those in parentheses are close.

<table>
<thead>
<tr>
<th>Language a₂: A-BC-DE</th>
<th>Language b₂: AB-CD-E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>A B</td>
<td>(AB)</td>
</tr>
<tr>
<td>(BC)</td>
<td>B C</td>
</tr>
<tr>
<td>C D</td>
<td>(CD)</td>
</tr>
<tr>
<td>(DE)</td>
<td>D E</td>
</tr>
<tr>
<td>(ABC)</td>
<td>(AB) C</td>
</tr>
<tr>
<td>(BC) D</td>
<td>B (CD)</td>
</tr>
<tr>
<td>C (DE)</td>
<td>(CDE)</td>
</tr>
<tr>
<td>(ABC) (DE)</td>
<td>(AB) (CDE)</td>
</tr>
</tbody>
</table>

In the ‘Cue Combination’ languages, the sentence types are the same as the
Statistical Regularities languages (a₂, b₂) or the d₁ language from Experiment 1 (c₂,
d₂). This results in all of the sentences being composed of whole phrases only, so
semantic groupings are always present within a phrase. The visual world sentence
types are shown in Table 3.4. In these languages there is a combination of statistical
and semantic (and prosodic, see below) cues, consistently pointing the learner towards
a multi-tiered hierarchical structure in each language.
Prosody and Auditory materials

The auditory stimuli for the ‘Statistical Regularity’ languages are identical to those in Experiment 1, involving an evenly spaced concatenation of individual word sounds. Prosody in the ‘Semantics and Prosody’ and ‘Cue Combination’ conditions, however, involved prosodic groupings that were matched to the semantic groupings in these languages. Phrases were spoken faster, at a higher pitch, with less pause between words, and in a single prosodic envelope. Words not forming a phrasal group still relied on separate sound files, with longer pauses between the words and a list-like prosody. Thus in a sentence such as “zami kufes lod boke”, a (BC) (DE) sentence from language CueCombine-a, the auditory stimuli would involve two sound stimuli, zami-kufes, pronounced as a phrase, and lod-boke, pronounced as a phrase. The overall sentence sounds like a list of these two phrases concatenated together. Thus prosody acts as a reinforcing cue with semantic reference worlds but does not add any

---

15 This means that the grouping diagrams in tables 3.3 and 3.4 also represent prosodically grouped phrases. Three word phrases were pronounced as a single unit.
additional grouping information beside what is present in the semantic sentence-to-world match. In order to prevent learners from relying on familiar prosody at the time of test, all test items were presented with list-like prosody and no prosodic phrase groupings. The test items were identical to those of Experiment 1. Learners of languages that involved prosodic cues were warned that the test items might sound slightly different than the exposure items and to ignore these differences.

3.2 Results

3.2.1 Sentential Word Order test

Across the eight languages in Experiment 2, participants averaged 89% correct (SD = 10%) on the test of sentential word order. This is nearly identical to learners in Experiment 1. Figure 3.3 shows the performance of participants on this task. A One-Way ANOVA comparison of sentential word order tests between the eight languages.
revealed no significant difference, F(7,72) = 1.48, p > .05, and all languages were learned significantly better than chance, p < .001. Regardless of which grammar learners were exposed to, they all learned the correct word order of the five word sentences equally well and significantly above chance.

### 3.2.2 Two-Word Phrase Grouping Preference test

Two-Word phrase grouping preference tests comparing every pair of neighboring categories (AB, BC, CD, and DE) were conducted. For each language, this permits the calculation of predilection for two-word phrases that are supported by statistical and/or semantic and prosodic cues. The specific tests used for each of the four hierarchical structures are listed in Table 3.5. Results of these analyses can be seen in Figure 3.4. While all groups of learners performed at greater than chance rates on this test (lowest average score (SemPro-b\textsubscript{2}) = 62.1%, t(9) = 3.041, p < .05), the different languages and cue groups were not equal. Collapsing across languages within the specific cue types (Statistical Regularities, Semantics and Prosody, Cue Combination), a One-way ANOVA revealed a significant effect of Cue type on performance, F(2,77) = 22.86, p < .001. Post-hoc analyses localized the cause of the

<table>
<thead>
<tr>
<th>a\textsubscript{2}: (A-BC)-DE</th>
<th>b\textsubscript{2}: AB-(CD-E)</th>
<th>c\textsubscript{2}: (AB-C)-DE</th>
<th>d\textsubscript{2}: AB-(C-DE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC - AB</td>
<td>AB - BC</td>
<td>AB - BC</td>
<td>AB - BC</td>
</tr>
<tr>
<td>BC - CD</td>
<td>AB - DE</td>
<td>AB - CD</td>
<td>AB - CD</td>
</tr>
<tr>
<td>DE - AB</td>
<td>CD - BC</td>
<td>DE - BC</td>
<td>DE - BC</td>
</tr>
<tr>
<td>DE - CD</td>
<td>CD - DE</td>
<td>DE - CD</td>
<td>DE - CD</td>
</tr>
</tbody>
</table>
result to greater performance by the ‘Cue Combination’ language learners (Average 79.4%) when compared with the ‘Statistical Regularities’ (61.5%) and ‘Semantics + Prosody’ (64.9%) learners, p < .001. The two single-cue conditions did not differ from each other, indicating that whether statistical regularities or semantics and prosody are present, learning is similarly moderate. One potential concern with this analysis is that it is not comparing equivalent sets of languages because there are four ‘Cue Combination’ languages (a₂, b₂, c₂, and d₂), but only two possible ‘Statistical Regularities’ and ‘Semantics and Prosody’ languages (a₂ and b₂). To control for this, the ANOVA and post-hoc testing were performed again without the c₂ and d₂ Cue Combination languages, and results were nearly identical (Cue Combination Average = 77.4%, post-hoc testing p < .001). Thus, it can be concluded that while learners are capable of acquiring the basic two-word phrase groupings from either statistical or
semantic and prosodic information, they learn much better when presented with both
cues simultaneously.

Another concern with the set of comparisons involved in this preference test is
the fact that it includes within “unsupported” bigrams those that are within the
expected three-word phrase. Given that the results of Experiment 1 showed evidence
of secondary learning in forming hierarchical structures, such scoring might result in a
deleterious effect on the ‘Statistical Regularities’ and ‘Semantics and Prosody’
language learners who use secondary learning to form three-word phrases. This could
cause these learners to acquire three-word groupings and then treat any bigram within
the larger group as equally good. To control for this potential problem, another
preference statistic was calculated, considering only the difference between cue-
supported two-word phrases versus bigrams that crossed the three-word phrase
boundary. The specific contrasts used in this ‘Narrow’ analysis are listed in Table 3.6,
and the results are depicted in Figure 3.4. Comparing these results with those in
Figure 3.3, above, it is quite obvious that the overall pattern of preferences is very
similar. On average, this ‘Narrow’ version of the analysis yielded preference scores
only 4% better (75.3%) than the prior, and more inclusive, analysis (71.3%). A Two-
way mixed design ANOVA revealed a significant main effect of Analysis type, F(1,72)

Table 3.6: Trial types for Narrow Two-word Phrase Grouping Preference tests:
Phrases supported by cues vs. phrases not supported by cues which cross the three-
word phrase boundary. Each test is listed as “Supported - Unsupported”.

<table>
<thead>
<tr>
<th>a₂: (A-BC)-DE</th>
<th>b₂: AB-(CD-E)</th>
<th>c₂: (AB-C)-DE</th>
<th>d₂: AB-(C-DE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC - CD</td>
<td>AB - BC</td>
<td>AB - CD</td>
<td>AB - BC</td>
</tr>
<tr>
<td>DE - CD</td>
<td>CD - BC</td>
<td>DE - CD</td>
<td>DE - BC</td>
</tr>
</tbody>
</table>
= 28.551, p < .001, and a main effect of Language, F(7,72) = 8.67, p < .001. These are qualified by a significant interaction between Analysis type and Language, F(7,72) = 2.37, p < .05. One-way Anovas within each Analysis type were both significant, Broad: F(7,72) = 7.90, p < .001; Narrow: F(7,72) = 8.27, p < .001. Language-by-language paired tests revealed that the locus of the improvement in test scores was isolated to languages CueCombine-a, CueCombine-b, and CueCombine-c, t(9)\textsubscript{a} = 3.70, t(9)\textsubscript{b} = 5.09, t(9)\textsubscript{c} = 2.76, p < .05, while all other languages were not significantly different on the two analyses. On the whole, these differences are quite small, and do not effect the conclusion that learners in the ‘Cue Combination’ languages perform far
better on two-word preference tests\textsuperscript{16}. No matter which analysis is explored, the presence of multiple cues does lead to greater learning of phrase preferences.

Because Experiment 1 revealed an Edge bias in hierarchical groupings, it is critical to explore whether the two-word grouping preferences in this experiment are subject to the same bias. This is achieved by separately analyzing bigrams at the edges of sentences (AB and/or DE) and two-word groups that are in the middle of sentences (BC or CD). For each language the trial types of the Narrowly defined two-word phrase grouping test were split into preference for the cue-supported Edge two-word group and preference for the cue supported non-edge two word group. While this produces an even split in language a\textsubscript{2} and b\textsubscript{2} across all cue types, as shown in Table 3.7, there are no non-edge cue-supported groups for language c\textsubscript{2} and d\textsubscript{2}, because the only cued phrases are the AB and DE edge phrases. Instead of simply leaving these languages out of the analysis entirely, their non-edge preferences were calculated as the percentage of time they chose the middle two-word group that was within the three-word group as opposed to the middle two-word groups that crosses the three-word group boundary. The results of these analyses can be seen in Figures 3.6 and 3.7.

<table>
<thead>
<tr>
<th>a\textsubscript{2}: (A-BC)-DE</th>
<th>b\textsubscript{2}: AB-(CD-E)</th>
<th>c\textsubscript{2}: (AB-C)-DE</th>
<th>d\textsubscript{2}: AB-(C-DE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge</td>
<td>DE - CD</td>
<td>AB - BC</td>
<td>AB - BC</td>
</tr>
<tr>
<td>Non-Edge</td>
<td>BC - CD</td>
<td>CD - BC</td>
<td>BC - CD</td>
</tr>
</tbody>
</table>

Table 3.7: Trial types for Edge and Non-Edge Two-word Phrase Grouping Preference tests: Phrases supported by cues vs. phrases not supported by cues which cross the three-word phrase boundary. Each test is listed as “Supported - Unsupported”.

\textsuperscript{16} In fact, given that Cue Combination languages were the only to show significant improvement with the more narrow analysis, the differences between cue types are even stronger in this second version.
The $c_2$, (AB-C)-DE, and $d_2$, AB-(C-DE), languages show their original learning in the Edge condition (87.3% and 79.4%, respectively) because the only phrases included in the original two-word phrase preference tests of these languages were edge phrases. In the non-edge test, involving the comparison of BC and CD, one of which is within the three-word phrase and the other of which crosses the three-word phrase boundary, the learners of $c_2$ and $d_2$ do not show any significant preference ($c_2$: 52.4%, $t(9) = .564$, $p > .05$; $d_2$: 45.9%, $t(9) = .851$, $p > .05$). This, however, is not necessarily indicative of poor learning, as neither of these pairs ought to be learned as two-word phrases. If the learners of these languages did properly learn the multi-tiered hierarchical structure, we would predict a lack of preference, as is seen here. These results further support the strong learning of these languages already shown by

Figure 3.6, 3.7: Edge and Non-Edge Two-word Phrase Grouping Preference test results. Phrases supported by cues vs. unsupported phrases that cross the phrase boundary.
previous analyses, but are quite different from the other languages in this experiment, so they will not be included in the subsequent analysis.

A 2x3 Two-way ANOVA comparing the Analysis type (Edge, Non-Edge) and Cues present (Statistical Regularities, Semantics and Prosody, Cue Combination) revealed significant main effects for both Analysis type, F(1,57) = 22.03, p < .001, and Cue, F(2,57) = 22.31, p < .001, but no interaction between them, F(2,57) = .901, p > .05. In order to explore the pattern of results within each of the analyses, One-way ANOVA's with Bonferroni post-hoc tests were performed on each of the analyses. The Edge test showed significant differences between cue types, F(2,57) = 8.00, p < .001, as did the Non-edge test, F(2,57) =13.58, p < .001. In each analysis, only the ‘Cue Combination’ learners performed significantly better than the ‘Statistical Regularities’ and ‘Semantics and Prosody’ learners, all p < .01. When separating the languages, however, a striking pattern emerges. While both the a\textsubscript{2} and b\textsubscript{2} languages for all three cue types are learned significantly better than chance in the Edge test, t\textsubscript{Stat-a\textsubscript{2}}(9) = 5.23, t\textsubscript{Stat-b\textsubscript{2}}(9) = 4.77, t\textsubscript{SemPro-a\textsubscript{2}}(9) = 14.08, t\textsubscript{SemPro-b\textsubscript{2}}(9) = 4.16, t\textsubscript{CueCombine-a\textsubscript{2}}(9) = 12.50, t\textsubscript{CueCombine-b\textsubscript{2}}(9) = 8.13, p < .01, the pattern is heterogeneous for the Non-Edge test. All three cue types perform significantly above chance for the Non-Edge b\textsubscript{2}, AB-(CD-E), languages, t\textsubscript{Stat-b\textsubscript{2}}(9) = 2.91, t\textsubscript{SemPro-b\textsubscript{2}}(9) = 2.81, t\textsubscript{CueCombine-b\textsubscript{2}}(9) = 6.99, p < .05. For the Non-Edge a\textsubscript{2}, (A-BC)-DE, languages, however, when only one cue is present learners fail to show a preference of BC over CD, t\textsubscript{Stat-a\textsubscript{2}}(9) = 0.15, t\textsubscript{SemPro-a\textsubscript{2}}(9) = 1.65, p > .05, but when both cues are present in combination, they succeed, t\textsubscript{CueCombine-a\textsubscript{2}}(9) = 8.65, p < .001. Thus, single cues permit learners in some languages, AB-(CD-E) specifically, to pass both the Edge and Non-Edge tests, but only Edge tests in the (A-BC)-DE
language. Cue combinations, on the other hand, produce greater learning of all languages, and successful learning of both Edge and Non-Edge two-word groups across different languages.

The two-word phrase grouping tests for this experiment have clearly shown that a combination of statistical information with semantic and prosodic cues is helpful in acquiring a hierarchical structure and overcoming biases. When presented with either statistical or semantic and prosodic information, learners did acquire the phrase structure preferences at greater than chance levels, and were strikingly similar to each other. This indicates that these cues individually function in similar ways for learners. When the cues are combined, however, there is an additive effect, producing excellent learning of hierarchical structure groupings even in the middle of sentences which cause trouble for the single-cue learners. It appears that the combination of cues is able to overcome structural biases in a way that single cues are not.

3.2.3 Three-Word Phrase Grouping Preference test

Each of the four grammars in Experiment 2 has a single three-word phrase that is supported by the cues or might be built via secondary learning as a combination of the two-word cue-supported phrase and a neighboring non-phrasal category. By testing learners' preferences for three-word phrases across the different cue types and languages, it will become apparent whether they have acquired a full multi-tiered hierarchical phrase structure or just a mono-stratal structure. In languages a₂, (A-BC)-DE, and c₂, (AB-C)-DE, the cues and secondary learning ought to lead to a preference for ABC as a three word phrase over the other possible combinations, BCD and CDE. In language b₂, AB-(CD-E), and d₂, AB-(C-DE), learners should acquire a
preference for CDE over all other alternatives. The results of the three-word phrase grouping preference test are shown in Figure 3.8. While all of the ‘Cue Combination’ language learners show strong learning for the appropriate three word phrase, $a_2 = 77.4\%$, $b_2 = 74.3\%$, $c_2 = 67.1\%$, $d_2 = 70.1\%$, all $t(9) \geq 4.81$, $p < .001$, the learners in the ‘Statistical Regularities’ and ‘Semantics and Prosody’ languages only show significant cue-consistent three-word phrase preferences in the $b_2$, AB-(CD-E) languages, Stat-$b_2 = 71.8\%$, SemPro-$b_2 = 74.3\%$, all $t(9) \geq 6.27$, $p < .001$, but not in the $a_2$, (A-BC)-DE, languages, Stat-$a_2 = 51.7\%$, SemPro-$a_2 = 59.5\%$, $p > .05$. Focusing only on languages $a_2$ and $b_2$, a 3x2 Two-way Cue type by Language ANOVA revealed a significant main effect of Cue type, $F(2,54) = 6.25$, $p < .01$, a significant main effect of Language, $F(1,54) = 10.45$, $p < .01$, and a significant Cue type by Language interaction, $F(2, 54) = 4.55$, $p < .05$. A One-way ANOVA comparing cue types for language $a_2$ was
significant, F(2,27) = 9.80, p < .001, with the locus of effect being due to better performance by ‘Cue Combination’ learners over both the ‘Statistical Regularities’, mean difference = 25.7%, p < .01, and ‘Semantics and Prosody’, mean difference = 18.1%, p < .05. No significant effect of Cue type was found in language b₂, F(2,27) = 0.14, p > .05.

To further explore why both of the single cue (A-BC)-DE languages failed to pass the three-word phrase grouping preference test, the tests were separated into preferences for ABC and preferences for CDE as seen in Figures 3.9 and 3.10. This separation reveals that the problem for the ‘Statistical Regularities’ and ‘Semantics and Prosody’ a₂ learners was two fold: They failed to form a preference for the cue-appropriate ABC group, and they instead formed a preference for the cue-inappropriate CDE group. The same problem did not occur for the single-cue b₂, AB-(CD-E) learners, as they properly learned a preference for the CDE group and not the...
ABC group. The ‘Cue Combination’ learners also show consistent cue-appropriate learning in both the $a_2$ and $b_2$ languages, preferring ABC and CDE respectively. There is also an interesting effect in the ‘Cue Combination’ $c_2$, (AB-C)-DE and $d_2$, AB-(C-DE), language groups, where there is a trend for the cue-appropriate group to be preferred, but both ABC and CDE are significant in both languages at greater than chance levels, representing a dispreference for BCD as a phrase. A direct comparison of ABC against CDE in these two languages does reveal a significant difference in the cue-appropriate direction, $t(18) = 2.20, p < .05$. This suggests that while the hierarchical structures of these languages were properly learned, the three word phrases were more weakly disambiguated than in the $a_2$ and $b_2$ languages because of the potential for an alternate statistically based parse.

Overall, performance on three-word phrase preference tests was very similar to performance on the two-word phrase preference tests. Learners exposed to languages with both statistical regularities and semantic and prosodic cues performed better than those learning languages with only one of these two cue types. Especially problematic in the three-word phrase tests was the singly-cued (A-BC)-DE language, which may be related to the difficulties that arose on the Non-Edge two-word tests described earlier in these same languages. Secondary learning also seems to show an interaction with cue use in forming three-word phrases, as evidenced by the weaker effects in the (AB-C)-DE and AB-(C-DE) languages as compared to the other languages.

3.3 Discussion

The results of Experiment 2 demonstrate that learners can acquire a multi-tiered hierarchical structure from exposure to a combination of cues in the input.
When these cues agree, they are sufficient to outweigh the structural biases that learners come to these tasks with. The fact that learners show preferences for three-word phrases as well as specific two-word phrases within these three-word groups indicates that they have not just learned a simple mono-stratal grouping, but a set of nested hierarchical structures. This result integrates quite well with the growing body of literature on multiple cue combinations in a variety of cognitive domains including phrase structure (Thompson and Newport, 2007), word segmentation (Christiansen, Allen & Seidenberg, 1998), vision (Jacobs, 2002), and multi-sensory perception (Driver & Spence, 2000). What is unique here, however, is the manifestation of a multi-tiered hierarchical mental representation as a result of this cue combination.

One conclusion to be drawn from the data here is that the Edge bias seen in Experiment 1 is quite consistent across languages. Looking at all of the languages cued either by statistical regularities alone or semantics and prosody alone in both experiments reveals a consistent preference among learners to treat two- and three-word groups that occur at the edges of sentences as more likely. This effect is not symmetric, leading learners to prefer both sentence initial and sentence final groups equally, but instead favors sentence final bigrams and trigrams. The fact that learners of the a2, (A-BC)-DE languages with only one cue to hierarchical structure pass the Edge two-word preference tests (DE vs. CD), but fail the Non-edge preference test (BC vs. CD), indicates that the success of these languages in the general two-word phrase grouping preference test was driven by learning the DE bigram that was supported by the statistical or semantic and prosodic cues and which also matched the End-Edge bias. Learners of the single-cued b2 languages, AB-(CD-E), did not show
the same split between the Edge (AB vs. BC) and Non-edge (CD vs. BC) tests, which means that their performance on the general two-word tests was not selectively driven by an overly strong preference for the Initial-Edge pair of AB. Instead, these learners show significant, but weak, learning of both the AB and CD phrases, seemingly unassisted by, and only partially hindered by, the End-Edge bias.

This bias is also present in the three-word phrase grouping tests for the singly-cued a$_2$, (A-BC)-DE languages. In these languages, if there is a consistent End-Edge bias, then learners will need strong enough cues to overcome a preference to treat both DE and CDE as a likely phrase. In the singly-cued a$_2$ languages, however, the cue points strongly to a pair of two-word groupings, while the three-word groupings need to be acquired from either secondary learning based on the position of the non-grouped category ‘A’ in the ‘Statistical Regularities’ condition, or by using a weak and unreliable cue as in the ‘Semantics and Prosody’ condition. Learners here do not seem to acquire a preference for this ABC group, and instead are guided by the End-Edge bias towards CDE. If learners are as attentive to rules of constituency, however, as they seemed to be in Experiment 1, this poses a dilemma: How can a CDE three word phrase exist if BC is one of the two word phrases, as this would involve breaking up an existing lower-tier constituent. One possible explanation that brings together the two-word and three-word phrase grouping preferences in these singly-cued a$_2$ languages, is that this dilemma is solved by participants by ignoring the constituency of the BC phrase, and it is this solution that drives the poor performance on the Non-Edge two-word tests. Put another way, learners of these languages seem to be parsing them as having the hierarchical structure of A-B-(C-DE).
This same problem does not hold for the cue-combination languages, as shown by the extremely strong performance of the ‘Cue Combination’ $a_2$ and $b_2$ learners on every single test run in this analysis. The $c_2$ and $d_2$ learners, exposed to both statistical regularities and semantic and prosodic cues, also showed thorough learning of two- and three-word groups, along with a tendency to prefer the cue appropriate complete phrase structure of the languages. This allows them to disambiguate the phrase structure of these languages in a way that learners of the $d_1$, AB-C-DE, language in Experiment 1 were not capable of. As neither cue type was sufficient on its own to overcome the End-Edge bias and produce full multi-tiered hierarchical learning, these cues must act in an additive fashion. Once these cues are combined, learners acquire the appropriate multi-tiered hierarchical structure with great flexibility and little influence from biases. The potential implication for natural language is that they must contain a set of reliable cues to hierarchical phrase structure for learners to be able to acquire them. This suggests that some of the universalities seen amongst human languages may be due not to an innate universal grammar, but instead to the fact that languages must satisfy a number of constraints to make cues available in order to be learnable. This same argument has been made by Morgan, Meier, and Newport (1987), in regards to the general domain of phrase grouping, but the results here make a clear case for the role of multiple combined cues for multi-tiered hierarchical phrase structures.

Taking Experiments 1 and 2 together, there are some important concerns about the design of the languages and the degree to which conclusions can be extended from these small miniature languages to natural language acquisition. One possible
explanation for learners showing an Edge bias throughout both experiments is that they are treating the sentences like a set of word lists and therefore are prone to seek out patterns in the serial order and serial position of words and categories. This is made all the more likely by the fact that all of the sentences are sampled from a maximum length sentence of five words always occurring in fixed positions within the sentence frame. No language so far has included sentences where words, categories, and phrases occur in different location in the sentence, a potential cue to phrase structure, and a potential cue against positional information that has proven useful in some miniature language experiments (Thompson and Newport, 2007).

A second general concern is that these languages are all too simple, in the variety of sentence types, the hierarchical structures involved, and the relationships between words and categories. This could create overly easy learning of some parts of the language, resulting in ceiling effects, while perhaps not creating enough of a varied stimulus to drive learning of other components of the language. Both the statistical regularities and semantic visual world groupings, while effective for learning in these languages, are not adaptable enough to function in more complicated languages. It is not clear that these relationships would be helpful enough to allow learning of more natural and elaborate linguistic systems. Experiment 3 is designed to address these concerns.
Chapter 4. Experiment 3

In the previous experiments, learners were able to successfully acquire a multi-tiered hierarchical structure from a combination of statistical, semantic, and prosodic cues. While learning is significant with less information in these languages, learners do not seem to form a full multi-tiered hierarchical structure without the combined cues. The languages used so far, however, are extremely simple, relying on only a single fixed sequence of 5 word categories as the maximal sentence. The conclusions reached about the importance of the combination of cues may, paradoxically, not in fact hold for more complicated languages. Throughout all of the analyses, care was taken to examine and control for an apparent preference to form groups of the bigrams and trigrams which occurred at the edges of sentences. This is especially problematic in the earlier experiments because there is a great deal of consistency in these edge groups due to the simplicity of the languages. In all of the languages, with the exception of the $c_1$, A-BCD-E language, which was poorly learned, at least one bigram and one trigram was a consistent edge group. This could have enhanced learning and provided an easy anchor from which learners could start forming a hierarchical structure.

The fixed word order of sentences in these languages also results in the same categories consistently occurring as neighbors. For instance, in the prior experiments, if an $A$ category word is present and followed by another word in a sentence, the subsequent word must be a member of the $B$ category. This is quite unlike natural languages, where lexical categories are preceded and followed by a variety of other categories, depending on the structure of the current sentential context. Since the only
hierarchical relationships in these languages occur between immediately neighboring categories, the only logical statistical regularities that could be calculated were based on these neighbors, what we will refer to as ‘Adjacent φ’. The other possible statistical relationships, for instance the predictive relationship between category A and category D, would be an instance of a ‘Non-Adjacent φ’. This was illogical, however, because these categories could not be hierarchically related. Prior research (Newport and Aslin, 2004) has shown that some non-adjacent, or non-local, statistical relationships are in fact learnable as long as the non-adjacencies occur between naturally related categories (such as Consonants or Vowels).

The fixed word order of the languages in Experiment 1 and 2 also led to using the semantic and prosodic cues to simply mark groupings. Since hierarchical structures necessarily involved sets of immediately neighboring categories, the simplest semantic relationship was an analogous category-location clustering. Similarly, prosodic groupings presented phrasally related and neighboring sets of categories as an auditory cluster. While there does not seem to be an obvious prosodic cue to non-adjacent categories\textsuperscript{17}, semantic cues can come in a number of different forms. In natural languages, hierarchically related words do not generally describe a set of objects simply co-occurring near each other in the world (with the exception of some coordinate structures). Instead, most express semantic dependency relationships, where a category will fill a required or optional role with respect to a hierarchically related word or act as a modifier of a hierarchical relative. These relationships are

\textsuperscript{17} While not prosodic, some morpho-phonological cues such as agreement may fill the role of auditorily marking hierarchically related non-adjacent categories.
asymmetric and complex, but form the backbone of a large number of syntactic theories as discussed in section 1.4.3.

Morgan and Newport (1981) tested whether the presence of a more complex and inherent dependency semantics would aid learners in acquiring a phrase structure grammar from a miniature language. Contrasting this with a simple constituent grouping, much like the semantic cue used here in Experiment 2, revealed little difference in learning. The authors took this to mean that semantics was providing a critically additional cue to phrase structure, but that the form of the semantics was unimportant to constituency and hierarchy learning as long as it supplies this additional cue. While the grammar of the language in that study was more complex than those in Experiment 1 and 2, it lacked a multi-tiered hierarchical structure. Considering the evidence from the Cue Combination languages in Experiment 2, it seems evident that this non-dependency constituent grouping semantics is also sufficient in combination with predictive dependencies to induce a multi-tiered hierarchical structure, at least in a very simple language.

The present experiment seeks to answer a number of important questions raised by previous work and the previous experiments reported here. First, is a dependency semantics more helpful to learners when the language they are exposed to is more complicated and involves multi-tiered hierarchical structures? Second, are learners capable of acquiring the hierarchical structures of languages with occasionally non-adjacent statistical regularities among word categories that are sometimes non-adjacent, as they are for languages with only adjacent regularities, and what cues are necessary for this acquisition? Third, how do these different forms of semantics
interact with adjacent and non-adjacent statistical regularities? Finally, are learners capable of acquiring a multi-tiered hierarchical structure that does not depend on phrase groupings consistently occurring at the edges of sentences? Four new miniature languages were devised to attempt to answer these questions. All of the languages share a common vocabulary, but vary on two dimensions: whether the statistical regularities hold between adjacent or non-adjacent words and categories, and whether the semantic reference world contains a constituency grouping cue or dependency and modification relationships. This combination permits us to measure the influence of adjacency in statistical regularities, the effect of different semantic mappings, and the interaction between these cues as learners attempt to acquire a more complex multi-tiered hierarchical structure. The languages are shown in Figure 4.1.

4.1 Methods

4.1.1 Participants

Forty adult monolingual English speaking undergraduate students participated in the experiment. Each participant was randomly assigned to one of four language conditions. All participants gave informed consent prior to participating and were paid for their participation upon completion of the experiment.

4.1.2 Description of the linguistic systems

Vocabulary

The same fifteen word forms were used in this language as in Experiments 1 and 2. There were seven lexical categories, however, instead of five, resulting in a different distribution of word-category matching. The new lexicon is shown in Table 4.1.
Figure 4.1: Languages for Experiment 3

<table>
<thead>
<tr>
<th>Constituent Grouping Semantics</th>
<th>Non-Adjacent Statistical Regularities</th>
<th>Adjacent Statistical Regularities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BP</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>CP</td>
</tr>
<tr>
<td></td>
<td>DP</td>
<td>DP</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependency Semantics</td>
<td>BP</td>
<td>BP</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>CP</td>
</tr>
<tr>
<td></td>
<td>DP</td>
<td>DP</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Grammar and Statistics:

Three sentence types exist in these languages, each containing 5 words. All sentences start with an A word and contain a pair of two-word sets, \{B, C\}, \{D, E\}, and \{F, G\}. There are two grammars in this experiment, one which contains Adjacent Statistical Regularities, and the other which contains partially Non-Adjacent Statistical Regularities. The order and placement of these sets in the sentences is determined by the following phrase structure grammars:

<table>
<thead>
<tr>
<th>Non-Adjacent Statistical Regularities</th>
<th>Adjacent Statistical Regularities</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → A {BP, DP, FP}</td>
<td>S → A {BP, DP, FP}</td>
</tr>
<tr>
<td>BP → CP B</td>
<td>BP → B CP</td>
</tr>
<tr>
<td>CP → C (DP)</td>
<td>CP → C (DP)</td>
</tr>
<tr>
<td>DP → EP D</td>
<td>DP → D EP</td>
</tr>
<tr>
<td>EP → E (FP)</td>
<td>EP → E (FP)</td>
</tr>
<tr>
<td>FP → GP F</td>
<td>FP → F GP</td>
</tr>
<tr>
<td>GP → G (BP)</td>
<td>GP → G (BP)</td>
</tr>
</tbody>
</table>

What is most striking about these grammars is their similarity; the only differences are marked in italics. Put simply, the difference between the languages is that the ‘Adjacent’ languages involves a consistent Head-Complement order with heads always preceding their complement, while the ‘Non-Adjacent’ languages have an inconsistent (across different phrases) Head-Complement order. This inconsistency creates linear
discontinuities, as in the sentence A-C-E-D-B. Because the BP requires a CP to the left of B, and the CP takes a DP to its right, this places the DP, D-E, in between the hierarchically related B and C. There is a restriction on these grammars, however, that all sentences must be 5 words long, and this creates the sets of three sentences seen in Figure 4.1, above. The different sentence types can be described as containing an A word followed by any two out of the three possible two-word phrases: BC, DE, and FG. If the sentences are thought of as sets of categories, where order does not matter, then the grammars for all the languages are identical. What distinguishes the grammars is the linear order in which categories and two-word phrases occur.

In order to calculate the statistical relationships between categories across both languages, the notion of $\phi$ has to be taken to mean the likelihood of co-occurrence within an entire sentence, and not just the likelihood of immediately neighboring co-presence\textsuperscript{18}. This can be supplemented by an adjacency statistic, which simply represents the percentage of the time that co-occurring pairs are immediately adjacent to each other. Calculations of all possible pairwise $\phi$ statistics along with their adjacency percentages can be found in Table 4.2. This table is organized into groups of possible bigrams: Group 1 represents hierarchically related categories that must occur together, Group 2 are hierarchically related categories which optionally occur together, Group 3 and 4 are both sets of categories which are never immediately hierarchically related, being separated by either one hierarchical phrase (Group 3), or

\textsuperscript{18} The reason for this limit is that $\phi$ requires that all possible sentences are counted only once. If we only compute $\phi$ for immediately adjacent pairs only, and a pair occurs non-adjacent, as in the relationship between A and B and ACB, then it is impossible to count this as an occurrence of A and not B, and also B and not A, but it is also unjustified to choose between these two options as it creates an artificial asymmetry in the statistic.
two (Group 4). What is apparent from this calculation is that the statistical regularities between categories, $\phi$, is identical between the two languages, but that the adjacency is dramatically different between them. In other words, the sentences share the same set of predictiveness relationships, but differ on whether these predictiveness relationships hold always between just immediately neighboring words or rather between words that are can be either adjacent or more distant relatives depending on the context. These relationships are also represented graphically in Figure 4.2. It is obvious here that the hierarchically related categories, those in Group 1 and 2, form a natural class in the Adjacent Statistical Regularities languages, but do not form such a natural cluster in the Non-Adjacent Statistical Regularities language. Adjacency is also a poorer cue in the Non-Adjacent languages to distinguish different bigram groups due to a reduced range of Adjacency percentages (0-50%).

**Semantic Reference World**

There are two types of semantic reference worlds used in these languages, each with both the Adjacent and Non-Adjacent Statistical Regularity grammars, and there

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\phi$</td>
<td>% Adjacent</td>
</tr>
<tr>
<td>1</td>
<td>BC, DE, FG</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>2</td>
<td>CD, EF, GB</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>3</td>
<td>BD, BF, CE, CG, DF, EG</td>
<td>0</td>
<td>50%</td>
</tr>
<tr>
<td>4</td>
<td>BE, CF, DG</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>
is no variation in the visual reference world between the two grammars. In the Constituent Grouping Semantics languages, each category picks out a location on the screen and each word determines the shape that appears in its category-appropriate location. Hierarchical relationships are reflected by visual location groupings on the screen, much as they were in Experiment 2, and category members form natural visual classes. Examples of the three sentence types can be seen in Figure 4.3.

Figure 4.2: $\phi$ statistics and adjacency percentage for all possible category bigrams.

Figure 4.3: Semantic Reference World examples for Constituent Grouping Semantics. Sentences represented from left to right are ABCDE, ADEFG, and AFGBC.
In the Dependency Semantics languages, each sentence is accompanied by a brief animation, and categories represent more natural semantic functions such as Nouns, Verbs, Adjectives, and Determiners. The approximate natural language categories and the meanings of individual lexical items are listed in Table 4.3, and a sequence from one sentence’s animation is shown in Figure 4.4. Each sentence describes a smiley-faced character moving from an initial position on the left of the screen in some manner (identified by the category A word) to a specific object or location on the right of the screen (identified by categories B-G). Just as natural language sentences do not describe everything in the world at a given moment or everything that existed around an event, included in these visual reference worlds is a set of visual shapes in varying locations that are not described in the sentences. Because categories in this language identify different features of the visual world, they are in a sense less confusable, and the lexical learning task ought to be made easier by

Table 4.3: Semantic reference word-to-world mappings for Dependency Semantics

<table>
<thead>
<tr>
<th>Category</th>
<th>Part of Speech</th>
<th>Lexical Item</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Verb</td>
<td>siz</td>
<td>Walk</td>
</tr>
<tr>
<td></td>
<td></td>
<td>heta</td>
<td>Bounce</td>
</tr>
<tr>
<td></td>
<td></td>
<td>folub</td>
<td>Teleport</td>
</tr>
<tr>
<td>B</td>
<td>Determiner (count)</td>
<td>mun</td>
<td>One</td>
</tr>
<tr>
<td></td>
<td></td>
<td>zami</td>
<td>Two</td>
</tr>
<tr>
<td>C</td>
<td>Noun (shape)</td>
<td>gesor</td>
<td>Square</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tav</td>
<td>Circle</td>
</tr>
<tr>
<td>D</td>
<td>Preposition (Location)</td>
<td>vigo</td>
<td>Left-of</td>
</tr>
<tr>
<td></td>
<td></td>
<td>kufes</td>
<td>Right-of</td>
</tr>
<tr>
<td>E</td>
<td>Noun (object)</td>
<td>lod</td>
<td>House</td>
</tr>
<tr>
<td></td>
<td></td>
<td>nehu</td>
<td>Tree</td>
</tr>
<tr>
<td>F</td>
<td>Adjective (Intensity)</td>
<td>pabif</td>
<td>Dark</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rup</td>
<td>Light</td>
</tr>
<tr>
<td>G</td>
<td>Adjective (hue)</td>
<td>boke</td>
<td>Blue</td>
</tr>
<tr>
<td></td>
<td></td>
<td>diral</td>
<td>Red</td>
</tr>
</tbody>
</table>
The addition of unmentioned visual shapes provides a supplemental challenge to equate the two semantic systems.

The three sentence types, ABCDE (or ACEDB), ADEFG (or AEGFD), and AFGB (or AGCBF), involve specific combinations of the semantic categories based on their hierarchical structure. The ABCDE sentence, such as “beta zami gesor vigo nebu”, shown in Figure 4.4, has the gloss “Bounce two squares left-of tree”, or in other words, ‘(Smiley) bounces to the two squares which are to the left of the tree’. An ADEFG sentence, such as “siz kufu lod pabif boke”, translates to “walk right-of house

---

19 Pilot testing of these languages for lexical word learning showed strong evidence of this.
dark blue and would be paired with a visual scene in which the Smiley-faced character walked to the right of the house which was dark blue. Finally an AFGBC sentence, “folub rap diral mun tav”, meaning “teleport light red one circle”, would occur with a scene of the character turning a light red color and then teleporting (disappears in a puff of smoke near the left of the scene and reappears moments later at the right of the scene) to a single circle. What is unique about this semantic system is that the location in the sentence of a given word changes its exact function in a very natural way, modifying its hierarchical neighbors instead of expressing a fixed and context independent meaning.

Prosody and Auditory materials

In all of these languages, each word in the sentence is pronounced separately and then the words are concatenated together in the same method used in Experiment 1 and the non-prosodically cued languages in Experiment 2. The reason for this was to provide consistency across the Adjacent and Non-Adjacent grammars. In the Adjacent Statistical Regularities languages, a prosodic clustering could have been used, similar to Experiment 2, but the same prosodic clustering could not be used in the Non-Adjacent languages. Because hierarchically neighboring categories, such as $B$ and $C$, sometimes occur quite distantly from each other, any prosodic grouping would necessarily include any intervening words. Instead of having separate prosodic systems for these two language types, which would potentially confound the results, a list-like pronunciation was utilized for all four languages.
4.1.3 Procedure

The general organization of exposure and testing was identical to Experiments 1 and 2. New test items were devised for these languages because of the new grammars.

Sentential Word Order Test

Tests for knowledge of correct sentential word order were very similar to those used in prior experiments, but with some minor modifications for the new grammars. A list of the sentence types used can be found in Table 4.4

Two-word Phrase Grouping Preference test

The two-word phrase grouping preference tests are split into three types for this experiment. First, in the “A tests”, we compared learners’ preferences for bigrams that presented the A category followed by the categories that occur immediately after it in the Adjacent Statistical Regularity languages (B, D, F) versus those that follow it in the Non-Adjacent languages (C, E, G). Second, in the “Phi tests”, bigrams that had a

Table 4.4: Sentential Word Order test types. Violations of correct word order are bolded.
predictive dependency between them, $\phi = 1$, or ‘Group 1’ from Table 4.2 (BC, DE, FG), were compared to the most similar possible bigrams that did not contain a predictive relationship at all, $\phi = 0$, from ‘Group 3’ (BD, DF, FB). Finally, in the “Hierarchy without statistics tests,” bigrams of categories that share a close hierarchical relationship but no statistical regularity (CD, EF, GB), “Group 2”, were compared to matches also drawn from ‘Group 3’ (CE, EG, GC). All test items were arranged so that the words within the bigrams occurred in the same order as in full sentences involving both categories. Thus, in the Adjacent languages, where $B$ preceded $C$, a test item using them would involve the bigram $BC$, but in the Non-Adjacent languages, where $B$ follows $C$, the same bigram would appear as $CB$. A full list of the test items is shown in Table 4.5.

Table 4.5: Two-word Phrase Grouping Preference test types.

<table>
<thead>
<tr>
<th>Non-Adjacent Statistical Regularities</th>
<th>Adjacent Statistical Regularities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Type</strong></td>
<td><strong>Test Type</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Test Items</strong></td>
</tr>
<tr>
<td>A test</td>
<td>AB - AC</td>
</tr>
<tr>
<td></td>
<td>AD - AE</td>
</tr>
<tr>
<td></td>
<td>AF - AG</td>
</tr>
<tr>
<td>Phi test</td>
<td>CB - DB</td>
</tr>
<tr>
<td></td>
<td>ED - FD</td>
</tr>
<tr>
<td></td>
<td>GF - BF</td>
</tr>
<tr>
<td>Hierarchy w/o Statistics</td>
<td>DE - EC</td>
</tr>
<tr>
<td></td>
<td>FE - GE</td>
</tr>
<tr>
<td></td>
<td>BG - CG</td>
</tr>
</tbody>
</table>
Three-word Phrase Grouping Preference test

Three possible types of three-word phrase groupings were generated and contrasted to determine learner’s preferences. The first set of these were three-word groups of a single category followed by a two word phrase, with a $\phi = 1$ relationship in it, such as CDE. We will refer to these as “Phi final” trigrams. The second set was the combination of a two-word phrase that contained a $\phi = 1$ relationship with the next category that this phrase took optionally, as in BCD; these will be called “Phi Initial”. Finally, to measure the preference of the Non-adjacent language learners for linearly occurring but not hierarchically related sequences, a set of trigrams that occurred directly in the input to these learners was also created and tested in all languages against the Phi final and Phi initial phrases; this set included groups such as BDE, and will be referred to as “Non-Adjacent Linear”. Word order was matched to the exposed language here, as it was for the two-word phrase grouping tests. A full list of the test items can be found in Table 4.6.

Table 4.6: Three-word Phrase Grouping Preference test types.

<table>
<thead>
<tr>
<th>Non-Adjacent Statistical Regularities</th>
<th>Adjacent Statistical Regularities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group type</strong></td>
<td><strong>Test Items</strong></td>
</tr>
<tr>
<td>Phi final</td>
<td>CED</td>
</tr>
<tr>
<td></td>
<td>EGF</td>
</tr>
<tr>
<td></td>
<td>GCB</td>
</tr>
<tr>
<td>Phi initial</td>
<td>CDB</td>
</tr>
<tr>
<td></td>
<td>EFD</td>
</tr>
<tr>
<td></td>
<td>GBF</td>
</tr>
<tr>
<td>Non-Adjacent Linear</td>
<td>BDE</td>
</tr>
<tr>
<td></td>
<td>DFG</td>
</tr>
<tr>
<td></td>
<td>FBC</td>
</tr>
</tbody>
</table>
4.2 Results

4.2.1 Sentential Word Order test

Learners of all four languages performed quite well on the test of sentential word order, averaging 82.3% correct. Results of this test for the four languages are shown in Figure 4.5. A Two-way between subjects ANOVA of the Statistical Regularities by type of Semantics revealed a nearly significant effect of Semantics, F(1,36) = 4.10, p = .05, with no significant effect of Statistical regularities, F(1,36)=.79, p > .05, and no Statistics by Semantics interaction, F(1,36) = 164, p > .05. Analyzing this effect of Semantics within each type of Syntactic regularities, however, revealed no significant differences. In the Non-Adjacent Statistical regularities languages, the Dependency Semantics learners performed slightly better, 84.2%, than the Constituent Grouping Semantic learners, 75.8%, but this difference was not significant, t(18)
Similarly, in the Adjacent Statistical regularities languages, the Dependency Semantics learners were better, 90.83%, than the Constituent Grouping Semantics learners, 78.34%, but this was also not significant, t(18) = 1.95, p > .05.

Overall, there is a trend for the learners of the Dependency Semantics languages to perform better on the test of sentential word order, but this is a small effect, and all learners acquired the word order at much greater than chance levels, t(9) = 4.49 to 9.71, p < .01.

### 4.2.2 Two-Word Phrase Grouping Preference test

In the first test of two-word phrase grouping, the ‘A test’, learners were asked to choose between a pair of bigrams involving an A category word followed by an immediately subsequent category word, or an A category word followed by the next category after that within a statistically reliable phrase (for instance AB vs. AC).

Scoring on this test was based on which of the bigrams occurred within the language, so correct answers were AB, AD, and AF in the Adjacent Statistical Regularities languages, and AC, AE, and AG in the Non-Adjacent Statistical Regularities languages; while each of these sets were the incorrect answers in the other condition. This test may be measuring a hierarchical relationship between the A category and either the categories \{B,D,F\} or \{C,E,G\}, but it may also be a measure of sensitivity to just linear order. In Figure 4.6, it can be seen that all groups tended to prefer linear order appropriate bigrams, and all but the Non-Adjacent/Constituent Grouping learners were significantly better than chance on this task, Non-Adjacent/Constituent Grouping = 59.17%, t(9) = 1.19, p > .05, Non-Adjacent/Dependency = 75.8%, t(9) = 3.27, p < .01, Adjacent/Constituent Grouping = 65.83%, t(9) = 2.35, p < .05, Adjacent/Dependency =
78.33%, t(9) = 4.63, p < .001. A Two-Way ANOVA of Statistics and Semantics revealed a marginal effect of Semantics, F(1,36) = 4.007, p = .053, but no effect of Statistics and no Interaction, F_{Stat}(1,36) = .396, F_{Interaction}(1,36) = .082, p > .05. As in the Sentential Word Order test, languages with Dependency Semantics are learned slightly better than those with Constituent Grouping Semantics, but this is a small overall effect. In general, learners prefer the local bigrams involving the A category that occur in the exposure set of the language over those that do not. The fact that there is no consistent preference for \{AB, AD, AF\} or \{AC, AE, AG\} suggests that learners are not making a choice based on hierarchical groupings, but instead based on linear order.
The second, and more important test of two-word phrase group learning involves the ‘Phi test’, in which bigrams with $\phi = 1$ were contrasted against alternatives that had $\phi = 0$. In the Adjacent Statistical Regularities languages, the $\phi = 1$ bigrams also had a locality score of 100%, while the contrasting alternative had a locality score of 0%. In the Non-Adjacent language, however, both the $\phi=1$ and $\phi=0$ bigrams had a locality score of 50%, making locality a non-distinctive cue. In purely statistical terms, this means that there are two cues to the preferred bigram in the Adjacent languages (Phi and locality) but only one (Phi) in the Non-Adjacent languages. Both the Constituency Grouping Semantics and the Dependency Semantics also provided potential cues to the correct bigram, although of different sorts, either as visual neighbors or visual incorporation. The results of this test can be seen in Figure 4.7. The four language groups showed a great diversity of preference for the $\phi=1$ bigrams.

![Figure 4.7: Two-Word Phrase Grouping Preference test: “Phi test”](image-url)
with the Non-Adjacent/Constituent Grouping learners doing no better than chance at 52.50%, $t(9) = .42$, $p > .05$, while the other three groups all showed a consistent preference, Non-Adjacent/Dependency, 70.83%, $t(9) = 4.16$; Adjacent/Constituent Grouping, 84.17%, $t(9) = 6.08$; Adjacent/Dependency, 92.50%, $t(9) = 14.65$; all $p < .01$. A Two-way ANOVA revealed a main effect of Statistics, $F(1,36) = 28.27$, $p < .001$, and a main effect of Semantics, $F(1,36) = 7.07$, $p < .05$, but no interaction between the two, $F(1,36) = .993$, $p > .05$. This indicates that both Statistical Adjacency and Semantic Dependencies aid learners in acquiring the basic hierarchical groupings in these languages and interact in a linear additive fashion.

The final two-word test, ‘Hierarchy without Statistics’, pitted bigrams that were hierarchically related through an optional phrase structure rule but carried no $\phi$ relationship, against other bigrams with no hierarchical relationship or $\phi$ regularity. For instance, taking two of the phrase structure rules from the Adjacent languages:

\[
\begin{align*}
CP &\rightarrow C \ (DP) \\
DP &\rightarrow D \ EP
\end{align*}
\]

the appropriate contrast for this test would be a $CD$ bigram against a $CE$ bigram. In the Non-Adjacent language, both bigrams have $\phi = 0$, and a 0% locality, while in the Adjacent languages, $\phi = 0$ still, but the hierarchically related bigram has a locality score of 100%, while the contrast has a 0% locality. Thus if a difference is found between the Adjacent and Non-Adjacent languages on this test it would indicate that learners are sensitive to locality of co-occurrence rather than statistical reliability. This does not appear to be the case, however, given the results shown in Figure 4.8. None of the languages actually show a preference on this test: Non-Adjacent/Constituent
Grouping, 54.17%, $t(9) = .557$, $p > .05$; Non-Adjacent/Dependency, 50.83%, $t(9) = .097$, $p > .05$; Adjacent/Constituent Grouping, 46.91%, $t(9) = .73$, $p > .05$; Adjacent/Dependency, 59.17%, $t(9) = 1.41$, $p > .05$. A Two-way ANOVA also, unsurprisingly, revealed no effect for Statistics, $F(1,36) = .421$, Semantics, $F(1,36) = .006$, or their interaction, $F(1,36) = 1.28$, all $p > .05$, so learners universally fail to form a preference in this test.

The results of the two-word phrase preference tests reveal that learners are sensitive to statistical and semantic cues to hierarchical structure and that these two cues seem to combine in an additive fashion. Additionally, simple locality relationships of bigrams, no matter what type of semantics is present, does not appear sufficient to result in learning of two-word groups that have a low $\phi$ relationship.
4.2.3 Three-word Phrase Grouping Preference test

The three-word phrase grouping tests involve three types of choices that are compared in all possible combinations. The ‘Phi final’ trigrams are sequences of 3 words that form a natural hierarchical constituent consisting of a word and its optional complement two-word phrase. For example, an EFG group such as “nehu rup diral” is an instance of an EP with a hierarchical structure:

\[ EP[\text{nehu } FP[rup GP[\text{diral}]]] \]

In the Dependency Semantics languages, this would translate to “tree (that is) light red”, a natural semantic group. The ‘Phi initial’ trigrams form a high likelihood group, but break up a constituent in a way that violates traditional rules of grammars and semantics. For example, a DEF group such as “kufes nehu rup” is contained within a DP, but DP also requires a GP as a complement of F, and nothing is present to fill this role:

\[ DP[kufes EP[\text{nehu } FP[rup GP[\emptyset]]]] \]

In the Dependency Semantics languages, this would translate to “Right-of tree (that is) light”, not a complete constituent. The final group, called “Non-Adjacent Linear” because it involves three categories that make up a sequence in the Non-Adjacent languages, fully violates the hierarchical constituencies because it skips a category that is necessary within the structure. It is only a possible three-word group in the Non-Adjacent languages, where it is a likely local sequence of words that do not form a hierarchical constituent. For instance, a DFG group such as “kufes rup diral” would be closest to the phrase structure:

\[ DP[kufes \text{EP}[\emptyset \ FP[rup GP[\text{diral}]]]] \]
There are two natural ways to present the results of these trials. First we will examine learners’ preferences for a given type of three-word group over the other possibilities averaged together, and then we will explore the individual contrasts between specific pairs of trigram types. The results of the three-word phrase grouping preference for the ‘Phi final’ trigrams is shown in Figure 4.9a, ‘Phi initial’ in Figure 4.9b, and ‘Non-Adjacent Linear’ in Figure 4.9c. What is most striking from these results is the general lack of a preference for most of the three-word groups. With the exception of the Adjacent/Dependency language, all of the other language learners show no preference whatsoever for any type of three word phrase, (maximum across 9 tests: t(9)=.338, p>.05). For the Adjacent/Dependency language, there is a consistent and significant preference for Phi final groups, 72.50%, t(9) = 4.39, p < .01, and a significant preference against the Non-Adjacent Linear groups, 28.13%, t(9) = 3.70, p < .01, but no preference for Phi initial groups, 49.38%, t(9) = .287, p > .05. Because these tests are orthogonal, involving all possible combinations of pairwise contrasts, i.e. the Phi initial tests include Phi initial vs. Phi final and Phi initial vs. Non-Adjacent Linear, there is no mathematical way for there to be significant preferences in any specific comparison for the three languages that show null results above. It is possible, however to have different patterns of pairwise preferences for the Adjacent/Dependency language. Graphs of each of the three possible individual comparisons for all four languages are shown in Figure 4.10a-c, while all three contrasts for the Adjacent/Dependency language are collectively shown alone in 4.10d for direct comparison. What is of greatest importance here is that the learners of the Adjacent/Dependency language have actually formed what appears to be a hierarchy of preferences for the three-word
groupings, preferring Phi final over Phi initial, 72.50%, $t(9) = 5.014$, $p < .01$, and over Non-Adjacent Linear, 72.50%, $t(9) = 3.139$, $p < .05$, but also preferring Phi initial over Non-Adjacent Linear, 71.25%, $t(9) = 3.597$, $p < .01$. This means that the reason for the lack of an effect in the collapsed Phi initial test for this language, shown in Figure 4.9b, was due to it balancing a positive preference for Phi initial over Non-Adjacent Linear.
with a negative preference when compared with \textit{Phi final}. Thus, the preference hierarchy learners seem to be forming in this language is \textit{Phi final} > \textit{Phi initial} > \textit{Non-Adjacent Linear}. 

Figure 4.10: Three-Word Phrase Grouping Preference tests: (a) Phi final > Phi initial, (b) Phi final > Non-Adjacent Linear, (c) Phi initial > Non-adjacent Linear, (d) Tests for Adjacent/Dependency language.
4.3 Discussion

The previous two experiments revealed that learners are capable of acquiring a hierarchical grouping organization from statistical regularities or semantic constituency grouping, and a multi-tiered hierarchical structure from a combination of both cues. However, this was all within reasonably simple languages which had a single fixed linear word order and therefore very strong positional cues to phrase structure. The current experiment expanded on these findings by making the languages more complicated and more carefully varying the nature of the statistical and semantic cues. The language in this experiment on which learners performed the worst and failed to show significant hierarchical structure learning of any kind, Non-Adjacent Statistical Regularities with Constituency Grouping Semantics, actually contains the same sort of statistical $\phi = 1$ information and visual reference world as the best learned Cue Combination languages from Experiment 2. The only significant differences between these are the complexity of the grammar and the dissociation of the statistical predictive dependency relationship from a linear adjacency relationship. The vast difference between the successful results of the prior experiment and the failure of subjects in this experiment indicates that these are potentially quite important issues that were not accounted for by the previous investigations.

The first important conclusion to be drawn from this study is that, even in this more complicated language, with varied five-word long sentence types, learners were able to successfully acquire the basic word order of sentences. Furthermore, the sentential word order learning for different language groups is not highly correlated with hierarchical structure learning. This suggests that word order learning and
hierarchical structure learning are two separate but parallel aspects of acquiring a language. These two facets need not be entirely dissociated, however, as local regularities of word order do seem to influence hierarchical structure learning. Instead, the lack of correlation means that word order learning likely precedes hierarchical structure learning and that just learning the appropriate order of words in a sentence is not sufficient to discover the hierarchical structure.

The analyses of two- and three-word phrase grouping preferences provide evidence regarding which language learning groups acquired the hierarchical structure of the language they were exposed to. Results of the two-word tests focus on a basic mono-stratal grouping of words into binary phrases and point toward a cue addition model of Adjacency and Semantic Dependency. The failure of learners in the Non-Adjacent Statistical Regularities with Constituent Grouping Semantics language indicates that the presence of a distinct statistical cue, \( \phi \), combined with a reference world which contained visual grouping of hierarchically related words, was insufficient to induce hierarchical structure learning. In contrast, the success of learners on the Phi test who during exposure had either adjacency information or a dependency semantics in the reference world, suggests that it is possible to overcome this more difficult learning task with stronger cues to phrase groupings. The supremacy of learning two-word phrase groupings in the Adjacent Statistical Regularities with Dependency Semantics language group further suggests that these cues act separably, each providing potentially helpful but separate information for learners. Had this group learned only as well as the Adjacent/Constituency Grouping and Non-Adjacent/Dependency learners, it would have suggested an interaction between Semantics and
Statistical Adjacency, with each cue leading to the same learning and a confluence of
cues offering no better information than a single cue.

This $\Phi$ test is primarily a test of whether learners can choose two-word groups
that have a strong $\phi$ relationship, and this cue is constant across all of the languages.
Since the $\phi$ cue is identical across the four languages, only the differences in the type
of Semantic relationships and Locality of statistical relationships could explain the
main effects of Semantics and Adjacency on whether learners choose the bigrams with
a strong $\phi$ relationship. This conclusion, however, is complicated by the lack of
effects for Semantics and Adjacency cues in the Hierarchy without Statistics test, even
though the bigrams used in this test share similar semantic dependency relationships
and identical locality statistics to those in the $\Phi$ test. This poses the challenge of
explaining why Semantic and Locality cues have an effect for learning bigrams with a
statistical regularity between them but have no effect on bigrams that lack such a
relationship. The most obvious conclusion is that these cues draw learners’ attention
to potentially related bigrams, but learners then only treat them as phrases if a
statistical dependency relationship actually holds. It is not the cues alone that are
causing phrase formation for two-word groups; rather, it is the combination of the
cues with statistical regularities.

While the importance of Adjacency in learning hierarchical structure is perhaps
unsurprising and fits smoothly with previous results, it is important to note that the
Constituent Grouping semantics that seems to be reasonably unhelpful here is the very
same cue that was so powerful in combination with statistical regularities in
Experiment 2. One interpretation is that the key difference between the Constituency
Grouping Semantics and the Dependency Semantics is the symmetry, or asymmetry, of the relationship between categories. The Constituency Grouping Semantics is a symmetric cue: Words that are in an immediate and required hierarchical relationship occur next to each other, but neither takes any sort of precedent or dominant position over the other. The Dependency Semantics, on the other hand, is clearly asymmetric: One word requires another, or modifies another, but the relationship is always directional. In the Adjacent Statistical Regularities languages, both types of semantics permit successful two-word group learning, which is logical given that either a symmetric or asymmetric relationship could create a successful, although different, parse of these languages. A symmetric cue could lead learners to a phrase structure like \( AP[A_{BP}[BC]_{DP}[DE]] \), quite similar to some of the hierarchical structures seen in Experiment 2. An asymmetric cue, on the other hand, would likely lead to more multi-tiered and nested structures such as \( AP[A_{BP}[B_{CP}[C_{DP}[DE]]]] \). These clearly make different predictions about three-word groups, but are equally good at predicting two-word group performance.

In the Non-Adjacent Statistical Regularities languages, however, the different sorts of relationships create very different hierarchical structures. While the Dependency Semantics would create the same hierarchical structure, with only changes in the linear order of words in the structure, the Constituency Grouping Semantics, with a symmetric grouping would not be able to create a full parse for any of the sentences. The core of this problem is that the most embedded phrase in each sentence occurs between the words of the phrase that dominates it. Thus, in a sentence like ACEDB, there is no symmetric way to combine an ED phrase with C
and B, unless the three form a flat ternary structure C+ED+B. This would be quite unusual and still does not permit B and C to form a natural constituent on their own. This incompatibility between a symmetric constituency grouping approach and Non-Adjacent, or “infixed,” hierarchical structures could explain the failure of learners in the Non-Adjacent/Constituency Grouping language, even while this semantic cue is successful in languages with Adjacent relationships between hierarchically related words. It is not that a dependency semantics is a priori better than a constituency grouping semantics, or a priori a stronger cue to phrase structure; rather, the important fact is that they create different types of relationships between words and are therefore compatible with different types of hierarchical structures.

The results of the three-word phrase grouping tests address the question of multi-tiered hierarchies, as they are based on combinations of words that already have hierarchical two-word grouping relationships. Just as predicted in the discussion of the two-word phrase grouping tests above, learners exposed to a Constituency Grouping Semantics do not form preferences for three-word groups, regardless of the Adjacency of the language or their performance on the two-word grouping tests. This suggests that these learners are applying a symmetrical analysis to their two word groups, and due to the overall grammar of the language, are not able to form three word groups without breaking the constituency of these existing bigrams. Learners exposed to a Dependency Semantics, on the other hand, show a divide between those who experience an Adjacent Statistics and those who learn a Non-Adjacent Statistics, with only the former group succeeding in forming consistent preferences. The success of learners exposed to this type of asymmetric semantic relationship, including their
distinction between Phi-initial and Phi-final three-word groups, fits well with the
discussion of two-word phrase preferences, above. What is not explained is why the
Adjacency of these languages should matter at all, as the hierarchical structures are so
similar. One possible explanation is that the Non-Adjacent languages are simply
harder to learn, but this seems unlikely, as some of the three-word groups tested for,
specifically the Phi-final groups such as CDE (or CED) actually do occur as adjacent
groups consistently in the language, as in the ACEDB sentences. If adjacency itself
were important, then the learners of the Non-Adjacent/Dependency language should
show a strong preference for Phi-final three-word groups over Phi-initial groups, just
like the Adjacent/Dependency learners did, but the results do not support this.

If adjacency is not inherently harder for this particular test, then the source of
the problem must lie in the hierarchical structures learners are acquiring. An
interesting possibility, which could lead learners to show no preference for three-word
phrases, is if the Non-Adjacency languages actually were hierarchically ambiguous,
much like the languages of Experiment 1. Although the relationship between
hierarchically related words is asymmetric in the dependency semantics, creating
either a role filling or modifying relationship, it might be possible for the hierarchical
structure to reflect these asymmetries in multiple ways. For instance, a BC phrase
such as “mun gesor”, meaning “one square”, is currently modeled as having the
hierarchical structure \( BP[B CP[C]] \), like the natural language phrase structure
\( DP[det NP[Noun]] \). But, just as there has been debate over whether phrases such as
“one square” should be classified as a Determiner Phrases or as a Noun Phrases, it
might be possible for learners to treat a BC bigram as having the structure \( CP[BP[B]C] \).
In the Non-Adjacent Statistical Regularities language, this would create a second possible hierarchical structure with the same two word phrase groups, \{BC\}, \{DE\}, and \{FG\}, but with different three-word constituents, \{BDE\}, \{DFG\}, and \{FBC\}^{20}. Thus instead of the hierarchical structure seen in Figure 4.11a, learners might acquire, or at least consider, the hierarchical structure shown in 4.11b. In the Non-Adjacent/Dependency language, learners might find the hierarchical structure unambiguous for two-word phrase groupings, but extremely ambiguous for three-word constituents, leading to the pattern of excellent two-word phrase learning and no significant three-word preferences. The same is not true, however, for the Adjacent Statistical Regularities language, as an alternative order of the hierarchical relationships in the two-word phrases would not permit a full and complete phrase structure parse for these languages. The problem lies in the fact that in an ABCDE sentence, if the structure of the BC phrase was $\text{CP}[\text{BP}[B] C]$, then there is no way for the DE phrase to attach as a sister to B without creating a crossed dependencies situation (see Section 1.1.4 and Figure 1.2). Simply put, in the Adjacent languages, the asymmetric semantic dependencies create only one possible multi-tiered hierarchical structure, and learners’

Figure 4.11: Alternative multi-tiered hierarchical phrase structures for Non-Adjacent Statistical Regularities with Dependency Semantics language

(a) ![Diagram](a)    (b) ![Diagram](b)

---

^{20} These would be expressed as EDB, GFD, and CBF in the linear structure of the language.
preferences match this structure quite well. In the Non-Adjacent languages, the asymmetric semantic dependencies could create two possible multi-tiered hierarchical structures for each sentence, and learners’ lack of preferences match this ambiguity as well.

The results of this experiment show that the exact nature of the statistical and semantic cues, as well as the complexity of the language, all have influences on learners abilities to acquire the hierarchical structure from a brief amount of exposure. Small two-word groupings are possible in a variety of conditions as long as there are some strong cues to the hierarchical relationships. These cues act in an additive way to increase learner’s attention to places where statistical regularities might hold. Larger three-word groupings can also be acquired, but only when cues are consistent with each other and unambiguous about how the sentences might be parsed into full multi-tiered hierarchical structures.
Chapter 5. General Discussion and Conclusions

5.1 Plato’s Problem

This dissertation began with a discussion of Plato’s problem for language acquisition, specifically how learners come to know the hierarchical structures of syntax from an imperfect language environment. The evidence presented here suggests that the solution to this problem may not be a single explanation, but instead a dynamic interaction among a set of underlying biases, cue sensitivities, learning flexibilities, and a structured environment. The solution to Plato’s problem is not entirely built into the learner, nor is it built into the language, but it is the interface between these two which seems likely to solve the problem.

In the experiments presented here, learners revealed a set of biases that allow them to utilize information in a language environment and extract a structured representation. These include a bias to seek out structure and learn rapidly from regularities near the edges of sentences and a bias to represent sentences as unambiguous multi-tiered hierarchies that are compatible with the information available. Learners of many of the languages extended beyond the cue-based Primary information to develop preferences for groups that are compatible with a broad hierarchical structure across the sentences, what I have called Secondary learning, which must be based in a bias to seek out a cohesive structural analysis. Combined with the appropriate cues in the environment and the appropriate sensitivities to these cues, a bias to seek out structure could take a small amount of information and generate a complex hierarchy.
Since there are numerous potential cues to hierarchical structure in natural languages, we chose to control and minimize the availability of cues to test learners’ sensitivities. Cue selection was based on past research and large bodies of linguistic and psycholinguistic work, making it highly likely that these cues would be meaningful and support structure learning. Each cue that was chosen - Statistical Regularities between words, Adjacency, Prosodic Groupings, and two types of Semantic relationships - did prove helpful to learners in acquiring a hierarchical structure. This means that learners came into the task with a variety of cue sensitivities and enough flexibility to learn from the information that was available in the language they were presented with. Not all cues, however, were equally successful, as some led to only mono-stratal groupings, while other cues, individually and in combinations, led learners to multi-tiered hierarchy learning. This indicates that language learners are not just attending to cues using entirely separable systems, given the necessity for cue combinations, nor are they using a singular cue-attentional system, given the additive nature of some cues.

Since learners show a great number of cue sensitivities, it might be expected that all of these cues would be used to support a very limited set of hierarchical arrangements, in order to simplify the process of mapping from cues to structures. This does not seem to be the case, however, as learners were actually quite flexible in the type of hierarchical structure they were willing to posit in response to cues. When statistical information was strong but equal across three neighboring categories, learners treated these three as a phrase, but did not show a preference for any binary grouping within the phrase. When the same information pointed to only two
categories grouping together, learners’ grouping preferences reflected this and only expanded to three word groups when a neighboring group was not a member of any other phrase. Learners also showed flexibility in the location of phrases within the sentence, and in the final experiment, with the linear order of relationships between words. All of this flexibility suggests that learners, even adults who are native speakers of a strongly word order based language such as English, do not bring extremely strong biases about the shape of syntactic structures to the language acquisition task.

Finally, it is important to note that these biases, cue sensitivities, and flexibilities combine in intelligent and surprisingly complex ways. The nuanced results of the various experiments, especially Experiment 3, seem to suggest that learners are mixing their own inherent preferences with varied information in a way that looks very much like a careful syntactic analysis of a new language. What is most striking about this, though, is that learners do not seem to be conscious of their analysis at all, as numerous post-experimental interviews confirmed that participants tried to determine the meaning of individual words but did not consciously attempt to figure out the underlying structure of the language. The combining of information here is occurring unconsciously through implicit learning, revealing a powerful, yet normally hidden mechanism. The foundations of this mechanism appear to be cues in the environment which are filtered through a set of information processing biases to allow learners to extract hierarchical structure from a set of linear strings of words, potentially solving Plato’s problem, at least for miniature language learning.
5.2 Implications for Natural Language

While the studies in this dissertation are limited to miniature artificial languages learned over a short period of time in a controlled environment, the learning system underlying the results is likely to be what is used during natural language acquisition. If this assumption is true, then the results reported here may have far reaching implications for natural language acquisition, linguistic theories, and questions about the evolution of language abilities in humans. Most central to the implications of this work is the notion that multiple cues need to be present in the language environment in order to induce the acquisition of multi-tiered hierarchical representations. Fortunately, all natural languages appear to share the types of cues (statistical regularities amongst categories, prosodic phrases, and semantic dependencies) used in these experiments, creating a rich environment for learners.

If learning does rely on a combination of these cues with information processing biases on how to map them to hierarchical structures, then this will also create limits on the complexity of the multi-tiered hierarchical structures that learners are able to acquire. Theories of language structure should reflect these limits, building hierarchies that are no more complicated than can be supported by the information available to learners. Fortunately, many recent syntactic theories as diverse as Chomsky’s *Minimalist Program* (1995) and Culicover and Jackendoff’s *Simpler Syntax* (2005) argue strongly for just this sort of limitation on hierarchical structures. There are some specific aspects of language structure that are predicted by the cues and biases described in this dissertation, especially in the domain of language universals. As Greenberg (1963/1966) observed, languages nearly universally have consistent
ordering of modifying elements relative to modified elements, and this also patterns closely with the ordering of the Subject, Verb, and Object. This corresponds closely with the *Adjacent Statistical Regularities* grammar that was used in Experiment 3, as it too had a consistent modifier-modified order that the *Non-Adjacent Statistical Regularities* languages lacked. Given that learners seem especially sensitive to adjacent statistical regularities, relying on a consistent phrase structure ordering helps to increase the likelihood that words which are hierarchical neighbors will also be linear neighbors. In a grammar where heads consistently precede complements (a right branching structure) or consistently follow complements (a left branching structure), there is strong adjacency among words within the same constituent. On the other hand, in a language where complement-head ordering is inconsistent, there will be many situations in which a head and its complement may be linearly separated by another complement of the first complement (i.e. X and its complement Y can have Y’s complement Z in between them, [X [[Z] Y]]).\(^{21}\) Thus, an important universality of language typology, regarding consistency of head and complement position across different types of phrases in the same language, could simply be a derivative of a cue sensitivity to adjacent information.

Since multi-tiered, hierarchically organized, natural language is restricted only to humans, the approach and model described here may also be relevant to explaining the species specificity of this ability. While some species do show attention to cues such as adjacency and statistical regularities (Hauser et.al., 2001; Newport, et.al.,

\(^{21}\) This analysis does not address the role of specifiers or adjuncts, both of which might intervene between a head and the head of its complement. One possible solution is that the pressure on languages for Head-Complement Adjacency also puts pressure on specifiers and adjuncts to be moved to the beginning or ends of sentences, a common syntactic analysis.
2004), it is not clear that these cues induce other species to form hierarchical representations. Additional cues used here, such as semantic relationships, are untested in non-humans, so some of these cue sensitivities may be quite species specific. Even if further research reveals a general sensitivity to these cues, different possible systems for combining them may account for the vast species variability in language learning abilities. Furthermore, the biases seen throughout these experiments may be species unique. Finally, what might set us apart could be our willingness and ability to take in complex language input and represent it as a multi-tiered hierarchical structure. To date, what little work has been done exploring animal’s representations of hierarchies has generally revealed very simple organizational structures for family relationships (Seyfarth, et.al. 2005), or for the ability to manipulate objects (Johnson-Pynn, et.al., 1999). While the results of this dissertation do not solve the problem of the evolution of language in any direct way, they do offer another alternative for approaching the question, and suggest a variety of comparative studies that may prove useful for addressing this vexing issue.

5.3 Future Directions

This research is significant in that it represents one of the first attempts to carefully investigate multi-tiered hierarchical learning from several cues in controlled miniature language environments. While the results here support a clear model of cue combinations and hierarchical structure learning, they raise many questions for further research. Nearly every design decision made in these experiments could have taken a different form, creating many more languages and potentially testing a variety of other
theoretical issues. These include the types of hierarchical relationships, the form of the Statistical and Semantic cues, and the nature of the lexicon.

In each of the languages, the hierarchical structures used were reasonably simple, consisting of binary or ternary combinations that each functioned to form an entire phrase. There was no attempt to create different types of hierarchical relationships within a given phrase. Natural languages, however, utilize at least three potentially different hierarchical relationships, those between a Head and its Complements, Specifiers, and Adjuncts ($X^0$ attachment, X-bar attachment, and XP attachment). Since these varied hierarchical relationships sit at the core of much linguistic theory and create a great deal of multi-tiered hierarchical structure, they are a logical and necessary component for future work with miniature languages. One potential extension of the current work could be an attempt to ask whether these different hierarchical relationships interact with the Adjacency/Non-Adjacency findings from Experiment 3. Just as Newport and Aslin (2004) showed that Non-Adjacent regularities between a set of Consonants or a set of Vowels can assist learners in word segmentation, it may be possible for learners to accommodate Non-Adjacent relationships between a set of Heads with their Complements or a set of Specifiers in a miniature language hierarchy learning task.

While two types of semantic regularities were used in these languages, only one type of statistical regularity was controlled and studied here, that of $\phi$. As discussed earlier, $\phi$ is a symmetric measure of correlation which combines the notion of forward and backward transitional probability in a natural way without being bound by issues of word order or adjacency. While this was clearly helpful for clarity's sake, it also
masked some potential issues. There are a number of languages which might have a very strong $\phi$ cue and yet no natural hierarchical structure, as in the set: $\{AC, CA, ABC, ACB, CAB, CBA\}$. While the correlation between $A$ and $C$ might be prominent here, it may not lead to hierarchical structure learning because of the inconsistencies in word order between the two categories. Further study of grammars which distinguish different statistical cues to hierarchical structure may prove helpful. A second issue which relates the semantic cues to this issue of varying statistical regularities is whether different types of semantics cue learners to attend to different types of statistical relationships. While the symmetric Constituent Grouping semantics logically fits with a symmetric $\phi$ cue, the asymmetric Dependency Semantics may cue learners preferentially to encode forward or backward transitional probabilities, depending on the direction of the dependency. One way this might function is that when a modifier precedes the category it modifies, learners may seek out backwards transitional probabilities, to reflect the regularity with which the modifier is present given the presence of the modified word. A modified-modifier order, on the other hand, might cue learners to attend more strongly to forward transitional probability. Controlled studies which distinguish forward and backward transitional probabilities and pair them with varied orders of dependency semantics and constituent grouping may prove fruitful in further developing the analysis of the relationship between semantic and statistical cues to hierarchical structure.

The third component of these languages deserving of future study is the nature of the lexicon. In every language used here, words were entirely interchangeable within their respective categories, a property that simplifies statistical and semantic
relationships, but does not reflect natural language categories. In natural languages words have specific lexical properties such as subcategorization, as well as more category-general properties. These properties are often placed on an inheritance hierarchy, with some lexical items being more prototypic of their categories and others having unique restrictions. Subcategorization creates patterns in the statistical regularities between words and their semantic dependencies on a word-by-word basis, but across a corpus of sentences would also affect the regularities holding between categories. While such lexical variations seem likely to increase the difficulty for language learners, it is also possible that the particular variability and consistency across category members might act as additional information from which to extract hierarchical structure. There is a body of literature (Gomez, 2002; Gerken, 2006; Wonnacott, Newport & Tanenhaus, 2008) which has explored the acquisition of generalization and maintenance of individual patterns across categories in miniature languages, but none of these studies has explored the interaction between this property and multi-tiered hierarchical learning.

Finally, the work here has been done entirely with adults, and previous research has shown that children learn languages quite differently (Hudson Kam & Newport, 2005; Johnson & Newport, 1989). The present experiments, in their current form, require such a lengthy exposure period that they are unlikely to be viable if performed with young children. The underlying constructs, however, ought to be testable with modified and simplified methodologies. Along with this extension across age groups, it is of interest to ask whether this type of multi-tiered hierarchical learning also extends to other non-language domains, both in adults and children, as
hierarchical organization seems to be so universal across cognition and throughout development. Just as previous work on transitional probabilities has revealed commonalities across varied domains including sequences of tones, visual shapes, and motor planing (Saffran, et.al., 1999; Kirkham, et.al., 2002; Hunt & Aslin, 1998), subsequent work utilizing statistical regularities and other cues should help to reveal how learners acquire hierarchical representations in non-language domains. One difficulty for this extension is to determine what types of cues would be appropriate in different domains and how they would equate with the statistics, semantics, and prosody used here.

Clearly there are numerous avenues to pursue with this method of research, but the complexity, power, and universal application of multi-tiered hierarchical structure make all of these valuable potential contributions to understanding cognition.

5.4 Conclusions

This dissertation has presented evidence that multi-tiered hierarchical structures are learnable from a reliably constructed input language. This is the first test of multi-tiered hierarchy acquisition in a controlled environment, and elucidates a number of components of this process. First, learners acquire preferences for hierarchical groups in accordance with a variety of potential cues including statistical regularities between words, prosodic auditory groupings, and semantic relationships of visual grouping or visual dependency relationships. These cues are each influential on the learning of grouping preferences, but appear to interact with, and are sometimes overwhelmed by, information processing and structure building biases when the cues are only present individually. These biases do not appear to be pre-existing structural
preferences or an innate grammar system, but rather simple biases to group elements near the edges of sentences and to seek out hierarchical structural analyses. Learners also appear to utilize their biases to perform Secondary analyses of the language environment. This allows them to form preferences for potential hierarchical structure in addition to, but not in conflict with, those supported by the direct environmental cues. Finally, when multiple cues are present and congruent, learners achieve better acquisition of multi-tiered hierarchical structures and are better able to overcome their existing grouping biases. Overall, the results presented here have shown that learners come to the acquisition task with a variety of potent tools to extract information from the language environment and build complex representation of it that reach beyond the simple information that is available. This may be the core of how people solve Plato’s problem for hierarchical structures in language.
References


Cognition, 26, 269-286.

Language Learning and Development, 1, 51-195.


