ABSTRACT

Title of Document: BEYOND STATISTICAL LEARNING IN THE ACQUISITION OF PHRASE STRUCTURE

Eri Takahashi, Ph.D., 2009

Directed By: Associate Professor Jeffrey Lidz, Department of Linguistics

The notion that children use statistical distributions present in the input to acquire various aspects of linguistic knowledge has received considerable recent attention. But the roles of learner's initial state have been largely ignored in those studies. What remains unclear is the nature of learner's contribution. At least two possibilities exist.

One is that all that learners do is to collect and compile accurately predictive statistics from the data, and they do not have antecedently specified set of possible structures (Elman, et al. 1996; Tomasello 2000). On this view, outcome of the learning is solely based on the observed input distributions.

A second possibility is that learners use statistics to identify particular abstract syntactic representations (Miller & Chomsky 1963; Pinker 1984; Yang 2006). On this view, children have predetermined linguistic knowledge on possible structures and the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone.

This dissertation examines how the environment interacts with the structure of the learner, and proposes a linking between distributional approach and nativist approach to language acquisition. To investigate this more general question, we focus on how infants, adults and neural networks acquire the phrase structure of their target language.

This dissertation presents seven experiments, which show that adults and infants can project their generalizations to novel structures, while the Simple Recurrent Network fails. Moreover, it will be shown that learners' generalizations go beyond the stimuli, but those generalizations are constrained in the same ways that natural languages are constrained. This is compatible with the view that statistical learning interacts with inherent representational system, but incompatible with the view that statistical learning is the sole mechanism by which the existence of phrase structure is discovered.

This provides novel evidence that statistical learning interacts with innate constraints on possible representations, and that learners have a deductive power that goes beyond the input data. This suggests that statistical learning is used merely as a method for mapping the surface string to abstract representation, while innate knowledge specifies range of possible grammars and structures.

BEYOND STATISTICAL LEARNING IN THE ACQUISITION OF PHRASE STRUCTURE

By

Eri Takahashi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Associate Professor Jeffrey Lidz, Chair
Professor Norbert Hornstein
Associate Professor William Idsardi
Professor Howard Lasnik
Professor James Reggia

Dedication

This dissertation is dedicated to my grandmother, Kikuno Takahashi, who I've lived with all my life and who kept me company all those after-school afternoons so my only-child self wasn't lonely at all.

Acknowledgements

First and foremost, I would like to thank my advisor Jeff Lidz for his guidance and support over the last four years. Without his patience and encouragement, I wouldn't have been able to start my research as a language acquisitionist and complete this project. Thank you!

I also wish to thank all my committee members: Norbert Hornstein, Bill Idsardi, Howard Lasnik and Jim Reggia for taking time to read this dissertation and giving me insightful and helpful comments. Without them, I wouldn't have been able to get this far.

I am particularly grateful to people who have helped me with this research. Rebecca Baier for all her constant support in the infant lab. Without her, I wouldn't have known first thing about babies! Also, thank you to all the wonderful research assistants in the infant lab who are so helpful and hardworking. In particular, thanks to Shannon McDaniel for her help in recording the stimuli.

Tim Hunter and Brian Dillon, who helped me with computational side of this project, also deserve thanks. Without them, I wouldn't have been able to even start this project. Another special thanks to Bob Frank who helped me design the network simulations. Without him, there would be no Chapter 5.

I would also like to thank the Heiwa Nakajima Foundation for their generous scholarship that supported me for the first two years here.

Big thank you to Kathi Faulkingham, Rob Magee and Kim Kwok for all their help and support in all things administrative.

I really want to thank Peggy Antonisse and Tonia Bleam. They taught me how to teach and I really enjoyed teaching. Their passion made me want to become a teacher. Without them, I wouldn't have discovered my true passion for teaching.

Thank you to Ariane, Ellen, Maki, Rebecca and So-One for baking amazing "tree" cakes for my defense; Akira for the champagne; Maki for organizing the dinner; and Phil and Ellen for the flowers! Without them, my defense day wouldn't have been so great.

I would like to thank all my friends at Maryland. My classmates - Chizuru, Ellen, Phil, Rebecca, Stacey and Yuval - who stuck together throughout our five-year journey. Thanks to all my Japanese sempai and friends - Akira, Chizuru, Hajime, Maki, Masaya, Takuya, Tomo and Utako. My officemates - Chizuru, Rebecca and Tim. My unofficial officemates in 3416F: Ariane, Diogo, Jon, Masaya, So-One and Utako. And all other friends I've made here (you know who you are!) - without them, grad school wouldn't have been so rewarding.

And my parents who are my biggest support system and who always believed in me no matter what. Without them, I wouldn't have known to believe in myself.

And finally to Phil, my best friend and soulmate. Without you, I wouldn't be complete. Colon, dash, asterisk.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Syntactic knowledge that a child comes to acquire is highly complex. Syntactic rules are stated in abstract forms, in that they operate over variables (categories and phrases) rather than individual words. In order for a child to learn syntactic rules, it is necessary not only to generalize over categories but also to have a hierarchical structural representation. And yet, children acquire their native language in a relatively short period of time, without explicit teaching, irrespective of the language they are learning, and most of all, despite the fact that input data available to children seems imperfect and uninformative with respect to the complex syntactic patterns that are learned in the end (Chomsky 1975). For example, structure dependency of movement rules is argued to be unlearnable from the exposure to data (Chomsky 1975; Crain & Nakayama 1987; Crain 1991; Legate & Yang 2002). The acquired syntactic knowledge is a grammar that generates sentences that include the input data but also exceed them. That is, children's resultant grammar can produce sentences that were not necessarily learnable from exposure. Any approach to language acquisition needs to be able to account for this fact.

These observations led researchers in language acquisition to two different types of approaches to language acquisition. One (so-called nativist) approach is that children come equipped with the innate knowledge of possible structural representations (Chomsky 1959, 1975, 1980, 1981; Fodor 1966; Baker 1979; Pinker 1984; Crain 1991). It is proposed that much of the resultant syntactic knowledge is

not learned from the environment but is already built-in to the learner. This approach does not claim that everything is innate. The dimensions that are proposed to be innate are those that the input is uninformative for, e.g., formal categories (nounhood, verbhood), vocabulary of representations (subjects, complements, constituents) and constraints. This limits hypothesis space, which in turn restricts the number of possible interpretations of the input data. Under this view, learners look at the data and compare it against a class of possible models. One important consequence of this approach is it can account for the fact that children's grammar produces structures that they have never encountered before.

The other type is that the input contains sufficient statistical regularities that guide the learner to arrive at the abstract representations (Elman et al. 1996; Bybee 1998; Tomasello 2000). In this case, learners' task is to collect and compile accurately predictive statistics from the data, and learners would not have antecedently specified set of possible models. It is suggested "even in the total absence of [reliable] evidence, the stochastic information in data uncontroversially available to children is sufficient to allow for learning… [T]he correct generalization emerges from the statistical structure of the data" (Lewis & Elman 2002).

These two approaches are incompatible with each other as stated in the traditional terms. However, such nature/nurture distinction is a false dichotomy.

First, nature alone is not sufficient. It may be that the representational vocabulary needs to be inherently known, but that does not entail that the input data is irrelevant. For example, child might come with innate knowledge about formal categories, but it is not enough to know that "there exists verbs" to figure which word

in a given sentence is the verb. There must also be a mechanism for a child to find the verbs in the input (Fodor 1966; Pinker 1984; Grimshaw 1981; Chomsky 1981; Macnamara 1982). In this sense, the input is not trivial in determining how the surface strings map onto abstract representations. Innate knowledge about possible abstract representations helps restrict the hypothesis space, but learners still need a mechanism to identify which particular abstract representation best fits the surface form in any given sentence in the exposure language.

Second, nurture alone is not sufficient either. Simply showing that the input contains sufficient data for a child to learn a certain phenomenon and that the child is sensitive to them does not entail that there is no constraint from the learner on the set of possible representations. Additionally, there are countless number of distributions a learner can track in the input, and without a space of possible representations, it is impossible to determine which distributions are the relevant ones to build the representations from. Tracking distributions needs to feed into a decision process about representations, and having a predetermined space of possible representations helps that (Pinker 1984). Therefore, a learner must know in advance which statistical distribution they should pay attention to, and for what purpose.

In this way, both nature and nurture need each other. This dissertation proposes a linking between distributional approach and nativist approach to language acquisition. We need both nativism and distributional learning from the input, but the question is how the two interact, what is innate and what is learned from the input. The two might play different roles in language acquisition – innate knowledge specifies range of possible grammars and structures, while statistical learning is a

method for mapping the surface string to abstract representation. If statistical learning was the sole mechanism of acquiring a language, then it can only reproduce the statistical distribution of exposure sentences, but if the learner comes with innate constraints and knowledge on possible operations and structures, it can go beyond simply reproducing the exposure sentences. In any case, this dissertation is an attempt to examine how the environment interacts with the structure of the learner.

The notion that children use statistical distributions present in the input to acquire various aspects of linguistic knowledge has received considerable recent attention (Saffran, Aslin & Newport 1996a; Redington, Chater & Finch 1998; Maye & Gerken 2000; Gomez 2002; Maye, Werker & Gerken 2002; Mintz, Newport & Bever 2002; Mintz 2003; Swingley 2005; among others). In specific, it has been suggested that distributional information can play a role in the acquisition of phonemes (Maye, Werker & Gerken 2002; Maye & Gerken 2000), word segmentation (Saffran, et al. 1996a), word categories (Redington, et al. 1998; Mintz 2003), syntax-like regularities (Gomez & Gerken 1999) and non-adjacent dependencies (Gomez 2002; Gomez & Maye 2005). The roles of statistical distributions have traditionally been examined by those who put more emphasis on the roles of the environment, and the roles of learner's initial state have been largely ignored in those studies. What remains unclear is the nature of learner's contribution. At least two possibilities exist.

One is that learners use these acquired statistics to create an illusion of structure.[1] What a learner does is to track the surface distributions and carry forward a

[1] One possible explanation for why distributions seem to be informative for structure-like phenomena could be because many linguistic phenomena are dependent on structure.

summary of those distributions (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett 1996). According to this view, a learner does not come equipped with linguistic symbolic component. Under this view, the learner does not have domain-specific constraints on possible linguistic structures, but the learning may be restricted by constraints on general learning mechanism. As Saffran, Aslin & Newport (1996a) suggests, "some aspects of early development may turn out to be best characterized as resulting from innately biased statistical learning mechanisms rather than innate knowledge."

A second possibility is that learners use statistics to identify particular abstract syntactic representations (Miller & Chomsky 1963; Yang 2006; Pearl 2007). According to this view, the learner may come equipped with antecedently known knowledge and the statistical learning interacts with that knowledge by determining the questions that the statistical distributions are relevant for answering. Hence, the outcome of the learning is a combination of the generalization formed based on the observed input and innate knowledge.

To investigate this more general question, we focus on how infants (and adults and networks) acquire the phrase structure (PS) of their target language in this dissertation.

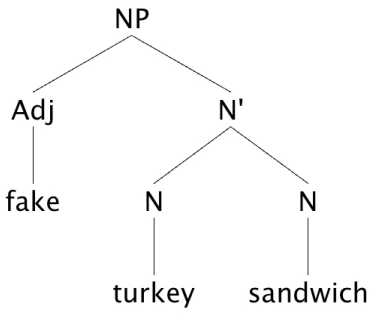
Traditionally in linguistic theory, it has been believed that sentences of human language are not simply linear strings of words, but words in a sentence constitute a hierarchical phrase structure (Chomsky 1957; Jackendoff 1977). It has been proposed, for example, that hierarchical structure is what is responsible for deriving different

meanings for the sentence with identical linear order. Imagine a phrase like the following.

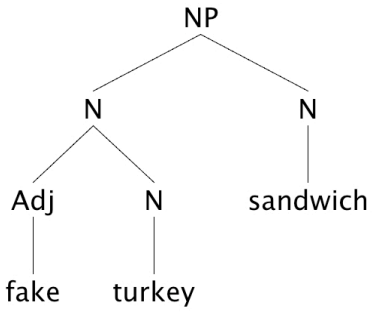
(1) Fake turkey sandwich

There are two possible meanings for this ambiguous phrase. One is that it is a turkey sandwich but is fake. It could be a toy or a plastic model of one, i.e., not edible. The other meaning is that it is a sandwich made with fake turkey like *Tofurky*. So it is edible, but may be vegetarian. These two distinct meanings are possible because the phrase can have two different structures.

(2)



(3)



(2) is the tree structure representation for the first meaning where the whole is a fake, and (3) is the tree representation for the second meaning where the sandwich is made with fake turkey. In this way, the fact that one identical linear string can be ambiguous supports that human language sentences have abstract hierarchical structure. Furthermore, this hierarchical constituency is what gives rise to recursive nature of language. In other words, because we have internal units within units, it is possible to achieve infinite creativity, which is a hallmark of natural language.

To illustrate the roles that constituents play, imagine a sentence like (4). This sentence is composed of higher-level groupings, such as [*the boy from the creek*], [*from the creek*], [*the creek*], and [*met Steven Spielberg*], and you can draw a phrase structure (PS) tree as in Figure 1.

(4) The boy from the creek met Steven Spielberg

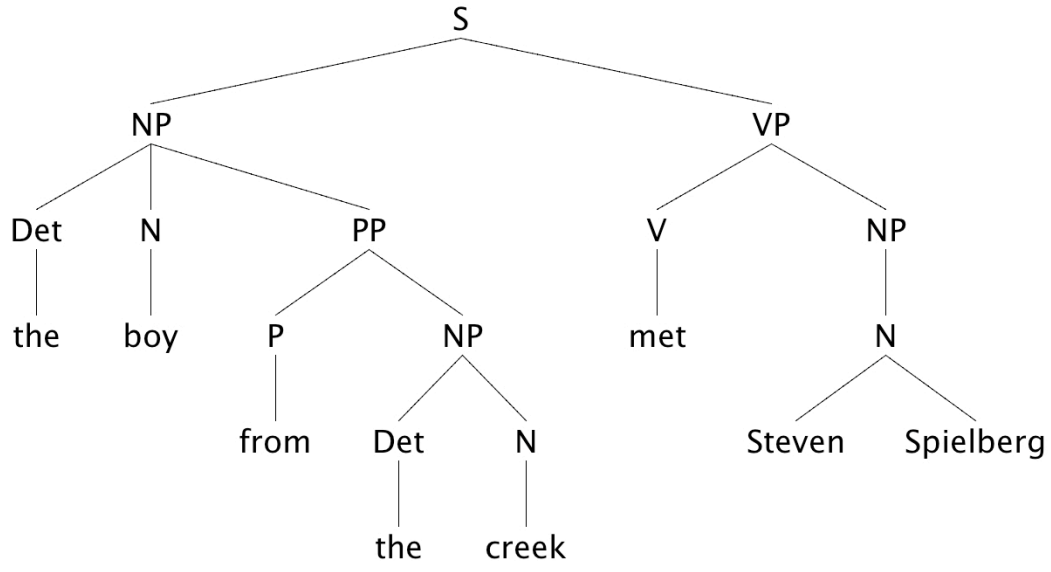


Figure 1: An example phrase structure representation

Each of those nodes in the tree forms a grouping called constituents. Notice that the constituent [*the boy from the creek*] contains two other constituents inside of it ([*from the creek*], [*the creek*]) yielding a nested hierarchical structure.

Syntactic constituency is important because all operations of the grammar make reference to them. Human language is a combinatorial system that operates not on linear strings of words but on those units called constituents.

Here are examples of roles constituents play in the grammar. First, you can substitute words with a proform (e.g., pronouns), but whatever being substituted must be a constituent. For example, you can replace the NP [*the boy from the creek*] with a pronoun *he,* and say (5). But you cannot just replace *the boy* and say (6).

(5) He met Steven Spielberg

(6) *He from the creek met Steven Spielberg

This is because the two words *the boy* do not form a constituent by themselves in this particular sentence. You can only substitute a syntactic constituent with a proform. Even if you know this rule (that only constituents are substituted by a proform), if you do not know the constituency of the sentence, you would not know which words can be replaced or not.

Similarly, if you have a sentence like (7), you can replace the VP [*met the director of E.T*] with a proform *did so*, but not just *met the director* as in (8).

(7) The boy from the creek met the director of *E.T.* and the girl did so too

(8) *The boy from the creek met the director of *E.T.* and the girl did so of *Notting Hill* too

Here, the proform *did so* must replace the whole VP [*met the director of E.T.*] and not just *met the director*. The sentence in (7) must mean, "The boy from the creek met the director of *E.T.* and the girl also met the director of *E.T.*" This is because in (7), the words *met the direct of E.T.* forms a constituent, but the words *met the director* does not form a constituent of their own. Proform substitution only replaces a constituent, and it cannot replace a non-constituent.

Second, only phrasal constituents can undergo movement operations such as *wh*-question and clefting as in (9)-(10).

(9) a. Steven Spielberg met the boy from the creek

b. Who did Steven Spielberg meet?

c. *Who from did Steven Spielberg meet the creek?

(10) a. It was the boy from the creek that met Steven Spielberg

b. *It was the boy that from the creek met Steven Spielberg

In (9)b and (10)a, what is being moved (in addition to being replaced by a *wh*-word in (9)) is the whole NP [*the boy from the creek*]. On the other hand, in (9)c and (10)b, what is being moved is *the boy from* and *the boy,* respectively. The sentences in (9)c

and (10)b are unacceptable in English because non-constituents are being moved. Again, even if you inherently know that only constituents can be moved, if you do not know the specific phrase structure for the given sentence, you would not know which words you can move.

Third, some constituents can be optional. In (12), the PP [*from the creek*] is missing, but the sentence is still grammatical.

(11) The boy [from the creek] met Steven Spielberg

(12) The boy met Steven Spielberg

Fourth, on the category level, the same phrasal type can appear more than once in the sentence. For instance in (13), there are two NPs.

(13) $_{NP}$[The boy from the creek] met $_{NP}$[Steven Spielberg]

Fifth, the constituents are interchangeable as long as they are of the same category.

(14) $_{NP}$[Steven Spielberg] met $_{NP}$[the boy from the creek]

In this way, phrasal constituents play a fundamental role in any syntactic operation. All syntactic operations refer to and manipulate constituents.[2] This fact is

[2] This, in turn, leaves statistical footprints that could be informative about the syntactic structure of sentences for learners to detect.

called “structure dependence,” and it is a term for the fact that “the rules operate on expressions that are assigned a certain structure in terms of a hierarchy of phrases of various types” (Chomsky, 1988; 45). In order to acquire a grammar, all a child has access to, as input data is sentences together with possible meanings. A scientist can try to discover what grammar a child has acquired or what grammar the child thinks generated the sentence by using those constituency tests above that reveal the posited structure. One potential problem for a child is the fact that constituency and phrase structure are highly abstract notions and the surface form does not come with visible labels or brackets to signal the constituency. At first glance, the input seems like simply linear sequences of sounds. A child might come with innate knowledge about constraints on possible phrase structure representations (e.g., binary branching, nested hierarchical structure, endocentricity), which would restrict the representational space for possible structures. However, even that does not guarantee that the learner will build the correct phrase structure representation. Since words vary from language to language, a child has to learn exactly which words go with which words in the particular language they are learning in order to arrive at the correct structural representation. In other words, it is not enough for a child to know that “there exists phrase structure” or that “a sentence is composed of a subject NP and a predicate” to figure out the constituency of a sentence. There must also be a mechanism that guides the child to the correct phrase structure representation of sentences for a particular language (Fodor 1966; Pinker 1984; Grimshaw 1981; Chomsky 1981; Macnamara 1982).

What kind of information might be readily available for a prelinguistic child, other than the linear strings of sounds? It has been proposed that prosodic, semantic and distributional information of sentences are perceptually available to a child (Gleitman & Wanner 1982; Gleitman, Gleitman, Landau & Wanner 1988; Morgan 1986; Pinker 1984; Grimshaw 1981; Macnamara 1982; Morgan, Meier & Newport 1989; Saffran 2001; Thompson & Newport 2007, among others). We will review studies that investigated infants' sensitivity to those properties in Chapter 2. Earlier research emphasized the necessary role of semantic (Pinker 1984) and/or prosodic (Gleitman & Wanner 1982; Gleitman, Gleitman, Landau & Wanner 1988; Morgan 1986; Peters 1983) cues in driving the acquisition of phrase structure. In a comparison of the utility of distributional and linguistic cues, Morgan, Meier & Newport (1987) found that adults were able to acquire the constituency of artificial languages only when the distributional information was augmented with correlated semantic, prosodic or morphological cues.

Recently, however, Thompson & Newport (2007) suggested that the distributional cues in those experiments were simply not strong enough, in and of themselves, to be informative. Instead, they showed that adult language learners could, indeed, exploit transitional probabilities in acquiring an artificial phrase structure grammar. In particular, it has been proposed that "transitional probabilities", which is a statistic that measures the predictiveness of the following element given one element, can be used by adults to successfully learn phrasal groupings of words (Thompson & Newport 2007) in miniature artificial languages. One problem with the artificial grammar used in Thompson & Newport (2007) is that it contained phrases

with no internal structure, but internally nested hierarchical structure is a hallmark of natural language syntax. Therefore, these findings leave unresolved whether learners can detect statistical cues to internally structured phrases.

As mentioned above, another question that still remains unclear is what exactly is learned via this statistical learning algorithm. Focusing on the phrase-structure learning problem, one possibility is that learners use the statistics to create phrase structure representations from scratch. This view holds that each child has to discover the existence of phrase structure and its characteristics on the basis of distributional information. Under this view, statistical learning does not interact with knowledge that the learners might already have, and the generalizations the learners form is entirely based on the observed input. There are two concepts within this view. One is that learners have no innate knowledge about possible representations prior to the exposure (Tomasello 2000). Then after being exposed to the target language, the learner would build and construct the representations. This concept acknowledges that what a child ends up with is abstract, but they get there not because of innate syntactic competence, but because of other usage-based mechanisms. The other concept is that the acquired grammar only has the properties that are derived from the observed distributions. Under this view, what the learner ends up with is not an abstract structure but only an illusion of one (Elman 1991).

A second possibility is that learners use the statistics to identify particular abstract syntactic representations. According to this view, each child uses the input distribution to determine how the particular language maps words to trees of a highly restricted character. Under this view, statistical learning interacts with knowledge that

learners may already have and the outcome of the learning is based on both the observed input and the antecedently known knowledge. Here, the antecedently known knowledge implies that the class of possible representations is predetermined (e.g., endocentricity, binary branching, proform substitution only replaces constituents and not non-constituents, etc.). If the range of possible representations is already known to a learner, then all a learner has to do is to select the correct grammar that derives the surface strings. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. Previous studies have failed to explicitly distinguish these possibilities and thus neither possibility was explicitly supported by past studies.

The predictions for these two approaches can be summarized as follows (excerpt from Lust, 2006).

(15) Predictions of a purely statistical approach

i. Learners have a direct relation to input data
ii. No universal linguistic constraints are predicted (e.g., no structure dependence)
iii. Only randomly, if at all, attend to parametric variations of language
iv. Not creative but highly imitative; generalizations should only be based on perceived forms or analogy
v. Learners do not evidence universal language principles or patterns

(16) Predictions of an approach in which nativism and statistics interact

i. Learners have an indirect relation to input data
ii. Be constrained in language acquisition
iii. Be structure dependent from the beginning, and attend to the parameters of language variation
iv. Be creative, i.e., go beyond the stimuli, and not simply copy
v. Not offend universals shown across natural languages

This dissertation presents seven experiments investigating the answers to the questions: (a) whether adults, infants and Simple Recurrent Networks (SRN) are sensitive to the distributional information as a cue to the hierarchical phrase structure, (b) whether adults, infants and SRN can learn the constituency of an artificial language without any prosodic or semantic cues, (c) whether the representations are part of the learning system prior to the experience, and (d) what the deductive consequences of distributional learning are. We show that human adults and infants can learn nested hierarchical phrase structure by using statistics alone, while the SRN fails. Specifically, it will be shown that the predictions of a nativist approach in (16) are borne out, in particular we will show that learners' generalizations go beyond the stimuli and they are not simply copies of the input, but those generalizations are constrained in the same ways that natural languages are constrained. More specifically, even when the input only contained the information for constituency, the learners not only learned the constituent structure but also inferred that non-constituents cannot be moved. In other words, our experiments suggest that learners

show knowledge of the constraint on movement even in the absence of movement in the exposure data, which suggests that the learners knew the constraint antecedently. Importantly, knowing the constituent structure alone does not give this result.

These results are compatible with the view that statistical learning interacts with inherent representational system, but it is incompatible with the view that statistical learning does not interact with innate linguistic knowledge as proposed by Saffran et al. (1996a) and Elman (1991), for example. In this dissertation, we propose a way in which the innate knowledge and the environment might interact. We suggest that innate knowledge supplies a range of possible representations and constraints (e.g., constraint on movement rules), while statistical learning is used as a method for mapping the surface strings to abstract representations.

This dissertation is organized as follows. Chapter 2 reviews past findings on what types of cues children are sensitive to and might employ when learning phrase structure, in particular, prosodic, semantic and distributional information. Chapter 3 presents the results from the experiments done with adults. Chapter 4 presents experiments with infants. Chapter 5 presents SRN simulations. Finally, Chapter 6 summarizes findings from the dissertation.

Chapter 2: An Overview of Previous Research in the Acquisition of Phrase Structure

2.1 Prosodic bootstrapping hypothesis

In this dissertation, we are interested in how children arrive at the correct phrase structural representation for their target language. Discovering the relevant syntactic units of a language is a fundamental step in language acquisition. All grammatical operations make reference to syntactic constituents, such as Noun Phrases and Verb Phrases. Without knowing which words in a particular sentence form constituents, it is impossible to perform any grammatical operations. Even given innate constraints on possible phrase structure representations and knowledge that sentences must be represented with nested hierarchy, some learning mechanism must be present for the learner to arrive at the correct representation in a particular language. In other words, innate knowledge is not sufficient for this task and some kind of information in the input is necessary. One of the most obvious information sources in the environment is acoustic information.

It has been suggested that the input speech signal comes with acoustic cues to syntactic organization and that children are sensitive to those cues as information about the syntactic structure of the sentence. This proposal is known as *the prosodic bootstrapping hypothesis* (Gleitman & Wanner 1982, Gleitman, Gleitman, Landau & Wanner 1988, Morgan 1986, Peters 1983). The prosodic bootstrapping hypothesis

proposes that acoustic information contains cues to syntactic boundaries that can be employed by learners. The majority of research on lexical, phrasal and clausal segmentation has been done in the framework of prosodic bootstrapping. Now, it might be worthwhile to note that what was intended by the original proposal of prosodic bootstrapping (Gleitman & Wanner 1982) was that prosodic information is *one of many kinds of cues* that could be used to discover or construct the phrase structure. In this section, we will review past findings to examine whether the prosodic bootstrapping is a real possibility for language learners.

2.1.1 Acoustic correlates at syntactic boundaries

Prosodic phonology (Selkirk 1984, Nespor & Vogel 1986) proposes that utterances are hierarchically organized with several layers. The highest constituent of prosodic hierarchy is the intonational phrase. An utterance is composed of one or more intonational phrases, which roughly corresponds to a clause. An intonational phrase is composed of one or more phonological phrases, which in turn is composed of one or more prosodic words. A prosodic word consists of one content word and some function words. The prosodic hierarchy is illustrated in the figure below.

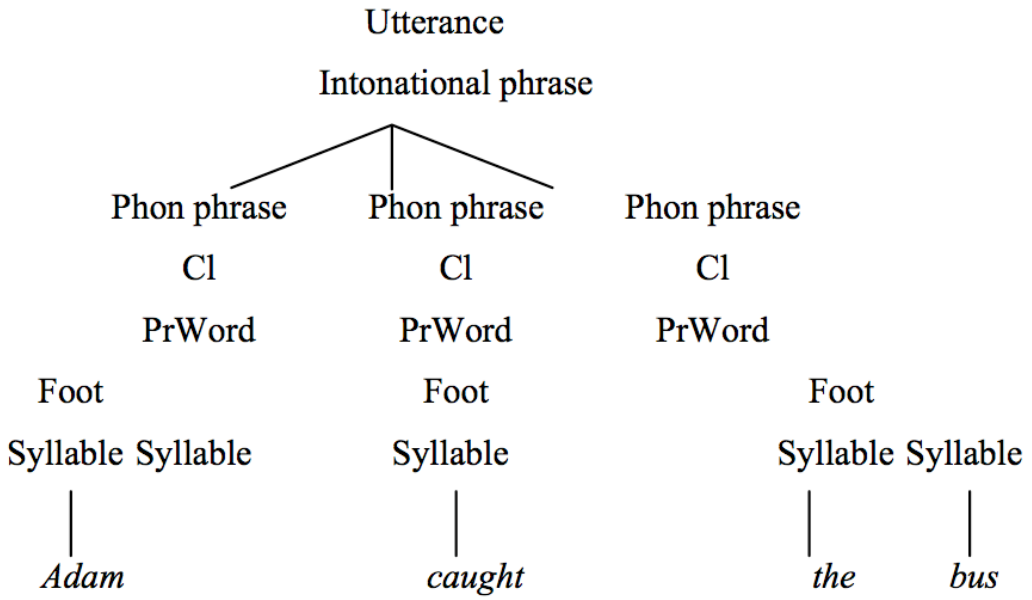


Figure 2: Prosodic hierarchy based on Nespor & Vogel (1986) (from Hicks 2006).

A phonological phrase boundary always corresponds with a syntactic boundary. Whenever there is a phonological phrase boundary, there is a syntactic phrase boundary there. But the opposite may not be true: a syntactic boundary does not always correspond with a phonological phrase boundary. In other words, one phonological phrase may consist of more than one syntactic phrase. For example, in the sentence $_{S}$[$_{NP}$[the dog] $_{VP}$[chased $_{NP}$[the cat]]], the VP *chased the cat* may form one phonological phrase, but it contains another syntactic boundary inside of it, which is the boundary for the object NP.

It has been widely observed that there are acoustic correlates that signal syntactic boundaries (Selkirk 1984, Nespor & Vogel 1986). These cues include preboundary lengthening (Beckman & Edwards 1990, 1992, Cooper & Paccia-Cooper 1980, Klatt 1975, Wightman, Shattuck-Hufnagel, Ostendorf & Price 1992),

pause duration (Cooper & Paccia-Cooper 1980, Scott 1982), change in pitch (Beckman & Pierrehumbert 1986), greater initial strengthening (Fougeron & Keating 1997, Keating, Cho, Fougeron & Hsu 2003) and reduction of coarticulation between phonemes across boundary (Byrd, Kaun, Karayanan & Saltzman 2000, Hardcastle 1985, Holst & Nolan 1995). Additionally, a phonological phrase typically contains one melodic contour. Moreover, the presence of these prosodic cues has been confirmed in a number of languages (Barbosa 2002 for Brazilian Portuguese; de Pijper & Sanderman 1994 and Quené 1992 for Dutch; Fisher & Tokura 1996b for Japanese; Hayes & Lahiri 1991 for Bengali; Keating, Cho, Fougeron & Hsu 2003 for English, French, Korean and Taiwanese; Pasdeloup 1990 and Rietveld 1980 for French; Wightman et al. 1992 for English).

The strength of those prosodic cues varies depending on the indicated boundaries. The boundaries that are higher on the prosodic hierarchy (i.e., intonational phrase) are indicated by stronger prosodic cues (Cooper & Paccia-Cooper 1980). In particular, there are longer pauses, stronger preboundary lengthening and increased intonation at intonational phrase (i.e., clause) boundaries than at phonological phrase boundaries (Cho & Keating 1999; Shattuck-Hufnagel & Turk 1996).

Although they may be weaker, there are prosodic cues at phonological phrase boundaries, including final lengthening and a single pitch contour (Wightman, Shattuck-Hufnagel, Ostendorf & Price 1992), greater initial strengthening (Fougeron & Keating 1997, Keating, Cho, Fougeron & Hsu 2003), reduced coarticulation

between phonemes that span across the phonological phrase boundary (Byrd, Kaun, Narayanan & Saltzman 2000, Hardcastle 1985, Holst & Nolan 1995).

One potential problem is that not all syntactic boundaries (17) correspond with prosodic boundaries (18).

(17) He was / eating an enormous apple

(18) [He was eating] [an enormous apple]

Often you observe no, or even misleading, prosodic changes at phrase boundaries ((19) vs. (20)).

(19) [[$_{NP}$ This] [$_{VP}$ is [$_{NP}$ the dog that chased [$_{NP}$ the cat that bit [$_{NP}$ the rat that lived in [$_{NP}$ the house that Jack built]]]]]]

(20) This is the dog / that chased the cat / that bit the rat / that lived in the house / that Jack built

(19) is the syntactic bracketing of the sentence, while (20) is how a speaker would produce prosodic boundaries. It has been long noted that prosody does not always mirror the hierarchical structure of syntax (e.g., Chomsky & Halle 1968). In the example above, the major syntactic constituents are the NPs embedded within the VPs as in (19). However, speakers tend to produce prosodic boundaries between the noun and the relative clause as in (20). If the learners simply take pauses to mean that

there is a syntactic boundary at that pause, then the learners would not incorrectly parse the sentence. That is because there is a syntactic boundary at the pauses in (20), namely the boundary between a noun and a relative clause. However, if the conclusion the learners make is that words on both sides of the pause form a constituent, then the sentence would not be correctly parsed. For example, the learners might conclude that *this is the dog* and *that chased the cat* both form constituents which is wrong. At this point, we do not know which conclusion (the stronger one or the weaker one) the learners might draw. In any case, the problem is that learners need to eventually figure out what inferences are licensed by the prosodic information.

2.1.2 Infants' sensitivity to acoustic cues at syntactic boundaries

We see that the acoustic cues are present in the input, but are infants sensitive to these cues? It has been found that infants as young as 6-months of age are sensitive to prosodic cues at clause boundaries (Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss & Kennedy 1987, Jusczyk 1989). In Hirsh-Pasek et al. (1987), two sets of passages were created. In the "coincident" version, 1-sec pauses were inserted at clause boundaries. In the "non-coincident" version, 1-sec pauses were inserted between words in the middle of a clause. 7-10 month-old infants listened longer to the coincident version. Accordingly, the authors claim that prosodic cues serve as a marker of clause boundaries and infants are sensitive to it.

In order to make sure that the infants were responding to prosody and not some other information, Jusczyk (1989) conducted a follow-up study. The samples

were low-pass filtered at 400Hz to remove most of the phonetic information. 6 month-old infants showed a preference for the coincident version, by listening significantly longer to it. This suggests that well before their first birthday, infants are sensitive to prosody as a marker of clause boundaries.

So, very young infants seem to be sensitive to the prosodic markings of clause boundaries, but what about phrases? Past research have suggested that the prosodic cues at phrasal boundaries are weaker than the ones at clausal boundaries (Beckman & Edwards 1990; Fisher & Tokura 1996a, 1996b; Gerken et al. 1994). Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward & Piwoz (1992) conducted a series of experiments to investigate whether infants are sensitive to the acoustic correlates at phrasal boundaries. Two sets of passages were created. In the coincident set, 1-sec pauses are inserted right before the predicate, as in (21). In the non-coincident set, 1-sec pauses were placed right after the verb, as in (22).

(21) Coincident version:

Did you / spill your cereal? Do you / want to pick it all up? That / looks great. You / want to put it back in your little container here?

(22) Non-coincident version:

Did you spill / your cereal? Do you want / to pick it all up? That looks / great. You want / to put it back in your little container here?

The idea is that, in the "coincident" sentences, a prosodic cue (in this case, a pause) coincided with a syntactic phrase boundary, and everything after the pause is the predicate which usually forms a phonological phrase. On the other hand, in the "non-coincident" version, the pause is always after the verb. The results showed that 9-month-old infants, but not 6-month-olds, listened longer to the "coincident" samples, which had pauses at the major phrase boundaries.

In these experiments, the coincident version always had a pause before the predicate and the non-coincident version after the verb. Therefore, one might raise concerns that there was something special about verbal predicates. In order to check this possibility, Jusczyk et al. (1992) created two new sets of passages.

(23) Coincident version:

Many different kinds of animals / live in the zoo. The dangerous wild animals / stay in cages. Some of the animals / are friendly and like to be petted.

(24) Non-coincident version:

Many different kinds / of animals live in the zoo. The dangerous / wild animals stay in cages. Some / of the animals are friendly and like to be petted.

The coincident version was created by placing 1-sec pauses immediately after subject NPs, which commonly form a phonological phrase. In the non-coincident version, 1-sec pauses were placed somewhere within the subject NPs, which is an unnatural place to have a prosodic boundary. The results showed that 9-month-olds listened

longer to the coincident version, indicating that the infants are sensitive to the prosodic information at phrasal boundaries.

The findings of Hirsh-Pasek et al. (1987) and Jusczyk et al. (1992) indicate that infants are sensitive to acoustic properties of clause and phrase boundaries when the prosodic cues are reliably available. But it has not been shown yet that infants actually use such sensitivity in processing of fluent speech. Nazzi, Kemler Nelson, Jusczyk & Jusczyk. (2000) investigated this issue with clauses. 6-month-old infants were familiarized with sequences such as *rabbits eat leafy vegetables,* and then tested with either well-formed "Rabbits eat leafy vegetables" or non-unit "… rabbits eat. Leafy vegetables …" Infants listened longer to the passages containing the well-formed familiar sequence.

However, Soderstrom, Kemler Nelson & Jusczyk (2005) argue that in Nazzi et al. (2000), infants did not have to segment the speech stream based on the prosodic cues, since they were presented with already segmented target sequences. Thus, it does not tell us whether prosodic cues help infants extract the relevant sequences, which is much more similar to what they actually have to do in language acquisition. Therefore, Soderstrom et al. (2005) conducted the following experiment.

(25) Familiarization

i. John doesn't know what rabbits eat. *Leafy vegetables taste so good*.

ii. *Rabbits eat leafy vegetables*. Taste so good is rarely encountered.

(26) Test

i. *Leafy vegetables taste so good*. Salad is best with dressing.

ii. Students like to watch rabbits eat. Leafy vegetables make them chew.

iii. Squirrels often feed on acorns. *Rabbits eat leafy vegetables*.

iv. Mothers must buy leafy vegetables. Taste so good helps their families.

The underlined sequences are "clause straddling" and the italicized ones are "clause coincident." During familiarization, infants heard either (25)i or (25)ii. At test, infants heard all four passages in (26). This experimental setup is claimed to be more realistic because target sequences are embedded in fluent speech throughout the experiment. The results showed that 6-month-old infants listened longer to the test passages that contained the sequences that were "clause coincident" (italicized) during familiarization. In addition, infants also listened longer to the test passages that matched the prosodic structure during familiarization (regardless of coincident or straddling). Therefore, the authors conclude that infants are able to use prosodic information to recognize the sequences in fluent speech and detect such sequences in different fluent speech. In sum, prosodic cues appear to be useful for infants to encode and recognize word sequences in fluent speech.

To investigate the same issue with phrases, Soderstrom et al. (2003) created a series of experiments. The following natural speech stimuli were used.

(27) At the discount store, **new watches for men** are simple and stylish. In fact, some **people # buy the whole** supply of them.

(28) In the field, the old frightened **gnu # watches for men** and women seeking trophies. Today, **people by the hole** seem scary.

The boldfaced phrases without # are the syntactically well-formed noun phrase (NP) target sequences. The boldfaced phrases with # are the syntactic non-unit (NU) target sequences. Infants were assigned to either the "watches" condition or "people" condition. During the familiarization, infants just heard the boldface target sequences. During the test, infants heard the entire passages. 6-month-olds, as well as 9-month-olds, listened longer to the NP version than the NU version at test. In other words, infants preferred to listen to the passage that contained the syntactically well-formed familiarized target sequence.

In order to examine whether the obtained effect was specific to NPs, the same experiment with VPs was carried out.

(29) Inventive people **design telephones** at home. A fresh idea with **promise # surprises** no-one who words there.

(30) The director of **design # telephones** her boss. New developments **promise surprises** for their old buyers.

Again, 6-month-old infants showed a preference for the passages containing the well-formed VP, by listening longer to it.

Even though this result suggests that infants can recognize the familiar sequences embedded in the larger passages, Soderstrom et al.'s (2005) criticism also applies here. Since infants were familiarized with already-extracted sequences (boldface), we do not know whether prosodic information helped infants to segment the speech stream. Therefore, a follow-up study one could do is to train infants with passages as in (31)-(32) and to test them with (27)-(28).

(31) I got these **new watches for men** from that store in Montreal. I saw **people # buy the whole** boxful of them.

(32) Did you know that **gnu # watches for men** in the field? Especially **people by the hole** are the excellent target for them.

In this way, we can examine whether children use prosodic markers for both extracting and recognizing sequences of words in fluent speech.

In any case, Soderstrom et al.'s (2003) results are the first evidence that infants recognize familiar phrases embedded in the fluent speech only when that phrase is prosodically well-formed. This is also the first evidence that infants as young as 6-month old are sensitive to the prosodic phrasal grouping.

As mentioned earlier, the main acoustic correlates at syntactic boundaries are preboundary lengthening, pause duration and change in pitch. But these cues do not seem to weigh the same. Seidl (2007) showed that pitch is an essential cue for successful segmentation of clauses, while neither pause duration nor preboundary lengthening was found to be necessary for English-learning 6-month-old infants. The

most significant finding was that none of these cues was sufficient on its own. Even pitch, which was found to be an essential cue, had to be paired with either pause or preboundary lengthening in order for 6-month-old infants to successfully segment clauses. This suggests that a combination of at least two acoustic cues is required for detecting syntactic boundaries. This finding is relevant to us because it strengthens our suggestion that in order to figure out phrase structure of a sentence, infants do not just use one kind of cue, but probably a combination of several kinds of cues are used, or that at least it is more helpful to have more cues than just one.

2.1.3 Using prosodic cues for lexical access

In addition to clausal and phrasal segmentation, a number of studies have investigated prosody's effect on lexical segmentation. 3-day-old infants discriminated sequences of syllables that contain a word boundary from those that do not, suggesting that newborns are sensitive to acoustic correlates at phonological phrase boundaries (Christophe, Dupoux, Bertoncini & Mehler 1994, Christophe, Mehler and Sebastián-Gallés 2001). However, we should remember that being sensitive to the acoustic correlates does not necessarily entail that babies actually use them for the purpose of lexical segmentation.

Gout, Christophe & Morgan (2004) investigated whether infants can use phonological phrase boundaries to constrain lexical access online. The stimuli of the following kind were created.

(33) a. [The scandalous *paper*] [sways him] [to tell the truth]

b. [The outstanding *pay*] [*per*suades him] [to go to France]

In (33)a, the bisyllable *paper* is contained within a phonological phrase, whereas in (33)b, the same bisyllable *pay#per* straddles a phonological phrase boundary. One of the prosodic differences between the two bisyllables was phrase-final vowel lengthening: the vowel [eɪ] in *pay#per* was longer than in *paper,* while the vowel [ə] in *paper* was longer than in *pay#per.* The consonant [p] in *pay#**p**er* was longer than in *pa**p**er* (phrase-initial consonant lengthening). There was a short pause between the two syllables in *pay#per,* but not in *paper.* During the familiarization phrase, the infants were presented with *paper*-type stimuli, and at test, they were presented with both *paper*- and *pay#per*-type sentences. It was found that 13-month-old English-learning infants listened longer to *paper*-type sentences than to *pay#per*-type sentences, while 10-month-old infants did not show any difference in looking times. This shows that 13-month-olds are sensitive to prosodic correlates at phonological phrase boundaries and can exploit it in segmenting words from fluent speech. Adults also have been shown to use the acoustic cues at phonological phrase boundaries to constrain online lexical access (Christophe, Peperkamp, Pallier, Block & Mehler 2004). These results suggest that both infants and adults use prosodic information to help them with segmentation of words.

Furthermore, Christophe, Nespor, Guasti & Van Ooyen (2003) proposed that infants make use of prosodic features to infer the syntactic head parameter. French and Turkish both have word-final stress. Syntactically, however, while French is a

head-initial language, Turkish is head-final. Therefore, the phonological phrase prominence is final in French and initial in Turkish. Thus, regarding the phonological properties, the phonological phrase prominence is the only thing that distinguishes the two languages. The French and Turkish materials were synthesized and all the phonemes were made identical. The only difference between the two materials was the phonological phrase prominence. Using the high amplitude sucking paradigm, 6-12-week-old French infants discriminated the French and Turkish sentences. This result suggests that young infants are sensitive to the difference in phonological phrase prominence.

2.1.4 Mismatch between syntax and prosody

As mentioned earlier in this chapter, not all syntactic boundaries are marked with a prosodic boundary. Often you observe no, or even sometimes misleading, prosodic markers. Can prosodic cues be useful for infants to detect syntactic boundaries in those non-isomorphic cases?

Gerken, Jusczyk & Mandel (1994) investigated this problem. They presented infants with one set of passages that had 1-sec pauses inserted immediately after the subject (e.g., *he / ate four strawberries*), while the other set had pauses immediately after the verb (e.g. *he ate / four strawberries*). Half of the 9-month-old infants heard passages with lexical subjects (e.g. *the caterpillar*). The other half heard passages with pronoun subjects (e.g. *he*). Only the infants in the lexical subject condition listened longer to the passages with pause after subject than the passages with pause after verb. Infants in the pronoun condition showed no preference for either version.

What Gerken et al. (1994) show is that when the prosodic boundaries match the syntactic boundary, infants are able to detect such cues, as in the lexical subject case. However, when the syntactic boundary and prosodic boundaries do not match, learners are not able to identify the syntactic boundaries from the prosodic cues. This suggests that there must be other cues, in addition to prosody, that infants can employ to help them bootstrap the syntactic structure of their language. In this disseration, we will investigate whether one type of statistical information, transitional probability, can be one of such cues.

2.1.5 Prosodic cues vs. statistical cues

A number of studies, including ones that are reviewed here, have shown that infants can use prosodic cues as one information source about where the word boundaries might be (Cutler & Norris 1988, Jusczyk, Houston & Newsome 1999b, Morgan 1996). Recently, it has also been shown that infants can use statistical cues for word segmentation (Saffran, Aslin & Newport 1996a). For example, the transitional probability from one sound to the next within a word (*pre-tty*) is usually higher than that of between words (pre*tty#ba*by). Given this, Saffran et al. (1996a) tested 8-month-old infants using the familiarization-preference procedure. Infants were exposed to auditory arrays of an artificial language for 2 minutes (*bidakupadotigolabubidaku*…). The transitional probability between syllables within a (artificial) "word" was 1.0 (e.g., *bida*) while the TP between syllables across words was 0.33 (e.g., *kupa*). The peaks and dips of the TPs were the only cue to the word boundaries, since there was no prosodic information about the word boundaries in the

auditory stimuli (e.g., no pauses, no stress changes). At test, the infants were presented with two types of test samples: one with "words" from that artificial language (e.g., *pabiku*) and "part-words" (e.g., *pigola*). The part-words were created by joining the last syllable of a word with the first two syllables of another word. The 8-month-old infants had significantly longer listening time for part-words than words, showing a novelty preference. These results indicate that 8-month-old infants are sensitive to the distributional information and they can use that information to discriminate between words and part-words.

Saffran, Newport & Aslin (1996b) explored the interaction of prosodic cues and statistical cues. A distributional cue (transitional probabilities) was combined with a prosodic cue (vowel lengthening). The experiment with adults showed that the vowel lengthening alone was not informative and it was informative only when it was combined with the distributional cue, which suggests that statistical cue prevailed over acoustic cue. However, in that study, the statistical cue and prosodic cue were not in conflict with each other. The potential "words" cued by transitional probability and the ones cued by vowel lengthening were the same.

When the two types of cues are in conflict, 8-month-old infants listened longer to (prosodically ill-formed) statistical words than (prosodically well-formed) statistical part-words (Johnson & Jusczyk 2001). Given that the infants in Saffran et al. (1996a) listened longer to part-words showing a novelty effect, we can also interpret this result as a novelty effect. In that case, it means that infants treated the statistical words as nonwords and prosodic words as real words. This indicates that they relied more heavily on stress than transitional probability as a cue.

But was it really novelty effect? What if it was a familiarity effect? In the infant literature, when the stimuli are relatively simple and easy for them to learn, infants are familiarized soon and they get bored, hence they tend to listen longer to the novel items at test (Hunter & Ames 1989, Aslin, Saffran & Newport 1998, Echols, Crowhurst & Childers 1997, Saffran et al. 1996a). On the other hand, when the stimuli are complex, it takes longer for infants to be familiarized, so they tend to listen longer to the familiar items at test (Houston, Santelmann & Jusczyk 2004, Jusczyk & Aslin 1995, Jusczyk, Hohne & Bauman 1999a, Jusczyk et al. 1999b, Mattys & Jusczyk 2001). Could it be that the stimuli in Johnson & Jusczyk (2001) were too complex, therefore infants listened longer to the familiar items? If so, it would mean that infants relied more on statistics than prosody.

To examine this possibility, Johnson & Jusczyk (2001) carried out a follow-up study, where the words cued by the statistics and prosody matched. 8-month-old infants listened longer to part-words, indicating that they did learn the words and showed a novelty preference. This result implies that the materials involving both statistical and prosodic cues were not too complex for the infants to learn. Therefore, it confirms the account for the previous experiment, which is that infants weigh the prosodic cues more heavily than statistical cues when the two cues are in conflict.

Thiessen & Saffran (2003) investigated infants' developmental reliance on conflicting cues. When the two cues – stress and transitional probability – signaled conflicting word boundaries, 6.5- to 7-month-old infants listened longer to the (prosodically well-formed) statistical part-words, showing a novelty preference, which indicates that they were paying attention not to stress but primarily to statistical

information. This suggests that 6.5-7-month-olds weigh statistical cues more heavily than stress cues.[3] However, a recent study reports that when provided with a list of segmented words prior to testing, 7-month-olds can use stress as a cue for lexical segmentation, which shows that experience with isolated words facilitates infants' learning of language specific rhythmic patterns (Thiessen & Saffran 2007, Gambell & Yang 2005).

These results let us begin to see a developmental trend of infants' attention to different cues. At around 7.5 months, infants begin to be sensitive to stress patterns (Jusczyk et al. 1999b). As a result, at 8-9 months, they start relying more heavily on stress cues than statistical cues. This might be because infants at this stage are not capable of integrating more than one type of cues (Morgan & Saffran 1995). Morgan & Saffran (1995) demonstrated that 6-month-olds were not able to integrate sequential and suprasegmental cues, while 9-month-olds were. Thus, it was suggested that the ability to integrate multiple kinds of information arises sometime between 6 and 9 months.

2.1.6 Summary of prosodic bootstrapping

One of the most obvious information sources for phrase structure available in the input is acoustic information. As the prosodic bootstrapping hypothesis proposes, prosody can provide information for some syntactic structures. When the prosodic cues are reliably available, young infants are sensitive to acoustic markers at as

[3] Nevertheless, whether infants can really use statistical cues in real life is debatable. Gambell & Yang (2005) have shown that speech to young children does not contain reliable TP boundaries, because most of the words are monosyllabic.

clausal, phrasal and lexical boundaries. However, the inferences about structures that are cued by prosody are not sufficient for building complete phrase structure representations. For example, when there is a mismatch between phonological and syntactic phrases, infants could be misled and misparse the sentence. More importantly, even when the learners are sensitive to the acoustic cues at syntactic boundaries, that does not entail that learners make the correct inferences about the relation between the surface cues and structures that they indicate. So, infants need additional information to help them infer the correct phrase structure. In the following sections, we will look at possibilities of semantic information and distributional information.

2.2 Semantic bootstrapping hypothesis

Another source of prelinguistic information that can be useful to a child in figuring out syntactic structure of a sentence is semantics. Pinker (1984) proposes what is called *the semantic bootstrapping hypothesis* (also based on Pinker 1982, Grimshaw 1981, Macnamara 1982). This hypothesis makes use of the fact that a lot of the times, nouns refer to physical objects, verbs refer to actions, adjectives refer to attributes, subject of a sentence is usually the agent of the action, object of a sentence is usually the patient or theme, and so on. In other words, there exist these syntax-semantics correspondences. It is hypothesized that the concepts such as physical objects, actions, attributes, and agent-of-action and patient-of-action are perceptible and perceivable information to a child. The main claim of the semantic bootstrapping

hypothesis is that, if these notions are perceptible by children, then children can use the fact that each of these notions corresponds on numerous occasions to a respective grammatical entity, such as nouns, verbs, subjects and objects. Coupled with basic notions of phrase structure rules (e.g., that S consists of a subject NP and a VP, VP consists of a V and an object NP), children can correctly parse a sentence using this syntax-semantics correspondence.

Let us see how this works. Imagine a child hears a sentence *The dog chased the cat.* The child uses the syntax-semantics correspondences to identify and label each word with its category. *Dog* and *cat* are physical objects, so the child would label them nouns. *Chased* is an action, so it will be labeled as a verb. *The* refers to definiteness in discourse, so it will be labeled as a determiner. Thus, we have the following result.

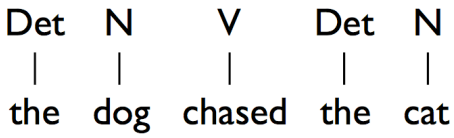


Figure 3: Labeling the categories

Next task is to group the words into phrases. Grimshaw (1981) and Pinker (1984) suggest that the constituents of each phrase (which is universal) is antecedently known to a child, and what a child has to learn is the linear order of its constituents (which varies cross-linguistically). For example, a child knows inherently that a sentence consists of a subject NP and a VP, a subject NP consists of an optional determiner and a obligatory noun, a VP consists of an optional object NP and an

obligatory verb, and so on. If a child knows that, then they will arrive at the intermediate structure below.

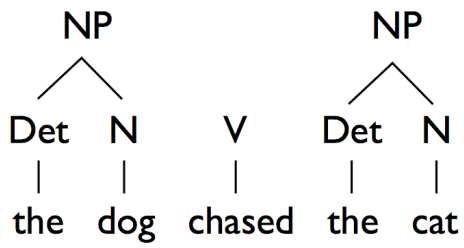


Figure 4: Forming Noun Phrases

Now, at this point, nothing prevents a child from incorrectly parsing the sentences as follows.

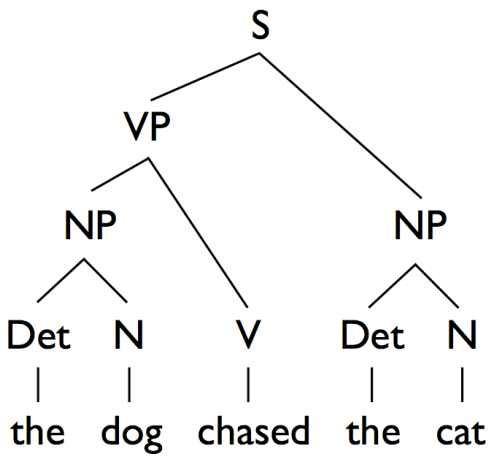


Figure 5: An incorrect tree

However, if the child can perceive from the discourse that *the dog* in this sentence is the agent of the *chasing,* and that *the cat* here is the patient of the *chasing*, then using the syntax-semantics correspondence (i.e., agent = subject, patient = object), the child

can infer that *the dog* must be the subject of the sentence and that *the cat* must be the object of the sentence. If the child figures that out and if they antecedently know that the subject NP is the immediate daughter of a sentence, and the object NP is the daughter of the VP, they can parse the sentence correctly.

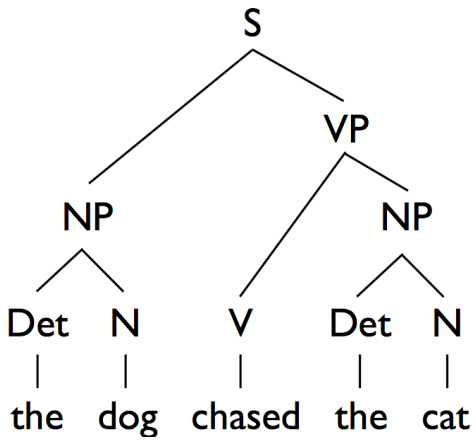


Figure 6: Correct phrase structure representation of an example sentence

One obvious problem with the semantic bootstrapping hypothesis is that the syntax-semantics correspondence is not perfect and in fact, often does not hold. For example, not all nouns denote a physical object (e.g., *a thought*), not all actions are verbs (e.g., *the reading, a flight*), and not all subjects denote an agent of an action (e.g., *John received a letter, John sustained an injury*). In case of passives, the agent-subject patient-object relation is reversed (e.g., *The pizza was eaten by John*). In case of topicalization, the subject appears at the end of the sentence (e.g., *Eats a lot of pizza, that guy*). In other words, the syntax-semantics correspondence only holds mostly in what is called "basic sentences" which are declarative, simple, affirmative, pragmatically neutral and minimally presuppositional (Keenan 1976).

These points are well noted by the researchers who proposed the semantic bootstrapping hypothesis and here is their solution to this problem. Pinker (1984) proposes that *at first*, a child only utilizes the basic, canonical examples in which the syntax-semantics correspondences hold. This could be achieved by either (a) the fact that first set of (very early) parental input rarely contains sentences that violates syntax-semantics correspondences; (b) the child would filter out and ignore non-basic, non-canonical examples at first, by using contextual cues such as special intonation, extra marking on the verb, presuppositions and interrogative or negative illocutionary force of an utterance. This suggestion is supported by a general observation that children's first words that are nouns are universally physical objects, children's first verbs usually denote actions, first adjectives are attributes, subjects are usually agent of an action, and objects are usually patients (e.g., Brown 1973, Bowerman 1973, Macnamara 1982, Nelson 1973, Slobin 1973).

It is proposed that a child first uses semantic bootstrapping for the basic sentences, and when they encounter non-basic sentences, they use other means to parse the sentence, in particular Pinker (1984) proposes a process called *structure-dependent distributional learning.* For instance, if a child encounters a sentence like *The situation justified the measures,* since the syntax-semantics correspondence does not hold in this sentence, the semantics alone cannot help the child to correctly parse it. Pinker (1984) suggests that in this case, the distributional information helps the learner. For example, a child would know by now that *the* is a determiner. And because they would also already know that the only phrase a determiner can be a part of is a NP. So they deduce that *situation* and *measures* must be nouns. Next, although

they may not know that the word *justified* is a verb because it does not denote an action, they can notice the *–ed* ending and if they know that the *–ed* ending signals past tense of verbs, then they can deduce that *justified* must be a verb. And if they know the PS rules of English, they can arrive at the correct parse for the sentence.

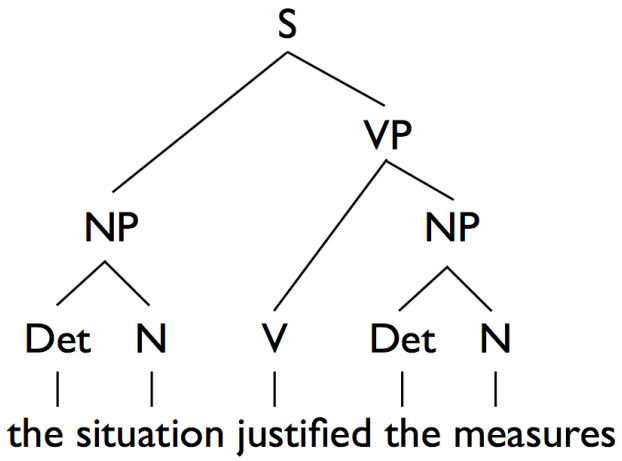


Figure 7: PS tree of a sentence in which syntax-semantics correspondence does not hold

In sum, similarly to the prosodic bootstrapping, the semantic bootstrapping is not sufficient by itself. Pinker (1984) specifically states that the semantic bootstrapping hypothesis does not claim that children fail to perform distributional analyses. On the contrary, Pinker (1984) claims that distributionally-based analyses override semantically-based analyses when the two are in conflict. Previous studies have shown that distributional information can aid learners in parsing a sentence when it is semantically ambiguous or when the semantics are misleading (Lebeaux & Pinker 1981, Katz, Baker & Macnamara 1974, Gelman & Taylor 1983). What is claimed by the semantic bootstrapping hypothesis is that semantically-driven analyses

interact with distributional analyses and that semantics can help learners determine which are the relevant distributional analyses to perform.

In other words, neither the prosodic or semantic bootstrapping hypotheses are in conflict with what I will propose in this dissertation, nor am I claiming that statistical distribution is the only cue that the learners use in figuring out the phrase structure of a language. All I am claiming is that statistical distribution may be one of the many cues to constituency. But are learners actually sensitive to distributional information? In the next section, we review previous studies that examined effectiveness of statistical cues for infants and adults.

2.3 Artificial language experiments

Recent studies have suggested that statistical distribution might be one of the information sources for acquisition of various features of language. For example, it has been suggested that distributional information can play a role in the acquisition of phonemes (Maye, Werker & Gerken 2002, Maye & Gerken 2000), word segmentation (Saffran et al. 1996a, Swingley 2005), word categories (Redington, Chater & Finch 1998, Mintz, Newport & Bever 2002, Mintz 2003) and syntax-like regularities (Gomez & Gerken 1999). The question then is, can learners use distributional information as a cue to constituency? Below, I will review a series of artificial language learning studies that investigated how phrasal groupings might be learned.

2.3.1 Learning constituency through reference

Morgan and Newport (1981) was the first of a series of studies that investigated what cues learners might employ to learn the constituent structure of a miniature artificial language. In the artificial language of this study, each word had a referent (shaped objects). Adult participants simultaneously heard the spoken sentence and saw the corresponding referents on a screen. At test, they were asked to judge which of two fragments formed a better group or unit. In each pair, one fragment constituted a syntactic phrase in the language, whereas the other fragment consisted of adjacent but syntactic non-constituent words. Another test asked asked which of the two sentences was preferable. Each pair had structure-preserving and structure-destroying transformations of a sentence. In the former, a syntactic constituent had undergone movement and in the latter, adjacent but non-constituent words had undergone movement. The results showed that only the participants who were given the input where the referents were spatially organized consistently with the syntactic constituency fully learned the language.[4] The participants who were not given the syntactically-consistent spatial organization of referents failed at the tests. It is claimed that these results show that you can induce a phrase structure tree

[4] Morgan & Newport (1981) note that it is surprising how well participants performed on the transformation test even though the input did not contain any transformation sentences and they were not given any criteria for choosing the structure-preserving answers. They state: "This is striking evidence that adults, given a brief exposure to a simple language, may be capable of developing sophisticated linguistic intuitions and that the acquisition of artificial languages by adults may be quite similar to the acquisition of natural languages by children." It is also suggested that "natural language structure is constrained by the acquisition process", rather than the acquisition process being constrained by the linguistic structure. Morgan & Newport (1981) does not discuss this issue any further, so it is difficult to comprehend their point, but it seems that they are suggesting that learners develop structure-dependent rules themselves, instead of learners being equipped with some inherent knowledge. In view of our Experiment 2 and our interpretation of its results, it is interesting that we came to different conclusions.

representation only if you receive an extra (in this case, semantic) cue in addition to the distributional cue. However, that might have been due to the fact that, as we will see below, the statistical cue in this study was not very reliable. At any rate, these results show that perceptual grouping of words facilitates the learning of hierarchical phrase structure.

2.3.2 Learning constituency through prosody

In a subsequent study, Morgan, Meier and Newport (1987) show that prosody is a helpful cue when learning phrasal groupings of an artificial language. Their claim was that even though distributional cues alone should be logically sufficient to deduce the syntactic structure, other redundant cues are necessary for successful learning of the language. The same artificial language and test items as Morgan & Newport (1981) were used. Only the adult subjects who were given the input sentences read with prosody that was consistent with syntactic bracketing fully learned the language. These results suggest that prosodic cues that are consistent with syntactic bracketing strongly enhance the learning of syntactic structure.

2.3.3 Morphological cues to phrase structure

Morgan et al. (1987) also claim that grammatical morphology can be a signal to phrase structure. In particular, function words are often placed at phrase boundaries, either at the beginning or the end of a phrase (Clark & Clark 1977, Kimball 1973).

Morgan et al. (1987) tested whether adults can bootstrap phrase structure based on morphological cues. The results showed that the adult subjects who were given the input sentences with function words at the edges of syntactic phrases performed significantly better than subjects who received no such information. This suggests that function words placed at phrasal boundaries help enhance the learning of syntactic constituency.[5]

Another type of morphological cue Morgan et al. (1987) suggest is concord morphology. In natural languages, words can agree in case, gender, number or definiteness (Morgan et al. 1987). As with the previous experiments, this experiment tested whether concord morphology could be a cue to phrasal grouping using an artificial language. Adult participants who were given the input that contained inflectional morphemes that matched the syntactic bracketing performed the best. These results suggest that the presence of concord morphology can significantly improve the learning of syntax. One caveat is that it is not always the case that agreement happens within a phrase. In case of adjective-noun agreement, for example, the agreement occurs within a phrase, but in case of subject-verb agreement, for instance, the agreement crosses phrase boundary. So the learners have to know that there might be a phrase boundary between agreeing elements.

[5] A similar idea was proposed in Christophe et al. (1997). It was proposed that young infants recognize function words early and that can help segmentation and categorization of neighboring words ("function word stripping hypothesis": Christophe et al. 2007, Hicks 2006). It has further been proposed that recognition of function words can help not only lexical but also syntactic (phrasal) segmentation and categorization (Christophe et al. 1997).

2.3.4 Cross-sentential cues to phrase structure

We reviewed studies that investigated what cues learners might employ to figure out the constituent structure of a language. The cues the previous studies looked at were overt and local in the sense that they were internal to the sentence themselves, such as prosody and function words. Morgan, Meier & Newport (1989) propose that non-local cues might also be available to the learner for figuring out the constituency. These are cross-sentential cues that can only be detected if compared with another sentence. They note that a number of transformational rules in natural language are structure-dependent. For example, only phrasal constituents can be substituted by pro-forms.

(34)

a. Paul likes to go to the movies and John does so too

b. *Paul likes to go to the movies and John does so to the concerts

(35)

a. The man with the glasses is tall

b. He is tall

c. *He with glasses is tall

Similarly, only phrasal constituents can undergo movement.

(36)

a. John likes that girl over there

b. That girl over there, John likes

c. *That girl, John likes over there

Accordingly, Morgan et al. (1989) created an artificial language that incorporated substitution and movement rules. The basic phrase structure rules and PS tree are given in (37) and Figure 8.

(37) S → AP + BP + (CP)

AP → A + (D)

BP → { E / CP + F }

CP → C + (D)

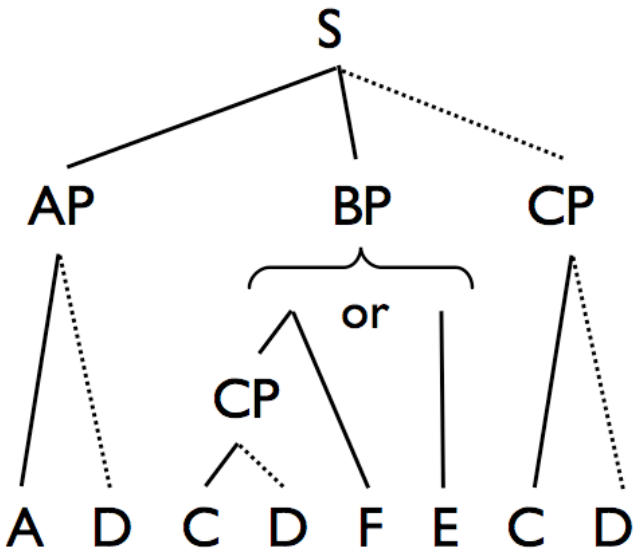


Figure 8: PS tree of the artificial language in Morgan et al. (1989)

Adult subjects were randomly assigned to one of the three input conditions. The input of Condition 1 included only the sentences drawn from the base language, which involved no substitution. The input of Condition 2, on the other hand, included two new transformational rules (38) that allowed for a constituent to be replaced by a proform.

(38) a. A + (D) → "ib"

b. C + (D) → "et"

In the input, the sentences with proforms were shown along with the base sentence as in Figure 9. By comparing the two sentences, subjects could figure out that "ib" substituted for BIF and PEL, in the example in Figure 9. Condition 3 included a transformational rule to allow a constituent to move to the front of the sentence.

(39) AP – BP – (CP) → BP + AP – (CP)

Again, the permuted sentence was shown alongside the base sentence as in Figure 9 in the input. In the example of Figure 9, by looking at how SOG FAC is moved to the front of the sentence, you could figure out that SOG FAC is a constituent.

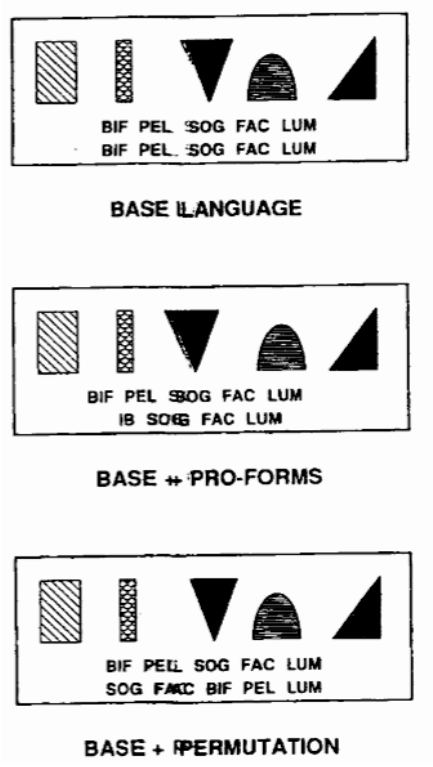


Figure 9: Examples of input stimuli

The subjects in Conditions 2 and 3 performed significantly above chance on the constituency tests, while subjects in Condition 1 were at chance. These results suggest that non-local cues such as substitution or permutation can be a cue to phrase structure. One caveat is that the transformed sentence (substituted or permuted) was always shown together with the corresponding base sentence, as in Figure 9. This made the comparison between the base and transformed sentences obvious. Morgan et al. (1989) also ran a pilot follow-up study where the base and transformed sentences were shown separately. In this case, the subjects who were exposed to transformed sentences in input did not learn any better than the subjects who were

only exposed to the base language in input. Morgan et al. (1989) therefore conclude that cross-sentential cues such as substitution and permutation serve as a cue to phrase structure only when the related sentences are presented side by side. Nevertheless, in the current paper, we are going to show that it is possible to learn phrase structure on the basis of distributional cues such as permutation and substitution, without presenting the related sentences side by side.

2.3.5 Predictive dependencies as a cue to phrase structure

Saffran (2001) and Saffran, Hauser, Seibel, Kapfhamer, Tsao & Cushman (2008) were concerned with a similar question as we are in this dissertation. Given the linear strings of words as input, how do children arrive at the hierarchical phrase structure representation? In the series of experiments we reviewed above (Morgan & Newport 1981, Morgan et al. 1987, 1989), the successful learning of phrase structure was achieved only when there were additional correlated cues, such as prosody and function words. Saffran (2001) and Saffran et al. (2008) propose that, in addition to those supplementary cues, other cues exist within the PS rules themselves, namely the dependencies between words. In these experiments, the artificial language used in the above experiments (Morgan & Newport 1981, Morgan et al. 1987, 1989) was slightly adapted as follows.

(40) S → AP + BP + (CP)

AP → A + (D)

BP → { E / CP + F }

CP → C + (G)

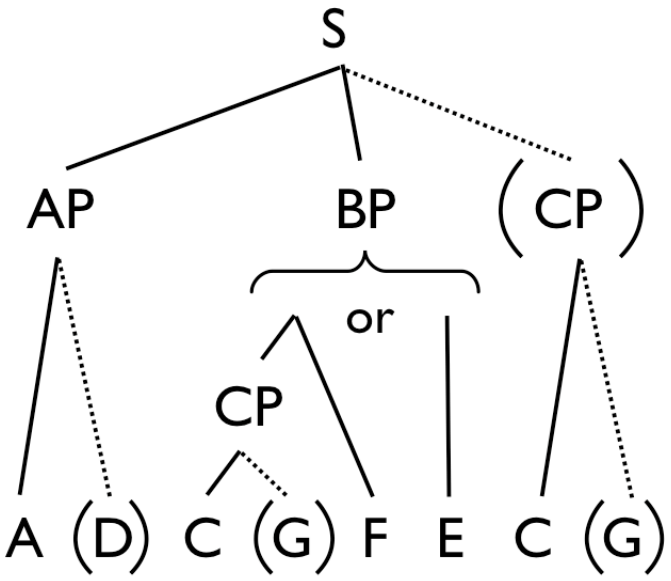


Figure 10: PS tree for the artificial language in Saffran (2001)

The only change was that in the previous grammar, CP consisted of a C and an optional D, which was also present in another phrase (i.e. AP), whereas in this grammar, CP consists of a C and an optional G, which is not present in other phrases. In this way, there were complete dependencies between A and D, and C and G. These dependency relations were the crucial predictive pattern in this language. The occurrence of a D word invariably predicts the occurrence of an A word, however, the occurrence of an A word does not necessarily predict the occurrence of a D word. Such unidirectional predictability is observed in natural languages. For example in English, whenever there is a determiner (e.g. *a, the*), there is a noun (*a man, the cat*),

while the existence of a noun does not always indicate the occurrence of a determiner (e.g. *men, cats*).

The stimuli were presented auditorily. Adult subjects were divided into three groups – two experimental groups and one control group. In one of the experimental groups, the intentional condition, subjects were given explicit instruction to learn the rules of the nonsense language. In the incidental condition, the primary task was to color on the computer while the nonsense language played in the background. On all test items, both intentional and incidental groups outperformed the control group.

Given the success of the incidental condition with adults, Saffran (2001) tested children between the ages of 6 and 9 (Mean = 7 years 7 months) on the same material. Again, the main task was coloring on the computer. The same test items were used. The children in the experimental group significantly outperformed the control group, although the effect was smaller compared with the adults' data.

Saffran et al. (2008) tested 12-month-old infants on the same artificial language, using the head-turn preference procedure. Half of the infants were familiarized with the artificial language described in (40) that exhibited predictive dependencies. For example, whenever there is a D word, there must be an A word preceding it, and whenever there is a G word, there must be a C word in front of it. Such predictive dependencies only occurred within a phrase, thus giving rise to the TP peaks and dips at phrase boundaries. This condition was called "predictive" condition. The other half of the infants were familiarized to a "non-predictive" language. The non-predictive language lacked the aforementioned predictive dependency within a phrase. In this language, any word in a phrase could be optional,

thus the occurrence of a word in a phrase did not predict the occurrence of another word. The 12-month-old infants in Saffran et al. (2008) listened longer to the ungrammatical test sentences than the grammatical test sentences, but only in the predictive condition and not in the non-predictive condition.

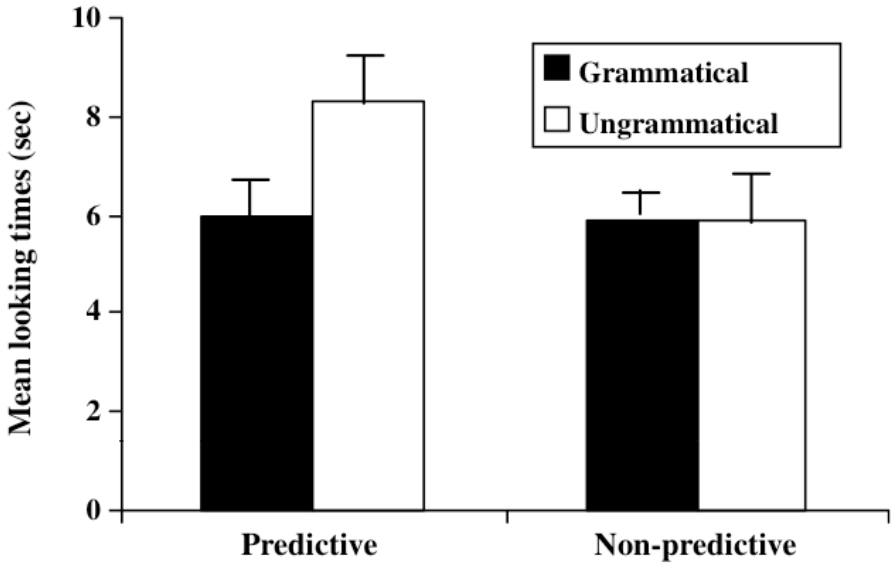


Figure 11: Results from the infant experiment in Saffran et al. (2008)

Saffran (2001) and Saffran et al. (2008) argue that these results suggest that the predictive dependencies within phrases alone are sufficient for learners to detect the phrasal units. When the presence of one element predicts the presence of another, phrases are easily detected. However, the languages learned by the subjects in these experiments could be represented by finite state grammars like the following.

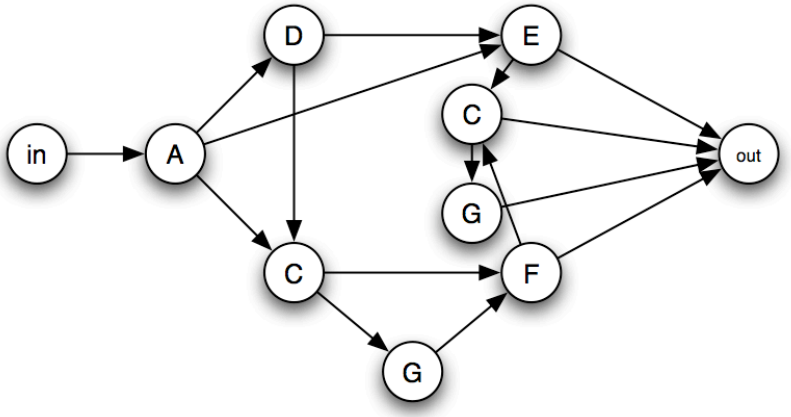


Figure 12: FSA of the predictive language in Saffran (2001)

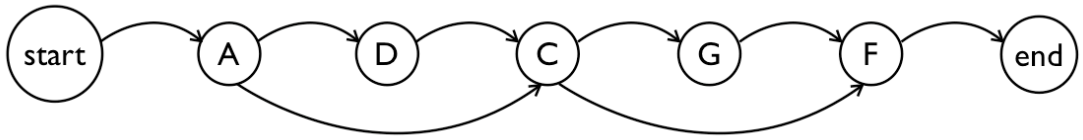


Figure 13: FSA of the predictive language in Saffran et al. (2008)

For example in Saffran (2001), in the fragment test, subjects chose *CG* over *FC*, and *CGF* over *GFC.* It is claimed that it was because learners formed a hierarchical phrase structure representation as in Figure 10. But subjects could have chosen *CG* over *FC,* simply because there was a strong backward dependency between *C* and *G,* but not between *F* and *C.* This kind of probabilistic FSA could explain the learning in Saffran (2001). Similarly, since the artificial languages in Saffran et al. (2008) can also be represented by a finite state grammar like in Figure 13, it has not yet been shown that infants learned a hierarchical phrase structure. All that the infants learned could be an FSA like above, which indicates that what they learned is the linear order

of words in this artificial language. This is because the grammatical test sentences obeyed the linear order of the language, whereas the ungrammatical test sentences violated the linear order. Furthermore, the ungrammatical test sentences involved novel structures that were never seen by the infants, whereas the grammatical test sentences involved the structures that were already exhibited in the familiarization period. Although the actual word strings were new, at the category level, the grammatical test sentences were not new, so it is impossible to conclude that the infants in Saffran et al. (2008) were extending their generalizations to a novel structure. If one of the test sentences is completely novel and ungrammatical, and the other has a very familiar structure, it is not surprising that infants were able to distinguish the two. If the test items included a transformational test like the ones in Morgan et al. (1989), it would have been a more powerful assessment for the subjects' knowledge of constituency, and of their deductive power.

2.3.6 Transitional probability as a cue to phrase structure

In Saffran et al. (1996a), it was shown that 8-month-old infants can detect word boundaries based on the transitional probabilities calculated over syllables. Transitional probability is the degree to which one element predicts the following element. The forward and backward transitional probabilities are calculated as follows.

(41) Forward transitional probability

$$\text{Probability of } Y \mid X = \frac{\text{frequency of } XY}{\text{frequency of } X}$$

(42) Backward transitional probability

$$\text{Probability of } X \mid Y = \frac{\text{frequency of } XY}{\text{frequency of } Y}$$

In a recent study, Thompson & Newport (2007) investigated whether learners can use such TPs, calculated over words, to discover phrasal boundaries in an artificial language. Now, one might wonder whether the transitional probability was at work in the above studies we just reviewed. Thompson & Newport calculated both forward and backward TPs between word classes in those studies. In Morgan & Newport (1981) and Morgan et al. (1987, 1989), neither forward nor backward TPs were informative as to the location of phrase boundaries. That is, the TP was neither higher within phrases nor lower across phrases. Even though the grammars in those studies exhibited key features such as optionality, repetition, substitution and movement, other factors such as optionality of elements within single phrases worked against them. For example, in the PS rules of Saffran (2001) in (40), the D word is optional within AP, whereas the A word is obligatory. This makes the backward TP between A and D always 1.0 (because whenever there is a D, there is an A). This is good, since A and D form a constituent. However, since D is optional, the forward TP between A and D is 0.5. It means that half the time A is directly followed by an element that is from another constituent. This kind of optionality within a single

constituent made the whole TP pattern uninformative. This means that no previous literature has shown whether the transitional probability can signal phrasal boundaries.

Accordingly, Thompson & Newport (2007) created a miniature artificial language that was made up of word classes, A, B, C, D, E, and F. Each word class had three lexical items. The words were further grouped into phrasal units [AB], [CD] and [EF]. If all the sentences in the language were canonical sentences as in (43), then both the TP within phrases (e.g. AB) and the TP across phrasal boundaries (e.g. BC) would be 1.0.

(43) A B C D E F

However, natural language has ways in which sentences differ from canonical ones. First, some phrases can be optional as follows.

(44) a. The box on the counter is red

b. The box is red

In the case of this artificial language, imagine that the phrase CD is optional and is dropped half of the time.

(45) A B E F

Now, the TP between phrases (e.g. BC) is 0.5, while within phrases is still 1.0. Second, some phrases appear more than once in a sentence.

(46) $[_{NP}$ The cat] chased $[_{NP}$ the mouse]

Imagine that the phrase AB appears twice, provided that we are computing TPs over word categories.

(47) A B C D E F A B

In (47), the TP between phrases (e.g. BC) is reduced to 0.5, while within phrases is kept constant. Third, some phrases undergo movement.

(48) a. [The cat] chased [the mouse]

b. [The mouse] is chased by [the cat]

Imagine that the phrase EF moves to the front half of the time.

(49) E F A B C D

Again, the TP across the phrase boundary (e.g. DE) is reduced to 0.5, while that of within phrases is still 1.0. In this way, because the rules of syntax manipulate constituents, there is a statistical byproduct of these manipulations, which are TP

peaks, and dips within and across phrases, respectively. The three key features taken up in Thompson & Newport (2007) are optionality, repetition and movement.

A female speaker read the sentences aloud with a list intonation. Adult subjects were divided into an experimental group and a control group. The input to the experimental condition incorporated either one of the three key features – optionality, repetition and movement of phrasal constituents – or all the three. These manipulations served to create TP peaks within phrases and TP dips between phrases. The input to the control condition also incorporated the three features, only in this case, any adjacent elements were allowed to be optional, repeated and moved. This served to flatten the TP peaks and dips. These manipulations resulted in the TP patterns in Table 1.

Table 1: TPs when all three key features were incorporated (Thompson & Newport 2007)

	A-B	B-C	C-D	D-E	E-F
Experimental Condition	1.00	0.33	1.00	0.22	1.00
Control Condition	0.67	0.71	0.58	0.59	0.47

The experiment extended for 5 days. Each day, subjects were exposed to the auditory input for 20 mins. Tests were administered on Days 1 and 5. The phrase test consisted of pairs of words, one of which was a phrasal constituent in the language (e.g. AB) and the other was a legal sequence in the language but not a constituent (e.g. BC). The task was a forced-choice on a computer and the subjects were told to choose the sequence that sounded "more like a group or unit from the language."

The adult subjects in the experimental condition chose the constituents over non-constituents significantly more often as early as Day 1 and on Day 5 (the testing was administered only on Days 1 and 5). The control group performed at chance. The effect was much larger when all the three key features were incorporated in the input (Figure 14) than when only one of the features was used, for example, when only the optionality was included (Figure 15).

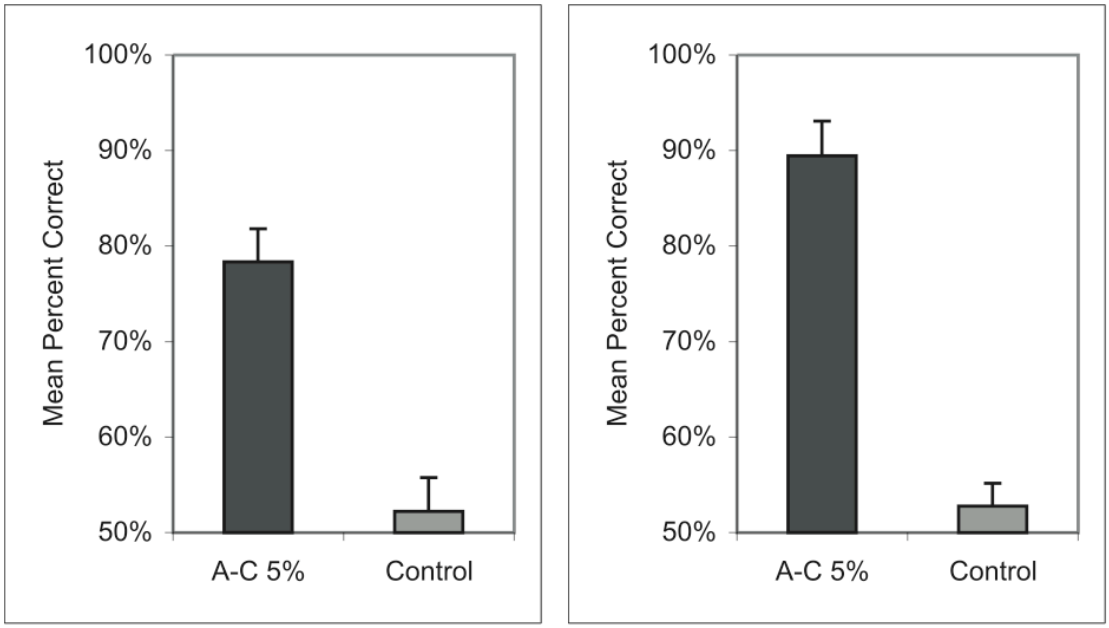


Figure 14: Results of the Phrase Test on Day 1 (left) and Day 5 (right), with all features incorporated (Thompson & Newport 2007)

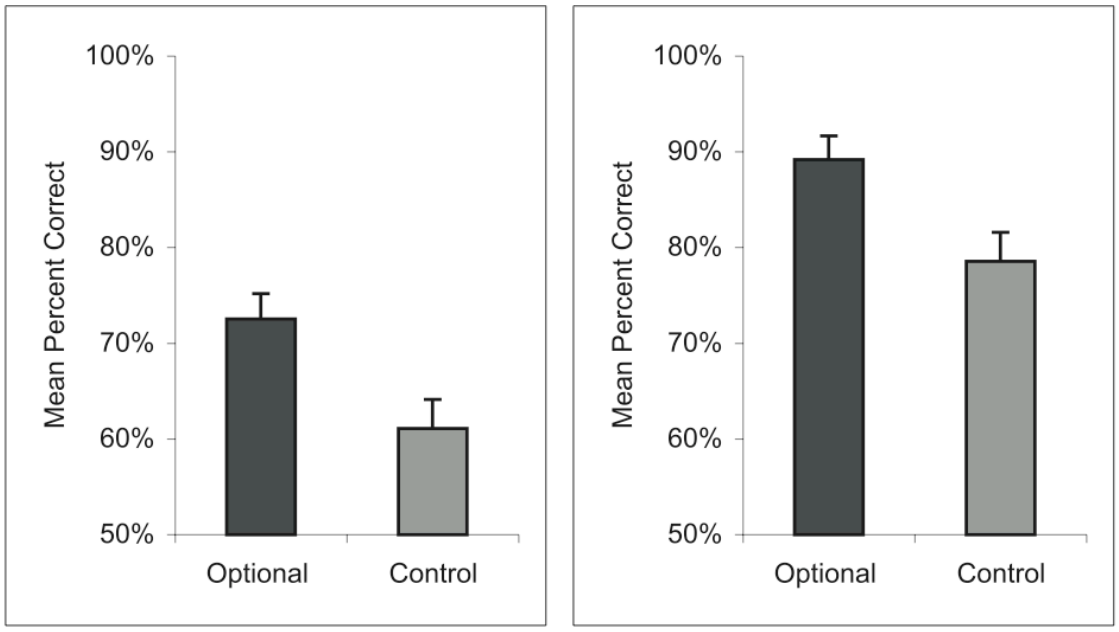


Figure 15: Results of the Phrase Test on Day 1 (left) and Day 5 (right), when incorporating only the optionality (Thompson & Newport 2007)

Based on these results, Thompson & Newport (2007) conclude that the subjects successfully learned the phrasal groupings based on the TP statistics and that distributional information can help learning of phrase structure. However, one could ask whether it was really the transitional probability that was in play, or some frequency effect. Subjects could have been choosing a word sequence over another simply because it appeared more frequently together than the other one. Thompson & Newport (2007) examined this and found no positive correlation between the right answers and the co-occurrence frequency. To illustrate, imagine the phrase test consisted of a pair SOT FAL and FAL SIB, and that the right answer (the syntactic constituent) is FAL SIB. In the presentation set, the sequence SOT FAL (the wrong

answer) appeared four times more frequently than FAL SIB. This shows that there was no frequency effect.

What Thompson & Newport (2007) showed is that the computation of transitional probability statistics can help learners with phrasal segmentation. However, what was learned in that study was phrasal bracketing that had a flat structure as in Figure 16, not a hierarchical structure.

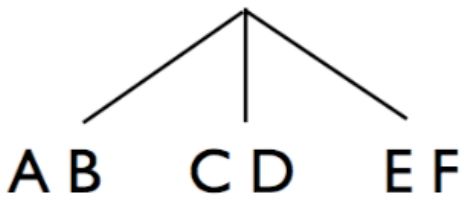


Figure 16: Sentence structure used in Thompson & Newport (2007)

But, having the correct *hierarchical* phrase structure is crucial for any later syntactic or semantic development. Then, can the transitional probability also be a cue to the hierarchical phrase structure?

2.4 The present experiments

Above, we reviewed studies that showed that learners can use various cues to learn artificial languages. Generally in artificial language studies, words are not associated with particular meanings. One question then is whether phrase structure can be learned independent of semantics. If the meanings of the words in a sentence

are given, at least some parts of the phrase structure should be attained for free. For example, if you have a sentence like (50),

(50) That boy likes this little puppy

and if you knew the meaning of "this little puppy", since these three words together denote a single object, you will naturally consider them as a unit. However, this mechanism does not seem to work for the VP. Without knowing whether this language is an SVO or OVS language, you cannot tell whether the object of the verb "likes" is "that boy" or "this little puppy". In this way, even though knowing the meanings of words would be helpful for building the structure, it is unlikely that you can learn the meanings of most words without learning the syntactic structure first. Therefore, we propose that there must be some way to acquire phrase structure that does not require prior acquisition of the word meanings. This paper examines one such mechanism.

One issue that has not been brought up in previous artificial language studies is whether the statistical learning mechanism interacts with anything other than the input the learners receive. Here, we present three possibilities. We realize that these three possibilities might be extremes of a spectrum and that there is probably a range of possibilities in between. However, we will present those three for the sake of brevity and clarity of the argument.

One possibility is that what is learned through statistical learning is solely based on the input signal (in this case, distributional information), and that statistical

learning does not interact with other constraints that the learners might already have. We will call this first possibility "Limited" Hypothesis. Second possibility is that what is learned through statistical learning is not limited to what is observed in the input, but the generalizations the learners form are bounded by some constraints in a predictable way. We will call this "Beyond and Constrained" Hypothesis. According to this hypothesis, statistical learning interacts with knowledge that was not obtained from the observed input. An example of being "constrained in a predictable way" would be movement of a constituent which is a natural rule in languages. In other words, under this view, the generalizations the learners form are compatible with what is possible in natural languages. Finally, a third possibility is that learners generalize beyond what they see in the input, but their generalizations are not necessarily constrained in a predictable way. We will call this third possibility "Beyond and Unconstrained" Hypothesis. This view would predict that the generalizations the learners form can go beyond the observed input and do not necessarily have to be compatible with what is allowed in natural languages. An example of this might be something like movement of a non-constituent, which is unnatural in natural languages, but if a learner is unconstrained, this is a logical possibility.

In order to find out which of these hypotheses' predictions would be borne out, we conducted seven original experiments with human adult subjects, infants and computational network simulations.

Chapter 3: Adult Experiments

In this chapter, we present two experiments with adult participants that investigated whether the representations are part of the learning system prior to the experience, and what the deductive consequences of distributional learning are. We look into whether statistical learning interacts with antecedently known constraints, or whether learners create an illusion of structure entirely based on observed input alone. By manipulating what is included or excluded from the exposure set, we can control to see whether certain information is necessary in the input for a learner to deduce the target structure or not. If adults cannot generalize beyond the received input, then it would suggest that the deductive power of a learner is limited to the observed distributions, whereas if adults can generalize beyond the input, then it would suggest that the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone.

A more immediate question in this chapter is, given that previous studies (Thompson & Newport 2007) have only shown that transitional probabilities serve as a cue to phrasal groupings, whether the statistical cues to the multiply embedded hierarchical structures can be detected by learners. In addition, these experiments ask whether adults can learn the constituency of an artificial language without any prosodic or semantic cues.

3.1 Experiment 1 (Adult 1)

3.1.1 Description of the linguistic systems

Two miniature artificial languages – Grammar 1 and Grammar 2 – were created. While the artificial language in Thompson & Newport (2007) contained phrases with a flat structure as in Figure 17, the artificial language in Morgan & Newport (1981), Morgan et al. (1987, 1989) and Saffran (2001) did contain phrases with internal hierarchical structures. Thus, that language was adapted here as our Grammar 1.

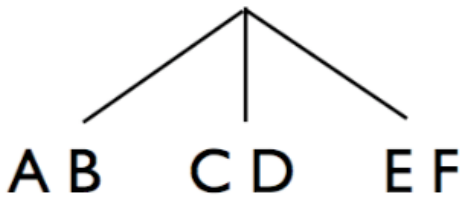


Figure 17: Sentence structure used in Thompson & Newport (2007)

The control group in Thompson & Newport (2007) failed to learn the phrasal bracketing. But was this because the transitional probabilities were not an informative cue for the phrase structure, and so there was no statistical cue? In the control condition in Thompson & Newport (2007), not only constituents but also non-constituents could undergo operations such as movement, substitution, repetition and optionality. This led to lack of TP peaks in dips in the control familiarization set (e.g., mean of TPs within a phrase = 0.57, mean of TPs across phrases = 0.65). In this way, the TPs were not an informative cue for phrase boundaries. On the other hand, one

could argue that the failure to learn in the control group could be due to the fact that learners could not find any grammar that would generate the sentences. Since the presented sentences were so random in that both constituents and non-constituents were operated on, there was no single grammar that could generate all the sentences. Such lack of grammar may have caused the control subjects to fail. In our experiments, we wanted to avoid this confound of causes. Therefore, we created a second grammar, Grammar 2, to serve as our control. In this way, both groups (people who hear Grammar 1 as input and people who hear Grammar 2 as input) would have a grammar that can generate the sentences. So failure to learn the language would not be due to lack of grammar that generates the sentences.

The two grammars share the identical word classes and lexical items, which were adapted from Thompson & Newport (2007). Each word class contained three nonsense lexical items.

Table 2: Nonsense words assigned to each word class

Word Class	A	B	C	D	E	F
	KOF	HOX	JES	SOT	FAL	KER
	DAZ	NEB	REL	ZOR	TAF	NAV
	MER	LEV	TID	LUM	RUD	SIB

The basic phrase structure rules and phrase structure trees for Grammar 1 and Grammar 2 are given below.

(51) PS rules for Grammar 1

S' → S + (CP)

S → AP + EP

$$AP \rightarrow \left\{ \begin{array}{l} A + B \\ ib \end{array} \right\}$$

$$EP \rightarrow \left\{ \begin{array}{l} CP + E \\ F \end{array} \right\}$$

$$CP \rightarrow \left\{ \begin{array}{l} C + D \\ et \end{array} \right\}$$

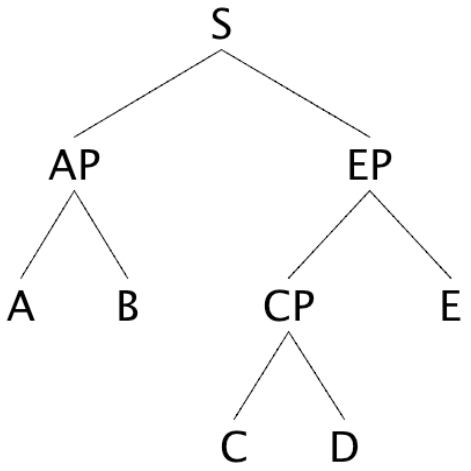


Figure 18: PS trees of the basic sentence in Grammar 1

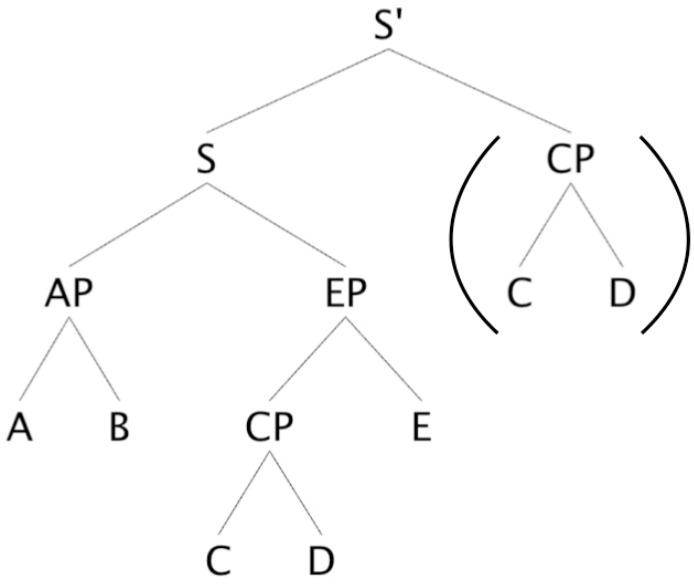


Figure 19: PS tree in Grammar 1 showing optionality and repetition

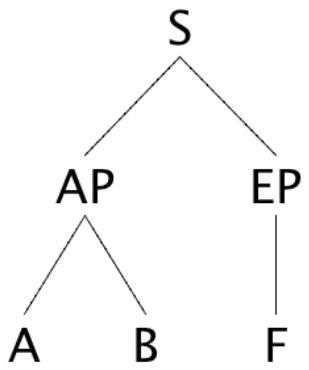


Figure 20: PS trees in Grammar 1 showing substitution

(52) PS rules for Grammar 2

S' → S + (BP)

S → AP + DP

AP → { A + BP / F }

DP → { D + E / ib }

BP → { B + C / et }

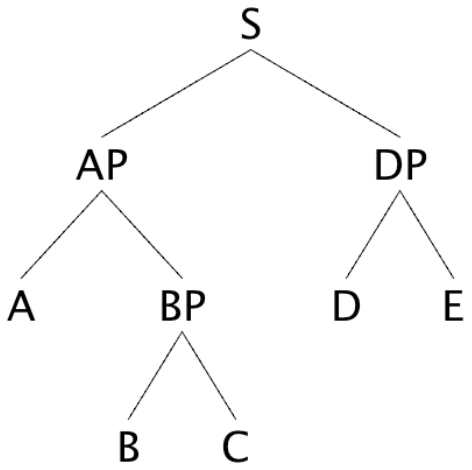


Figure 21: PS tree of the basic sentence in Grammar 2

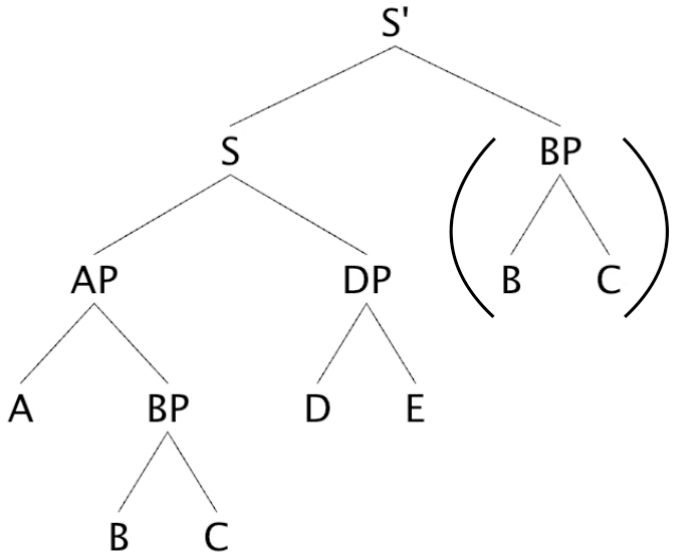


Figure 22: PS tree in Grammar 2 showing optionality and repetition

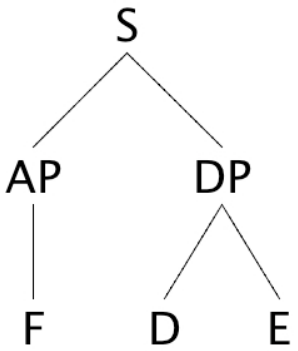


Figure 23: PS trees in Grammar 2 showing substitution

Grammars 1 and 2 are maximally similar and minimally different – contrasting only in constituent structure. In particular, the canonical sentences in both grammars are identical – *A B C D E*. The only difference is the phrase structure. For example, while AB is a constituent in Grammar 1, it is not in Grammar 2. CD is a constituent in Grammar 1 but not in Grammar 2. On the other hand, BC and DE are both constituents in Grammar 2 whereas they are not in Grammar 1.

In addition, the grammars also display nested hierarchical structure. In Grammar 1, a phrasal unit EP consists of an E word and another phrase CP, which in turn consists of C and D. The whole EP can also contain just an F word. Likewise in Grammar 2, the phrase AP contains an A word plus a BP, which contains B and C words. The whole AP can simply be represented by an F word.

These grammars incorporate the manipulations featured in Thompson & Newport (2007) such as repetition and optionality. The optional CP in Grammar 1 and BP in Grammar 2 bring about the repetition and optionality. Take Grammar 1 as an example. The basic sentence structure of Grammar 1 is ABCDE. If all the

sentences in the language had the structure canonical sentences as in ABCDE, then the TPs would not be a very informative cue for constituency, because both the TP within phrases (e.g., AB) and the TP across phrasal boundaries (e.g., BC) would be 1.0. Natural language has ways in which sentences differ from canonical ones, however. First, some constituents can be optional as follows.

(53) a. The boy [from the creek] met Steven Spielberg

b. The boy met Steven Spielberg

In the case of Grammar 1, imagine that the phrase CD is optional and is dropped half of the time as in (55), and there are two types of sentences in this language.

(54) A B C D E

(55) A B E

Now, the forward TP between phrases (e.g., BC) is 0.5, while within phrases (e.g., AB) is still 1.0. Second, some constituents (at the category level) appear more than once in a sentence and can be repeated.

(56) [$_{NP}$ The boy from the creek] met [$_{NP}$ Steven Spielberg]

Imagine that the phrase AB appears twice in Grammar 1, provided that we are computing TPs over word categories.

(57) A B C D E A B

Now, the backward TP between phrases (e.g., BC) is reduced to 0.5, while TP within phrases (e.g., AB) is still kept constant. In this way, repetition and optionality in natural languages create TP peaks and dips that serve as informative cues to phrase structure.

The artificial grammars used in our experiments also display another feature observed in natural languages, which is substitution of a phrasal constituent by a proform, just like in Morgan et al. (1989). For example in Grammar 1, the constituent AP, which usually consists of A and B words, can also be replaced with a pronoun-like element *ib*. Similarly, the CP in Grammar 1, which normally consists of C and D words, can be substituted by a profrom *et*. The proforms are borrowed from Morgan et al. (1989). In Grammar 2, the same proforms *et* and *ib* substitute for different constituents, BP and DP respectively. Substitution creates TP peaks and dips too. For example in Grammar 1, the constituent CD can be replaced by a proform *et,* as in (59).

(58) A B C D E

(59) A B et E

If there are these two types of sentences in the input, then the TP across phrases is lower (e.g., BC = 0.5) than TP within phrases (e.g., AB = 1.0). This makes TP pattern very informative .

Finally, the grammars also incorporate movement rules just like the ones in Morgan et al. (1989). The movement operation is be captured by the addition of following optional phrase structure rules.

(60) Optional PS rules added for Grammar 1

S' → EP + S

S → AP

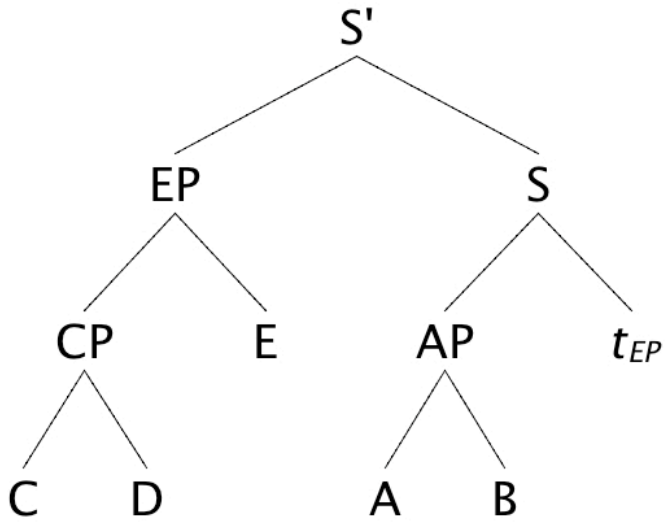


Figure 24: PS trees involving movement in Grammar 1

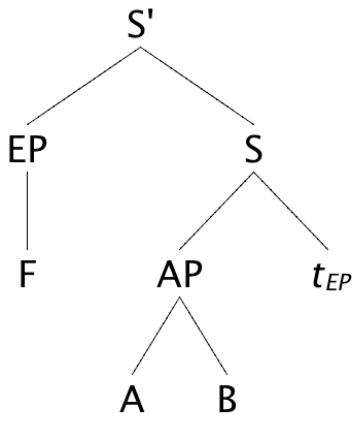


Figure 25: PS trees involving movement and substitution in Grammar 1

(61) Optional PS rules added for Grammar 2

S' → DP + S

S → AP

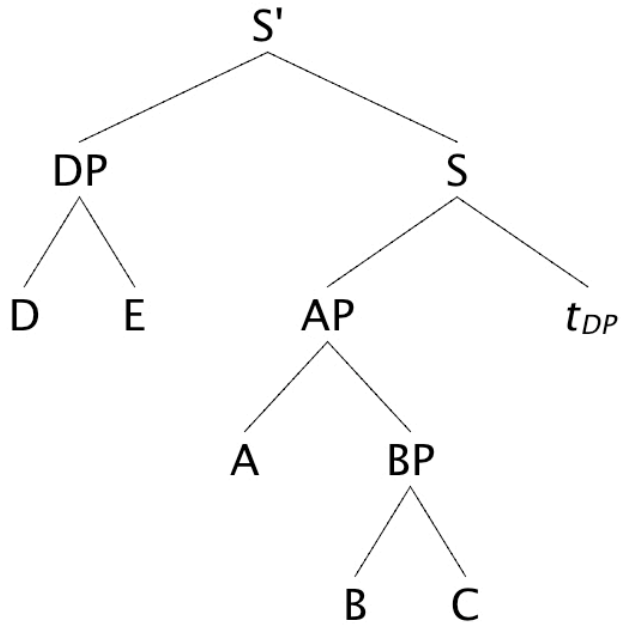


Figure 26: PS trees involving movement in Grammar 2

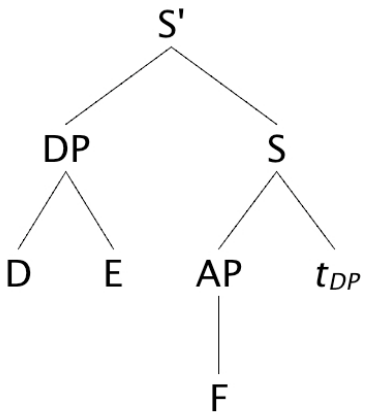


Figure 27: PS trees involving movement and substitution in Grammar 2

In Grammar 1, the EP can be moved to the front, while in Grammar 2, the DP can be moved. Again, this creates peaks and dips in TP that signal phrase boundaries. Imagine you have following two types of sentences in Grammar 1 input.

(62) A B C D E

(63) C D E A B

Now the TP across phrases (e.g., BC) is 0.5, but TP within phrases (e.g., AB) is 1.0.

In this way, these grammars included four types of manipulations which (a) made certain constituents optional, (b) allowed for the repetition of certain constituents, (c) substituted proforms for certain constituents and (d) moved certain constituents. Incorporating all these manipulations resulted in the higher TPs between words within phrases compared with the TPs across phrases. Within a phrase, the TP

is always 1.00. The patterns of TP based on all the possible sentences generated by the grammars are given in tables below.

Table 3: Transitional probabilities for all sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.81	1.00	0.51
Backward TP	1.00	0.45	1.00	0.90

Table 4: Transitional probabilities for all sentences in Grammar 2

	A-B	B-C	C-D	D-E
Forward TP	0.90	1.00	0.21	1.00
Backward TP	0.51	1.00	0.42	1.00

3.1.2 Method

Participants

Forty-four native speakers of English participated in Experiment 1. The participants were undergraduate students at the University of Maryland, gave informed consent prior to participating and received monetary compensation. Twenty-two participants were randomly assigned to hear Grammar 1 during the familiarization and the other 22 were assigned to Grammar 2.

Material

Both Grammars 1 and 2 generate finite languages without recursion that generate a total number of 7260 possible sentences each. 80 sentences were picked as the presentation set. Two sentences (2.5%) were the canonical sentence type (ABCDE) in both grammars. The TP patterns of the presentation set are given below.

All 80 sentences were randomized. The sentence types and 80 sentences that appeared in the presentation set are shown in Appendix A.

Table 5: Transitional probabilities for 80 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.24	1.00	0.25
Backward TP	1.00	0.19	1.00	0.34

Table 6: Transitional probabilities for 80 input sentences in Grammar 2

	A-B	B-C	C-D	D-E
Forward TP	0.33	1.00	0.15	1.00
Backward TP	0.18	1.00	0.16	1.00

Recording

Each word token was individually recorded into a Marantz PMD660 portable solid state recorder with a head-mounted Sennheiser HMD 280-13 microphone. A female speaker, who was blind to the nature of the experiment, read each word token in a list intonation. The recorded audio files were transferred into the Audacity sound editor. The words were concatenated into sentences with a 30 ms inter-word interval. The sentences lacked any prosodic cues to phrase boundaries. All the 80 sentences were then transferred into Psyscope 1.2.5 PPC program (Version X B45Dep) and concatenated with an intersentence interval (isi) of 1400 ms. The recorded block of 80 sentences lasted approximately 6 min. The 80 sentences were then randomized and repeated 6 times in a random order to form an input sound file of approx. 36 min duration. A sample sound file is available at http://ling.umd.edu/~eri/expt1_sound_sample.wav.

Procedure

The experiment was administered individually using a Psyscope 1.2.5 PPC program inside a small soundproof room with an iMac and Sennheiser HD 580 precision headphones. Given the success of the incidental condition in Saffran (2001), a similar procedure was adopted. Participants were asked to draw using colored pencils and paper, while listening to a nonsense language. They were told nothing about the structure of the language. They were informed that they would be tested on the nonsense language later, but not told about the aspects of the language that would be tested.

Participants were randomly assigned to either Grammar 1 or Grammar 2. The participants assigned to Grammar 1 heard the Grammar 1 input sentences during the familiarization. The Grammar 2 participants heard the Grammar 2 input sentences. Each participant heard the 80 sentences six times during the 36-min familiarization period. The 80 familiarization sentences were randomized each time by the Psyscope program. They then went through a practice period, where they were asked three practice questions, to familiarize themselves with the question-answering process. All the tests were forced-choice tests. Both Grammar 1 and Grammar 2 subjects received the identical 56 test items. There were 4 types of test items: *Fragment test, Movement test, Substitution test* and *Movement-plus-substitution test*, all of which are described in detail below. All the test items were randomized each time in the Psyscope program. Participants saw the following instruction on the computer screen.

(64) "In each trial, you will hear a pair of word-sequences - 1 and 2.

Your task is to respond, as accurately as you can, which of the two sequences belongs to the artificial language you just heard.

Press F if the 1st sequence belongs to the language.

Press J if the 2nd sequence belongs to the language."

Fragment Test

The Fragment Test was designed to assess the extent to which participants represented the input language in terms of phrasal groupings. The test was forced-choice and consisted of 16 items, 4 items testing each of the four fragment types. Each trial consisted of two fragments, one that was a phrasal constituent in the input language and the other that was often a legal sequence but not a constituent in the input language. The four fragment types that were tested are given in (65). The first two types are 2-member fragments and the last two are 3-member fragments.

(65) Fragment test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	AB	BC
2	CD	DE
3	CDE	ABC
4	ABF	FDE

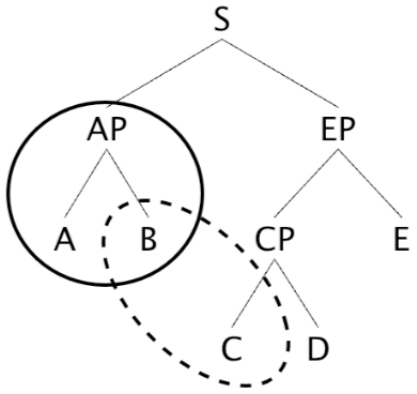


Figure 28: Grammar 1 (AB vs. BC)

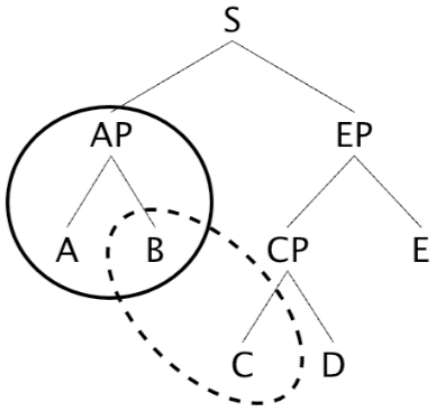


Figure 29: Grammar 2 (AB vs. BC)

A constituent in Grammar 1 (e.g. AB) is not a constituent in Grammar 2. Similarly, a non-constituent in Grammar 1 (e.g. BC) is a constituent in Grammar 2, as in the figures above. Consequently, the correct answer for Grammar 1 was always the incorrect answer for the Grammar 2 condition, and vice versa. Each fragment type contained 4 items. All the test items are given in Appendix C.

If the participants learn that the 2-member fragment (e.g. *CD*) is a constituent and that the 3-member fragment (e.g. *CDE*) is also a constituent, then, they must have learned a nested hierarchical structure as in Figure 30.

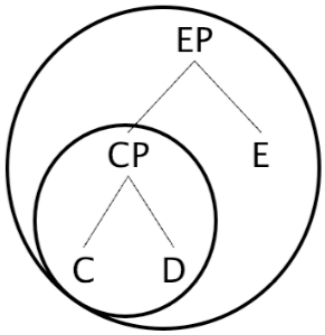


Figure 30: Internally nested hierarchical structure

To ensure that the performance on this test is a result of phrasal knowledge rather than frequency effects, we controlled the frequencies with which both groups of fragments appeared in the input. Specifically, *none* of the test items appeared in the input. Thus, the frequency with which both groups occurred was 0. They were all novel sequences. In this way, if learners attended only to sequential frequency, they would perform at chance. If their performance exceeds chance, it would indicate that they formed a higher-order phrasal representation.

The test items were concatenated using the same individual word token recordings as the input sentences. The pairs were presented with 1400 ms of silence between them. The test items were randomized each time and the correct answer was the first or the second equally often.

Movement Test

The Movement Test was designed to assess the extent to which participants allowed phrasal constituents to undergo a movement operation as opposed to non-constituents. This test was modeled on the transformational constituent test in Morgan & Newport (1981) and Morgan et al. (1987, 1989). The test was forced-choice and consisted of 16 items, 4 items testing each of the four sentence types. Each trial consisted of two sentences, one in which a constituent of the input language had been subjected to movement, and the other one in which a non-constituent of the input language had been subjected to movement, as in figures below. The correct answer for the Grammar 1 condition was always the incorrect answer for the Grammar 2 condition, and vice versa. The four sentence types that were tested are given in (66). All the test items are given in Appendix C. Again, none of the test sentences occurred during familiarization.

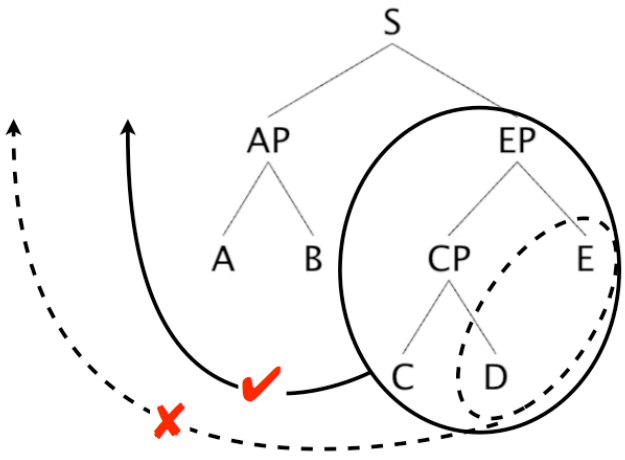


Figure 31: Grammar 1 (CDEAB vs. DEABC)

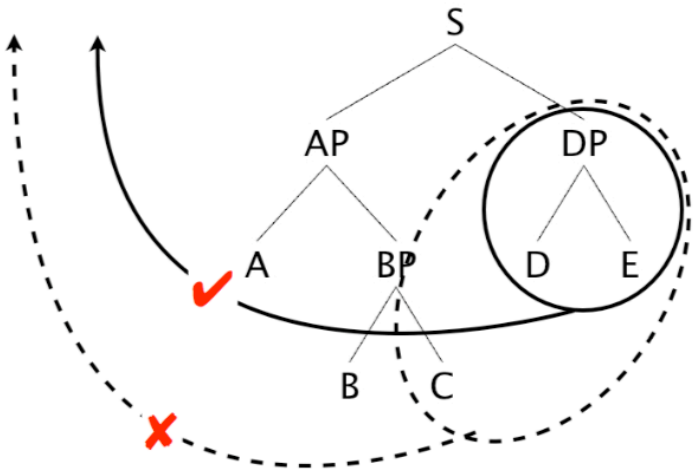


Figure 32: Grammar 2 (CDEAB vs. DEABC)

(66) Movement test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	CDEAB	DEABC
2	FAB	DEF
3	CDEABCD	DEABCBC
4	FABCD	DEFBC

Substitution Test

The Substitution Test was designed to assess the extent to which participants allowed phrasal constituents to be replaced by proforms *ib* and *et.* The test was forced-choice and consisted of 12 items, 4 items testing each of the three sentence types. Each trial consisted of two sentences, one in which a constituent of the input language was substituted for by a proform, and the other in which a non-constituent of the input language was substituted by a proform, as in Figure 33-Figure 34. The

correct answer for the Grammar 1 condition was always the incorrect answer for the Grammar 2 condition, and vice versa. The three sentence types that were tested are given in (67). All the test items are given in Appendix C. None of the test sentences occurred during familiarization.

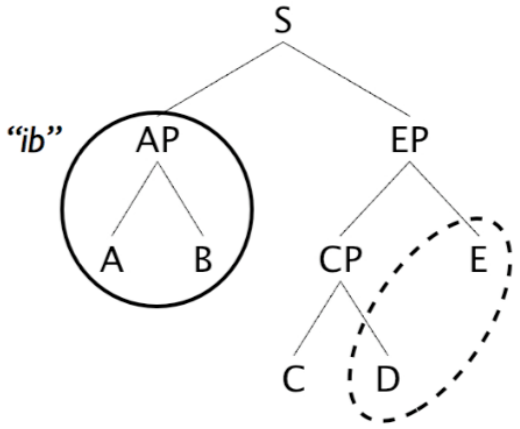


Figure 33: Grammar 1 (ib CDE vs. ABC ib)

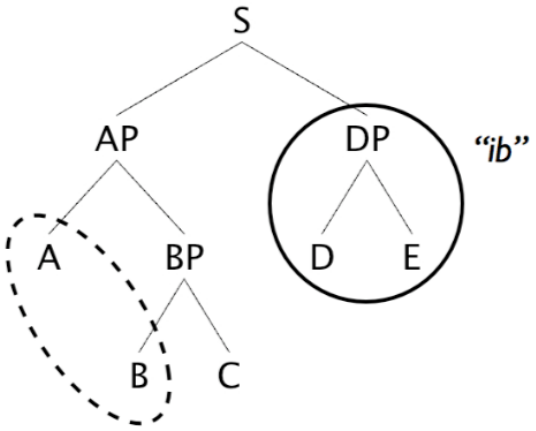


Figure 34: Grammar 2 (ib CDE vs. ABC ib)

(67) Substitution test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	ib CDE	ABC ib
2	AB et E	A et DE
3	ib et E	A et ib

Movement-plus-substitution Test

The Movement-plus-substitution Test was designed to assess the extent to which participants allowed phrasal constituents to be replaced by proforms, *ib* and *et,* and undergo movement. The test was forced-choice and consisted of 12 items, 4 items testing each of the three sentence types. Each trial consisted of two sentences, one in which a constituent of the input language was substituted for by a proform and moved, the other in which a non-constituent of the input language was substituted by a proform and moved, as in Figure 35-Figure 36. The correct answer for the Grammar 1 condition was always the incorrect answer for the Grammar 2 condition, and vice versa. The three sentence types that were tested are given in (68). All the test items are given in Appendix C. None of the test sentences occurred during familiarization.

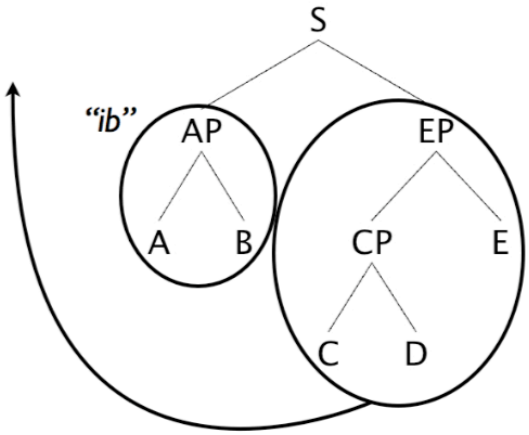


Figure 35: Grammar 1 (CDE ib)

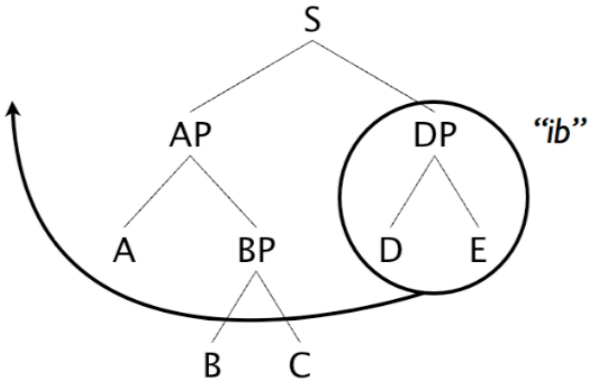


Figure 36: Grammar 2 (ib ABC)

(68) Movement-plus-Substitution test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	CDE ib	ib ABC
2	et EAB	DEA et
3	et E ib	ib A et

3.1.3 Hypotheses and predictions

Limited Hypothesis

We considered two distinct theories of learning. One was a learning theory in which the deductive power of a learner is limited to the observed distributions. Under this theory, learners do not come with a pre-determined set of possible structures or rules, and what learners do is to track the distributions and build an illusion of a structure entirely based on them, without any preconception of what is and what is not a possible structure. This learning theory can be expressed as a more concrete hypothesis with respect to our experiments, which is that when learners get certain input, they make generalizations based on only the input they get. We will call this "Limited" Hypothesis. According to this hypothesis, learners do not allow new structures that were not displayed in the input.

Beyond and Constrained Hypothesis

Another learning theory we considered is that a learner already knows an antecedently-specified range of possible representations, and statistics is merely used as a source of information that helps a learner select the correct grammar that derives the matching surface strings. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. This theory can be expressed as a more concrete hypothesis, which states that learners generalize beyond the input but the generalizations they form are bounded by some constraints in a predictable way. We

will call this "Beyond and Constrained" Hypothesis. This hypothesis proposes that learners' generalization extends to novel structures, as long as they are compatible with antecedently known constraints. An example of an antecedently known constraint would be something like movement of a constituent which is a natural rule in languages.

Beyond and Unconstrained Hypothesis

A third possibility is that learners generalize beyond what they see in the input but their generalizations are not necessarily constrained in a predictable way. We will call this third hypothesis "Beyond and Unconstrained" Hypothesis. An example of this might be something like movement of a non-constituent, which is unnatural in natural languages, but if a learner is unconstrained, this is a logical possibility.

Table 7: Table of hypotheses

	Deductive power of learner	*Nature of predetermined representations*
Limited Hypothesis	Limited to observed distributions	None
Beyond and Constrained Hypothesis	Beyond what can be derived from observed distributions	Limited by constraints found in natural language
Beyond and Unconstrained Hypothesis	Beyond what can be derived from observed distributions	Unlimited by constraints found in natural language

In order to simplify the argument, let us take the case of the movement test to talk about predictions that the above hypotheses make. In this experiment, the

familiarization input the participants receive includes movement sentences. And at test, they have a choice between a sentence which moved a constituent in their language and a sentence which moved a non-constituent. The structure of "correct" answer was already seen in the input although the actual strings of words of the test sentences were novel. And the structure of the "incorrect" answer was not seen in the input. Limited Hypothesis would predict that learners will correctly choose the consistent answer, since that has the structure that they have seen. According to this hypothesis, they do not allow new structures that they did not see in the input.

Beyond and Constrained Hypothesis would also predict that the learners will correctly choose the correct answer, since the consistent test item moved a constituent which is a natural operation in language. On the other hand, the participants would reject the incorrect answer, since it moved a non-constituent which is an impossible operation in natural languages.

Lastly, Beyond and Unconstrained Hypothesis would predict that the learners might allow both test sentences, since they can allow something that they did not see in the input and they are not bounded by a constraint that says you cannot move a non-constituent. It is possible that the generalization the learners form based on the input they get would be that you can move any two neighboring elements. For example, if you heard Grammar 1 as input, and if the generalization you make from that is you can move any two neighboring elements, then both test items CDEAB and DEABC would be licit, since both have neighboring two words moved. Consequently, at test, the participants would not choose one test item over the other, so the performance should be at chance.

In this way, the first two hypotheses (Limited Hypothesis and Beyond and Constrained Hypothesis) predict the same outcome, although the outcome would be caused by different reasons. The only hypothesis that make a unique prediction in this experiment is Beyond and Unconstrained Hypothesis.

Table 8: Predictions for Experiment 1

	Views	*Predictions*
Limited Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Adults will perform at chance

3.1.4 Results and discussion

The question that we were interested in in Experiment 1 was whether participants can learn the hierarchical phrase structure representation on the basis of transitional probability. If the subjects did acquire their input grammars, the Grammar 1 subjects should have learned constituency consistent with Grammar 1. On the other hand, the Grammar 2 subjects should have learned constituency consistent with Grammar 2, which is incompatible with the Grammar 1 constituency, since the two grammars have inconsistent constituent structures. In each trial for every test, one of the pair was the correct answer for Grammar 1, while the other was the correct answer for Grammar 2. Thus, if subjects learned the constituency, we predict that subjects in Grammar 1 would choose the correct answer for Grammar 1 significantly more often than the subjects in Grammar 2. Below, we report the percentage of times

Grammar 1 subjects chose the Grammar 1-compatible answers in contrast with the percentage of times Grammar 2 subjects chose the Grammar 1-compatible answers.

Grammar 1 vs. Grammar 2

Fragment Test

The participants in Grammar 1 chose the Grammar 1-consistent answers for the 2-member fragment tests 56% of the time, while the participants in Grammar 2 chose them 46% of the time. This difference was significant in a one-tailed independent samples *t*-test: $t(42) = 1.81$, $p = 0.039$. The one-tailed significance value is reported, because we have a specific directional prediction: we predicted that if subjects learned their input grammar, they would choose the answers that are compatible with their learned grammar. As for the 3-member fragment tests, the participants in Grammar 1 did not choose the Grammar 1-consistent answers (mean = 44%) significantly more often than the Grammar 2 participants (mean = 48%): $t(42) = -0.637$, $p = 0.26$.

Movement Test

As for the Movement Test, the Grammar 1 participants did choose the Grammar 1-consistent answers (mean = 55%) significantly more often than the Grammar 2 participants (mean = 47%): $t(42) = 1.84$, $p = 0.037$.

Substitution Test

On average, the participants in the Grammar 1 condition chose the Grammar 1-compatible answers (mean = 54%) more often than the participants in the Grammar 2 condition (mean = 48%). However, this difference did not reach significance in a one-tailed independent samples *t*-test: $t(42) = 1.30$, $p = 0.10$.

Movement-plus-substitution Test

On the Movement-plus-substitution Test, the Grammar 1 subjects chose the Grammar 1-compatible answers (mean = 50%) more often than the Grammar 2 subjects (mean = 44%). This difference was marginally significant: $t(42) = 1.64$, $p = 0.054$.

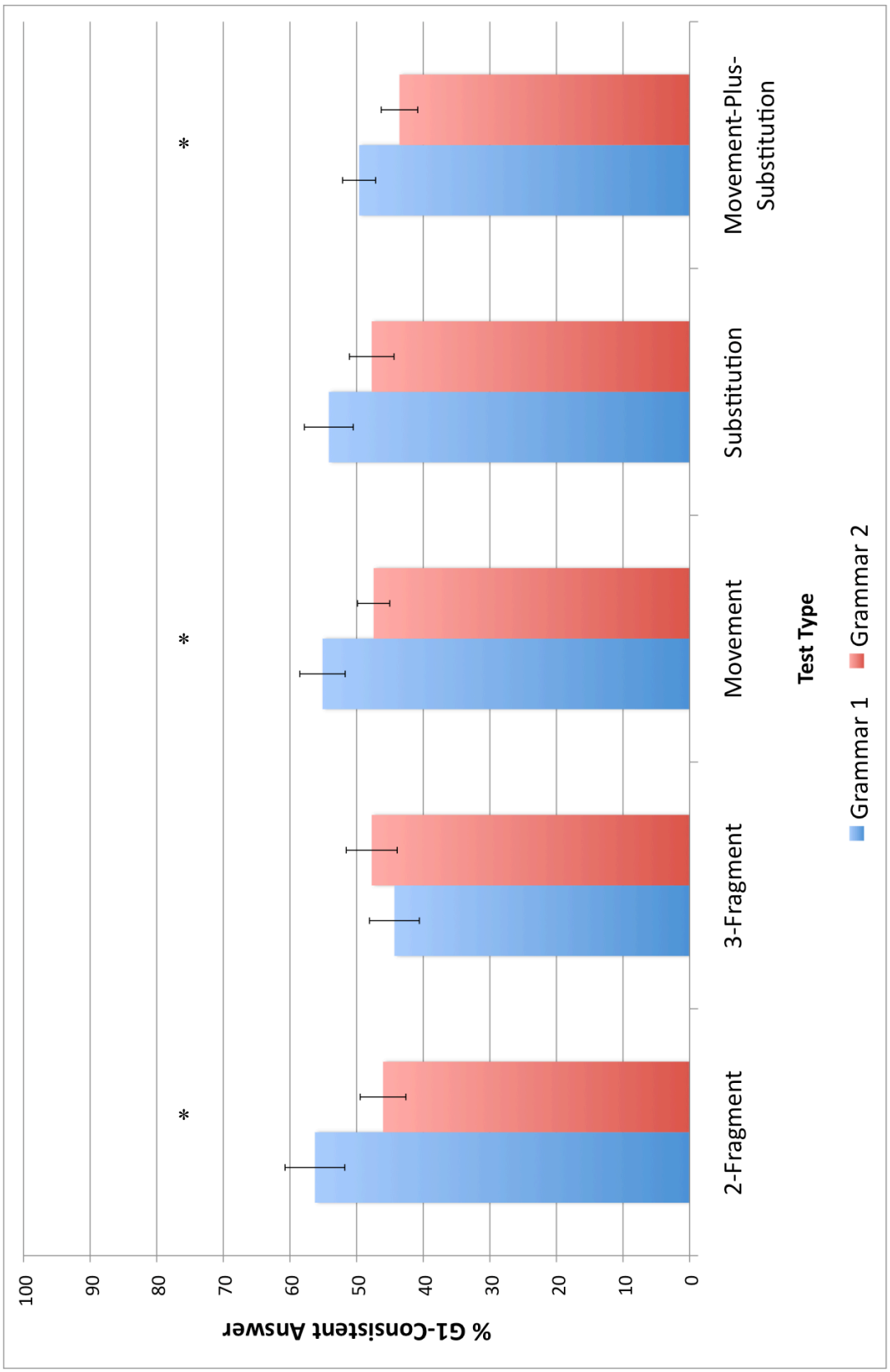


Figure 37: Experiment 1 results. Comparison between Grammar 1 vs. Grammar 2

Against chance

The next analysis tested the experimental groups' performance against chance. If subjects did learn their input grammars, regardless of which grammar they were exposed to, they should have chosen answers consistent with their corresponding grammar, significantly more than chance. Hence for the next set of analyses, we collapsed together the data of the Grammars 1 and 2.

Overall Result

On the whole, subjects in both experimental groups (Grammars 1 and 2) chose the corresponding consistent sentence over the inconsistent sentence significantly more than chance, in a one-tailed independent samples *t*-test (mean = 53%, standard error = 0.01, $t(86) = 2.48$, $p = 0.0075$). Below, we report results from individual tests.

Fragment Test

The participants in both Grammar 1 and Grammar 2 chose the corresponding consistent 2-member fragments significantly more often than chance (mean = 55%, SE = 0.028, $t(86) = 1.83$, $p = 0.036$). As for the 3-member fragments, the participants did not choose the consistent answers significantly more often than chance (mean = 48%, SE = 0.027, $t(86) = -0.628$, $p = 0.27$).

Movement Test

As for the Movement Test, the participants chose the consistent answers significantly more often than chance (mean = 54%, SE = 0.021, $t(86) = 1.85$, $p = 0.034$).

Substitution Test

On average, the participants performed at chance (mean = 53%, SE = 0.025, $t(86) = 1.31$, $p = 0.097$).

Movement-plus-substitution Test

On the Movement-plus-substitution Test, the participants chose the consistent answers (mean = 53%, SE = 0.019) more often than chance. The difference between subjects' performance and chance was marginally significant: $t(86) = 1.60$, $p = 0.057$.

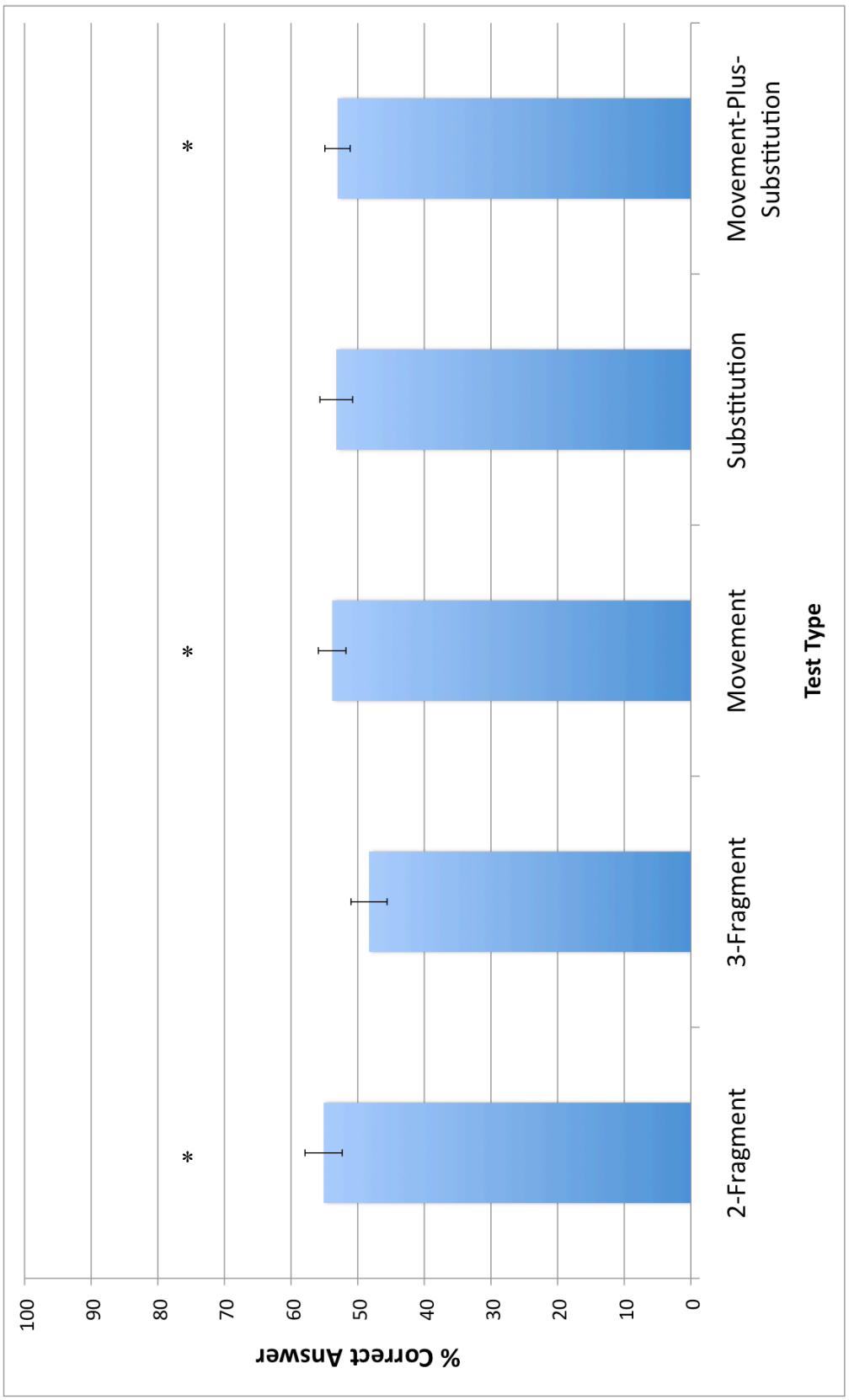


Figure 38: Experiment 1 results. Comparison against chance

Discussion

The results of the 2-member Fragment Test, Movement Test and Movement-plus-substitution Test show a significant difference between the two input groups. The participants who heard Grammar 1 as input chose the Grammar 1-consistent answers significantly more often than the participants who heard Grammar 2 during familiarization. Put another way, participants in both groups chose the answers that were consistent with their input grammar significantly more often than chance. On the other hand, we did not observe any learning on the 3-member Fragment test and Substitution Test. This implies that participants could not learn 3-member fragments and what the proforms stood for, based on 36-min exposure to the artificial language.

Nevertheless, the fact that participants succeeded on the 2-member Fragment Test tells us that they formed the correct phrasal groupings based on the input of 36 min of exposure. Subjects in the Grammar 1 condition chose AB and CD, which are constituents in Grammar 1, to be consistent with their learned grammar, over BC and DE, which are not constituents in Grammar 1. Similarly on the Movement-plus-substitution Test, subjects in Grammar 2 preferred BC and DE, which are constituents in Grammar 2, to be substituted by proforms and moved, to AB and CD, which are not constituents in Grammar 2, to be substituted and moved.

On Movement Test, subjects seem to have chosen the sentences in which constituents in their learned grammar were moved over the sentences in which non-

constituents were moved. For example, from a canonical sentence ABCDE, the Grammar 1 subjects seem to have allowed CDE to move (as in CDEAB), but not DE to move (as in DEABC). In Grammar 1, CDE is a constituent whereas DE is not. Recall that the result of the Fragment Test suggested that the Grammar 1 subjects seem to know that CD is a constituent. If they know that CDE and CD are constituents, but not DE, that means that they formed a structural representation in Figure 30, where CDE has a nested hierarchical structure with an embedded constituent CD.

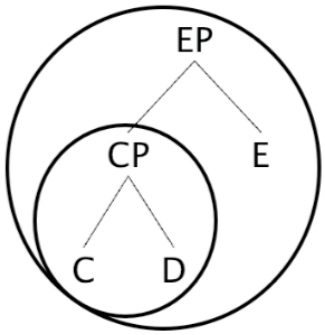


Figure 30**: Internally nested structure**

Moreover, these results are not due to frequency effects, because none of the test items appeared in the input. Therefore, the TP between words in the tests (e.g. KOF HOX, DAZ NEB) were always 0. In order to arrive at the correct answer, participants had to, first, have categorized lexical items into word classes (e.g. KOF, DAZ = A, HOX, NEB = B), then compute the relevant statistics (e.g. TP between AB).

One objection to our conclusions may be that the adults did not really have a hierarchical tree representation like we argue, but that the subjects were simply noticing the chunks of constituents in the consistent (grammatical) test sentences. That is, in the consistent test sample, "good" transitions exist, meaning transitions from a category to another category that has been observed (i.e., constituents), whereas in the inconsistent test sample, "bad" transitions exist, meaning the transition from a category to another category that was not observed in the data (i.e., non-constituents). One could argue that the results in this experiment could be attained if the participants were merely noticing the "good chunks" (constituents) versus "bad chunks" (non-constituents). While this is a relevant concern, it cannot have been the case. Take a look at the movement test sentences that were used.

(69) Movement test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	CDEAB	DEABC
2	CDEABCD	DEABCBC
3	FAB	DEF
4	FABCD	DEFBC

In the first test sentence type (CDEAB vs. DEABC), there are two good chunks in CDEAB if the familiarization language was Grammar 1, namely CD and AB. If your familiarization language was Grammar 2, then there are two good chunks in DEABC, namely DE and BC. What is important, however, is that there are good chunks of one grammar in the other grammar's consistent test sentences. Put another way, the inconsistent test sample in your grammar contains good chunks in your grammar too.

For instance, if your grammar was Grammar 1, there is a good chunk (i.e., AB) in the G2-consistent answer. If your grammar was Grammar 2, there is a good chunk (i.e., DE) in the G1-consistent answer. This is illustrated in Figure 39 below. Solid lines represent good chunks (i.e., constituents) in G1, and dotted lines represent constituents in G2.

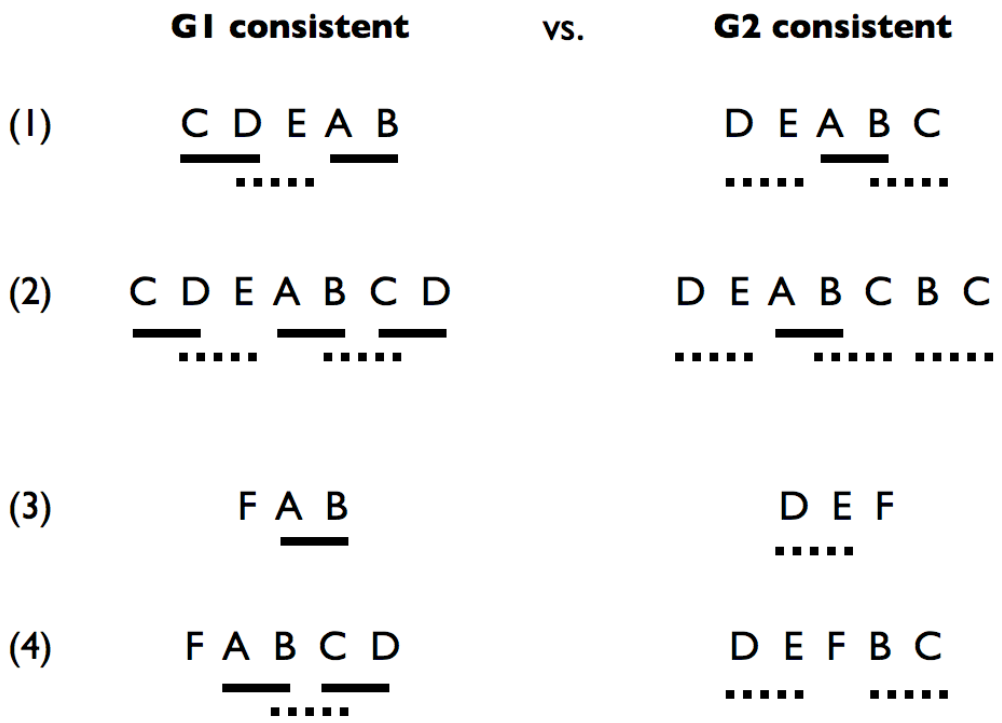


Figure 39: Number of "good chunks" vs. "bad chunks"
Solid line represents good chunks for G1 and dotted line represents good chunks for G2

In the second test sentence type (CDEABCD vs. DEABCBC), there are three good chunks of G1 in the G1-consistent test sentence (CDEABCD), namely CD, AB and CD, but there still is one good G1 chunk in the G2-consistent test sentence

(DEABCBC), namely AB. If you were familiarized in G2, there are three good chunks in the G2-consistent answer (DE, BC, BC), but there are two good chunks in the G1-consistent answer too (DE, BC).

In this way, since there always are good chunks in the "wrong" answer too, simply noticing good chunks does not grant you the correct answer. One might argue that the *number* of good chunks is always higher in the correct answer than in the incorrect answer. There are two and three good G1 chunks in the G1-consistent answers, while there is only one good G1 chunk in the G2-consistent answers. As for G2, there are two and three good G2 chunks in the G2-consistent answers, while there are one and two good G2 chunks in the G1-consistent answer. If the number of good chunks makes a difference, then there should be a difference in participants' performance within the consistent answers too, because there are only two good chunks in the first test type (CDEAB and DEABC) in both grammars, while there are three good chunks in the second test type (CDEABCD and DEABCBC). However, this difference in participants' performance on the first movement test type (CDEAB and DEABC) and the second type (CDEABCD and DEABCBC) was not significant in the independent-samples *t*-test (mean = 53%, 59%, respectively; $t(42) = -0.942$, $p = 0.352$).

One might look at the third test type (FAB vs. DEF) in Figure 39 and argue that in that particular test, there is no good chunk in the wrong answer, and as a result, you can choose the correct answer by merely noticing the good chunks. This is a valid concern. If this is the case, one would expect participants to perform better at this test type than at first (CDEAB vs. DEABC) and second test types (CDEABCD vs.

DEABCBC), because while the third test type does not contain good chunks in the wrong answers, the first and second test types do. Nevertheless, the participants' performance on the third test type (FAB vs. DEF) was not significantly different from the performance on the first type (mean = 49%, 53%, respectively; $t(42) = -0.768$, $p = 0.447$) or the second type (mean = 49%, 59%, respectively; $t(42) = -1.581$, $p = 0.121$) in independent-samples *t*-tests.

In sum, we can now reject the hypothesis that the participants' success was due to a greater number of good transitions in the consistent test sentences than the number of good transitions in the inconsistent sentences. Simply detecting the good chunks in the consistent test sentences cannot have achieved the results of this experiment. If it was not the number of good transitions that differentiated the correct and incorrect test sentences, then what was it? The difference is what is being moved. In the consistent test sentences, constituents are moved, while in the inconsistent test sentences, non-constituents are moved. Hence, our conclusion that that was the distinguishing factor still holds.

It is worth noting that the learning achieved here is not as robust as previous studies. For instance, the experimental group in Thompson & Newport (2007) achieved almost 80% accuracy as early as Day 1, after only 20 min of exposure. On the other hand, the highest success rate in our Experiment 1 was 55%. There are two responses to this observation. First, the artificial language in Thompson & Newport (2007) was much simpler than our artificial languages. The canonical sentence was ABCDEF with a flat structure like the following:

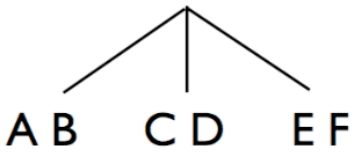


Figure 40: Phrase structure in Thompson & Newport (2007)

The constituents were AB, CD and EF and such grouping is very intuitive. If you are given a sentence with 6 words, it seems very natural and obvious to divide them into 3 groups of two. In fact, even the control group, who were not given any statistical cue, scored well above chance, achieving 60% accuracy on Day 1 and almost 80% accuracy on Day 5, as in Figure 15. Thompson & Newport (2007) speculate that perhaps that was because native English-speaking participants had tendency to break up input strings into binary groupings or to impose trochaic foot structure even when there was no prosodic information.

In contrast, our grammars are much more complex and display nested hierarchy. Since our canonical sentence is a 5-word string (ABCDE), it is impossible to impose a binary grouping. Therefore, the low success rates in this study could be due to the complexity of our grammars.

Second, our familiarization period was relatively short compared with previous studies. In Saffran (2001), it was 30 min for 2 days, accumulating a total of 60 min of exposure. In Thompson & Newport (2007), it was 20 min for 5 consecutive days, accumulating a total of 100 min of exposure. In our experiment, it was 36 min and just one day. As a result, the task was very hard and this could have led to the

large error rates.[6] In any case, our main finding is that, even though the performance in this experiment was not as robust as previous literature, our participants did perform significantly above chance on 3 of 5 tests.

The results of Experiment 1 offer an answer to one of our questions, which was whether the TP can be a cue to not only the phrasal groupings but also hierarchical constituent structure. And the answer seems to be positive. By including features of natural languages such as optionality, repetition, substitution and movement, there emerge TP peaks and dips. We found that not only can learners infer phrasal groupings on the basis of such statistical pattern, but they can also infer nested hierarchical structure.

In Experiment 1, the Movement Test yielded a significant effect of learning. The subjects chose the sentences where a constituent had undergone movement over sentences where a non-constituent had undergone movement. Going back to our three hypotheses, the results from this experiment indicate that the predictions made by Beyond and Unconstrained Hypothesis were not borne out, since this hypothesis predicted that the performance on the movement test would be at chance. On the other hand, the predictions made by both Limited Hypothesis and Beyond and Constrained Hypothesis were borne out, because the learners correctly chose the consistent answer. However, the results of this experiment do not differentiate these two hypotheses, since both hypotheses predicted the identical outcome.

[6] We are considering a follow-up study where the familiarization last for 2 or 3 days to boost the learning. If the limited familiarization time was causing the low success rate, increasing the familiarization period (and sleeping) should lower the error rates.

Table 9: Predictions and outcomes for Experiment 1

	Views	*Predictions*	*Outcome*
Limited Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers	✓
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers	✓
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Adults will perform at chance	✗

The Limited Hypothesis is the view that the deductive power of a learner is limited to the observed distributions, and statistical learning does not interact with innate constraints. Beyond and Constrained Hypothesis, on the other hand, is the view that the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions. The results of Experiment 1 do not support one or the other of these two views. But given that this is the critical question in this dissertation, we would want to find a way to tell them apart.

One might argue that the success of the movement test, for example, in this experiment was due to the abundance of movement sentences in the input, and that subjects were simply choosing the ones that they were most familiar with. In fact, the presentation set in this experiment did include a large number of sentences that had undergone movement operation. In the Grammar 1 input, 40% (32/80) of the whole presentation set was movement sentences. In Grammar 2, it was 48% (38/80).

If we remove all the sentences generated via movement (and substitution by proform) rules and the participants still succeed at the movement test, it would indicate that the participants were not merely compiling predictive statistics from the

data, because the correct "answer" does not appear in the exposure set. It would imply that participants were acting on the knowledge that was not available in the input, specifically the knowledge that you cannot move a non-constituent. Accordingly, in Experiment 2, we will remove all the sentences generated via movement rules and substitution rules and we will test them on movement and substitution tests, in the hope of being able to tell apart the two hypotheses. In this way, we should be able to tease apart the two hypotheses because they would make different predictions. This way, we hope to explore whether the representations are part of the learning system prior to the experience, and what the deductive consequences of distributional learning are.

3.2 Experiment 2 (Adult 2)

Experiment 2 tries to answer one of our main questions of this dissertation, which is what the deductive consequences of distributional learning are. In this experiment, we remove all the sentences generated by movement (and substitution by proforms) rules from the input and examine whether subjects can succeed under such condition. There are two possible outcomes. Under a learning theory where the deductive power of a learner is limited to the observed distributions, learners should not allow new structures that were not displayed in the input. Therefore, learners would consider both consistent and inconsistent test samples to be ungrammatical, since both samples involve novel structures. According to this view, learners do not

come with a pre-determined set of possible structures or rules, in this case, learners would not know in advance that you can only move constituents. So, the subjects would fail to choose the correct test sentences in which constituents are moved. If this were the case, it would suggest that what learners do is to track the distributions and build an illusion of a structure entirely based on them, without any preconception of what is and what is not a possible structure.

On the other hand, under a learning theory where a learner already knows an antecedently-specified range of possible representations, statistics is merely used as a source of information that helps a learner select the correct grammar that derives the matching surface strings. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. On this view, another possible outcome is that the subjects succeed in this condition, and they can correctly choose the test sentences in which constituents are moved, over test sentences in which non-constituents are moved. If so, it would suggest that learners' generalization extends to novel structures, as long as they are compatible with antecedently known constraints.

3.2.1 Description of the linguistic systems

The same artificial grammars, Grammar 1 and Grammar 2, were used. The only difference was that all examples generated via movement and substitution-with-proform rules were excluded from the familiarization. Just like in Experiment 1, 80 sentences were picked as the presentation set. Three sentences (3.8%) were the

canonical sentence type (ABCDE) in both grammars. There were four sentence types, which is shown below.

(70) Familiarization sentence types in Experiment 2

Grammar 1		*Grammar 2*	
A B F	(9)	A B C D E	(3)
A B C D E	(3)	F D E	(10)
A B C D E C D	(19)	A B C D E B C	(16)
A B F C D	(49)	F D E B C	(51)

While the input lacked movement rules, it still included other manipulations such as repetition and optionality. These features contributed to make the TPs between words within phrases higher than the TPs across phrases. The resulting TP patterns of the presentation set are given below. All 80 sentences were randomized. The sentence types and 80 sentences that appeared in the presentation set are shown in Appendix B.

Table 10: Transitional probabilities for 80 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.28	1.00	0.24
Backward TP	1.00	0.24	1.00	1.00

Table 11: Transitional probabilities for 80 input sentences in Grammar 2

	A-B	B-C	C-D	D-E
Forward TP	1.00	1.00	0.22	1.00
Backward TP	0.22	1.00	0.24	1.00

3.2.2 Method

Participants

Forty-four native speakers of English participated in Experiment 2 as subjects. The participants were undergraduate students at the University of Maryland, gave informed consent prior to participating and received monetary compensation. Twenty-two participants were randomly assigned to hear Grammar 1 during the familiarization and the other 22 were assigned to Grammar 2.

Recording, Procedure, Tests

The recording and the procedure for Experiment 2 were identical to those for Experiment 1. Participants were exposed to the presentation set of 80 sentences six times, for a total of 36 min of exposure. The administered tests were identical to the ones in Experiment 1.

3.2.3 Hypotheses and predictions

Recall that while the results of Experiment 1 were not compatible with the Beyond and Unconstrained Hypothesis, they were compatible with both Limited Hypothesis and Beyond and Constrained Hypothesis. This was mostly because all of the test structures were included in the familiarization. In this experiment, we remove all the sentences generated by movement and substitution rules, which means that the test sentences have novel structures that were not seen in the input. Now the three hypotheses make distinct predictions.

For convenience' sake, let us take the case of the movement test to discuss different hypotheses and predictions. According to the first hypothesis, which we call the "Limited" Hypothesis, learners do not generalize beyond what was observed in the input. So at test, when they see two novel structures – one that moved a constituent and one that moved a non-constituent – they would consider both to be illicit, because neither was seen in the input. Thus, the performance should be at chance.

According to the second hypothesis ("Beyond and Constrained Hypothesis"), learners can generalize beyond the observed input, but their generalizations are restricted in principled way. For instance, learners might have the knowledge that you cannot move a non-constituent in natural languages. If this were the case, on the movement test, the participants would allow the consistent test sentence in which a constituent was moved, but they would not allow the inconsistent test sentence in which a non-constituent was moved, because while the former is a possible movement, the latter is an impossible operation in language. Thus, the participants should show a preference towards the consistent test items over the inconsistent test items.

According to the third hypothesis ("Beyond and Unconstrained Hypothesis"), learners' generalizations could go beyond what was observed in the input and those generalizations do not have to be constrained by some principles. For example, one generalization the learners could form is that you can move any neighboring elements. If this were the case, on the movement test, learners might allow both test structures even though they are both novel, because both test sentences move

neighboring words. If so, both test sentences would be licit for the learners and the performance at test would be at chance, that is, the learners would not choose one over the other.

In this way, the three hypotheses make distinct predictions. Both "Limited" and "Beyond and Unconstrained" Hypotheses predict that the performance on the movement test would be at chance although for different reasons. The only hypothesis that predicts a different outcome is Beyond and Constrained Hypothesis, which predicts that learners would choose the consistent test sample over the inconsistent test sample.

Table 12: Predictions for Experiment 2

	Views	*Predictions*
Limited Hypothesis	Both test sentences are ungrammatical	Adults will perform at chance
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Adults will perform at chance

3.2.4 Results and discussion

Our question in Experiment 2 was whether removing the movement sentences from the familiarization would nonetheless license the inference that only constituents can be moved.

Grammar 1 vs. Grammar 2

Fragment Test

On the 2-member fragment test, the participants in Grammar 1 did not choose the Grammar 1-consistent answers (mean = 53%) reliably more than participants in Grammar 2 (mean = 50%) in a one-tailed independent samples *t*-test: $t(42) = 0.689$, $p = 0.248$. On the 3-member fragment tests, the participants in the Grammar 1 condition did not choose the Grammar 1-compatible answers (mean = 47%) significantly more often than the participants in the Grammar 2 condition (mean = 44%, $t(42) = 0.654$, $p = 0.259$) either.

Movement Test

As for the Movement Test, the Grammar 1 participants did choose the Grammar 1-consistent answers (mean = 61%) significantly more often than the Grammar 2 participants (mean = 44%): $t(42) = 3.675$, $p = 0.0005$.

Substitution Test

On the Substitution Test, the participants in Grammar 1 did not choose the Grammar 1-consistent answers (mean = 47%) more often than the Grammar 2 participants (mean = 52%): $t(42) = -0.868$, $p = 0.196$.

Movement-plus-substitution Test

On the Movement-plus-substitution Test, the Grammar 1 subjects chose the Grammar 1-compatible answers (mean = 49%) less often than the Grammar 2 subjects (mean = 56%) in a one-tailed *t*-test: $t(42) = -1.925$, $p = 0.031$.

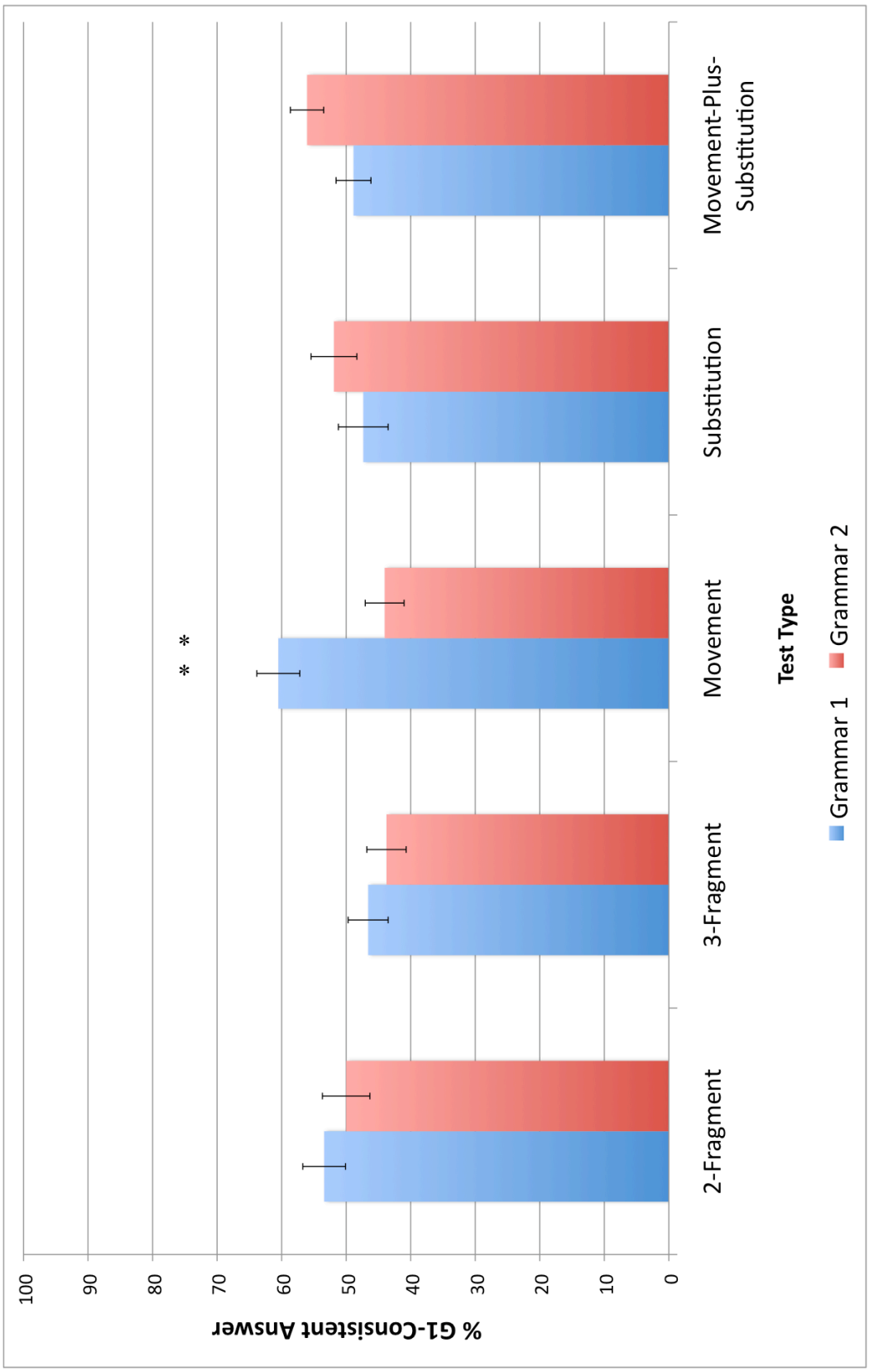


Figure 41: Experiment 2 results. Comparison between Grammar 1 vs. Grammar 2

Against chance

The next analysis tested the experimental groups' performance against chance. If subjects learned their input grammars, they should have chosen answers consistent with their corresponding grammar more often than chance, no matter which grammar they were exposed to. Thus for the next set of analyses, we collapsed together the data from the two grammars.

Overall Result

On the whole, subjects in both groups (Grammars 1 and 2) did not choose the corresponding consistent sentence over the inconsistent sentence significantly more than chance (mean = 51%, SE = 0.01. One-tailed independent samples *t*-test: $t(86) = 0.928$, $p = 0.178$). Below, we report results from individual tests.

Fragment Test

On average, the participants in both Grammar 1 and Grammar 2 did not choose the corresponding consistent 2-member fragments reliably more often than chance (mean = 52%, SE = 0.025, $t(86) = 0.693$, $p = 0.245$). Similarly, for the 3-member fragments, the participants did not choose the consistent answers reliably more often than chance (mean = 51%, SE = 0.023, $t(86) = 0.626$, $p = 0.267$).

Movement Test

As for the Movement Test, the participants chose the consistent answers more often than chance, and this difference was highly reliable (mean = 58%, SE = 0.022, $t(86) = 3.674$, $p < 0.001$). [7]

Substitution Test

The participants did not choose the corresponding consistent sentences more often than chance (mean = 48%, SE = 0.026, $t(86) = -0.878$, $p = 0.192$).

Movement-plus-substitution Test

On the Movement-plus-substitution Test, the participants did not choose the consistent answers significantly more often than chance (mean = 46%, SE = 0.019) $t(86) = -1.91$, $p = 0.03$).

[7] At first glance, the participants in Experiment 2 appear to have performed better on the Movement Test (mean = 58%) than the participants in Experiment 1 (mean = 54%). However, this difference was not significant ($t(86) = -1.44$, $p = 0.153$) in an independent samples *t*-test.

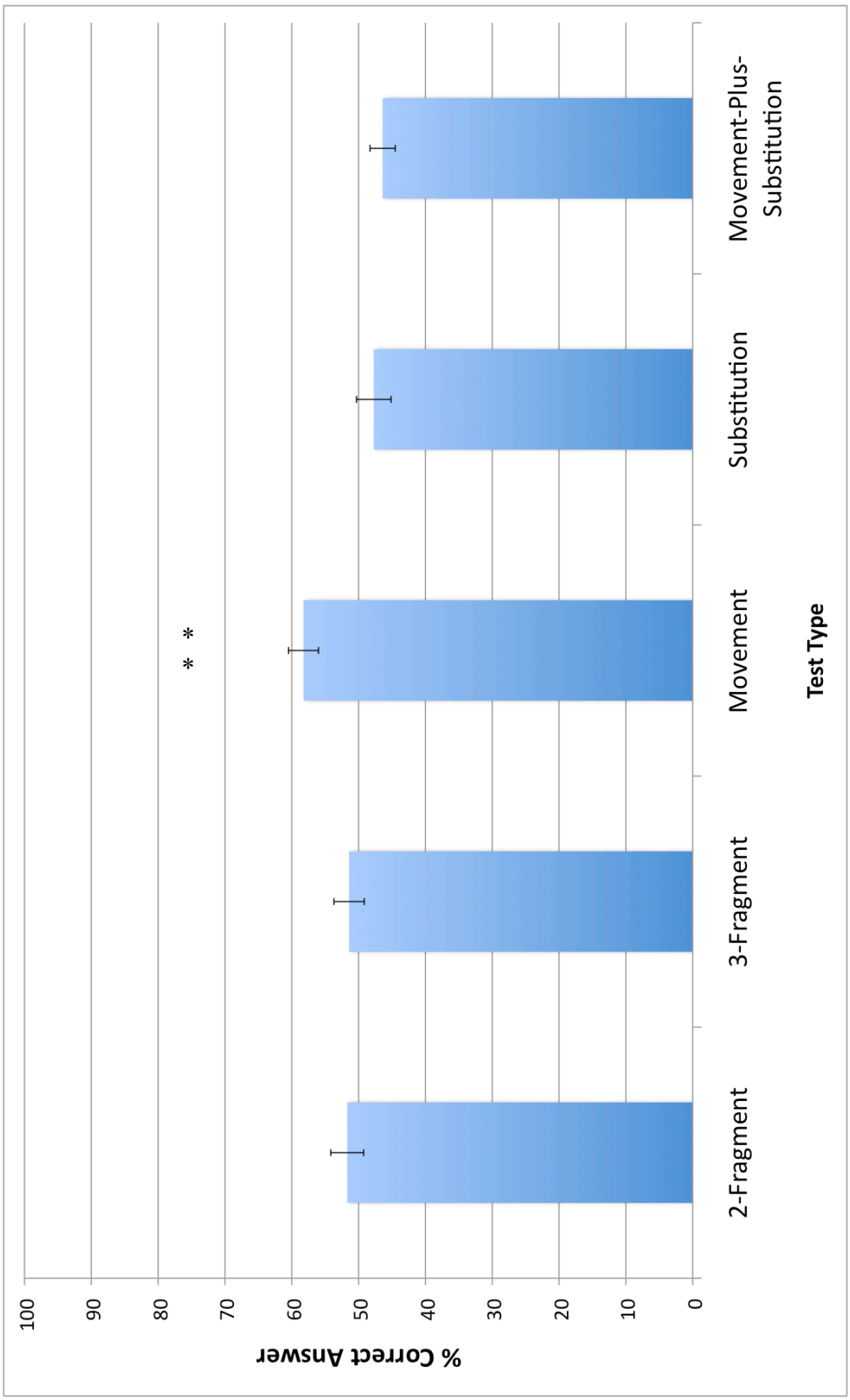


Figure 42: Experiment 2 results. Comparison against chance

Discussion

Except for the Substitution Test and Movement-plus-substitution Test, there was a general trend for choosing the input-consistent answers in all tests. The effect was highly reliable in the Movement Test. This result is especially striking because the performance of participants was most successful on the tests that involved movement, even though there was no movement sentences in the input. This confirms that the participants' success on the Movement Test in Experiment 1 was not due to the abundance of movement sentences in the familiarization. Even when the input lacked movement sentences, adults chose the sentences in which constituents underwent movement as "grammatical" sentences in the artificial language.

In contrast with the results of the Movement test, participants did not successfully learn that only constituents can be replaced by a proform. This could be due to the fact that substitution rules were not introduced during the familiarization in Experiment 2. Since the proforms were excluded from the input, the participants saw them for the first time during the test. As a result, that probably confused the participants. It is interesting that in the absence of movement and substitution rules in the input, people can infer that only constituents can be moved, but not that only constituents can be substituted. There could be several reasons for such asymmetry. One possibility is that, while you do not need input to infer that only constituents can be moved, but you need sufficient information to infer that only constituents can be replaced by proforms. In other words, although the constraints on movement and

substitution may be innate and universal, you need some kind of input as a trigger to set the constraint on substitution to work, but you do not need any trigger to set off the movement rule. This line of possibility is certainly compatible with the results of Experiment 2. Another possibility is that learning substitution rules requires some kind of reference. For example, when replacing *red bottle* with "one" in a sentence like, *The boy likes the red bottle and the girl likes that one,* you have to know that "one" refers to *red bottle.* However, in an artificial language learning experiment, no semantic information that corresponds with the sentences is given. When learning that DAZ HOX is replaced by a proform *ib,* you do not even know what "DAZ HOX" refers to. All you have is the distributional information that DAZ and HOX often appear together. This suggests that statistical information that signals constituency is not adequate for deducing that only constituents can be substituted. It might be that it also requires semantic information for the referent of the proform. The current experiments do not answer these questions, but nonetheless, it is worth noting that we found a contrast between movement-rule learning and substitution-rule learning.

Going back to our three hypotheses, the results of the movement test in Experiment 2 are only compatible with Beyond and Constrained Hypothesis, since this was the only hypothesis that predicted this outcome. The other two hypotheses (Limited Hypothesis and Beyond and Unconstrained Hypothesis) predicted that the learners would not choose one test sentence over the other and that the performance would be at chance. However, the results show that the adults preferred the consistent test samples to the inconsistent test samples.

Table 13: Predictions and outcomes for Experiment 2

	Views	*Predictions*	*Outcome*
Limited Hypothesis	Both test sentences are ungrammatical	Adults will perform at chance	✗
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Adults will choose consistent answers	✓
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Adults will perform at chance	✗

The information about constituency was contained in the input, but the participants in Experiment 2 were not given any movement sentences in the input. In that situation, why should the participants choose the sentences that moved constituents over the sentences that moved non-constituents? If the generalization you form is entirely based on the input, both test structures should be equally illicit, since both are new. If the generalization you form is not restricted to what was observed in the input, then you might equally allow both test structures, since both moved neighboring elements. But that is not what happened. What happened was that the participants chose the new structures in which constituents, but not non-constituents, were moved. Since the information that constituents can be moved was not included in the input, the inference must have come from some constraints that were known to the learners. And that is what was predicted by the Beyond and Constrained Hypothesis. In other words, the generalizations that were formed by the learners based on the input in Experiment 2 were not restricted to just the input, but they were restricted in a way that is predictable considering what is possible and what is impossible in natural languages. More specifically, the participants seem to have

behaved in the way that was compatible with possible operations in natural languages.

Here, let us consider and examine alternative accounts for the results obtained in this experiment. First alternative account can be dubbed something like "strange first word" account. If you look at the sentence types of the Movement test in (71), you see that all the Grammar 2-compatible test sentences begin with a D word.

(71) Movement test

	Grammatical in Grammar 1	*Grammatical in Grammar 2*
1	CDEAB	DEABC
2	FAB	DEF
3	CDEABCD	DEABCBC
4	FABCD	DEFBC

In contrast, none of the Grammar 1-compatible test sentences begin with a D word. This is due to the fact that CD is a constituent in Grammar 1. Since this study did not have optionality of an element within a single phrase, C and D always appear together, which is why no sentence began with a D word in the input of Grammar 1. One could argue that participants' success was due to such serial position effects. The participants who heard Grammar 1 during the familiarization might know that the Grammar 2 test sentences are not from the language they were familiarized to, simply because no sentence had begun with a D. The results of Experiment 2 would be undermined if participants were simply noting such linear pattern. It would mean that participants rejected the incorrect sentences not because non-constituents underwent

movement, but because they never saw a sentence begin or end with a particular word class.

This is a valid objection; however, this cannot have been the case. That is because the Grammar 1-compatible test sentences did not appear in the familiarization either. The Grammar 1-compatible movement test sentences start with either C or F. See the list of familiarization sentence types of Experiment 2 in (72).

(72) Familiarization sentence types in Experiment 2

Grammar 1	*Grammar 2*
A B F	A B C D E
A B C D E	F D E
A B C D E C D	A B C D E B C
A B F C D	F D E B C

No input sentence of Grammar 1 in this experiment starts with C or F. Therefore for the participants who had been familiarized with Grammar 1, both groups of test sentences (G1-compatible and G2-compatible) are equally unfamiliar and unseen. Even if they kept track of the serial positions of some elements, it would not help, since neither test sentence type appeared in the input. Hence, we can back up our interpretation of the results, which is that participants chose the sentences that moved constituents instead of non-constituents.

Similarly, the G2-compatible movement test sentences all begin with a D word, while the G1-compatible test sentences begin with a C or F word. This is because DE and BC are constituents in Grammar 2, so no sentence in G2 starts with a C, since C is always preceded by B. So, one could argue that, if you were familiarized

with Grammar 2, you could have chosen the correct answer by simply choosing the test sentences that start with D, instead of C. This objection is a valid concern, but it cannot have been the case. That is because none of the familiarization sentences in Grammar 2 started with a D word either (see (72)), because no movement sentences were included in the input in this experiment. Interestingly, however, some familiarization sentences of Grammar 2 actually start with F. And half of the G1-compatible test sentences start with F. Therefore, if you were only paying attention to the first word, you could actually be misled, and choose the wrong answers instead. But this was not attested. The participants who heard G2 as input did not perform significantly worse on the second and fourth test types (in which the G1-compatible answers start with F) (mean = 42%) than the first and third test types (mean = 46%; paired-samples *t*-test: $t(21) = 0.668$, $p = 0.511$). In this way, we can reject the alternative account that subjects were merely taking note of the good and bad first words.

Second alternative account is similar to the first alternative, but it can be dubbed "strange last word". If you look at the sentence types of the Movement test in (71), you see that most of the Grammar 2-compatible test sentences end in a C word. In contrast, none of the Grammar 1-compatible test sentences ends in a C word. This is due to the fact that CD is a constituent in Grammar 1. Since this study did not have optionality of an element within a single phrase, C and D always appear together, which is why no sentence ends with a C word in the input of Grammar 1. One could argue that participants succeeded by simply noticing that ending a sentence with a C word is strange in Grammar 1, thus rejecting test sentences that ended with a C word.

This is a relevant concern, but it cannot have been the case either. Notice that one of the G2-compatible test sentences ends with F. It should also be noted that in the familiarization set, none of the input sentences of Grammar 2 ends with an F word, while some of the Grammar 1 input sentences end with F (see (72)). If the participants were simply paying attention to good and bad last words, then this could be misleading. The participants who heard Grammar 1 as input could be misled to think that the G2-compatible test sentence DEF is the correct answer, since they have seen sentences ending with F. If this is the case, then it would predict that participants perform better at test sentences ending with a C word than test sentences ending with F, because rejecting test sentences that end with C would be easier than rejecting test sentences that end with F. Nevertheless, this prediction was not borne out. The participants in Experiment 2 who heard Grammar 1 during the familiarization phase did not perform significantly better or worse on the second test type (FAB vs. DE**F**) (mean = 57%) than the first type (CDEAB vs. DEAB**C**) (mean = 59%; $t(21) = -0.326$, $p = 0.747$), third type (CDEABCD vs. DEABCB**C**) (mean = 69%; $t(21) = -1.498$, $p = 0.149$) or fourth type (FABCD vs. DEFB**C**) (mean = 57%; $t(21) = 0$, $p = 1.0$), in paired-samples *t*-tests. In sum, we can reject the alternative hypothesis that participants succeeded by simply paying attention to good and bad last words.

Third alternative account can be dubbed "number of good chunks", which is the hypothesis that the subjects were simply noticing the chunks of constituents in the consistent (grammatical) test sentences, and that they did not have a hierarchical tree representation like we argue. It could be that in the consistent test sample, "good" transitions exist, meaning transitions from a category to another category that have

been observed (i.e., constituents), whereas in the inconsistent test sample, "bad" transitions exist, meaning the transition from a category to another category that was not observed in the data (i.e., non-constituents). One could argue that the results in this experiment could be achieved if the participants were merely noticing the "good chunks" (constituents) versus "bad chunks" (non-constituents). While this is a relevant concern, it cannot have been the case. Take a look again at the movement test sentences that were used.

(73) Movement test

	Grammatical in Grammar 1	Grammatical in Grammar 2
1	CDEAB	DEABC
2	CDEABCD	DEABCBC
3	FAB	DEF
4	FABCD	DEFBC

In the first test sentence type (CDEAB vs. DEABC), there are two good chunks in CDEAB if the familiarization language was Grammar 1, namely CD and AB. If your familiarization language was Grammar 2, then there are two good chunks in DEABC, namely DE and BC. What is important, however, is that there are good chunks of one grammar in the other grammar's consistent test sentences. Put another way, the inconsistent test sample in your grammar contains good chunks of your grammar too. For instance, if your grammar was Grammar 1, there is a good chunk (i.e., AB) in the G2-consistent answer. If your grammar was Grammar 2, there is a good chunk (i.e., DE) in the G1-consistent answer. This is illustrated in Figure 43 below. Solid lines

represent good chunks (i.e., constituents) in G1, and dotted lines represent constituents in G2.

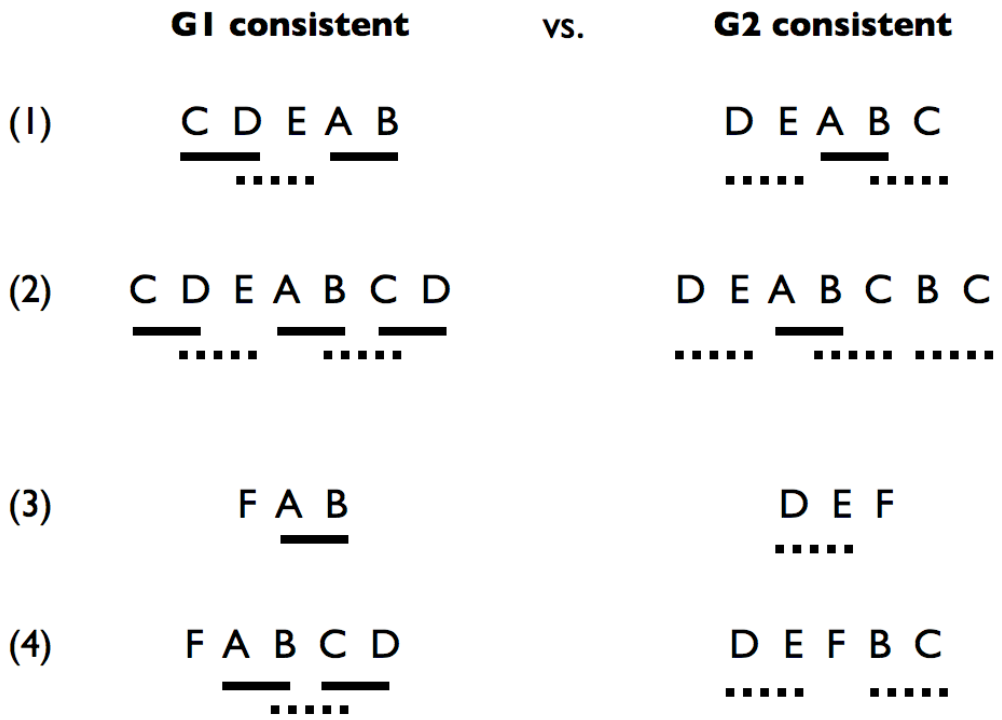


Figure 43: Number of "good chunks" vs. "bad chunks". Solid line represents good chunks for G1 and dotted line represents good chunks for G2

In the second test sentence type (CDEABCD vs. DEABCBC), there are three good chunks of G1 in the G1-consistent test sentence (CDEABCD), namely CD, AB and CD, but there still is one good G1 chunk in the G2-consistent test sentence (DEABCBC), namely AB. If you were familiarized in G2, there are three good chunks in the G2-consistent answer (DE, BC, BC), but there are two good chunks in the G1-consistent answer too (DE, BC).

In this way, since there always are good chunks in the "wrong" answer too, simply noticing good chunks does not grant you the correct answer. One might argue that the *number* of good chunks is always higher in the correct answer than in the incorrect answer. There are two and three good G1 chunks in the G1-consistent answers, while there is only one good G1 chunk in the G2-consistent answers. As for G2, there are two and three good G2 chunks in the G2-consistent answers, while there are one and two good G2 chunks in the G1-consistent answer. If the number of good chunks makes a difference, then there should be a difference in participants' performance within the consistent answers too, because there are only two good chunks in the first test type (CDEAB and DEABC) in both grammars, while there are three good chunks in the second test type (CDEABCD and DEABCBC). However, this difference in participants' performance on the first movement test type (CDEAB and DEABC) and the second type (CDEABCD and DEABCBC) was not significant in the independent-samples *t*-test (mean = 59%, 60%, respectively; $t(42) = -0.193$, $p = 0.848$).

One might look at the third test type (FAB vs. DEF) in Figure 43 and argue that in that particular test, there is no good chunk in the wrong answer, and as a result, you can choose the correct answer by merely noticing the good chunks. This is a relevant concern. If this is the case, one would expect participants to perform better at this test type than at first (CDEAB vs. DEABC) and second test types (CDEABCD vs. DEABCBC), because while the third test type does not contain good chunks in the wrong answers, the first and second test types do. Nevertheless, the participants' performance on the third test type (FAB vs. DEF) was not significantly different from

the performance on the first type (mean = 58%, 59%, respectively; $t(42) = -0.107$, $p = 0.916$) or the second type (mean = 58%, 60%, respectively; $t(42) = -0.298$, $p = 0.767$).

In sum, we can reject the third alternative account that the participants' success was due to a greater number of good transitions in the consistent test sentences than the number of good transitions in the inconsistent sentences. Simply detecting the good chunks in the consistent test sentences cannot have achieved the results of this experiment. The critical factor that helped the participants distinguish the consistent and inconsistent test samples must have been that, in the consistent test sentences, constituents are moved, while in the inconsistent test sentences, non-constituents are moved. And the results of this experiment are compatible with this conclusion.

The results of Experiment 2 are only compatible with the idea that the role of a learner is to identify the mapping between the surface forms and one of a range of possible grammars that generated them. On this approach, the role of the statistics is to drive inferences about which grammar out of the set of possible grammars is responsible for the input data. More specifically, only grammars that allow movement of a constituent but not of non-constituents are considered as a possibility and grammars that move non-constituents must not have been an option. Otherwise, it is impossible to explain why the subjects were able to identify the correct answers in the absence of movement sentences in the input. It is worth noting here that constituency alone does not give this result. Constituency is a necessary condition to drive this result, and yet, it alone does not imply that non-constituents cannot be moved. For example, the following type of finite-state automata could generate the familiarization

sentences, and together with some kind of probabilistic table, the FSA could learn constituency.

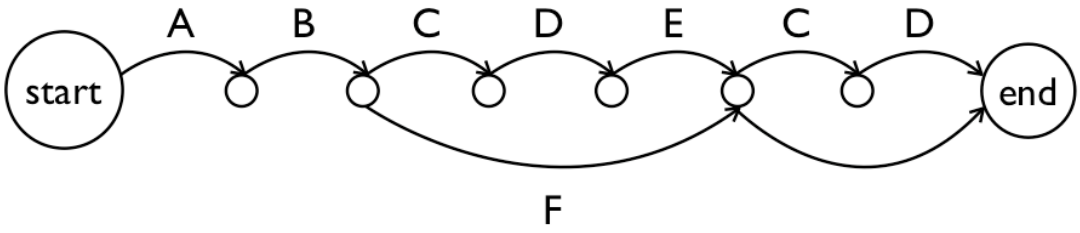


Figure 44: FSA for familiarization sentences of Grammar 1 in Experiment 2

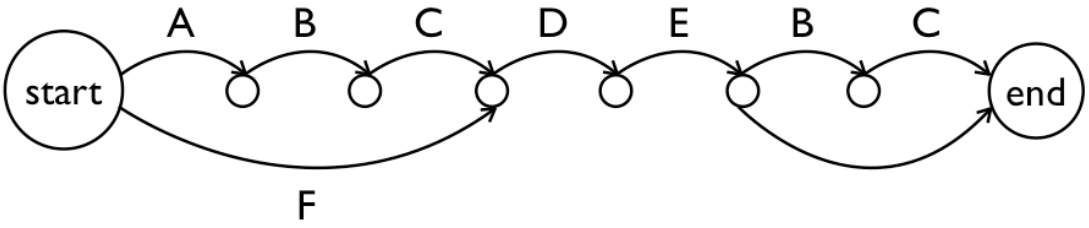


Figure 45: FSA for familiarization sentences of Grammar 2 in Experiment 2

However, even having a representation of constituency is not sufficient to achieve the results of Experiment 2. If you only have FSA like above, you cannot choose the correct consistent answers at test. For example, it is impossible to choose between CDEAB vs. DEABC, sentences you have never seen before. One might argue that with probabilistic FSA, you can simply choose the one with a greater number of good transitions. But we already discussed above that this does not work. What is needed in order to choose the correct answer is the knowledge that you can only move

constituents. The fact that the participants in our experiment seemed to know that without being told that constraint suggests that they already knew that prior to the exposure. The results from this experiment are compatible with the view that this constraint is linguistic in nature. In any case, the constraint could not have been formed by simply being exposed to the artificial language during the experiment, thus must have been contributed by the learners themselves.

Experiment 2 was an attempt to answer one of our main questions of this dissertation, which is what the deductive consequences of distributional learning are The results of Experiment 2 suggest that learners' acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. This implies that statistics are merely used as a source of information that helps a learner select the correct grammar that derives the matching surface strings. Furthermore, it also suggests that the representations the learners form are limited in the same ways that natural language is constrained.

Chapter 4: Infant Experiments

In this chapter, we extend our investigation to testing infants. In the classic artificial language phrase structure learning studies (Morgan and Newport 1981; Morgan, Meier and Newport 1987; Morgan et al. 1989) and in Thompson and Newport (2007), only adults have been tested. Saffran et al. (2008) tested infants, but we argued in Chapter 2 that what the infants acquired could have been a finite-state grammar as in Figure 13, and not necessarily a hierarchical phrase structure. The results of Saffran et al. (2008) could have been achieved by infants simply learning the linear order of word categories.

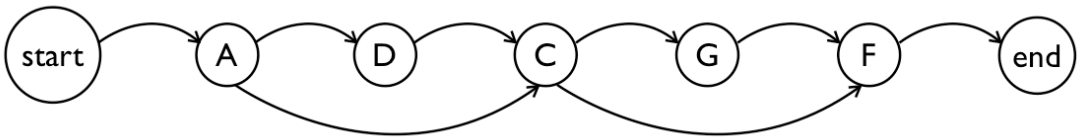


Figure 46: FSA of the predictive language in Saffran et al. (2008)

In other words, whether infants can learn the hierarchical phrase structure of an artificial language is yet to be shown. To this end, we tested infants to see whether they can learn on the basis of statistical information.

We concluded in Experiment 2 that the adults knew that only constituents are allowed to move prior to the exposure. Nonetheless, one could potentially argue that the fact that adults succeeded on the movement test in the absence of movement in the input is because they already knew a natural language. And natural languages only allow movement of constituents. So one could argue that adults extended that knowledge in learning the artificial language. In order to examine this possibility, we tested infants. If our results from Experiments 1 and 2 model what happens in language acquisition, infants might perform the same way as adults did. Experiment 3 is the first of a series of infant experiments and it is a replication of our Experiment 1.

We chose to test mainly 18-month-old infants in this dissertation because this is around the age that infants begin to show their knowledge of syntax. Gomez & Gerken (1999) showed that, by the age of 12 months, infants are sensitive to statistical distributions in an artificial language generated by a finite state grammar. By 14-months of age, infants begin to demonstrate sensitivity to properties of their native language syntax (Hirsch-Pasek & Golinkoff 1996), even though infants at this age are hardly producing two- or three-word sentences themselves. Finally, Santelmann & Jusczyk (1998) showed that 18-month-old infants are sensitive to non-adjacent morpho-syntactic dependencies, but not 15-month-olds. Additionally, Gomez (2002) also showed that by 18 months, infants are able to detect non-adjacent dependencies in an artificial language. We chose to test 18-month-old infants because we supposed that although they are not producing "sentences", they are able to comprehend and sensitive to syntactic and statistical characteristics of language.

4.1 Experiment 3 (Infant 1)

4.1.1 Method

Participants

Infants were recruited via a mailing list. Fourteen infants, approximately 18 months of age were tested (age range: 17 months 15 days to 19 months 9 days; mean: 18 months 17 days). Eight additional infants were tested but excluded from analyses for the following reasons: crying (n = 4), inattentiveness (n = 3) and equipment failure (n = 1). The infants were randomly divided between two familiarization conditions. Half of the infants (n = 7) heard Grammar 1 as input during the familiarization period and the other half (n = 7) heard Grammar 2. Parental consent was obtained prior to testing, in accordance with the NIH standards for the ethical treatment of human subjects.

Material

The artificial languages used in this experiment were identical to the ones in Experiments 1 and 2. Just like in Experiment 1, the familiarization input included movement. The only difference was that 30 sentences, instead of 80 sentences, were picked as the presentation set. Two sentences (6.7%) were the canonical sentence type (ABCDE) in both grammars. The TP patterns of the presentation set are given in tables below. The sentence types and 30 sentences that appeared in the presentation set are shown in Appendix D.

Table 14: Transitional probabilities for 30 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.19	1.00	0.24
Backward TP	1.00	0.24	1.00	0.26

Table 15: Transitional probabilities for 30 input sentences in Grammar 2

	A-B	B-C	C-D	D-E
Forward TP	0.28	1.00	0.17	1.00
Backward TP	0.21	1.00	0.15	1.00

Following Gomez & Gerken (1999), the 30 sentences were randomly grouped into six sets of 5 (henceforth "samples"). Using the same word tokens recorded for Experiments 1 and 2, the five sentences of each sample were concatenated in the Audacity sound editor with an isi of 1000 ms in a random order. Each familiarization sample was approximately 18 s in duration.

Given its success in Experiments 1 and 2 and given the short attention span of infants, only the Movement Test was used here. In particular, CDEAB vs. DEABC was used. The test consisted of 4 items, which are shown below.

(74) Movement test

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Type	Sentences	Type	Sentences
Movement test	1	CDEABCD	JES SOT FAL KOF HOX	DEABCBC	SOT FAL KOF HOX JES
	2		REL ZOR TAF DAZ NEB		ZOR TAF DAZ NEB REL
	3		TID LUM RUD MER LEV		LUM RUD MER LEV TID
	4		TID ZOR RUD MER NEB		ZOR RUD MER NEB TID

Two random orders were generated for each type (i.e., CDEAB and DEABC), resulting in four test samples (two Grammar 1-consistent and two Grammar 2-consistent). The test sentences were concatenated in the same way as the presentation set in the Audacity sound editor with an isi of 1000 ms. Each test sample was approximately 14.6 s in duration.

Procedure

We used the head-turn preference procedure (Jusczyk & Aslin 1995, Kemler Nelson, Jusczyk, Mandel, Myers, Turk & Gerken 1995). Each infant was held on their parent's lap. The parent was seated in a chair in the center of the test booth. Throughout the experiment, the parent listened to music on an iPod over Sennheiser PXC 250 noise canceling headphones with Sennheiser NoiseGard. There was a TV screen in the center front of the room and two flashing lights on each side of the sidewalls. There was also a loudspeaker under each sidelight.

In order to familiarize the infant with the head-turn procedure, the experiment began with a practice music trial. Each trial began by showing a colorful picture on

the TV screen in the front. When the infant looked at the TV screen, the picture disappeared and one of the sidelights began to flash. The side of the flashing light was determined randomly by a computer program each time. When the infant made a head turn of at least 30° in the direction of the flashing sidelight, the audio sample began to play and continued until its completion or until the infant failed to maintain the 30° head turn for 2 consecutive seconds. If the infant turned away briefly, but looked back again within 2 s, although the time spent looking away was not included in the count, the audio continued playing. The light kept flashing whenever a sample was playing. When the sample completed or the infant looked away for more than 2 s, the audio and the flashing light stopped and the centering picture appeared on the TV again. And the same procedure was repeated.

A camera placed on top of the TV videotaped the infant. The experimenter watched the infant on a TV screen in the adjacent control room, but they could not hear any audio. The experimenter recorded the actions of the infant by pressing the buttons (*center*, *right*, *left* or *away*) using the computer program. Since the computer program randomly picked which sample to play each time and the experimenter could not hear any audio, they were always blind as to which sample was playing on a particular trial.

The practice music trial lasted about 1 min. The familiarization phase began right after the music trial. During the familiarization, the maximum amount of time an infant was allowed to keep looking at a particular side was 40 s ("maximum block length"). The six acquisition samples were played in a random non-repeating order each time. If the infant kept looking past the length of a sample (18 s), another sample

started without a break. If the infant looked away for more than 2 consecutive seconds, the language sample terminated even if this meant truncating a string in midstream. In that case, the same sample was played from where it was cut off in the next trial. Every infant accumulated a minimum of 70 s familiarization ("switch criterion") before going on to the test phase. This amounts to approximately 19.5 sentences.

During the test phase, the four test samples were played in a random non-repeating order. Here, the maximum block length was 90 s. The four samples were divided into two groups – Group 1 (Grammar 1-consistent) and Group 2 (Grammar 2-consistent). During the test, there was no minimum length a child had to accumulate (i.e. no switch criterion), instead, Group 1 was played once and Group 2 was played once. Half of the infants heard Group 1 first and the other half heard Group 2 first. If the infant kept looking past the length of a sample (14.6 s), another sample from the same group started without a break. Unlike the acquisition phase, if the infant looked away for more than 2 consecutive seconds, that particular sample terminated and was never played again.

The cycle of a familiarization phase and a test phase were repeated up to 3 times. The procedure for the first cycle is as stated above. The second and third cycles were shorter in length in that the switch criterion was 35 s instead of 70 s. The test phase remained the same. If the infant got fussy or started crying, the experiment was stopped. If it stopped before it got to the test phase of the first cycle, that infant's data was not included in the analysis. If it stopped after the first test or second test phase,

the data was included up to that point. Therefore, among the included infants, the accumulated familiarization time could vary approximately from 70 s to 140 s.

4.1.2 Hypotheses and predictions

Unlike the adult experiments, because we cannot ask infants whether they think the test sentence is grammatical in the artificial language, what we have as a measure is looking times to the two test samples. Here, let us review our three hypotheses.

Table 16: Table of hypotheses

	Deductive power of learner	*Nature of predetermined representations*
Limited Hypothesis	Limited to observed distributions	None
Beyond and Constrained Hypothesis	Beyond what can be derived from observed distributions	Limited by constraints found in natural language
Beyond and Unconstrained Hypothesis	Beyond what can be derived from observed distributions	Unlimited by constraints found in natural language

The first hypothesis, Limited Hypothesis is the most conservative hypothesis out of the three, since the generalization that is formed from the input is solely based on what was observed. In case of the movement test, the infants would assume that you can only move something that you have seen moved in the input. For example, if you had heard Grammar 1 during the familiarization, you have seen structures ABCDE and CDEAB among others. At the test phase, when you are presented with two

sentences (CDEAB vs. DEABC), you would be able to tease them apart, because one of them (CDEAB) you have already seen and the other one has a structure that was never seen before.

Hypothesis Two – Beyond and Constrained Hypothesis: The generalization the infants form is that you can move more than just what you saw moved in the input but it has to be a constituent in the artificial language. You cannot move non-constituents. Under this hypothesis, infants would be able to tell apart the two test samples, since one moves a constituent while the other moves a non-constituent. This hypothesis predicts the same outcome as predicted by Limited Hypothesis but for different reasons. This hypothesis presupposes antecedently known knowledge, whereas the Limited Hypothesis does not involve such knowledge.

Hypothesis Three – Beyond and Unconstrained Hypothesis: What infants learn from the input is that you can move any neighboring words. If this is the case, both test samples would be allowed since both have neighboring words moved to the front (CDE and DE). So at test, infants would equally listen to the two test samples.

All of these three hypotheses are compatible with the input data that infants receive. Out of the three hypotheses, the only one that predicts a different outcome from the other two is Beyond and Unconstrained Hypothesis. The predictions of each hypothesis are summarized in the table below.

Table 17: Predictions for Experiment 3

	Views	*Predictions*
Limited Hypothesis	Only the observed test sentences are grammatical	Infants will show a difference in looking times
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times

4.1.3 Results and discussion

The time that each infant oriented to the loudspeaker on each trial was recorded. Infants accumulated an average of 114.08 s acquisition time during the familiarization phase (range: 70.84 – 195.77 s). Four infants completed 3 cycles, another four infants completed 2 cycles and the rest did not complete more than 1 cycle due to fussiness. Means of infants' looking times during the test phase were computed separately for Group 1 and Group 2. For the infants who heard Grammar 1 during the familiarization, samples in Group 1 were consistent with their learned grammar. Likewise, for infants who heard Grammar 2 as input, Group 2 was consistent with their input grammar. Consequently, for Grammar 1 infants, we coded the looking times to Group 1 as "consistent" and the looking times to Group 2 as "inconsistent". For Grammar 2 infants, looking times to Group 1 was coded "inconsistent" and looking times to Group 2 was coded "consistent".

We first provide the data from just the first trial, since everyone completed the first trial, whereas not everyone completed the other two trials. The mean looking time at either side during the test was 6.79 s. The standard deviation was 7.14 s. The

data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated the trial from one infant who looked at a side for over 37 s. The remaining infants in both conditions looked longer to the group that was inconsistent with their input grammar (mean = 8.50 s) than the group that was consistent with the input grammar (mean = 4.12 s). This difference was significant in a two-tailed Paired Samples t-test ($t(12) = -2.423$, $p = 0.032$, $r = 0.57$) and in a two-tailed Wilcoxon Signed Ranks test ($Z = -2.20$, $p = 0.028$). 11 out of 13 infants had longer average looking times for the inconsistent samples, which is significant by Sign Test ($p = 0.0225$).

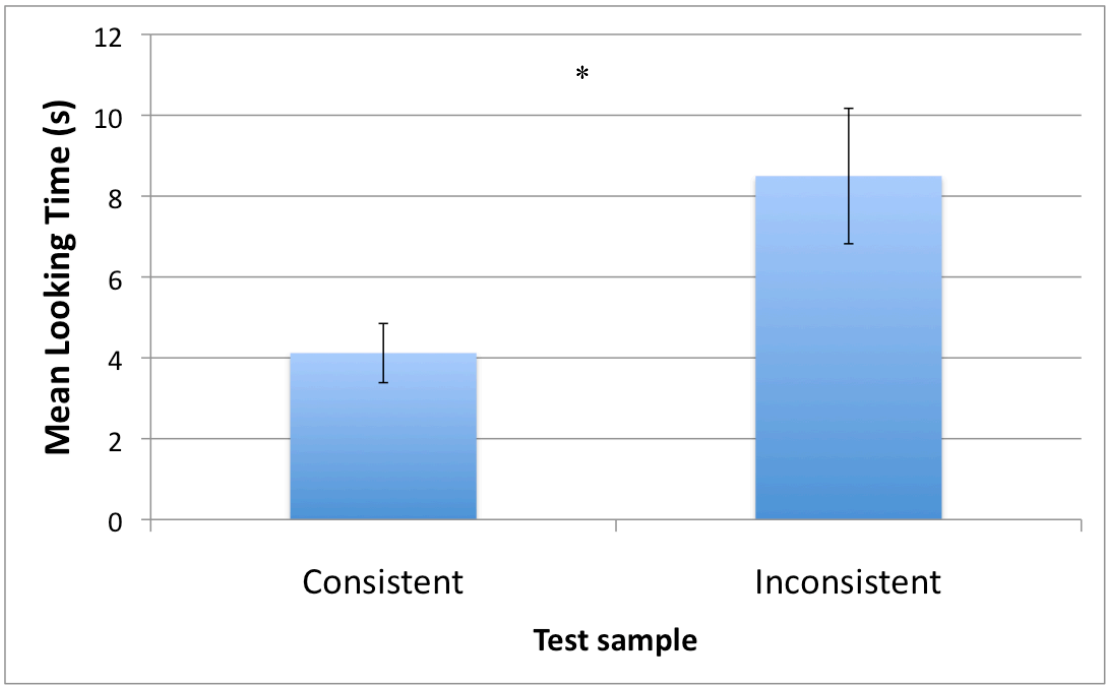


Figure 47: Experiment 3 results of the first trial

Now we provide the results from all trials. For infants who completed more than one trial, the looking times were averaged. The mean looking time at either side during the test was 8.63 s (SD = 10.43). Again, the data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated three trials of two infants. The results show that infants in both conditions looked longer to the inconsistent samples (mean = 8.64 s) than the group that was consistent with the input grammar (mean = 4.62 s). This difference was significant in a two-tailed Paired Samples t-test ($t(13) = -2.541$, $p = 0.025$, $r = 0.58$) and in a two-tailed Wilcoxon Signed Ranks test ($Z = -2.23$, $p = 0.026$). 11 out of 14 infants had longer average looking times for the inconsistent samples.

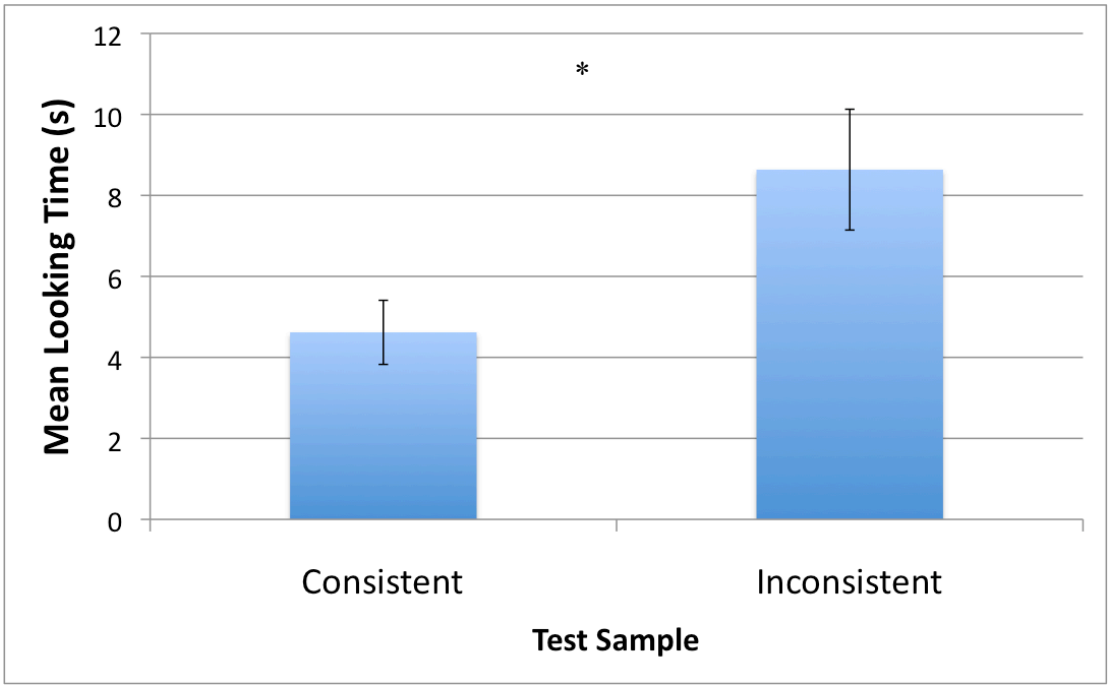


Figure 48: Experiment 3 results of all trials

Specifically, Grammar 1 infants listened longer to test sentences of the type DEABC (inconsistent) that they have never seen before than to CDEAB (consistent) which was already familiar to them. In the latter, CDE, which is a constituent in Grammar 1 is moved to the front of the sentence, while in the former, DE, which is not a constituent in Grammar 1 is moved. And vice versa for the infants in the Grammar 2 condition. This result suggests that infants showed a novelty preference.

The purpose of Experiment 3 was to determine whether infants can learn the phrase structure on the basis of statistical information. After only 114 s (less than 2 min) of exposure, infants distinguished samples that were consistent with their input grammar from those that were inconsistent with their input grammar, as reflected by significantly longer looking times to the inconsistent samples. Moreover, the fact that infants showed a novelty preference by listening longer to the inconsistent samples than the consistent sample at test tells us that they were familiarized with the artificial language very quickly and, by the time they got to the test phase, that the infants already had a well-established representation of the grammar so that new instances that were consistent with the grammar were not as interesting as new instances that were inconsistent with their established representation. Finally, the results of Experiment 3 indicate that 18-month-old infants can learn the phrase structure of an artificial language on the basis of statistical distribution without any prosodic or semantic information. This supports our claim that statistics can be one of the information sources for bootstrapping phrase structure.

One alternative account for this result could be that the infants did not really have a hierarchical tree representation like we argue, but that they were simply noticing the chunks of constituents in the consistent (grammatical) test sentences. That is, in the consistent test sample, "good" transitions exist, meaning transitions from a category to another category that has been observed (i.e., constituents), whereas in the inconsistent test sample, "bad" transitions exist, meaning the transition from a category to another category that was not observed in the data (i.e., non-constituents). One could argue that the results in this experiment could be attained if the participants were merely noticing the "good chunks" (constituents) versus "bad chunks" (non-constituents). While this is a relevant concern, it cannot have been the case. Take a look at the movement test sentences that we used.

(75) Movement test

Grammatical in Grammar 1	Grammatical in Grammar 2
CDEAB	DEABC

There are two good chunks in CDEAB if the familiarization language was Grammar 1, namely CD and AB. If your familiarization language was Grammar 2, then there are two good chunks in DEABC, namely DE and BC. What is important, however, is that there are good chunks of one grammar in the other grammar's consistent test sentences as well. Put another way, the inconsistent test sample in your grammar always contains good chunks of your grammar too. For instance, if you were familiarized with Grammar 1, there is a good chunk (i.e., AB) in the G2-consistent

answer, too. If you were familiarized with Grammar 2, there is a good chunk (i.e., DE) in the G1-consistent answer. This is illustrated in Figure 39 below. Solid lines represent good chunks (i.e., constituents) in G1, and dotted lines represent constituents in G2. In this way, simply noticing the "chunks" would not achieve the results of this experiment, thus we can reject that alternative account.

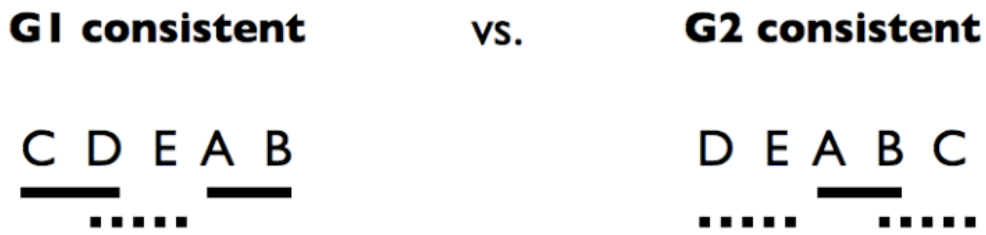


Figure 49: Number of "good chunks" vs. "bad chunks"
Solid line represents good chunks for G1 and dotted line represents good chunks for G2

Going back to our three hypotheses, the results from Experiment 3 are compatible with both Limited Hypothesis and Beyond and Constrained Hypothesis, but not Beyond and Unconstrained Hypothesis.

Table 18: Predictions and outcomes for Experiment 3

	Views	*Predictions*	*Outcome*
Limited Hypothesis	Only the observed test sentences are grammatical	Infants will show a difference in looking times	✓
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times	✓
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times	✗

Only the Beyond and Unconstrained Hypothesis predicted that infants would show no difference in looking times to the two test samples, however the infants did show a significant difference in looking times. Both Limited Hypothesis and Beyond and Constrained Hypothesis predicted this outcome, so at this point, we cannot determine which of these two hypotheses might be correct. The "consistent" test sample had a familiar structure to the infants and it had a moved constituent. The "inconsistent" test sample had an unfamiliar structure and it had a moved non-constituent. This is why the two hypotheses (Limited Hypothesis and Beyond and Constrained Hypothesis) predict the identical outcome. In the following experiments, we hope to tease these two hypotheses apart by having a different kind of test items.

4.2 Experiment 4 (Infant 2)

Experiment 4 is an attempt to replicate Experiment 3 (Infant 1) with 12-month-old infants instead of 18-month-olds. Given that by the age of 12 months, infants are sensitive to statistical distributions in an artificial language generated by a finite state grammar (Gomez & Gerken, 1999; Saffran et al., 2008), we wanted to see whether infants younger than 18 months are also sensitive to the distributional information signaling the hierarchical phrase structure.

4.2.1 Method

Participants

Fifteen infants, approximately 12 months of age were tested (age range: 11;29 – 13;13; mean: 12;22). Eighteen additional infants were tested but excluded from analyses for the following reasons: fidgeted and did not complete test (n = 16) and equipment failure (n = 2). The infants were randomly divided between two familiarization conditions. Half of the infants (n = 8) heard Grammar 1 as input during the familiarization period and the other half (n = 7) heard Grammar 2. Parental consent was obtained prior to testing, in accordance with the NIH standards for the ethical treatment of human subjects.

Material and procedure

The material (familiarization and test items) and the procedure (head-turn preference procedure) used in this experiment were identical to Experiment 3 (Infant 1).

4.2.2 Results and discussion

The time that each infant oriented to the loudspeaker on each trial was recorded. Infants accumulated an average of 131.18 s of acquisition time during the familiarization phase (range: 78.78 – 183.41 s). Nine infants completed 3 cycles and six infants completed 2 cycles.

For the infants who heard Grammar 1 during the familiarization, we coded the looking times to test items that were G1-consistent as "consistent" and the looking times to test items that were G2-consistent as "inconsistent". Likewise, for infants who heard Grammar 2 during the familiarization, looking times to G2-consistent test items were coded "consistent" and looking times to G1-consistent test items were coded "inconsistent".

We first provide the data from just the first trial, since everyone completed the first trial. The mean looking time at either side during the test was 7.99 s. The standard deviation was 7.53 s. The data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated the trial from two infants who looked at a side for over 30 s. The remaining thirteen infants looked at the "consistent" sample for the average of 4.64 s and the "inconsistent" sample for the average of 6.40 s. This difference was not

significant in a two-tailed paired samples t-test ($t(12) = -1.484$, $r = 0.39$, $p = 0.164$). 9 out of 13 infants had longer average looking times for the inconsistent samples.

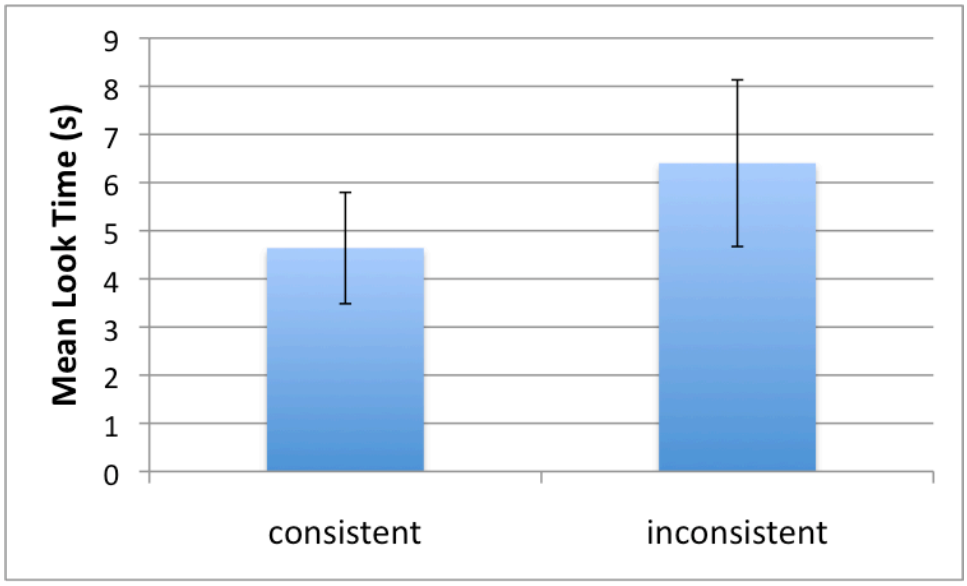


Figure 50: Experiment 4 results of the first trial

Now we provide the results from all trials. For infants who completed more than one trial, the looking times were averaged. The mean looking time at either side during the test was 8.42 s (SD = 9.70). Again, the data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated one trial of an infant. The results show that infants looked to the consistent samples for the average of 6.36 s, and to the inconsistent samples for the average of 7.06 s. This difference was not significant in a two-tailed Paired Samples t-test ($t(14) = -0.49$, $r = 0.13$, $p = 0.631$). 9 out of 15 infants had longer average looking times for the inconsistent samples.

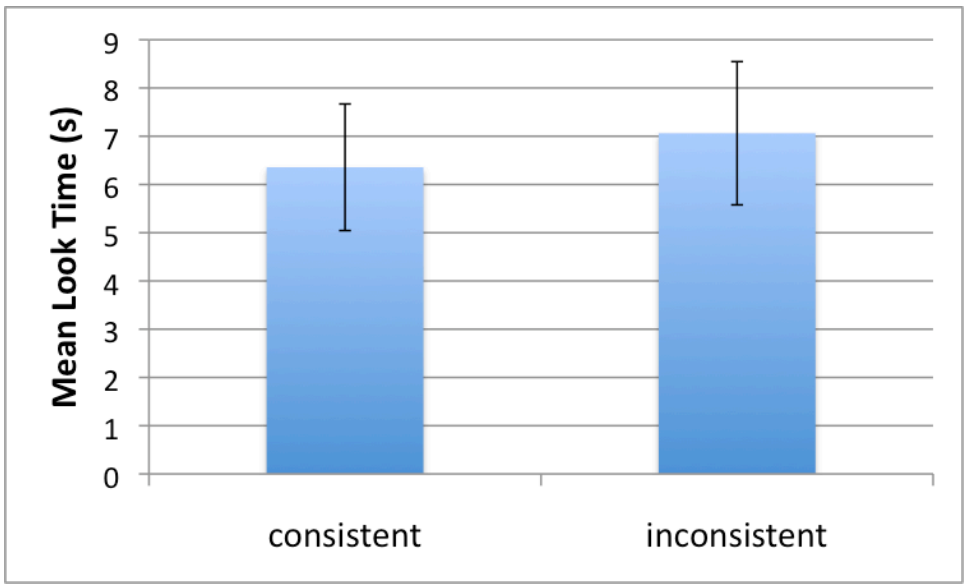


Figure 51: Experiment 4 results of all trials

This experiment tried to replicate Experiment 3 (first infant experiment), using the same stimuli and procedure, with 12-month-old infants, however, the results were inconclusive. On average, the 12-month-olds had a slight tendency to look longer at the inconsistent samples just like the 18-month-olds, but the difference did not reach statistical significance. There could be a few possible causes for this null result.

First possibility is that we did not run enough subjects, and had we run more infants, the difference might have reached significance. A power analysis revealed that in order to achieve a power of 0.8, we need to run 76 subjects in total. Thus, it is possible that if we run more subjects, we will obtain an effect.

Second possibility is that the head-turn preference procedure might not have been suited for the infants of this age. We did lose more than half of the infants we ran because they were too fidgety and did not pay attention. It could have been that

the artificial languages used in our study were too boring or too complex for them to stay attentive. However, given the fact that past research (Gomez & Gerken 1999; Shady 1996; Saffran et al. 2008) was successfully conducted using this method indicates that this probably was not the biggest issue.

Third, the stimuli could have been too complex and it might have been impossible for 12-month-olds to track this type of distributional information. The only cue to the structure was varying transitional probabilities between words within phrases and across phrases. Past studies have showen that much younger infants (8 months) are capable of tracking transitional probabilities between syllables to learn word boundaries (Saffran et al. 1996a). In Saffran et al. (1996a), however, all that the infants had to do was figure out the word boundaries – there was no hierarchy involved in the input or the test items. All they had to do was track the linear distribution. Similarly, although the 12-month-olds successfully learned the syntactic system of an artificial language in Gomez & Gerken (1999) and Saffran et al. (2008), the artificial languages in those past studies did not necessarily involve hierarchy – the system could have been learned through tracking the linear word order. In both Gomez & Gerken (1999) and Saffran et al. (2008), the experimental results could have been achieved by infants learning finite state grammars. The infants in those studies did not need to have had a hierarchical representation of structure. On the other hand in our study, not only do infants have to track the distributions, they also have to build or map the abstract hierarchical phrase structure based on them. It could have been that the 12-month-old infants were too young to do that mapping, because their computational capacities do not yet allow them to make inferences from this

type of statistics. What we speculate is that 12-month-olds require more exposure to allow them to draw relevant syntactic conclusions.

4.3 Experiment 5 (Infant 3)

In the previous infant experiments in this dissertation, the familiarization input included sentences created via movement rules. And at test, they were given two kinds of sentences – sentences in which a constituent in the input grammar was moved and sentences in which a non-constituent was moved. We are calling these *consistent* and *inconsistent* test samples respectively. The infants had seen the structure of the consistent test sentences in the input, although the exact word sequences were new. And the results of the first infant experiment indicate that the 18-month-olds could at least differentiate the consistent and inconsistent test samples. However, because the structure of the consistent test items was already seen during the familiarization, it does not tell us whether infants can extend what they have learned to novel structures. More specifically, the results of Experiment 3 (Infant 1) were compatible with both Limited Hypothesis and Beyond and Constrained Hypothesis, since both predicted the identical outcome. This was because the "consistent" test sample in Experiment 3 (Infant 1) had a familiar structure to the infants and it had a moved constituent. The "inconsistent" test sample had an unfamiliar structure and it had a moved non-constituent. The Limited Hypothesis is the view that the deductive power of a learner is limited to the observed distributions,

and statistical learning does not interact with innate constraints. Beyond and Constrained Hypothesis, on the other hand, is the view that the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions. The results of Experiment 3 (Infant 1) do not support one or the other of the two. But given that this is the critical question in this dissertation, we would want to find a way to tell them apart.

In this experiment, we will use something the infants have not seen in the input as test sentences. In particular, the input will include movement of some constituents, but during the test phase, we will present sentences that move different constituents than in the familiarization. If all they learn is entirely based on the input, they might not be able to differentiate the two test samples, because neither structure was seen in the input. On the other hand, if they could generalize beyond what was observed, then they may be able to differentiate what is linguistically possible, but novel, and what is linguistically impossible and novel.

4.3.1 Method

Participants

Twenty-four infants, approximately 18 months of age were tested (age range: 17 months 6 days to 19 months 18 days; mean: 18 months 28 days). Seven additional infants were tested but excluded from analyses because of fussiness (n = 5) and equipment failure (n = 2). Parental consent was obtained prior to testing, in accordance with the NIH standards for the ethical treatment of human subjects.

Material

The artificial language and words used in this experiment were identical to Grammar 2 in the preceding infant experiments in this study. We only used Grammar 2 in this experiment and not Grammar 1. The reasons for this will be explained in the following section where we talk about test sentences. The TP patterns of the presentation set are given in the table below.

Table 19: Transitional probabilities for 30 input sentences in Experiment 5

	A-B	B-C	C-D	D-E
Forward TP	0.28	1.00	0.17	1.00
Backward TP	0.21	1.00	0.15	1.00

Just like in previous infant experiments, the familiarization input included operations on constituents seen in natural languages like optionality, repetition, substitution by a pro-form, and most importantly, movement. However, only one constituent was moved in the familiarization set, namely DE. No other constituents were moved in the input. The following are PS trees for sentences in the artificial grammar used in this experiment with and without movement.

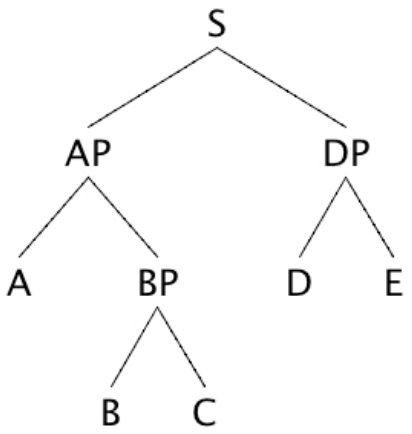


Figure 52: PS tree for a familiarization sentence without movement

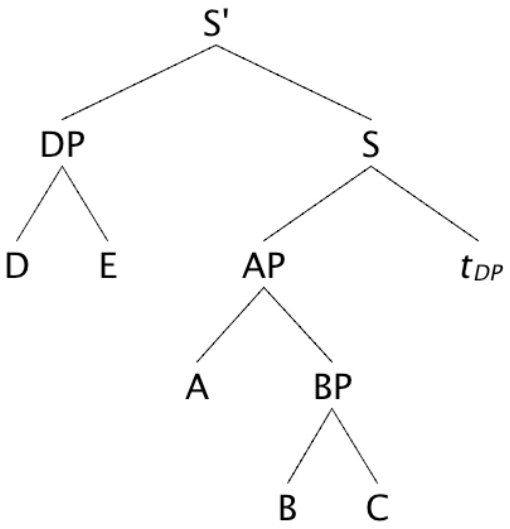


Figure 53: PS tree for a familiarization sentence with movement

Test items

Notice that the movement rules in the familiarization front the constituent DE to the front (Figure 53). At test, we moved a *different* constituent (and non-constituent) to the front. Test sentences are: BCADE and CDABE. The test consisted

of 4 test items, which are shown below. None of the word sequences in all the test items appear in the familiarization set.

(76) Test sentences in Experiment 5

Grammatical		*Ungrammatical*	
Type	Sentences	Type	Sentences
BCADE	HOX JES KOF SOT FAL	CDABE	JES SOT KOF HOX FAL
	NEB REL DAZ ZOR TAF		REL ZOR DAZ NEB TAF
	LEV TID MER LUM RUD		TID LUM MER LEV RUD
	NEB TID MER ZOR RUD		TID ZOR MER NEB RUD

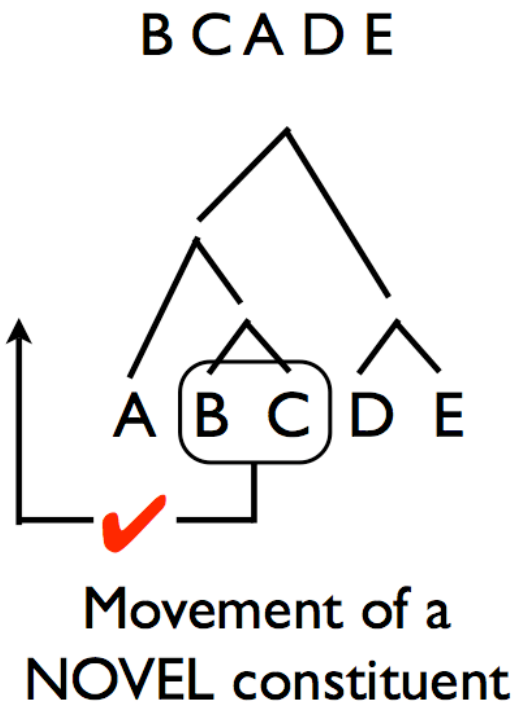


Figure 54: "Consistent" test sample

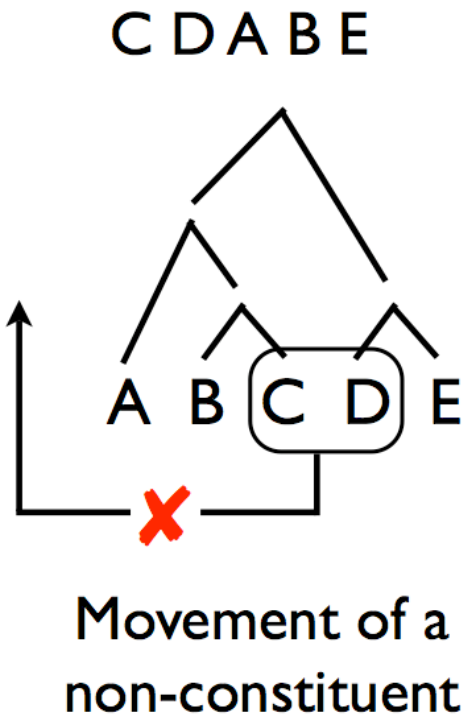


Figure 55: "Inconsistent" test sample

In the test sentence of the structure BCADE, what is moved to the front is BC, which is a constituent in this artificial language. Since movement of a constituent is a possible rule in natural languages, this test sentence is consistent with constraints of natural language, and we will call this the "consistent" test sample. On the other hand, in the test sentence of the structure CDABE, what is moved is CD, which is not a constituent in this language. Because movement of a non-constituent is an impossible rule in natural languages, this test sentence violates constraints of natural languages, and we call this the "inconsistent" test sample.

The test sentences were concatenated in the same way as the presentation set in the Audacity sound editor with an isi of 1000 ms. Each test sample was approximately 12.7 s in duration.

It should also be pointed out that in this experiment, we used only one artificial language (which is identical to Grammar 2 in previous experiments in this dissertation). The reason for this is because we needed a grammar in which two different constituents can be moved to the front. In Grammar 2 (the grammar used in this experiment), the canonical sentence is [[A[BC]][DE]]. There are two separate constituents that can be moved around – BC and DE. We needed to have two distinct constituents to be moved to the front, so that in the familiarization set, we can have one of them to move to the front, and at test, we can move the other constituent to move to the front.

In Grammar 1, the canonical sentence is [AB][[CD]E]. In this grammar, CD (and CDE) can be moved to the front, but AB cannot. More specifically, you could move AB to the front, but it will not make any difference in terms of linear sound sequences and it is impossible to let the participants know that we intended to have moved AB to the front. That is why we did not use Grammar 1 in this experiment.

Procedure

A slightly modified version of the head-turn preference procedure (Jusczyk & Aslin 1995, Kemler Nelson, et al. 1995) was used for this experiment. The difference between this and the other procedure is that in this procedure, the familiarization period was fixed (2 min) with a silent movie playing on the TV screen. This is the same procedure as the one reported in Gerken (2004, 2006), Gerken, Wilson & Lewis (2005) and Gerken & Bolt (2008).

The reason we changed to this procedure is because we noticed when we ran previous experiments that a number of infants could not pay long enough attention to fully participate in the head-turn procedure during the familiarization. In regular head-turn preference experiments, infants must actively participate in the head-turn looking task even during the familiarization period, because the familiarization audio plays only when the infants look at the flashing light. Therefore, even though there is a minimum amount of familiarization time that all infants must accumulate, in a way, the familiarization phase is infant-controlled, and how much input they get depends on how attentive they are. Because of this, we noticed that a number of infants were not attentive enough to accumulate required amount of input and we were losing a lot of subjects this way.

In the procedure for this experiment, each infant was held on their parent's lap, while the parent was seated in a chair in the center of the testing booth. Throughout the experiment, the parent listened to music on an iPod over Sennheiser PXC 250 noise canceling headphones with Sennheiser NoiseGard. There was a TV screen in the center front of the room. During the familiarization period, infants watched a silent movie, while the audio input of the artificial language played continuously from the speakers for 2 solid minutes. The video used in this experiment was of dogs flying in slow motion. This way, every infant received an equal amount of familiarization time (2 min).

After the familiarization phase, the test phase began. The test phase was basically the same as that of regular head-turn procedure. The only difference was that instead of flashing lights on sidewalls, the TV screen was used as the attention

getter. The test phase began by showing a colorful picture on the TV screen in the front. When the infant looked at the TV screen, the audio of one of the test samples started to play from the speakers, and continued until the infant failed to maintain the look at the TV screen for 2 consecutive seconds. If the infant turned away briefly, but looked back again within 2 s, although the time spent looking away was not included in the count, the audio continued playing. The colorful picture remained on the TV screen whenever a sample was playing. When the infant looked away for more than 2 s, the audio stopped and the TV went blank. Then, the colorful picture appeared on the TV again and the same procedure was repeated for the other test sample. Which test sample (consistent or inconsistent with the input grammar) played first was randomly determined by a computer program each time.

A camera placed on top of the TV videotaped the infant. The experimenter watched the infant on a TV screen in the adjacent control room, but they could not hear any audio. The experimenter recorded the actions of the infant by pressing the buttons (*look* or *away*) using the computer program. Since the computer program randomly picked which sample to play each time and the experimenter could not hear any audio, they were always blind as to which sample was playing on a particular trial.

4.3.2 Hypotheses and predictions

There are a few possible outcomes in this experiment. Hypothesis One (Limited Hypothesis): The most conservative hypothesis would be that the generalization that is formed from the input is solely based on what was observed.

You can only move something that you have seen moved in the input. Specifically in this case, what infants learn from the familiarization input is that you can only move the constituent DE in this artificial language, and nothing else. Under this hypothesis, then both test samples would be considered illicit, because both have something that is not DE moved (BC and CD).

Hypothesis Two (Beyond and Constrained Hypothesis): A less conservative hypothesis would be that it involves a more abstract generalization that goes beyond what was observed in the input. For instance, it might be hypothesized that what infants learn from this familiarization is that you can move constituents in the artificial language, but not non-constituents, which is compatible with the input they get. Under this hypothesis, infants would consider the "consistent" test sample to be licit, while the "inconsistent" test sample would be illicit. This is the hypothesis that is most compatible with what natural languages are like. That is because natural languages allow movement of a constituent but not of a non-constituent.

Hypothesis Three (Beyond and Unconstrained Hypothesis): What infants learn from the input is that you can move any two neighboring words. This hypothesis is also compatible with the input data they receive. If this were the case, both test samples would be allowed since both have two neighboring words moved to the front (BC and CD). This is the most liberal hypothesis out of the three hypotheses in that it maximally allows what you can move. The three hypotheses are summarized in the table below.

Table 20: Table of hypotheses

	Deductive power of learner	*Nature of predetermined representations*
Limited Hypothesis	Limited to observed distributions	None
Beyond and Constrained Hypothesis	Beyond what can be derived from observed distributions	Limited by constraints found in natural language
Beyond and Unconstrained Hypothesis	Beyond what can be derived from observed distributions	Unlimited by constraints found in natural language

Because we cannot ask infants whether they think the test sentence is acceptable or grammatical in the artificial language or not, what we have as a measure is looking times to the two test samples. If Limited Hypothesis is correct, we should see no difference in looking times to the consistent or inconsistent samples, because both would be considered ungrammatical by the infants.

If Beyond and Constrained Hypothesis is true, we would see a difference in looking times although we do not have a prediction as to which sample the infants would look longer at. One possibility is that they would look longer at the consistent test sample, because they think that is the "grammatical" sentence and it is compatible with the input grammar. On the other hand, they might look longer at the inconsistent test sample because that is the "ungrammatical" sentence which violates the structure of the input grammar. Notice that both types of test samples are "novel" in this experiment, because both test samples involve a new structure that was not included in the input. Therefore, we cannot simply predict that the infants would show a novelty preference.

If Beyond and Unconstrained Hypothesis is correct, the infants should show no difference in looking times, since both test samples would be considered licit. So only if Beyond and Constrained Hypothesis is correct should we see a difference in infants' looking times. The predictions for each hypothesis are summarized in the following table.

Table 21: Predictions for Experiment 5

	Views	*Predictions*
Limited Hypothesis	Both test sentences are ungrammatical	Infants will not show a difference in looking times
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times

4.3.3 Results and discussion

The time that each infant oriented to the loudspeaker on each trial was recorded. All infants accumulated 2 min of acquisition time during the familiarization phase.

There were two types of test sentences – BCADE and CDABE. BCADE involved movement of a *novel* constituent, so we will call it the "consistent" test sample in the following analyses. CDABE involved movement of a non-constituent, so this will be called the "inconsistent" test sample.

First, we report the results in terms of raw looking times. The mean looking time at either test sample was 17.66 s (SD = 19.73 s). The data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated trials from two infants who listened to a test sample for over 67 s. The remaining 22 infants, on average, looked longer to the inconsistent test sample (mean = 18.97 s) than the consistent test sample (mean = 10.16 s). This difference was significant in a two-tailed Paired Samples t-test ($t(21) = -2.489$, $p = 0.021$, $r = 0.48$) and in a two-tailed Wilcoxon Signed Ranks test ($Z = -2.451$, $p = 0.014$). 17 out of 22 infants had longer average looking times for the inconsistent samples. This difference was significant in Sign Test ($p = 0.017$).

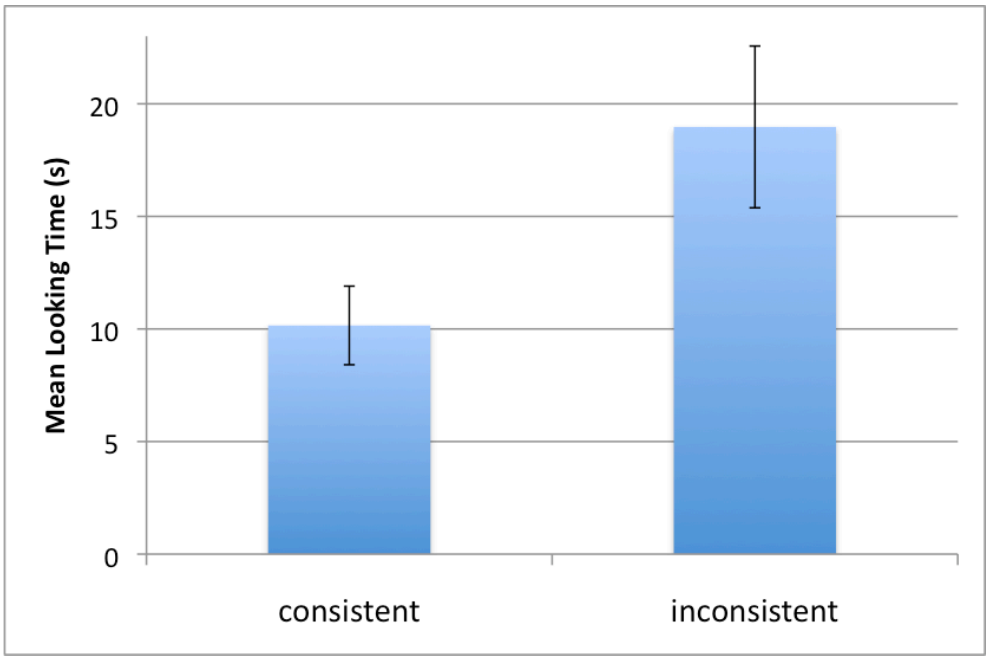


Figure 56: Experiment 5 results. Mean looking time (s)

Next, we report the proportion of looking time to each test sample. This is because we found a major tendency of infants looking to the first sample presented to them during the test phase much longer than the second sample presented to them, regardless of their consistency to the input grammar (consistent or inconsistent). In particular, infants listened to the first test item on average for 22.26 s and to the second test item for 13.07 s (two-tailed Paired Samples t-test: $t(23) = 1.784$, $p = 0.088$; two-tailed Wilcoxon Signed Ranks test: $Z = -1.943$, $p = 0.052$).

This tendency to look at the first sample presented to them longer is quite natural considering the setup. The infants first watch a silent movie for 2 minutes (while the artificial language plays from the speakers), then on the TV, a bright, colorful picture comes up which infants see for the first time. In this situation, it is expected that the infants get interested in the new picture and look at it for a long time the first time they see it.

To avoid this tendency to influence the data of the results, we took the mean of each infant's looking times to the first and second test items, then we calculated the looking times to the first and second test items as a proportion over the mean. The data from the infants whose looking time during the test phase was over 2.5 standard deviations from the mean was not included in the analyses. This eliminated trials from two infants who listened to a test sample for over 67 s. The remaining 22 infants, on average, looked longer to the inconsistent test sample (mean = 1.267) than the consistent test sample (mean = 0.733). This difference was significant in a two-tailed Paired Samples t-test ($t(21) = -3.226$, $p = 0.004$, $r = 0.58$) and in a two-tailed Wilcoxon Signed Ranks test ($Z = -2.711$, $p = 0.007$). 17 out of 22 infants had longer

average looking times for the inconsistent samples. This difference was significant in Sign Test ($p = 0.017$).

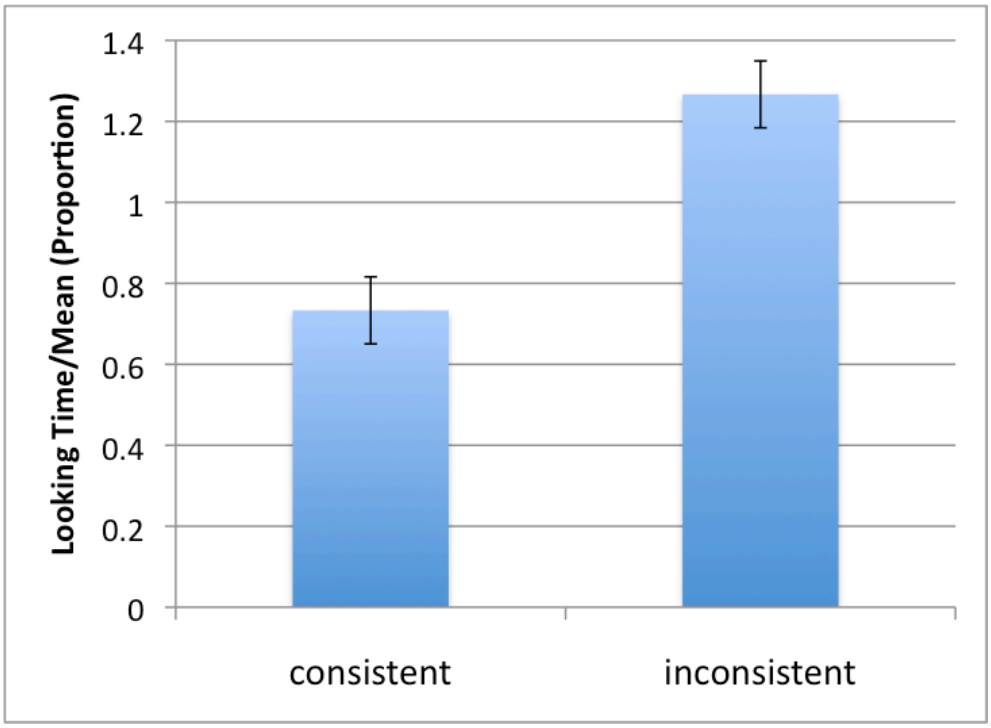


Figure 57: Experiment 5 results. Proportion of looking time over mean

The results show that 18-month-old infants looked longer at the inconsistent test sample, even with the consideration for the effect of longer look time towards the first test item. These results suggest that the distributional information can be used to cue phrase structure and that the infants can distinguish sentences which moved novel constituents vs. sentences which moved novel non-constituents.

Notice that the term "novelty" preference is not exactly accurate here, because both test samples were novel, and even the "consistent" test sample was never observed previously. More specifically, both test types involved a structure that was

not seen in the input (see Figure 54 and Figure 55 above). In this experiment, the familiarization set included sentences derived via a movement rule, but only one constituent, namely DE, was moved in the input (DEABC). The "consistent" test item (BCADE) was derived by movement of a constituent in the input language, but that constituent had never been moved in the input. So the structure was still novel. The "inconsistent" test item (CDABE) was derived by movement of a non-constituent in the artificial language. The results indicate that infants could distinguish the two types of test samples and they were not simply choosing what they had seen before, but at the very least, they were doing something new.

Furthermore, it should be noted here that this result could not have been obtained by a finite state grammar. The familiarization language in this experiment could be expressed by an FSA like the following.

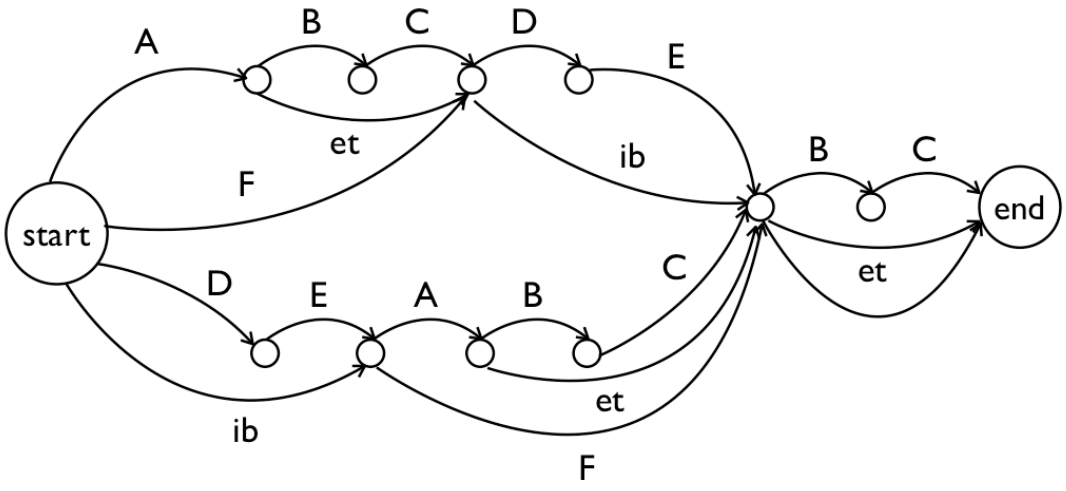


Figure 58: FSA for familiarization sentences in Experiment 5

However, if you only have the FSA like the one above, it is impossible to tell apart the two test sentences – BCADE and CDABE. Neither test structure (consistent or inconsistent) is represented in this FSA. Therefore, the infants must have had a phrase structure representation of the artificial language in order for them to succeed.

In terms of our three hypotheses, the predictions of the Limited Hypothesis and Beyond and Unconstrained Hypothesis were not borne out, since both hypotheses predicted no difference in infants' looking times. This suggests that the generalization the infants formed based on the received input is *not* that you can only move the constituent DE in this language. Furthermore, the generalization formed cannot be that you can move any two neighboring words either. The results of this experiment are only compatible with Beyond and Constrained Hypothesis, which predicted that infants would show a difference in looking times between the two test samples. This suggests that the generalization the infants formed was you can move any constituent in the language but you cannot move non-constituents. This is the only conclusion that is compatible with the results of this experiment.

Table 22: Predictions and outcomes for Experiment 5

	Views	*Predictions*	*Outcome*
Limited Hypothesis	Both test sentences are ungrammatical	Infants will not show a difference in looking times	✗
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times	✓
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times	✗

In other words, at least at the end of the experiment, the infants knew that moving constituents is a possible operation in language, but not moving non-constituents. If the infants have such general knowledge that moving constituents is allowed, they could extend that knowledge to allow movement of new constituents. But what is still unclear is where this knowledge came from. It could be that children knew this even before the familiarization period, or it could be that infants learned this general rule during the experiment. We will attempt to answer this question in our next experiment. More specifically, we will remove all the movement sentences from the familiarization input and test infants on the movement test (movement of a constituent vs. non-constituent). If the knowledge that you can only move a constituent is innate, then we would expect children to succeed at the task, whereas if the knowledge was acquired during the experiment and if the input information was necessary, then we would expect children to fail in the next experiment.

4.4 Experiment 6 (Infant 4)

The results of Experiment 5 (Infant 3) supported the Beyond and Constrained Hypothesis which indicates that infants formed a generalization that you can move a new element as long as it is a constituent. What is still unclear is that where that knowledge came from. Did infants form that generalization based on the knowledge they acquired during the 2-minute exposure to the artificial language? Or did they have that knowledge prior to the experiment? In this experiment, we will try to

answer this question by removing all the movement sentences (just as we did in Experiment 2 (Adult 2)) and test them on movement test.

There are a few possible outcomes. First possibility comes from a view that the input signal contains sufficient statistical regularities that guide the learner to arrive at the abstract representations (Elman et al. 1996; Bybee 1998; Tomasello 2000). On this view, a learner does not come with preexisting linguistic symbolic component, and learners' task is to collect and compile accurately predictive statistics from the data, thus the outcome of the learning is solely based on the observed input distributions. If this were the case, it is necessary to observe some constituent moving in the input to learn that you can move a constituent; consequently, we would expect infants to not be able to distinguish two test samples in this experiment.

Second possibility comes from a view that learners use statistics to simply identify particular abstract syntactic representations (Miller & Chomsky 1963; Yang 2006; Pearl 2007). On this view, the learner may come equipped with antecedent knowledge about possible linguistic structures and representations, and statistical learning interacts with that knowledge. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. If this were the case, it would not be necessary to have observed movement to learn that you can only move a constituent, then we would expect infants to be able to differentiate the two test samples.

4.4.1 Method

Participants

Thirty-one infants, approximately 18 months of age were tested (age range: 17 months 3 days to 19 months 13 days; mean: 18 months 10 days). Fourteen additional infants were tested but excluded from analyses because of fussiness (n = 11), crying (n = 2) and inattentiveness (n = 1). Parental consent was obtained prior to testing, in accordance with the NIH standards for the ethical treatment of human subjects.

Material

The same artificial grammars and words as in the previous experiments in this dissertation (Grammar 1 and Grammar 2) were used. The only difference was that all examples generated via movement rules were excluded from the familiarization. Just like in Experiment 3 (Infant 1), 30 sentences were picked as the presentation set. While the input lacked movement rules, it still included other manipulations such as repetition, optionality and substitution (by something other than pro-forms). In the Grammar 1 presentation set, one sentence (3.3%) was the canonical sentence type (ABCDE), three sentences (10%) involved substitution (ABF; i.e., the constituent CDE was replaced by F), six sentences (20%) involved repetition (ABCDECD) and twenty sentences (66.7%) involved both substitution and repetition (ABFCD; i.e., the constituent CDE was replaced by F and a constituent CD was added on the end). In the Grammar 2 presentation set, five sentences (16.7%) involved substitution (FDE; i.e., the constituent ABC was replaced by F), four sentences (13.3%) involved repetition (ABCDEBC) and twenty-one sentences (70%) involved both substitution

and repetition (FDEBC; i.e., the constituent ABC was replaced by F and a constituent BC was added on the end). These features contributed to make the TPs between words within phrases higher than the TPs across phrases. The resulting TP patterns of the presentation set are given below. All 30 sentences were randomized. The sentence types and 30 sentences that appeared in the presentation set are shown in Appendix E.

Table 23: Transitional probabilities for 30 input sentences in Grammar 1

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.23	1.00	0.21
Backward TP	1.00	0.21	1.00	1.00

Table 24: Transitional probabilities for 30 input sentences in Grammar 2

	A-B	B-C	C-D	D-E
Forward TP	1.00	1.00	0.14	1.00
Backward TP	0.14	1.00	0.13	1.00

Following Gomez & Gerken (1999), the 30 sentences were randomly grouped into six sets of 5 (henceforth "samples"). Using the same word tokens recorded for previous experiments, the five sentences of each sample were concatenated in the Audacity sound editor with an isi of 1000 ms in a random order. Each familiarization sample was approximately 17 s in duration.

The test sentences were identical to the ones in Experiment 3 (Infant 1). That is, only the movement test was used here as well, namely CDEAB vs. DEABC. The test consisted of 4 items, which are shown below.

(77) Movement test

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Type	Sentences	Type	Sentences
Movement test	1	CDEAB	JES SOT FAL KOF HOX	DEABC	SOT FAL KOF HOX JES
	2		REL ZOR TAF DAZ NEB		ZOR TAF DAZ NEB REL
	3		TID LUM RUD MER LEV		LUM RUD MER LEV TID
	4		TID ZOR RUD MER NEB		ZOR RUD MER NEB TID

Two random orders were generated for each type (i.e., CDEAB and DEABC), resulting in four test samples (two Grammar 1-consistent and two Grammar 2-consistent). The test sentences were concatenated in the same way as the presentation set in the Audacity sound editor with an isi of 1000 ms. Each test sample was approximately 14.6 s in duration.

It should be noted here that the exposure sentences could be generated by finite state grammars, like the following.

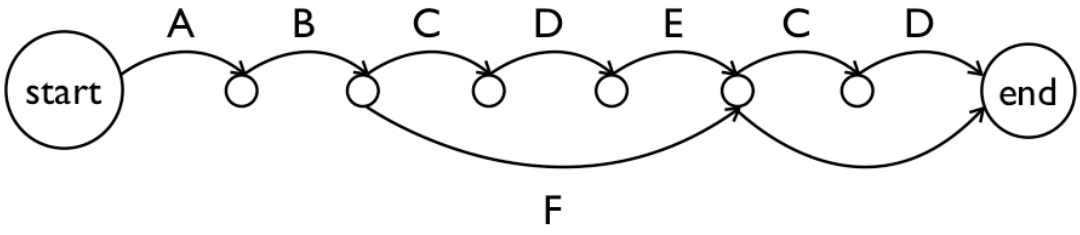


Figure 59: FSA for familiarization sentences of Grammar 1 in Experiment 6

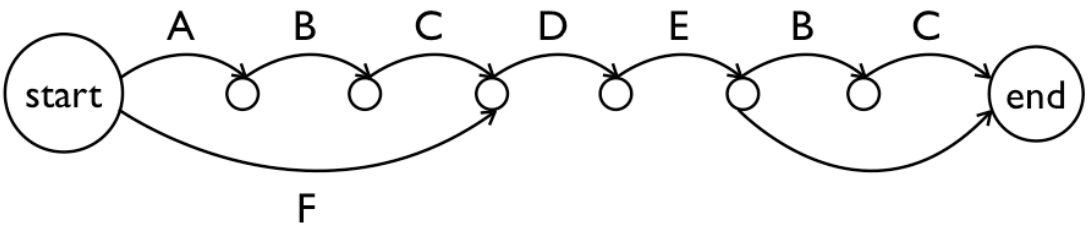


Figure 60: FSA for familiarization sentences of Grammar 2 in Experiment 6

But it should also be noted that these FSAs cannot generate the test sentences. Both types of test sentences (grammatical or ungrammatical) cannot be generated by those FSAs.

Procedure

The procedure we used in this experiment was identical to Experiment 5 (Infant 3) with fixed familiarization period. It was a slightly modified version of the head-turn preference procedure (Gerken 2004, 2006; Gerken, Wilson & Lewis 2005; Gerken & Bolt 2008) in that during the familiarization phase, the audio stimuli played continuously for 2 minutes while a silent movie (with moving laser lights) played on the TV screen. The main difference between this procedure and the regular head-turn procedure is that the familiarization phase is not infant-controlled here. The reason for this is so we would not lose so much data due to lack of attention and interest of the infants during the familiarization phase. The infants participate in the head turning only in the test phase, which was identical to the regular head-turn preference procedure.

4.4.2 Hypotheses and predictions

Let us review our three hypotheses. Hypothesis One (Limited Hypothesis): The generalization that the infants form is entirely based on the observed input, and the learners are not equipped with preexisting linguistic knowledge about possible structures (Elman et al. 1996; Bybee 1998; Tomasello 2000).

Hypothesis Two (Beyond and Constrained Hypothesis): A learner is equipped with preexisting knowledge about possible structures, and statistics is merely used as a source of information that helps a learner select the correct grammar that derives the matching surface strings. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. This hypothesis proposes that learners' generalization extends to novel structures, as long as they are compatible with antecedently known constraints. An example of an antecedently known constraint would be something like movement of a constituent, which is a natural rule in languages.

Hypothesis Three (Beyond and Unconstrained Hypothesis): Learners generalize beyond what they see in the input but their generalizations are not necessarily constrained in a predictable way. An example of this might be something like movement of a non-constituent, which is unnatural in natural languages, but if a learner is unconstrained, this is a logical possibility. The three hypotheses are summarized in the table below.

Table 25: Table of hypotheses

	Deductive power of learner	*Nature of predetermined representations*
Limited Hypothesis	Limited to observed distributions	None
Beyond and Constrained Hypothesis	Beyond what can be derived from observed distributions	Limited by constraints found in natural language
Beyond and Unconstrained Hypothesis	Beyond what can be derived from observed distributions	Unlimited by constraints found in natural language

In this experiment, we removed all the sentences generated by movement rules from the familiarization set, which means that the test sentences have novel structures that were not seen in the input. According to the Limited Hypothesis, learners do not generalize beyond what was observed in the input. So at test, when they see two novel structures – one that moved a constituent and one that moved a non-constituent – they would consider both to be illicit, because neither was seen in the input. Thus, the performance should be at chance.

According to Beyond and Constrained Hypothesis, learners generalize beyond the observed input, but their generalizations are restricted in a principled way. For instance, learners might have the knowledge that you cannot move a non-constituent in language. If this is the case, on the movement test, the participants would allow the "consistent" test sentence in which a constituent was moved, but they would not allow the "inconsistent" test sentence in which a non-constituent was moved, because while the former is a possible movement, the latter is an impossible operation in language.

According to Beyond and Unconstrained Hypothesis, learners' generalizations could go beyond what was observed in the input and those generalizations do not have to be constrained in a principled way. For example, one generalization the learners could form is that you can move any neighboring elements. If this is the case, on the movement test, learners might allow both test structures even though they are both novel, because both test sentences move neighboring words. If so, both test sentences would be licit for the learners and the performance at test would be at chance, that is, the learners would not choose one over the other.

Because we cannot ask infants whether they consider the test sentence to be grammatical or acceptable, what we measure is their looking times toward each test type. If Limited Hypothesis was correct, we should see no difference in looking times to the consistent or inconsistent samples, because both would be considered ungrammatical by the infants.

If Beyond and Constrained Hypothesis is correct, we would see a difference in looking times although we do not have a particular prediction as to which sample the infants would look longer at. One possibility is that they would look longer at the consistent test sample, because they think that is the "grammatical" sentence and it is compatible with the input grammar. On the other hand, they might look longer at the inconsistent test sample because that is the "ungrammatical" sentence which violates the structure of the input grammar. In Experiment 3 (Infant 1), the infants showed a novelty preference and looked longer at the "inconsistent" test sample. We suggested that that could have been due to the fact that infants were familiarized by the input and by the time of the test phase, they were already bored, so they preferred to listen

to the new, more surprising sentences. However, this could have been due to the fact that in Experiment 3 (Infant 1), the structures of the "consistent" test sentences had already appeared in the familiarization phase, which strengthens the possibility that the infants were bored with the familiar structures and were more intrigued by the new structures. In this experiment, however, both test structures are novel, so the infants might not behave the same way they did in Experiment 3.

If Beyond and Unconstrained Hypothesis is correct, the infants should show no difference in looking times, since both test samples would be considered licit. So only if Beyond and Constrained Hypothesis is correct, we should see a difference in infants' looking times. The predictions of each hypothesis are summarized in the following table.

Table 26: Predictions for Experiment 6

	Views	*Predictions*
Limited Hypothesis	Both test sentences are ungrammatical	Infants will not show a difference in looking times
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times

4.4.3 Results and discussion

The time that each infant oriented to the loudspeaker on each trial was recorded. All infants accumulated 2 min of acquisition time during the familiarization

phase. Just like in Experiment 3 (Infant 1), for the infants who heard Grammar 1 during the familiarization, we coded the looking times to "grammatical in G1" test sample as "consistent" and the looking times to "grammatical in G2" as "inconsistent". Likewise, for infants who were familiarized with Grammar 2, looking times to grammatical-in-G2 test sample were coded "consistent" and looking times to grammatical-in-G1 test sample were coded "inconsistent".

We report the results in terms of raw looking times. The mean looking time at either test sample was 13.5 s (SD = 11.3 s). The data from the infants whose looking time during the test phase was over 3 standard deviations from the mean was not included in the analyses. This eliminated trials from two infants who listened to a test sample for over 47 s. In addition, the data from the infants whose looking time during the test phase was shorter than 3 s was also excluded from the analyses, on the reasoning that less than 3 s is not an adequate amount of time to hear enough sentences to make a decision about the structure of the artificial language. This eliminated trials from four infants. The remaining 25 infants, on average, looked longer to the consistent test sample (mean = 15.08 s) than the inconsistent test sample (mean = 10.70 s). This difference was significant in a two-tailed Paired Samples t-test ($t(24) = 2.096$, $p = 0.047$, $r = 0.39$). 15 out of 25 infants had longer average looking times for the consistent samples.

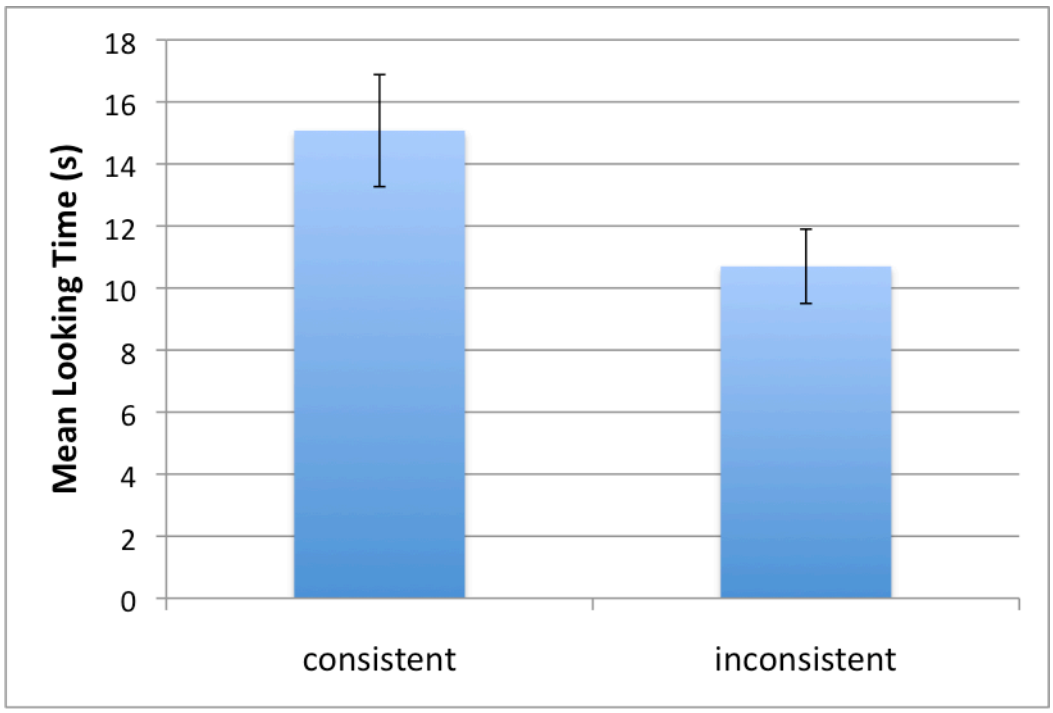


Figure 61: Experiment 6 results

The results show that 18-month-old infants looked longer at the consistent test sample than to the inconsistent test sample. These results suggest that the infants are sensitive to the transitional probabilities as a cue to the hierarchical phrase structure and that the infants can distinguish sentences that moved constituents vs. sentences that moved non-constituents in the input grammar. Just like in Experiment 5 (Infant 3), the term "novelty preference" is not appropriate here either because both test samples involved novel structures. In this experiment, the infants listened longer to the consistent sample, but even the consistent test sample involved a structure the infants had never seen before.

If we were to speculate about reasons for this result, one could argue that the infants listened longer to the consistent test sample because that was the grammatical

sentence in the artificial grammar. On the other hand, in the previous experiments (Experiments 3 & 5), the infants listened significantly longer at the inconsistent test samples, meaning they listened longer at the ungrammatical sentences. This experiment is the only experiment in this dissertation in which the infants showed a longer looking time to the consistent test sample. One thing that is different between this experiment and all the other infant experiments is that since the input lacked movement and substitution by proforms sentences, the number of sentence *types* of the input was much smaller than in other experiments. For example, in Experiment 3 (Infant 1) and Experiment 5 (Infant 3), there were 15 sentence types in the input, whereas in this experiment, there were only 4 sentence types in Grammar 1 input and only 3 sentence types in Grammar 2 input. Therefore, it is possible that infants in previous experiments were bored with their artificial grammar by the time that the familiarization period ended, but the infants in this experiment were more interested in seeing the different sentence types of their grammar.

Another possible reason for infants looking longer at the consistent sample is that, unlike previous experiments, even the consistent test structures were novel sample in this experiment. In Experiment 3 (Infant 1), the actual word strings of test sentences were novel, but the structures were not. But in this experiment, even the structures of grammatical test sentences were new, so that could be why infants were more interested in them.

Recall that the exposure sentences in this experiment could be generated by finite state grammars as in Figure 59 and Figure 60. Importantly, however, the results of this experiment could not have been obtained by those finite state grammars. That

is because those FSAs cannot generate the test sentences. If you only have the FSAs like the ones above, it is impossible to tell apart the two test structures – BCADE and CDABE. Neither test structure (consistent or inconsistent) is represented in this FSA. Therefore, the infants must have had a phrase structure representation of the artificial language in order for them to succeed. To the best of my knowledge, this is the first study showing infants learning of a hierarchical phrase structure instead of a finite state grammar.

One alternative account for this result could be that the infants did not really have a hierarchical tree representation like we argue, but that they were simply noticing the chunks of constituents in the consistent (grammatical) test sentences. That is, in the consistent test sample, "good" transitions exist, meaning transitions from a category to another category that has been observed (i.e., constituents), whereas in the inconsistent test sample, "bad" transitions exist, meaning the transition from a category to another category that was not observed in the data (i.e., non-constituents). One could argue that the results in this experiment could be attained if the participants were merely noticing the "good chunks" (constituents) versus "bad chunks" (non-constituents). While this is a relevant concern, it cannot have been the case. Take a look at the movement test sentences that we used.

(78) Movement test

Grammatical in Grammar 1	Grammatical in Grammar 2
CDEAB	DEABC

There are two good chunks in CDEAB if the familiarization language was Grammar 1, namely CD and AB. If your familiarization language was Grammar 2, then there are two good chunks in DEABC, namely DE and BC. What is important, however, is that there are good chunks of one grammar in the other grammar's consistent test sentences as well. Put another way, the inconsistent test sample in your grammar always contains good chunks of your grammar too. For instance, if you were familiarized with Grammar 1, there is a good chunk (i.e., AB) in the G2-consistent answer, too. If you were familiarized with Grammar 2, there is a good chunk (i.e., DE) in the G1-consistent answer. This is illustrated in Figure 39 below. Solid lines represent good chunks (i.e., constituents) in G1, and dotted lines represent constituents in G2.

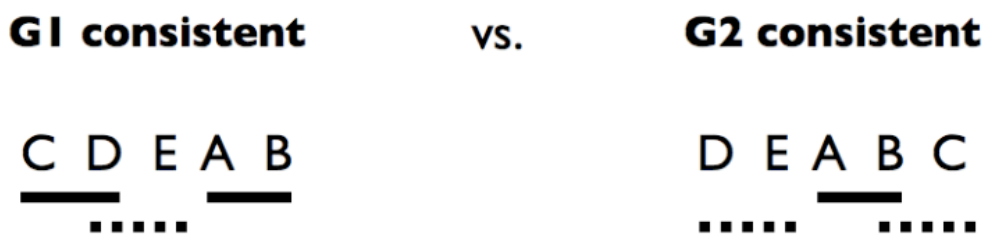


Figure 62: Number of "good chunks" vs. "bad chunks"
Solid line represents good chunks for G1 and dotted line represents good chunks for G2

In this way, simply noticing the "chunks" would not achieve the results of this experiment, thus we can reject that alternative account. Furthermore, even if you did notice "good chunks" (constituents), that knowledge alone cannot give this result. Simply having hierarchical structure and constituency does not give rise to the

asymmetry between moved constituents and moved non-constituents. You also need to know that only movement of constituents is possible and not non-constituents. In sum, you need two things to achieve the results in this experiment. One, you need to know constituency in the given sentence, and two, you need to have had a preexisting knowledge that moving non-constituents is an impossible rule.

In terms of the three hypotheses, again only the predictions of Beyond and Constrained Hypothesis were borne out. Both Limited Hypothesis and Beyond and Unconstrained Hypothesis predicted that the infants would show no difference in looking times, so neither of their predictions were borne out. Beyond and Constrained Hypothesis was the only hypothesis that predicted a difference in infants' looking times between the two test samples. This suggests that the generalization the infants form is not limited to what they saw during the familiarization and that the infants form a generalization that goes beyond the observed input. It also suggests that the generalization the infants form is not unconstrained in that they did not hypothesize that you can move any neighboring elements. The fact that the infants showed a difference in looking times (regardless of which one they listened longer to) indicate that they could at least distinguish the test samples that moved constituents vs. test samples that moved non-constituents. If statistical learning interacts with nothing but the presented distributional information, both test samples would be considered illicit, since both are new. The infants must have had a prior knowledge to help them distinguish the two test samples. But if that prior knowledge was that you can move any neighboring elements (which is not the case in natural language), then the infants would not have been able to distinguish the two test samples either, because both test

samples moved neighboring elements. One natural conclusion is that the infants knew that constituents can be moved, but not non-constituents, which is the case in natural language. And this is what was predicted by Beyond and Constrained Hypothesis, which states that infants' generalization is not restricted to the input distributions, but it interacts with innate knowledge on what is a possible operation and what is an impossible operation in natural language.

Table 27: Predictions and outcomes for Experiment 6

	Views	*Predictions*	*Outcome*
Limited Hypothesis	Both test sentences are ungrammatical	Infants will not show a difference in looking times	✗
Beyond and Constrained Hypothesis	Only the consistent test sentences are grammatical	Infants will show a difference in looking times	✓
Beyond and Unconstrained Hypothesis	Both test sentences are grammatical	Infants will not show a difference in looking times	✗

A question that was left unanswered by Experiment 5 (Infant 3) was where such knowledge came from. Was that knowledge known prior to the experiment or was it learned based on the familiarization set in the experiment? The results of this experiment suggest that it could not have been learned during the experiment because there were no movement sentences in the input in this experiment. In addition, simply representing constituency or simply having a hierarchical structure does not provide any information about what can and cannot be moved. The fact that infants were able to distinguish the two test structures that were both novel suggest that 18-month-old infants already have knowledge of structure dependent nature of movement. Hence,

this supports the view that learners come equipped with antecedent knowledge about possible linguistic structures and that the learners have deductive power that goes beyond what can be derived from the observed statistical distributions.

Chapter 5: Experiment 7 (Simple Recurrent Network Simulations)

In this chapter, we will present a series of neural network simulations with Simple Recurrent Networks (SRN) on the artificial language learning task. SRNs have been proposed to be able to learn a number of different aspects of human language, including syntax (Elman, et al. 1996; Elman 1991, 1993, Rohde & Plaut 1999). Recall that the results of Experiment 2 (Adult 2) and Experiment 6 (Infant 4) showed that even in the absence of movement sentences in the input, human adults and infants still could distinguish sentences in which constituents were moved vs. sentences in which non-constituents were moved. In other words, adults and infants were able to generalize beyond the observed input. From these results, we inferred that that knowledge must have been known antecedently because it could not have arisen from the input. Therefore, it would be interesting to see if learners can form the same generalization in the absence of such innate knowledge.

As Elman (1991) states:

> "[In the connectionist approach], tasks must be devised in which the abstract linguistic representations do not play an explicit role. The model's inputs and output targets are limited to variables which are directly observable in the environment. This is a more naturalistic approach in the sense that the model learns to use surface linguistic

> forms for communicative purposes rather than to do linguistic analysis… The value of this approach is that it need not depend on preexisintg preconceptions about what the abstract linguistic representations are. Instead, the connectionist model can be seen as a mechanism for gaining new theoretical insight."

Connectionism is an approach that tries to explain language acquisition without children having abstract linguistic knowledge (Elman et al. 1996). Simple Recurrent Network is a computational model that is claimed to reflect cognitive processing. SRNs have no inherent assumption about linguistic representation or structure, and yet, SRNs are proposed to successfully acquire language without innate, linguistic specific mechanisms (Rohde & Plaut 1999). We carried out the network simulations because we were interested in whether SRNs could learn to generalize beyond input, and whether structure could really follow from experience alone.

5.1. Method

We used the simulation software called LENS (Rohde 1999) for all the simulations reported below.

The architecture of the network

Simple Recurrent Network of the type proposed in Elman (1990) was used in these simulations. The basic structure of the network is shown in Figure 63, and the structure of the specific network used in our simulations is shown in Figure 64.

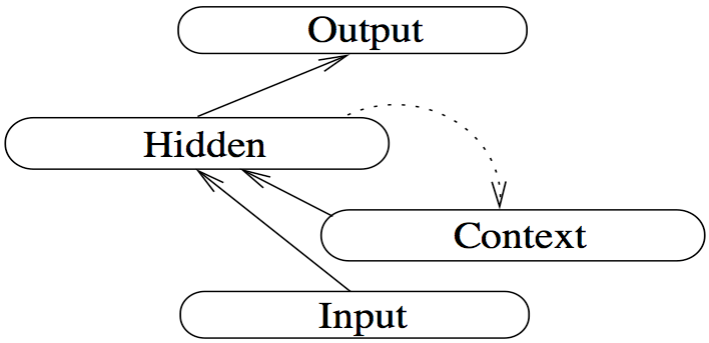


Figure 63: Basic structure of an Elman network, reprinted from Lewis & Elman (2001).

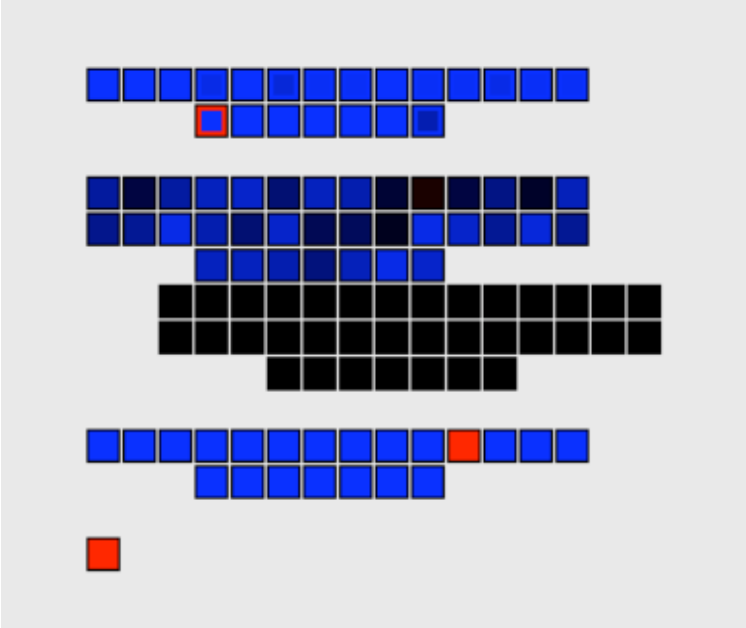


Figure 64: Structure of the network used in the simulaitons

The network had four layers: an input layer, a hidden layer, a context layer and an output layer. Our input and output layers consisted of 21 nodes each. The hidden and context layers consisted of 35 nodes each. The input, hidden and output

layers are feed-forward layers that are connected uni-directionally as in Figure 63. The hidden layer and the context layer are connected bi-directionally. That is, the activations of the hidden layer at one time step are copied into the context layer, which then can feed into the hidden layer as inputs at the next time step. In this way, the context layer works as a one-step state memory.

The activation levels of hidden and output nodes are computed as the sum of all the activations values of every unit, squashed by the logistic function. The training was done by adjusting the weights to minimize the sum of squared error between the word predicted by the network and the target next word, using the back-propagation learning procedure, similarly to what is reported in Elman (1991) and Rohde & Plaut (1999). The network's task at test was to correctly predict the next word.

The input grammars for the network were identical to Grammar 1 and Grammar 2 that were used for the adult and infant experiments. In particular, the same 80 sentences that were used for adult experiments were chosen here as input. Either Grammar 1 or Grammar 2 was fed as input. During the training of the network, one word was presented at a time. One epoch of 80 sentences were used during the training of the network, and the criteria for terminating learning was going through those 80 sentences. We chose the criteria of 80 sentences because those were the identical 80 sentences as the ones used in our adult experiments.

There were four input conditions: in the first condition, the network was trained on Grammar 1 as input and the input included movement sentences (just like in Experiment 1 (Adult 1)). We will call this condition G1-Train-Mvmt. In the second condition, the network was fed Grammar 2 and the input included movement

sentences (G2-Train-Mvmt). The TP patterns of the input for these two conditions are given below.

Table 28: Transitional probabilities for 80 input sentences in G1-Train-Mvmt

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.24	1.00	0.25
Backward TP	1.00	0.19	1.00	0.34

Table 29: Transitional probabilities for 80 input sentences in G2-Train-Mvmt

	A-B	B-C	C-D	D-E
Forward TP	0.33	1.00	0.15	1.00
Backward TP	0.18	1.00	0.16	1.00

In the third condition, the network was fed Grammar 1 as input but the input lacked any movement sentences (just like in Experiment 2 (Adult 2)). We will call this condition G1-Train-NoMvmt. Similarly, in the fourth condition, the network was fed Grammar 2 as input but the input lacked movement sentences (G2-Train-NoMvmt). The TP patterns of those conditions are listed below.

Table 30: Transitional probabilities for 80 input sentences in G1-Train-NoMvmt

	A-B	B-C	C-D	D-E
Forward TP	1.00	0.28	1.00	0.24
Backward TP	1.00	0.24	1.00	1.00

Table 31: Transitional probabilities for 80 input sentences in G2-Train-NoMvmt

	A-B	B-C	C-D	D-E
Forward TP	1.00	1.00	0.22	1.00
Backward TP	0.22	1.00	0.24	1.00

At test, we presented the movement test sentences (the same 16 sentences used for the movement test in Experiments 1 and 2 (Adult1 & 2)).

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Types	Sentences	Types	Sentences
Movement test	1	CDEAB	JES SOT FAL KOF HOX	DEABC	SOT FAL KOF HOX JES
	2		REL ZOR TAF DAZ NEB		ZOR TAF DAZ NEB REL
	3		TID LUM RUD MER LEV		LUM RUD MER LEV TID
	4		TID ZOR RUD MER NEB		ZOR RUD MER NEB TID
	5	FAB	KER KOF HOX	DEF	SOT FAL KER
	6		NAV DAZ NEB		ZOR TAF NAV
	7		SIB MER LEV		LUM RUD SIB
	8		NAV MER NEB		ZOR RUD NAV
	9	CDEABCD	JES SOT FAL KOF HOX JES SOT	DEABCBC	SOT FAL KOF HOX JES HOX JES
	10		REL ZOR TAF DAZ NEB REL ZOR		ZOR TAF DAZ NEB REL NEB REL
	11		TID LUM RUD MER LEV TID LUM		LUM RUD MER LEV TID LEV TID
	12		TID ZOR RUD MER NEB TID ZOR		ZOR RUD MER NEB TID NEB TID
	13	FABCD	KER KOF HOX JES SOT	DEFBC	SOT FAL KER HOX JES
	14		NAV DAZ NEB REL ZOR		ZOR TAF NAV NEB REL
	15		SIB MER LEV TID LUM		LUM RUD SIB LEV TID
	16		NAV MER NEB TID ZOR		ZOR RUD NAV NEB TID

We presented the network with 16 movement test sentences that are grammatical in Grammar 1 and further 16 movement sentences that are grammatical in Grammar 2, regardless of the input grammar and input condition. That is, no matter which grammar the network was fed as input (G1 or G2), both test types (G1-compatible or G2-compatible) were used for all the simulations. The network gets 16 test sentences as input, and they produce probability of the following word as output. The prediction was that if the network did learn the artificial language, it should assign higher probabilities to the test sentences that are consistent with their input grammar.

5.2. Results and discussion

We carried out a whole set of simulations with a range of training parameters, since we were unsure which parameter setting worked the best. The two parameters we varied are the batch size and the learning rate, because we had no a priori prediction as to which setting of these parameters would achieve the optimal learning. A batch size is the number of examples the network processes before it updates the weights of the links during the training. For example, a batch size of 10 means that the network updates weights of the links each time it process 10 examples. In this experiment, the batch size was varied from 19 to 59 with an interval of 10 (i.e., 19, 29, 39, 49, 59). Learning rate is the scale of how radical that weight change is. Bigger learning rate indicates that the weight change can be dramatic, while a small learning rate means the weight changes are small. The learning rate here was varied from 0.001 to 0.009 with an interval of 0.002 (i.e., 0.001, 0.003, 0.005, 0.007, 0.009).

Although these values seem smaller than what is generally used, they are close to the learning rate of 0.01 used in Lewis & Elman (2001). Additionally, smaller learning rates than ours have been used before, as in the learning rates ranging between 0.004 and 0.0003, reported in Rohde & Plaut (1999). In sum, we carried out 25 (5 x 5) different simulations for each condition, thus 100 (25 x 4) simulations altogether.

It should also be noted that in each simulation, we carried out 10 runs with the same parameter setting. The network outputs the probability of each word in test sentences. First, we took the product of the probabilities of all the words in each test sentence. Then we took the average probability of those 16 test sentences. Since it is the product of the probabilities, the numbers were extremely small, we therefore computed the log of those numbers, for all 10 runs. We then took the average of those 10 runs, giving an average probability for G1 test sentences and G2 test sentences. We then took the ratio of those two numbers (G1/G2), which is the dependent variable used in the analyses below.

In omnibus ANCOVA (ratio ~ batch size * learning rate * input condition), the covariate, batch size, was not significantly related to the ratio of probabilities ($F(1, 99) = 0.032$, $p = 0.858$), neither was the covariate, learning rate ($F(1, 99) = 0.03$, $p = 0.955$). There was no interaction of the batch size and the learning rate either ($F(1, 99) = 0.184$, $p = 0.669$). This suggests that varying the batch size or the learning rate did not have an effect on the output probabilities the network produced. The only effect that was significant was the main effect for input condition ($F(3, 99) = 3.361$, $p = 0.022$).

In a one-way ANOVA (ratio ~ input condition), there was a significant effect of input condition on the ratio of probabilities ($F(3, 99) = 3.431$, $p = 0.02$). Post-hoc tests revealed that the only pair that differed significantly from each other was G1-Train-Mvmt condition and G2-Train-Mvmt condition (Tukey HSD: $p = 0.023$; Bonferroni: $p = 0.026$). This effect is driven by the high ratio of the G2-Train-Mvmt condition (Mean = 1.026; Figure 65).

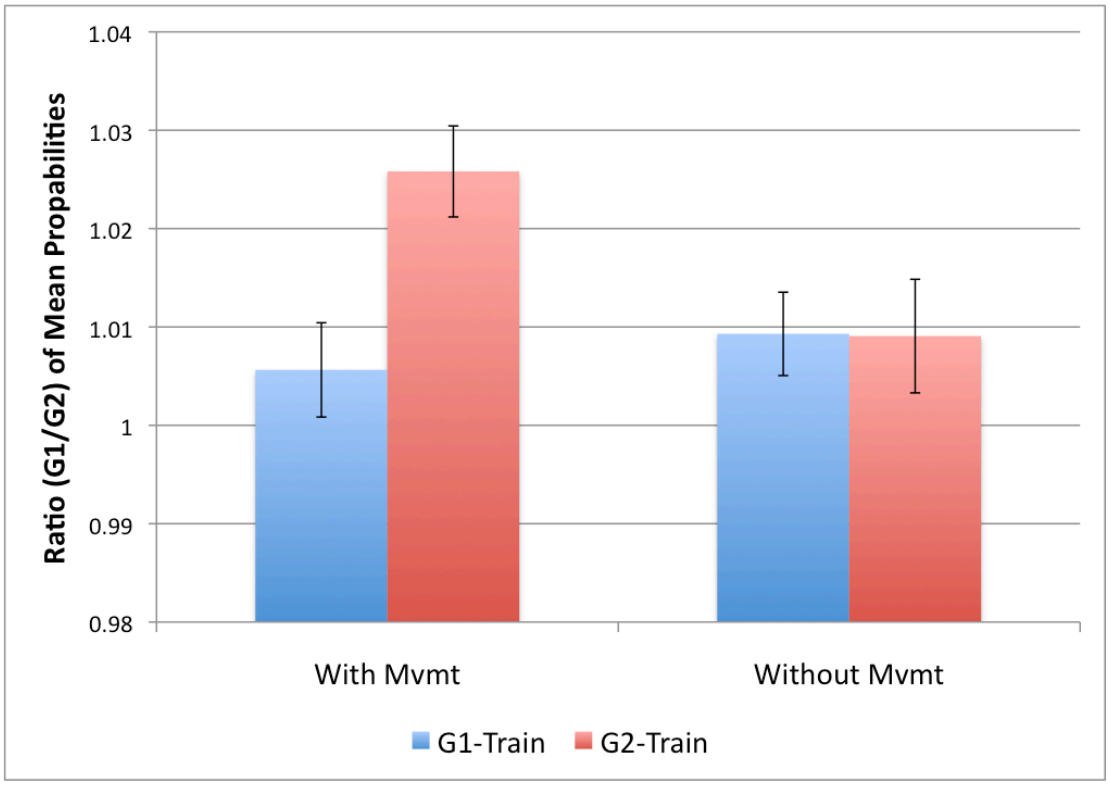


Figure 65: Overall mean ratio by condition

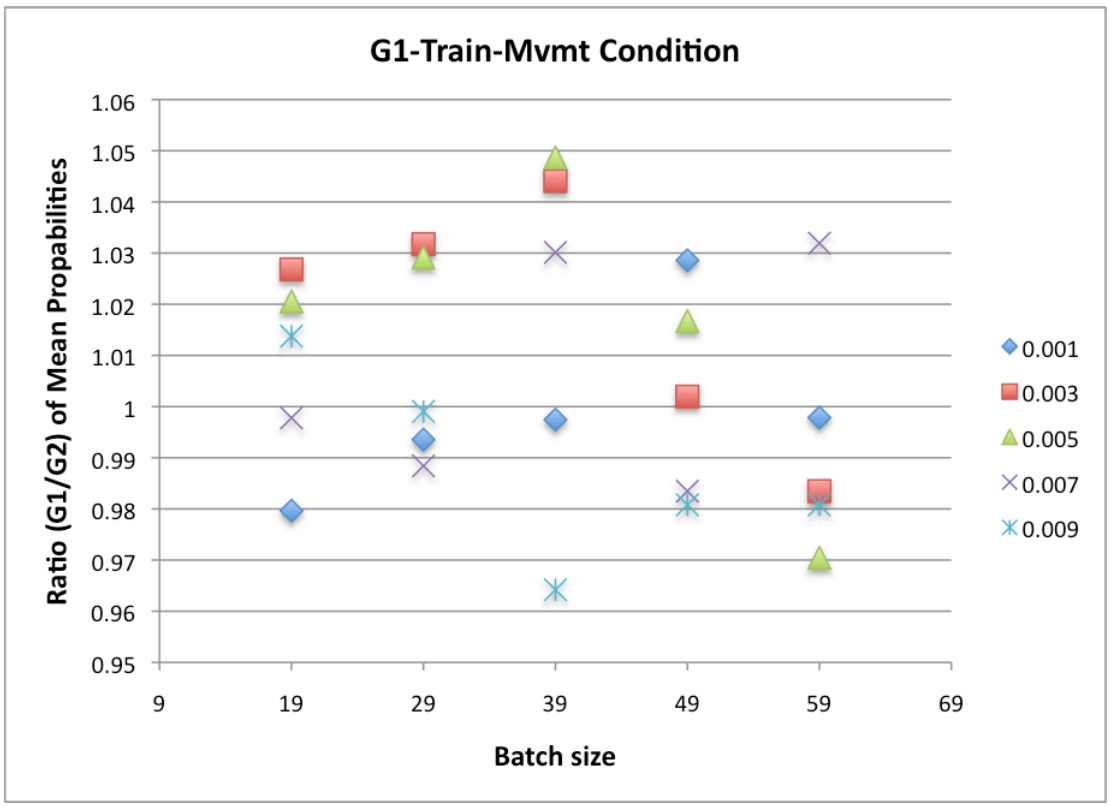


Figure 66: Simulation results of G1-Train-Mvmt condition

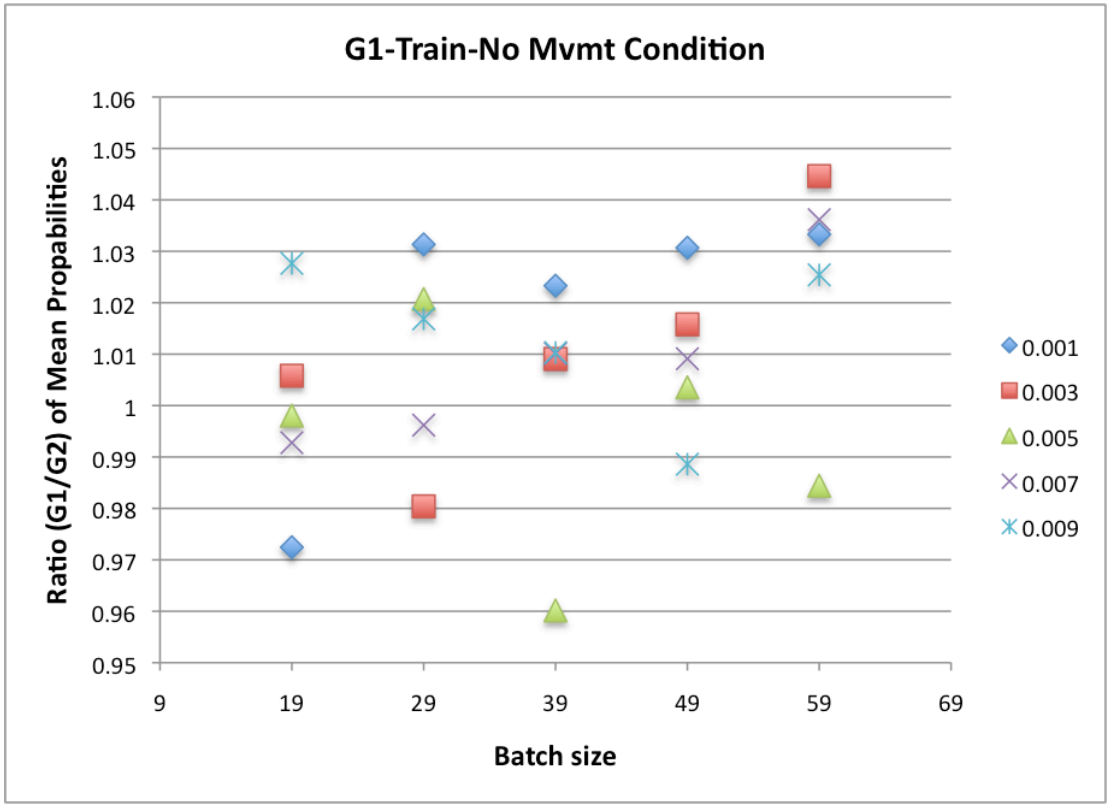


Figure 67: Simulation results of G1-Train-No Mvmt condition

Figure 68: Simulation results of G2-Train-Mvmt condition

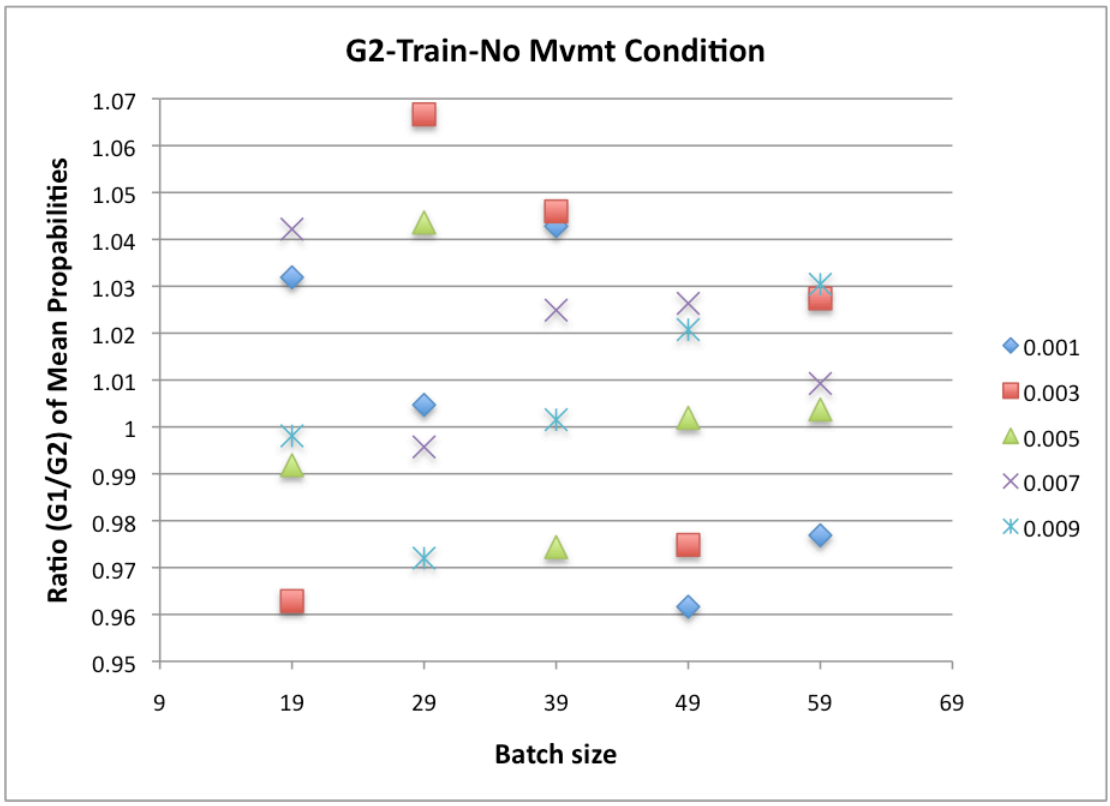


Figure 69: Simulation results of G2-Train-No Mvmt condition

G1-Train conditions

Looking within each combination of batch size and learning (i.e., each point in the above graphs), none of the simulations in the G1-Train-Mvmt condition and G1-Train-No Mvmt condition yielded a significant difference for the probabilities for G1 test sentences and G2 test sentences. In other words, in the simulations in which the network received G1 as input, the mean probabilities the network assigned for G1 test sentences were never significantly higher than the mean probabilities the network assigned for G2 test sentences, regardless of whether the input contained movement sentences or not (for an example, see Figure 70). That is, the SRN failed to learn the artificial language syntax completely when they received G1 as input. Note that since

these numbers are log probabilities, a larger negative log number indicates a larger probability (i.e., -4.3 signifies a larger probability than -4.45).

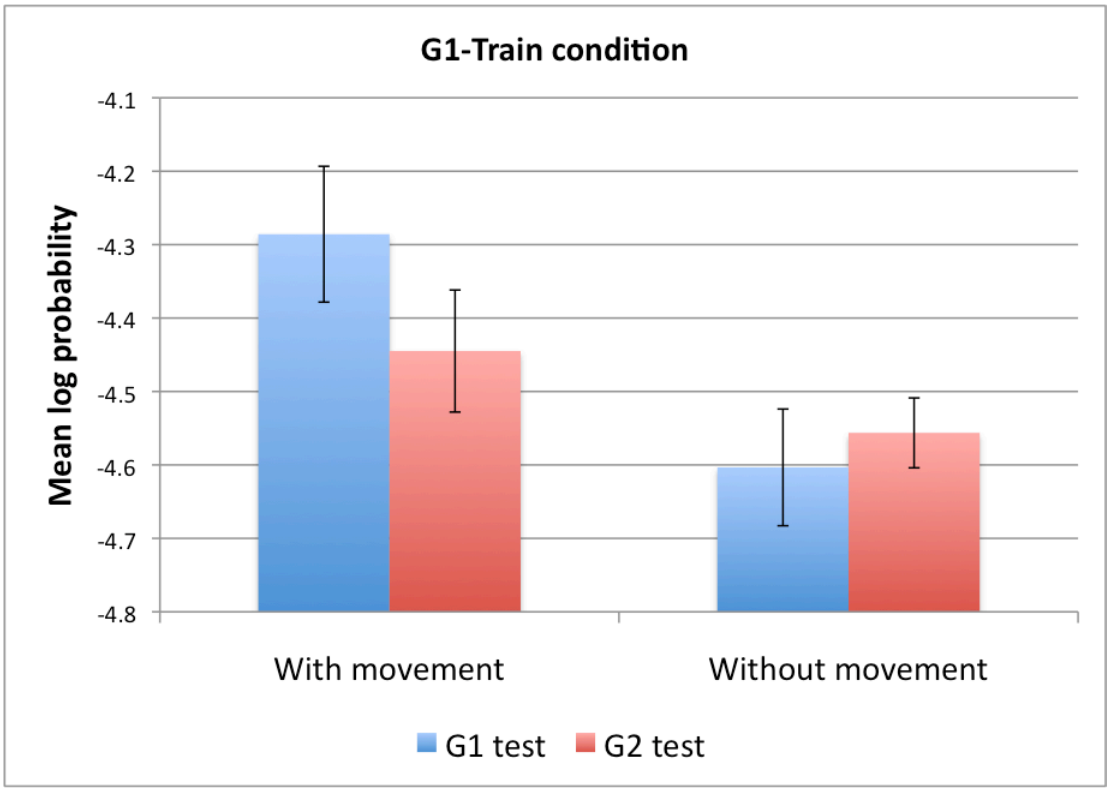


Figure 70: G1-Train condition simulations with batch size 39 and learning rate 0.009

G2-Train conditions

When the SRN received G2 with movement sentences as input (i.e., G2-Train-Mvmt condition; Figure 68), there were three settings of the two parameters in which the network achieved a successful learning. There are the three settings that achieved the most successful learning in the condition that included movement. Therefore, by choosing these settings for the following comparisons, we are giving the model the best chance to succeed. Importantly, however, the No-Movement counterparts (i.e.,

G2-Train-No Mvmt condition; Figure 69) of all of these three settings failed to learn the artificial grammar.

Specifically, in one combination with the batch size of 49 and the learning rate of 0.009 (Figure 71), in the G2-Train-Mvmt condition, the mean log probabilities for G2 test sentences (Mean = -4.220) were significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.527) in a two-tailed Paired Samples t-test ($t(9) = -2.651$, $p = 0.026$). On the other hand, with the same setting in G2-Train-No Mvmt condition, the mean log probabilities for G2 test sentences (Mean = -4.546) were not significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.640) in a two-tailed Paired Samples t-test ($t(9) = -1.401$, $p = 0.195$). Note again that since these numbers are log probabilities, a larger negative log number indicates a larger probability (e.g., -4.2 signifies a larger probability than -4.5).

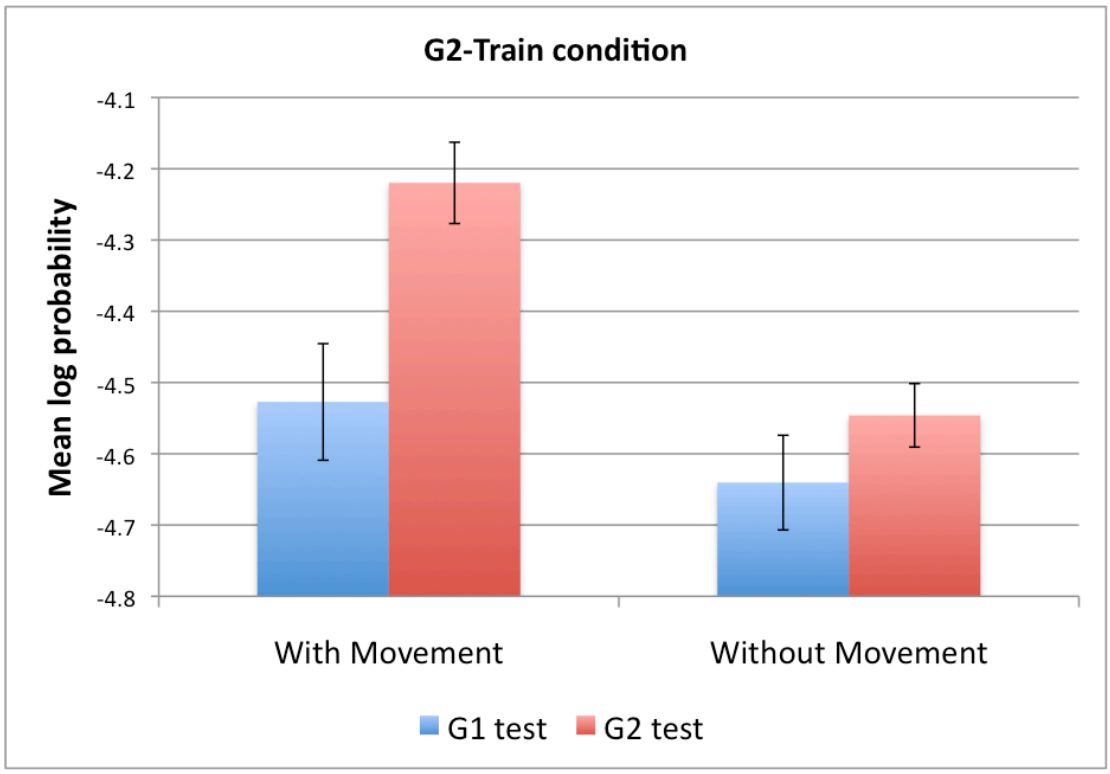


Figure 71: G2-Train condition simulations with batch size 49 and learning rate 0.009

With another setting of the batch size 19 and the learning rate 0.007, in the G2-Train-Mvmt condition, the mean log probabilities for G2 test sentences (Mean = -4.251) were significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.560) in a two-tailed Paired Samples t-test ($t(9) = -2.872$, $p = 0.018$). But again, with the same setting in G2-Train-No Mvmt condition, the mean log probabilities for G2 test sentences (Mean = -4.450) were not significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.637) in a two-tailed Paired Samples t-test ($t(9) = -2.149$, $p = 0.060$).

With the combination of batch size 59 and learning rate 0.009, in the G2-Train-Mvmt condition, the mean log probabilities for G2 test sentences (Mean = -

4.276) were significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.534) in a two-tailed Paired Samples t-test ($t(9) = -2.799$, $p = 0.021$). With the same setting in G2-Train-No Mvmt condition, however, the mean log probabilities for G2 test sentences (Mean = -4.583) were not significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.722) in a two-tailed Paired Samples t-test ($t(9) = -1.177$, $p = 0.269$).

In sum, all the parameter settings with which the network successfully learned the artificial grammar in the condition where the input included movement, the network with the identical settings failed to learn when the input lacked movement sentences. Even though we gave the models best chance to succeed by choosing the most successful settings, the model with the same settings still failed to learn in the No Movement condition. Furthermore, the opposite was also true. That is, the setting that achieved the most successful learning in the No Movement condition did not achieve a successful learning in the Movement condition counterpart. We will look at this analysis below.

There was only one setting in which the network was successful even when the input lacked movement sentences. That was combination of batch size 29 and learning rate 0.003 in the G2-Train-No Mvmt condition (Figure 72), and the mean log probabilities for G2 test sentences (Mean = -4.434) were significantly higher than the mean log probabilities for G1 test sentences (Mean = -4.729) in a two-tailed Paired Samples t-test ($t(9) = -2.642$, $p = 0.027$). However, with the same setting in the G2-Train-Mvmt condition, the network failed to learn the artificial language. The the mean log probabilities for G2 test sentences (Mean = -4.327) were not significantly

higher than the mean log probabilities for G1 test sentences (Mean = -4.340) in a two-tailed Paired Samples t-test ($t(9) = -0.141$, $p = 0.891$).

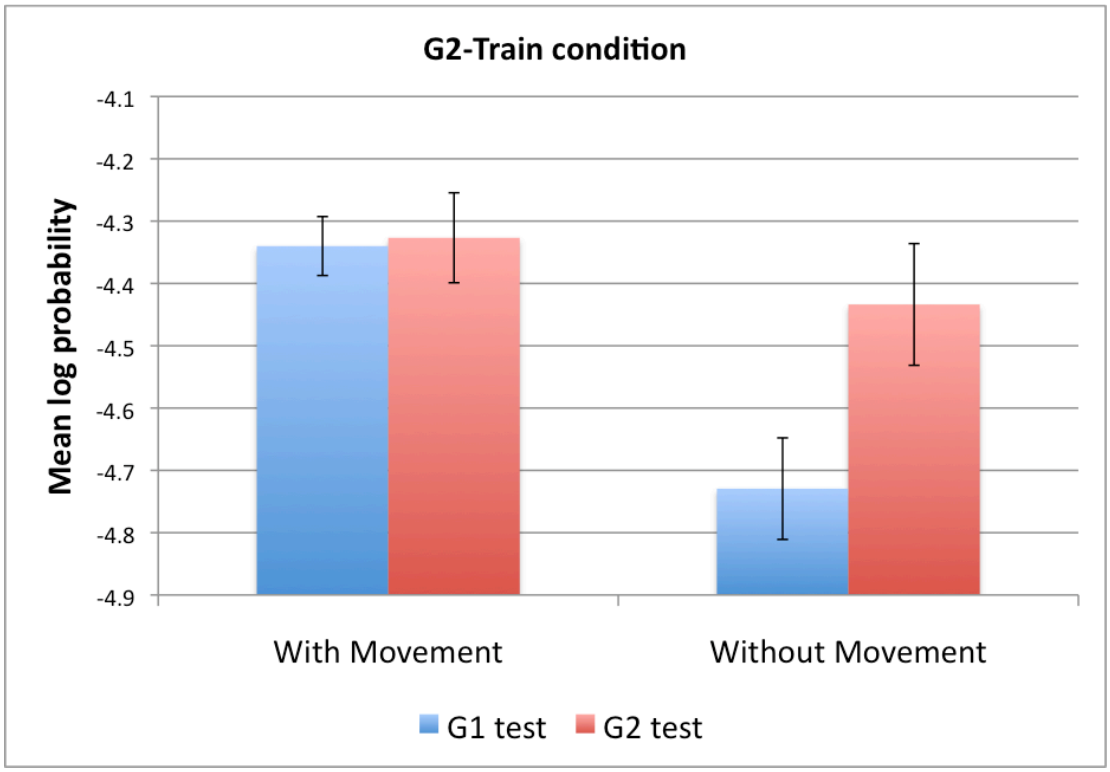


Figure 72: G2-Train condition simulations with batch size 29 and learning rate 0.003

To summarize, with the settings at which the SRN successfully learned the grammar with movement sentences in the input, it failed to learn when the input did not include movement. This indicates that based on the peaks and dips in the transitional probabilities, the SRN correctly figured out the constituency of the sentences, thus successfully predicting the consistent grammar at test in the movement-in-the-input cases. But, since SRNs do not have any assumption about what kind of structure is linguistically valid or what kind of operation is allowed (e.g.,

movement of constituents), when the input lacks movement, they failed to predict upcoming words at test. This suggests that networks cannot extend what they learned from the environment to novel structures, and that the generalization they form does not go beyond the observed input. This result strengthens our claim from Experiment 6 (Infant 4) that infants' knowledge that only constituents can be moved was known antecedently.

However, we also have results of the SRN seemingly succeeding in learning the artificial language when the input lacked movement sentences. But in this case, the network failed to learn the language when the input included movement. This result is harder to interpret, since it is unclear why the network can assign high probabilities to the consistent but never-seen-before structures while it fails to assign high probabilities to the consistent structures they have seen before. In the latter case, it should be so easy to distinguish the consistent and inconsistent grammars that we can only conclude that the network must not have learned the grammar at all in this case.

In any case, what is clear is that the neural networks do not act the same way as human infants do in our experiments. Human infants were able to distinguish the two different artificial languages with or without movement sentences in the input. The results from the simulations do not reflect how infants performed in our experiments. Infants were successful regardless of whether the input included movement or not. Because we have a single case where the network correctly assigned higher probabilities to the consistent test sentences than to the inconsistent ones when the input lacked movement, it is difficult to conclude, but what we suggest

from these simulations is that human infants must have some knowledge that the networks do not have, which helped the infants generalize beyond the input when faced with the unseen structures.

Chapter 6: Conclusions

Traditionally in language acquisition, nativism and empiricism have been characterized as two opposing views that do not need each other. But we suggested that both nature and nurture need each other. The question is how the two interact, what is innate and what is learned from the input. The two might play different roles in language acquisition – innate knowledge specifies range of possible grammars and structures, while statistical learning is a method for mapping the surface string to abstract representation. This dissertation was examined how the environment interacts with the structure of the learner.

The main questions in this dissertation were what the deductive consequences of distributional learning are, and whether the representations are part of the learning system prior to the experience. Is statistical learning entirely a product of tracking and summarizing the surface distributions? Or is it an interaction of tracking the distributions and innate knowledge that the learners already have? In order to investigate these questions, this dissertation focused on the acquisition of phrase structure as a case study.

Hierarchical representation is a hallmark of natural language syntax and phrasal constituency plays a fundamental role in any syntactic operation, since all syntactic operations refer to and manipulate it. A child might come with innate knowledge that "there exists phrase structure" or "a VP consists of a verb and an

optional NP" but that is not sufficient. That knowledge alone does not prevent a child from having an incorrect tree representation of a sentence, as in Figure 73.

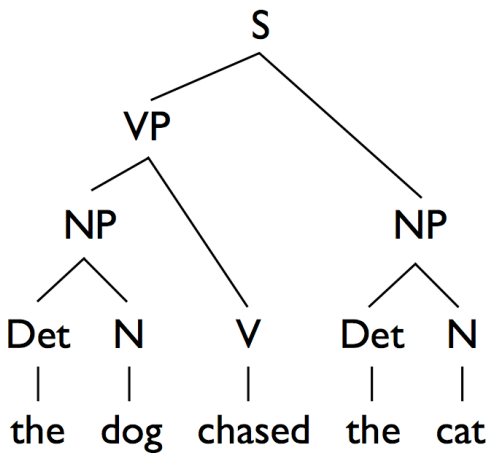


Figure 73: An incorrect tree

Since words, word order and grammatical rules differ from language to language, there must also be a mechanism that guides the child to the correct phrase structure representation of sentences for a particular language (Fodor 1966; Pinker 1984; Grimshaw 1981; Chomsky 1981; Macnamara 1982). A difficulty is constituency and phrase structure are highly abstract notions and the input to a child does not come marked with obvious labels or brackets signaling the constituency. Several different kinds of information were proposed to be perceptually available to a prelinguistic learner, including prosody (Gleitman & Wanner 1982; Gleitman, Gleitman, Landau & Wanner 1988; Morgan 1986), meaning (Pinker 1984; Grimshaw 1981; Macnamara 1982) and distribution (Morgan, Meier & Newport 1989; Saffran 2001; Thompson & Newport 2007).

In Chapter 2, we reviewed past studies that investigated infants' sensitivity to the prosodic, semantic and distributional information as a cue to syntactic structures. We saw that infants are in fact sensitive to these cues, but also that none of these cues is completely sufficient on its own. For example, there are cases where there exists a mismatch between phonological and syntactic phrases and between syntax and semantics. In those cases, learners could be misled and misparse the sentence. What we propose is that infants make use of a combination of all these cues. And one particular type of cue we investigated in this dissertation was distributional information.

Thompson & Newport (2007) showed that adults can learn the phrase structure of a miniature artificial language on the basis of transitional probability patterns. What has not been shown in Thompson & Newport (2007), however, is whether transitional probability can signal hierarchically nested structures. So our more specific question in this dissertation was: can infants infer hierarchical phrase structure on the basis of statistical distribution?

We can summarize our research questions as follows:

(79) What are the deductive consequences of distributional learning?

(80) Are representations a part of the learning system prior to the experience?

(81) Is statistical learning entirely a product of tracking and summarizing the surface distributions? Or is it an interaction of tracking the distributions and innate knowledge that the learners already have?

(82) Can infants learn internally nested hierarchical phrase structure on the basis

of statistical distribution?

(83) Can infants learn phrase structure of an artificial language without any prosodic or semantic information?

To answer those questions, we designed and carried out seven original experiments. We created two minimally different artificial languages that differed only in constituency. These two grammars (Grammar 1 and Grammar 2) were used throughout all seven experiments.

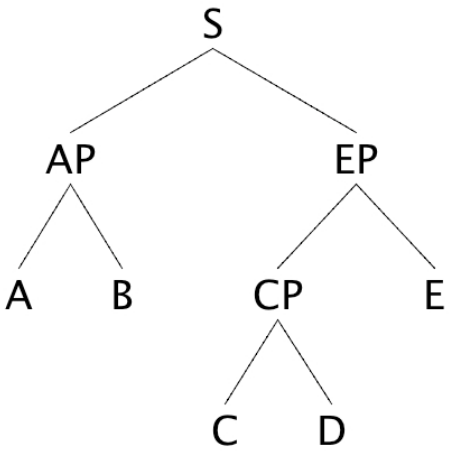


Figure 74: PS tree of the basic Grammar 1 sentence

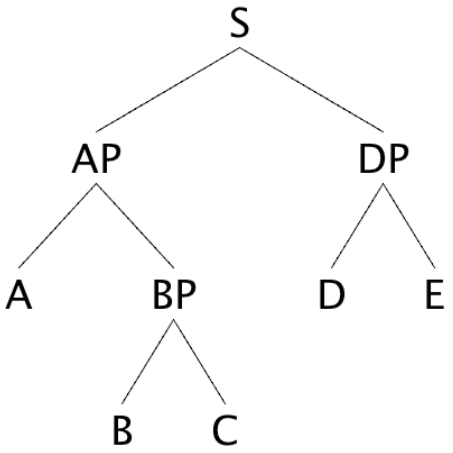


Figure 75: PS tree of the basic Grammar 2 sentence

In the experiments, the participants listened to either Grammar 1 or Grammar 2 during the familiarization/training period. At test, they were presented with two types of test sentence. One type was called the "consistent" sample and these were the test sentences that are grammatical in their input grammar. The other was called the "inconsistent" sample and they were the test sentences that are ungrammatical in their input grammar. One of the tests was movement test, whose consistent test sample had a constituent in the input language moved to the front, while the inconsistent test sample in the movement test would have a non-constituent moved to the front. What we were looking for was which test sample the participants would choose, listen longer to, or assign higher probabilities to (in case of network simulations).

There were three kinds of familiarization set. One had all kinds of sentences generated by all the operations, including movement and substitution rules (Experiment 1, Experiment 3, Experiment 4, Experiment 7). The other kind excluded sentences generated by some operations, such as movement and substitution by

proforms (Experiment 2, Experiment 6, Experiment 7). The final type included movement sentences but the structure of the movement sentences in the input and movement sentences of the test were different (Experiment 5).

Three specific hypotheses were presented for these experiments. Hypothesis One (Limited Hypothesis): The generalization that the infants form is entirely based on the observed input, and the learners are not equipped with preexisting linguistic knowledge about possible structures (Elman et al. 1996; Bybee 1998; Tomasello 2000). This hypothesis corresponds to the purely statistical learning theory.

Hypothesis Two (Beyond and Constrained Hypothesis): A learner is equipped with preexisting knowledge about possible structures, and statistics are merely used as a source of information that helps a learner select the correct grammar that derives the matching surface strings. Under this selective learning theory, the acquired representations have deductive consequences beyond what can be derived from the observed statistical distributions alone. This hypothesis proposes that learners' generalization extends to novel structures, as long as they are compatible with antecedently known constraints. An example of an antecedently known constraint would be something like movement of a constituent, which is a natural rule in languages.

Hypothesis Three (Beyond and Unconstrained Hypothesis): Learners generalize beyond what they see in the input but their generalizations are not necessarily constrained in a predictable way. An example of this might be something like movement of a non-constituent, which is unnatural in natural languages, but if a

learner is unconstrained, this is a logical possibility. The three hypotheses are summarized in the table below.

Table 32: Table of hypotheses

	Deductive power of learner	*Nature of predetermined representations*
Limited Hypothesis	Limited to observed distributions	None
Beyond and Constrained Hypothesis	Beyond what can be derived from observed distributions	Limited by constraints found in natural language
Beyond and Unconstrained Hypothesis	Beyond what can be derived from observed distributions	Unlimited by constraints found in natural language

General predictions for the purely statistical approach (Limited Hypothesis) and the approach in which statistics interacts with innate knowledge (Beyond and Constrained Hypothesis) are listed below (Except from Lust, 2006).

(84) Predictions of a purely statistical approach

i. Learners have a direct relation to input data
ii. No universal linguistic constraints are predicted (e.g., no structure dependence)
iii. Only randomly, if at all, attend to parametric variations of language
iv. Not creative but highly imitative; generalizations should only be based on perceived forms or analogy
v. Learners do not evidence universal language principles or patterns

(85) Predictions of an approach in which nativism and statistics interact

i. Learners have an indirect relation to input data
ii. Be constrained in language acquisition
iii. Be structure dependent from the beginning, and attend to the parameters of language variation
iv. Be creative, i.e., go beyond the stimuli, and not simply copy
v. Not offend universals shown across natural languages

The results of Experiment 1 and Experiment 3 (where the input included movement sentences) supported predictions made by both Limited Hypothesis and Beyond and Constrained Hypothesis, while rejecting the predictions made by Beyond and Unconstrained Hypothesis. From these results, we can rule out Beyond and Unconstrained Hypothesis, and we can be confident that a learner does not make a generalization that is impossible in natural languages, even if that generalization is compatible with the input data. The results of Experiment 5 (in which the exposure data included movement of one constituent and the subjects were tested on movement of a different constituent) revealed that Limited Hypothesis as well as Beyond and Unconstrained Hypothesis cannot be correct. The results of Experiment 5 indicate that the generalization the infants form is not entirely based on the statistical distribution observed, but it goes beyond that, and that it must be a combination of the observed input and the knowledge of some constraints. This further supports Beyond and Constrained Hypothesis. What was left unclear was where such knowledge came from – whether infants learned the constraint during the familiarization or whether

infants already knew it. The results of Experiment 2, Experiment 6 and Experiment 7 (in which the exposure data lacked movement sentences) support the idea that the infants already knew the constraint that you can only move constituents prior to the experiment, but the neural networks do not.

In other words, the predictions made by an approach that combines statistical learning with innate knowledge were born out.

(86) Predictions of an approach in which nativism and statistics interact

i. Learners have an indirect relation to input data
   - ✓ Learners learned the input data, but were also able to generalize the knowledge that was beyond the input data.

ii. Be constrained in language acquisition
   - ✓ The learners' generalization were constrained, not unconstrained, even when the input data were compatible with the unconstrained hypothesis.

iii. Be structure dependent from the beginning, and attend to the parameters of language variation
   - ✓ Learners knew that you can only move constituents and not non-constituents. What learners learned can only be described in a phrase structural representation, and cannot be attributed to learning of linear order or some other low-level regularities.

iv. Be creative, i.e., go beyond the stimuli, and not simply copy
   - ✓ The generalizations the learners formed went beyond the stimuli, in that the learners knew movement of a constituent is possible, but movement

of a non-constituent is impossible even in the absence of any movement in the stimuli.

v. Not offend universals shown across natural languages

- ✓ In all natural languages, movement of a non-constituent is not a possible rule. Learners in our experiments adhered to this.

Now, we can answer our research questions.

(87) What are the deductive consequences of distributional learning?

- ✓ Learners have a deductive power that goes beyond the input data.

(88) Are representations a part of the learning system prior to the experience?

- ✓ Learners have a preexisting knowledge about what is and is not possible in movement rules. This knowledge cannot have been learned discovered from the exposure data, therefore it must have been known prior to the experiment.

(89) Is statistical learning entirely a product of tracking and summarizing the surface distributions? Or is it an interaction of tracking the distributions and innate knowledge that the learners already have?

- ✓ The results of these experiments cannot have been explained if the learning is solely based on tracking and summarizing the surface distributions. They can only be explained if the statistical learning interacts with preexisting linguistic knowledge of the learner.

(90) Can infants learn internally nested hierarchical phrase structure on the basis

of statistical distribution?

- ✓ The results confirm that transitional probabilities can be a cue to not only phrasal groupings but also nested constituent structure. The only cue to constituency in the artificial languages was transitional probabilities. Participants were sensitive to the distributional information that signaled internally nested structure.

(91) Can infants learn phrase structure of an artificial language without any prosodic or semantic information?

- ✓ Artificial languages in these experiments lacked prosodic and semantic cues to phrasal boundaries. And yet, the participants were able to learn the structure of a sentence without relying on prosody or meanings.

To sum up, the current findings suggest that, in addition to cues such as prosody, morphology and semantics, transitional probability is another additional cue to phrase structure. The experimental results in this dissertation suggest that the transitional probability can be a cue to not only the phrasal bracketing but also hierarchical constituent structure. More importantly, the results of Experiments 2 and 6 showed that movement in the input is not required to learn that only constituents can undergo movement. Crucially, knowing the constituency of the language alone does not guarantee that you know only constituents can be moved. That is, the constraint on movement does not automatically follow from constituent structure. And our experiments showed that even when the learners were only given the evidence for constituent structure, they still knew that only constituents can be

moved. In other words, they did not need to have seen movement in the input to know the constraint on movement, which suggests that they knew the constraint already.

Finally, these results suggest that learners can project what they have learned based on the distributional information to novel structures they have not yet seen. Importantly, however, such projection to new structures occurred only within and not outside the realms of what is allowed in natural language. This provides novel evidence that statistical learning interacts with innate constraints on possible representations and rules. In particular, we wish to have shown a way in which the two (innate knowledge and statistical learning) interact. If learning was only based on the statistical distributions, it might help you correctly identify constituent structure of sentences, but it does not ensure the constraint on movement. Thus, we suggest that statistics are used only as a path into inherently known abstract representations. The learners have a deductive power that goes beyond the input stimuli, which suggests that statistical learning is used merely as a method for mapping the surface string to abstract representation, and that learners are constrained by innate knowledge that specifies range of possible grammars and structures.

Appendices

Appendix A: Familiarization sentences for Experiment 1 (Adult 1) and Experiment 7 (SRN Simulations)

Sentence types (numbers in parentheses indicate number of times used in the familiarization)

Grammar 1		*Grammar 2*	
et E ib C D	(3)	A B C D E et	(2)
A B C D E et	(1)	A et D E	(2)
C D E ib	(1)	D E F B C	(11)
A B C D E	(2)	A B C D E	(2)
F A B C D	(4)	D E F et	(3)
A B et E et	(2)	ib F et	(1)
ib et E et	(2)	ib A et B C	(1)
A B et E	(2)	ib A B C B C	(1)
ib et E C D	(2)	D E F	(2)
F ib	(1)	A B C D E B C	(7)
A B et E C D	(13)	ib A et	(1)
ib F C D	(5)	F D E et	(1)
et E ib	(1)	D E A et et	(1)
C D E A B et	(2)	ib F B C	(2)
A B C D EC D	(1)	D E A B C B C	(1)
et E A B	(2)	D E A et B C	(12)
F ib C D	(2)	ib F	(1)
A B F	(2)	A et D E B C	(6)
F ib et	(1)	A et D E et	(2)
ib F	(1)	D E A et	(1)
F A B	(2)	F ib B C	(3)
C D E ib et	(1)	F D E B C	(16)
C D E ib C D	(2)	A et ib et	(1)
C D E A B	(4)		
ib C D E C D	(3)		
A B F C D	(11)		
et E ib et	(1)		
A B F et	(1)		
et E A B C D	(5)		

Familiarization sentences

Grammar 1	*Grammar 2*
DAZ HOX REL LUM FAL	DAZ HOX REL LUM FAL
DAZ HOX REL LUM TAF	DAZ HOX REL LUM TAF
et TAF ib TID SOT	LUM FAL MER et LEV JES
et FAL ib JES LUM	LUM TAF NAV NEB JES
KOF LEV et RUD TID SOT	SOT RUD DAZ et
DAZ HOX NAV	ib KOF et HOX REL
NAV ib TID SOT	NAV LUM TAF LEV REL
TID SOT TAF ib JES LUM	NAV SOT RUD HOX REL
SIB MER HOX	SOT RUD DAZ et HOX TID
KOF LEV et RUD et	SOT RUD DAZ et et
DAZ LEV NAV et	NAV ZOR FAL LEV JES
NAV ib et	ib KER LEV REL
TID SOT TAF ib et	LUM TAF DAZ et LEV JES
et RUD MER HOX REL LUM	SOT TAF NAV
DAZ HOX NAV TID SOT	ZOR FAL KOF et HOX TID
ib KER TID SOT	SOT RUD SIB HOX REL
SIB DAZ HOX	SIB LUM TAF LEV JES
DAZ LEV NAV	ZOR FAL NAV LEV REL
ib NAV JES LUM	SOT TAF NAV et
KOF LEV JES ZOR FAL et	ZOR FAL KER LEV JES
MER HOX TID SOT RUD REL LUM	DAZ LEV JES ZOR FAL LEV REL
JES ZOR FAL MER HOX et	SOT TAF MER et HOX REL
MER HOX NAV JES LUM	KOF et LUM FAL
et RUD MER HOX REL SOT	MER et ZOR FAL
MER HOX et FAL	SOT TAF KER HOX REL
REL LUM TAF ib JES ZOR	ib KER
DAZ HOX et FAL	SOT RUD MER et LEV JES
KOF NEB et TAF REL SOT	ib MER et
KOF NEB et TAF JES LUM	ZOR FAL NAV
ib et TAF REL SOT	DAZ LEV JES ZOR FAL HOX REL
NAV ib REL SOT	SIB ZOR FAL et
et TAF KOF LEV	LUM TAF KOF et NEB JES
et FAL KOF NEB JES ZOR	DAZ et LUM FAL HOX TID
REL SOT RUD ib	KER ZOR FAL HOX REL
DAZ HOX et FAL REL LUM	NAV LUM FAL LEV REL
ib TID SOT RUD REL SOT	NAV SOT TAF LEV REL
REL SOT TAF KOF NEB	MER HOX TID SOT RUD HOX REL
DAZ LEV NAV TID SOT	LUM TAF KOF et LEV JES
REL LUM FAL KOF LEV	ZOR FAL NAV LEV JES
KOF LEV et TAF REL LUM	MER et ZOR FAL LEV REL

DAZ LEV SIB JES LUM	LUM FAL DAZ et LEV JES
ib JES LUM FAL TID SOT	SIB ib HOX REL
KER DAZ HOX REL LUM	ib NAV HOX TID
NAV DAZ LEV REL SOT	SIB ZOR FAL LEV REL
REL SOT TAF KOF NEB et	ZOR FAL SIB HOX REL
NAV ib	ib MER HOX TID HOX REL
ib et FAL REL SOT	KER SOT RUD HOX REL
MER HOX SIB JES LUM	MER HOX TID SOT TAF LEV REL
DAZ LEV et TAF TID SOT	KOF et LUM TAF NEB JES
DAZ LEV SIB REL LUM	ZOR FAL KER NEB JES
MER HOX et RUD TID SOT	SIB LUM FAL LEV REL
et TAF DAZ LEV	KOF et SOT RUD et
DAZ LEV NAV JES LUM	ZOR FAL NAV NEB JES
et RUD ib REL SOT	ZOR FAL DAZ et LEV REL
REL SOT TAF DAZ LEV	SOT TAF KER et
KOF NEB et TAF et	LUM FAL DAZ et HOX TID
KER MER HOX REL LUM	DAZ et ZOR FAL HOX REL
DAZ HOX NAV REL SOT	DAZ HOX TID SOT TAF LEV JES
ib KER REL SOT	SOT RUD KER NEB JES
et FAL ib et	KOF LEV REL LUM TAF HOX REL
DAZ LEV et FAL JES ZOR	DAZ et LUM TAF HOX REL
et TAF ib	LUM TAF MER et LEV REL
ib SIB	MER et LUM TAF et
DAZ LEV et FAL TID SOT	NAV SOT TAF HOX TID
et RUD KOF LEV JES ZOR	LUM FAL NAV HOX TID
et TAF MER HOX TID SOT	DAZ HOX REL SOT TAF LEV JES
DAZ HOX KER REL LUM	KOF LEV REL SOT TAF et
ib KER JES LUM	ib SIB et
KER DAZ LEV REL SOT	ZOR FAL DAZ LEV REL HOX REL
DAZ HOX et TAF JES LUM	NAV SOT TAF LEV JES
DAZ HOX KER REL SOT	NAV LUM FAL LEV JES
ib et TAF et	MER et LUM FAL NEB JES
DAZ LEV KER TID SOT	MER HOX REL SOT RUD et
ib JES LUM TAF JES LUM	MER et ib et
ib et RUD et	SIB LUM TAF HOX REL
DAZ LEV et RUD REL SOT	KER ib LEV JES
ib NAV REL LUM	SOT RUD SIB et
KOF NEB et FAL REL LUM	KER SOT TAF LEV REL
JES ZOR FAL DAZ LEV	NAV ib LEV REL
DAZ LEV et FAL REL LUM	KER SOT RUD HOX TID

Appendix B: Familiarization sentences for Experiment 2 (Adult 2) and Experiment 7

(SRN Simulations)

Sentence types (numbers in parentheses indicate number of times used in the familiarization)

Grammar 1		*Grammar 2*	
A B F	(9)	A B C D E	(3)
A B C D E	(3)	F D E	(10)
A B C D E C D	(19)	A B C D E B C	(16)
A B F C D	(49)	F D E B C	(51)

Familiarization sentences

Grammar 1	*Grammar 2*
DAZ HOX SIB JES LUM	KER LUM TAF HOX TID
DAZ LEV SIB REL LUM	KER SOT TAF HOX REL
MER HOX KER JES LUM	NAV ZOR FAL HOX TID
DAZ LEV KER REL LUM	NAV SOT RUD NEB JES
KOF LEV KER	NAV SOT TAF LEV JES
MER HOX KER REL SOT	NAV LUM TAF HOX TID
KOF NEB KER	SIB SOT TAF
MER HOX SIB TID SOT	NAV ZOR FAL
DAZ LEV SIB REL SOT	NAV LUM FAL NEB JES
MER HOX KER TID SOT	KER ZOR FAL LEV REL
DAZ HOX SIB TID SOT	SIB SOT TAF NEB JES
KOF NEB NAV JES LUM	KER ZOR FAL HOX REL
KOF NEB JES LUM TAF JES LUM	SIB SOT RUD HOX TID
KOF NEB KER JES ZOR	SIB LUM FAL LEV REL
KOF NEB SIB JES LUM	KER ZOR FAL
DAZ HOX KER REL LUM	NAV SOT RUD LEV JES
DAZ LEV SIB JES LUM	KER SOT RUD NEB JES
DAZ LEV KER	SIB SOT RUD LEV REL
DAZ HOX TID SOT TAF REL LUM	KER LUM TAF NEB JES
KOF NEB KER REL SOT	SIB ZOR FAL LEV REL
DAZ LEV REL SOT TAF REL SOT	NAV LUM FAL HOX REL
KOF NEB SIB REL SOT	KER SOT TAF LEV REL
DAZ HOX KER	DAZ LEV REL LUM TAF
MER HOX KER JES ZOR	NAV LUM FAL
MER HOX TID SOT TAF JES ZOR	KER LUM TAF HOX REL
KOF NEB NAV TID SOT	SIB SOT TAF HOX REL
DAZ HOX KER REL SOT	SIB SOT RUD LEV JES
DAZ LEV KER JES LUM	DAZ LEV JES LUM FAL NEB JES
MER HOX SIB JES ZOR	NAV LUM FAL LEV REL

KOF LEV NAV JES ZOR	KER LUM TAF LEV JES
DAZ LEV REL LUM TAF TID SOT	KER ZOR FAL HOX TID
DAZ HOX KER JES ZOR	KER LUM FAL LEV JES
KOF LEV SIB TID SOT	SIB LUM FAL LEV JES
KOF LEV KER JES ZOR	NAV SOT TAF
MER HOX TID SOT TAF REL SOT	SIB LUM TAF NEB JES
MER HOX NAV TID SOT	SIB SOT TAF LEV JES
KOF NEB NAV REL LUM	KER LUM FAL HOX TID
DAZ HOX REL LUM TAF JES LUM	NAV LUM TAF HOX REL
DAZ LEV SIB TID SOT	MER HOX REL LUM TAF HOX REL
KOF LEV REL LUM TAF	KOF LEV JES LUM FAL HOX TID
MER HOX NAV JES LUM	KER LUM FAL NEB JES
DAZ LEV KER TID SOT	DAZ HOX TID SOT RUD HOX TID
DAZ HOX KER TID SOT	KOF NEB JES LUM FAL HOX TID
MER HOX SIB REL LUM	KER SOT RUD HOX TID
KOF NEB SIB REL LUM	KER SOT TAF
KOF NEB SIB TID SOT	KER SOT RUD HOX REL
DAZ HOX REL SOT TAF	SIB LUM FAL HOX REL
DAZ HOX SIB REL LUM	KER SOT RUD LEV REL
KOF LEV SIB REL SOT	SIB LUM TAF HOX TID
KOF LEV NAV	NAV ZOR FAL LEV JES
MER HOX NAV JES ZOR	DAZ LEV REL SOT TAF LEV JES
MER HOX TID SOT RUD	SIB ZOR FAL HOX REL
DAZ LEV NAV REL SOT	SIB LUM TAF LEV REL
DAZ LEV NAV JES LUM	DAZ HOX REL SOT TAF LEV JES
KOF LEV KER REL SOT	SIB SOT RUD NEB JES
DAZ HOX NAV	NAV SOT RUD HOX REL
DAZ LEV NAV JES ZOR	NAV SOT TAF HOX TID
KOF LEV NAV REL SOT	KER ZOR FAL LEV JES
KOF NEB KER REL LUM	DAZ LEV REL SOT RUD LEV JES
KOF NEB JES ZOR FAL TID SOT	DAZ HOX TID SOT RUD HOX REL
KOF LEV KER JES LUM	SIB ZOR FAL
DAZ HOX REL SOT TAF JES LUM	SIB LUM FAL
DAZ LEV KER REL SOT	MER HOX REL SOT RUD HOX TID
MER HOX SIB	NAV SOT RUD LEV REL
MER HOX SIB REL SOT	KER ZOR FAL NEB JES
DAZ LEV REL LUM TAF JES ZOR	NAV LUM TAF LEV REL
KOF NEB JES LUM FAL JES ZOR	KOF LEV REL LUM FAL
KOF NEB JES LUM FAL TID SOT	KER LUM FAL
MER HOX NAV REL SOT	KOF LEV REL LUM TAF
DAZ LEV REL SOT RUD REL LUM	KOF NEB JES LUM FAL HOX REL
DAZ LEV JES LUM TAF REL LUM	KOF LEV JES LUM TAF HOX REL
KOF LEV NAV TID SOT	NAV LUM TAF LEV JES
DAZ LEV REL SOT RUD REL SOT	DAZ LEV JES ZOR FAL HOX REL
DAZ HOX SIB	DAZ LEV JES LUM TAF NEB JES
KOF LEV KER TID SOT	KER SOT RUD LEV JES

KOF LEV SIB	SIB SOT RUD
DAZ HOX REL LUM FAL JES ZOR	NAV ZOR FAL NEB JES
MER HOX TID SOT RUD REL LUM	NAV LUM TAF NEB JES
KOF LEV REL LUM FAL TID SOT	MER HOX REL LUM TAF NEB JES
MER HOX REL LUM FAL JES LUM	KOF LEV REL LUM FAL LEV JES

Appendix C: Test items for Experiments 1 & 2 (Adult 1 & 2)

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Types	Sentences	Types	Sentences
Fragment test	1	AB	KOF HOX	BC	NEB REL
	2		DAZ NEB		LEV TID
	3		MER LEV		HOX JES
	4		MER NEB		NEB TID
	5	CD	JES SOT	DE	SOT FAL
	6		REL ZOR		ZOR TAF
	7		TID LUM		LUM RUD
	8		TID ZOR		ZOR RUD
	9	CDE	JES SOT FAL	ABC	KOF HOX JES
	10		REL ZOR TAF		DAZ NEB REL
	11		TID LUM RUD		MER LEV TID
	12		TID ZOR RUD		MER NEB TID
	13	ABF	KOF HOX KER	FDE	KER SOT FAL
	14		DAZ NEB NAV		NAV ZOR TAF
	15		MER LEV SIB		SIB LUM RUD
	16		MER NEB NAV		NAV ZOR RUD
Movement test	17	CDEAB	JES SOT FAL KOF HOX	DEABC	SOT FAL KOF HOX JES
	18		REL ZOR TAF DAZ NEB		ZOR TAF DAZ NEB REL
	19		TID LUM RUD MER		LUM RUD MER LEV TID

			LEV		
	20		TID ZOR RUD		ZOR RUD
			MER NEB		MER NEB TID
	21	FAB	KER KOF	DEF	SOT FAL KER
			HOX		
	22		NAV DAZ		ZOR TAF
			NEB		NAV
	23		SIB MER LEV		LUM RUD
					SIB
	24		NAV MER		ZOR RUD
			NEB		NAV
	25	CDEABCD	JES SOT FAL	DEABCBC	SOT FAL KOF
			KOF HOX JES		HOX JES
			SOT		HOX JES
	26		REL ZOR TAF		ZOR TAF
			DAZ NEB		DAZ NEB
			REL ZOR		REL NEB REL
	27		TID LUM		LUM RUD
			RUD MER		MER LEV TID
			LEV TID LUM		LEV TID
	28		TID ZOR RUD		ZOR RUD
			MER NEB TID		MER NEB TID
			ZOR		NEB TID
	29	FABCD	KER KOF	DEFBC	SOT FAL KER
			HOX JES SOT		HOX JES
	30		NAV DAZ		ZOR TAF
			NEB REL ZOR		NAV NEB
					REL
	31		SIB MER LEV		LUM RUD
			TID LUM		SIB LEV TID
	32		NAV MER		ZOR RUD
			NEB TID ZOR		NAV NEB TID
Substitution test	33	ib CDE	ib JES SOT	ABC ib	KOF HOX JES
			FAL		ib
	34		ib REL ZOR		DAZ NEB
			TAF		REL ib
	35		ib TID LUM		MER LEV TID
			RUD		ib
	36		ib TID ZOR		MER NEB TID
			RUD		ib
	37	AB et E	KOF HOX et	A et DE	KOF et SOT

			FAL		FAL
	38		DAZ NEB et TAF		DAZ et ZOR TAF
	39		MER LEV et RUD		MER et LUM RUD
	40		MER NEB et RUD		MER et ZOR RUD
	41	ib et E	ib et FAL	A et ib	KOF et ib
	42		ib et TAF		DAZ et ib
	43		ib et RUD		MER et ib
	44		ib et RUD		MER et ib
Movement-plus-substitution test	45	CDE ib	JES SOT FAL ib	ib ABC	ib KOF HOX JES
	46		REL ZOR TAF ib		ib DAZ NEB REL
	47		TID LUM RUD ib		ib MER LEV TID
	48		TID ZOR RUD ib		ib MER NEB TID
	49	et EAB	et FAL KOF HOX	DEA et	SOT FAL KOF et
	50		et TAF DAZ NEB		ZOR TAF DAZ et
	51		et RUD MER LEV		LUM RUD MER et
	52		et RUD MER NEB		ZOR RUD MER et
	53	et E ib	et FAL ib	ib A et	ib KOF et
	54		et TAF ib		ib DAZ et
	55		et RUD ib		ib MER et
	56		et RUD ib		ib MER et

Appendix D: Familiarization sentences for Experiments 3, 4 & 5 (Infant 1, 2 & 3)

Sentence types (numbers in parentheses indicate number of times used in the familiarization)

Grammar 1		*Grammar 2*	
A B F	(2)	F D E	(2)
ib et E C D	(1)	A et D E B C	(8)
C D E A B et	(1)	A et D E	(1)
et E ib C D	(1)	D E F B C	(1)
A B et E C D	(6)	A B C D E	(2)
A B C D E	(2)	A et D E et	(1)
F A B C D	(2)	D E F	(3)
C D E ib C D	(1)	F ib B C	(1)
F A B	(1)	ib F B C	(1)
A B F C D	(4)	A B C D E B C	(1)
et E A B	(3)	F D E B C	(4)
ib F et	(1)	A B C D E et	(1)
F A B et	(1)	ib A et B C	(1)
C D E A B C D	(1)	D E A B C et	(1)
A B et E	(3)	D E A et B C	(2)

Familiarization sentences

Grammar 1	*Grammar 2*
DAZ HOX REL LUM FAL	DAZ HOX REL LUM FAL
DAZ HOX REL LUM TAF	DAZ HOX REL LUM TAF
DAZ HOX et FAL JES LUM	MER et SOT TAF NEB JES
et RUD DAZ HOX	NAV ZOR FAL
KOF LEV et FAL REL SOT	KOF et SOT RUD LEV REL
MER HOX et TAF	MER HOX REL SOT RUD NEB JES
KOF LEV SIB JES ZOR	NAV ZOR FAL HOX REL
KOF LEV KER REL SOT	SOT TAF KER
DAZ LEV et TAF	LUM FAL DAZ et LEV JES
KOF NEB et FAL JES ZOR	KER ib LEV JES
et FAL DAZ LEV	DAZ LEV REL LUM FAL et
et FAL KOF LEV	SIB SOT TAF HOX TID
ib et TAF JES LUM	SIB ZOR FAL NEB JES
DAZ HOX NAV JES ZOR	DAZ et LUM TAF LEV JES
DAZ LEV SIB	KOF et LUM TAF NEB JES
KOF LEV et RUD JES ZOR	ZOR FAL NAV NEB JES
DAZ HOX NAV JES LUM	SOT TAF MER HOX REL et
JES LUM FAL DAZ LEV JES LUM	KOF et LUM FAL HOX REL
KOF NEB et TAF	MER et SOT RUD
ib KER et	KOF et SOT TAF et

et FAL ib JES ZOR	ib MER et NEB JES
SIB MER HOX et	MER et LUM FAL LEV REL
MER HOX et FAL REL LUM	LUM TAF MER et LEV REL
KER DAZ HOX REL SOT	KOF et LUM FAL LEV JES
SIB KOF LEV JES LUM	SOT TAF NAV
REL LUM FAL KOF NEB et	SIB SOT RUD HOX TID
JES LUM FAL ib REL SOT	MER et SOT RUD NEB JES
KER KOF LEV	NAV SOT TAF
KOF NEB NAV	ib KER NEB JES
KOF LEV et TAF JES ZOR	SOT RUD KER

Appendix E: Familiarization sentences for Experiments 6 (Infant 4)

Sentence types (numbers in parentheses indicate number of times used in the familiarization)

Grammar 1		*Grammar 2*	
A B C D E	(1)	F D E	(5)
A B F	(3)	A B C D E B C	(4)
A B C D E C D	(6)	F D E B C	(21)
A B F C D	(20)		

Familiarization sentences

Grammar 1	*Grammar 2*
KOF LEV NAV TID SOT	SIB LUM TAF HOX TID
KOF NEB SIB JES ZOR	KER LUM FAL
MER HOX SIB REL SOT	SIB LUM FAL HOX TID
KOF LEV SIB TID SOT	SIB SOT RUD
DAZ HOX SIB JES LUM	NAV SOT TAF HOX REL
MER HOX REL SOT RUD	NAV SOT TAF NEB JES
DAZ HOX KER REL LUM	SIB LUM TAF LEV JES
DAZ HOX SIB REL SOT	KER LUM FAL NEB JES
MER HOX KER	KOF NEB JES LUM TAF NEB JES
KOF LEV NAV REL LUM	KER SOT TAF LEV JES
DAZ LEV SIB JES ZOR	SIB LUM FAL
DAZ HOX NAV TID SOT	KER ZOR FAL HOX TID
DAZ HOX REL LUM TAF REL LUM	DAZ LEV JES LUM FAL LEV REL
KOF LEV KER JES LUM	NAV SOT TAF LEV REL
MER HOX SIB JES LUM	KER ZOR FAL
MER HOX TID SOT RUD REL LUM	SIB LUM TAF LEV REL
DAZ HOX REL LUM FAL JES LUM	NAV SOT RUD LEV JES

MER HOX REL SOT RUD JES LUM	KER LUM FAL LEV REL
DAZ LEV REL SOT RUD JES ZOR	KER LUM FAL HOX REL
KOF LEV KER REL SOT	SIB SOT TAF LEV REL
MER HOX SIB JES ZOR	NAV ZOR FAL LEV REL
DAZ HOX KER	NAV LUM FAL
KOF NEB NAV REL LUM	KER LUM TAF HOX REL
DAZ HOX SIB TID SOT	KER LUM TAF NEB JES
KOF LEV NAV JES ZOR	SIB ZOR FAL NEB JES
KOF NEB KER	KOF LEV REL LUM FAL LEV REL
KOF LEV NAV REL SOT	KOF NEB JES LUM TAF LEV JES
MER HOX KER REL SOT	NAV ZOR FAL NEB JES
KOF NEB NAV REL SOT	NAV ZOR FAL LEV JES
DAZ HOX REL SOT RUD JES ZOR	KER SOT RUD HOX TID

Appendix F: Test items for Experiments 3, 4 & 6 (Infant 1, 2 & 4)

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Type	Sentences	Type	Sentences
Movement test	1	CDEAB	JES SOT FAL KOF HOX	DEABC	SOT FAL KOF HOX JES
	2		REL ZOR TAF DAZ NEB		ZOR TAF DAZ NEB REL
	3		TID LUM RUD MER LEV		LUM RUD MER LEV TID
	4		TID ZOR RUD MER NEB		ZOR RUD MER NEB TID

Appendix G: Test items for Experiments 5 (Infant 3)

		Grammatical in Grammar 1		*Grammatical in Grammar 2*	
		Type	Sentences	Type	Sentences
Movement test	1	BCADE	HOX JES KOF SOT FAL	CDABE	JES SOT KOF HOX FAL
	2		NEB REL DAZ ZOR TAF		REL ZOR DAZ NEB TAF
	3		LEV TID MER LUM RUD		TID LUM MER LEV RUD
	4		NEB TID MER ZOR RUD		TID ZOR MER NEB RUD

Bibliography

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

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

Barbosa, P. A. (2002). Integrating gestural temporal constraints in a model of speech rhythm production. In S. Hawkins & N. NGuyen (Eds.), *Proceedings of the ISCA workshop on Temporal Integration in the Perception of Speech* (pp. 54). Cambridge: Cambridge University Printing Service.

Beckman, M. & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and the physics of speech,* 152-178. Cambridge: Cambridge University Press.

Beckman, M. & Edwards, J. (1992). Intonational categories and the articulatory control of duration. In Y. Tohkura, Eric Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech, perception, production and linguistic structure,* 359-376. Tokyo: OHM Publishing Co.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

Bowerman, M. (1973). *Early syntactic development: a cross-linguistic study with special reference to Finnish.* Cambridge: Cambridge University Press.

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

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistics Society*, 34, 421-435.

Byrd, D., Kaun, A., Narayanan, S., & Saltzman, E. (2000). Phrasal signatures in articulation. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in Laboratory Phonology V: Acquisition and the lexicon* (pp. 70–87). Cambridge: Cambridge University Press.

Cho, T., & Keating, P. (1999). Articulatory and acoustic studies of domain-initial strengthening in Korean. *UCLA Working Papers in Phonetics*, *97*, 100–138.

Chomsky, N. (1956). Three models for the description of language. In *I.R.E. Transactions on Information Theory,* IT-2, 113-124.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35, 26-58.

Chomsky, N. (1975). *The Logical Structure of Linguistic Theory*. New York: Plenum Press.

Chomsky, N. (1980) *Rules and representations*. Oxford: Basil Blackwell.

Chomsky, N. (1981) *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. (1988). *Language and problems of knowledge.* Cambridge, Mass.: MIT Press.

Chomsky, N. & M. Halle (1986). *The Sound Pattern of English.* New York: Harper & Row.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, *95*, 1570-1580.

Christophe, A., Guasti, M. T., Nespor, M., Dupoux, E., & van Ooyen, B. (1997). Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition. *Language and Cognitive Processes*, *12*, 585-612.

Christophe, A., Mehler, J. & Sebastián-Gallés, N. (2001). Perception of prosodic boundary correlates by newborn infants. *Infancy*, *2*, 385-394.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2007: in press). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*.

Christophe, A., Nespor, M., Guasti, M. T., & van Ooyen, B. (2003). Prosodic structure and syntactic acquisition: The case of the head-complement parameter. *Developmental Science, 6,* 213-222.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language, 51*, 523-547.

Clark, H. & Clark, E. (1977). *Psychology and language: An introduction to psycholinguistcs.* New York: Harcourt Brace Jovanovich.

Cooper, W. & Paccia-Cooper, J. (1980). *Syntax and speech*. Cambridge, MA: Harvard University Press.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14, 597-650.

Crain, Stephen and Mineharu Nakayama (1987). Structure dependency in grammar formation. *Language* 63: 522-543.

Cutler, A., & Norris, D.G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception & Performance*, *14,* 113-121.

de Marcken, Carl. (1996). *Unsupervised Language Acquisition*. Doctoral dissertation. MIT.

de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America, 96*, 2037–2047.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36*, 202–225.

Elman, Jeffrey. (1990). Finding structure in time. *Cognitive Science,* 14, 179-211.

Elman, Jeffrey. (1991). Distributed Representations, Simple Recurrent Networks, and Grammatical Structure. *Machine Learning*, 7, 195-225.

Elman, Jeffrey (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71-99.

Elman, J., Bates, E., Johnson, M.H., and Parisi, D. (1996). *Rethinking innateness : A connectionist perspective on development.* MIT Press.

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

Fisher, C., & Tokura, H. (1996b). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 343–363). Mahwah, NJ: Lawrence Erlbaum Associates.

Fodor. J.A. (1966) How to learn to talk: Some simple ways. In F. Smith and G. Miller (Eds.), *The genesis of language* (pp. 105-122). Cambridge, MA: MIT Press.

Fougeron, C. & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728-3740.

Gelman, S. & Taylor, M. (1983). Semantic vs. syntactic clues: the proper-common distinction in 2-year-olds. Paper presented at the Boston University Conference on Language Development, Boston, October.

Gerken, LouAnn. (2004). Nine-month-olds extract structural principles required for natural language. *Cognition,* 93, B89–B96.

Gerken, LouAnn. (2006). Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition,* 98: 3, B67-B74.

Gerken, LouAnn, & Bolt, Alex. (2008). Three Exemplars Allow at Least Some Linguistic Generalizations: Implications for Generalization Mechanisms and Constraints. *LANGUAGE LEARNING AND DEVELOPMENT*, 4(3), 228–248.

Gerken, L.-A., Jusczyk, P. W., & Mandel, D. R. (1994). When prosody fails to cue syntactic structure: Nine month olds' sensitivity to phonological versus syntactic phrases. *Cognition, 51*, 237–265.

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

Gleitman, L., Gleitman, H., Landau, B. & Wanner, E. (1988). Where learning begins: initial representations for language learning. In F. Newmeyer (Ed.), *The Cambridge linguistic survery.* Cambridge: Cambridge University Press.

Gleitman, L. & Wanner, E. (1982). Language acquisition: The state of the state of the art. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art.* Cambridge: Cambridge University Press.

Gomez, Rebecca (2002). Variability and detection of invariant structure. *Psychological Science,* 13, 431-436.

Gomez, R. & Gerken, L.A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition, 70*, 109-135.

Rebecca L. Gomez & Jessica Maye (2005). The Developmental Trajectory of Nonadjacent Dependency Learning. *Infancy*. Vol. 7, No. 2, 183-206.

Gout, A., Christophe, A. & Morgan, J. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language, 51,* 547-567.

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

Hardcastle, W. (1985). Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. *Speech Communication*, *4*, 247–263.

Hayes, B., & Lahiri, A. (1991). Bengali intonational phonology. *Natural Language and Linguistic Theory*, *9*, 47–96.

Hicks, Jessica (2006). *The Impact of Function Words on the Processing and Acquisition of Syntax.* Doctoral dissertation. Northwestern University.

Hirsch-Pasek, Kathy and Roberta M. Golinkoff (1996). *The origins of grammar: evidence from early language comprehension*. MIT Press, Cambridge, MA.

Hirsh-Pasek, K., Kemler Nelson, D., Jusczyk, P., Cassidy, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26*, 269-286.

Holst, T., & Nolan, F. (1995). The influence of syntactic structure on [s] to [sh] articulation. In B. Connell & A. Arvaniti (Eds.), *Papers in Laboratory Phonology IV: Phonology and Phonetic Evidence* (pp. 315–333). Cambridge: Cambridge University Press.

Houston, Derek, Santelmann, Lynn & Jusczyk, Peter (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes, 19*, 97–136.

Hunter, M. A., & Ames, E. W. (1989). A multifactor model of infant preferences for novel and familiar stimuli. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (vol. 5, pp. 69–93), Norwood, NJ: Ablex.

Hyman, L. M. (1977). On the nature of linguistic stress. In L. M. Hyman (Ed.), *Studies in stress and accent* (*Southern California Occasional Papers in Linguistics,* No. 4, pp. 37–82). Los Angeles: University of Southern California.

Jackendoff, Ray. (1977). *X-bar syntax*. Cambridge, MA: MIT Press.

Johnson, E., & Jusczyk, P. W. (2001). Word segmentation by 8 month olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44,* 548–567.

Jusczyk, P. (1989). *Perception of cues to clausal units in native and non-native languages.* Paper presented at the biennial meeting of the Society for Research in Child Development, Kansas City, Missouri.

Jusczyk, P. W., & Aslin, R. N. (1995). Infant's detection of sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

Jusczyk, P., Hirsh-Pasek, K., Kemler Nelson, D., Kennedy, L., Woodward, A. & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology, 24*, 252-293.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999a). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, *61*, 1465–1476.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999b). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39,* 159–207.

Katz, B., Baker, G., and Macnamara, J. (1974). What's in a name? On the child's acquisition of proper and common nouns. *Child Development,* 45: 269-273.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2003). Domain-initial articulatory strengthening in four languages. In J. Local & R. Ogden & R. Temple (Eds.), *Papers in Laboratory Phonology, VI: Phonetic Interpretation*, 143-161. Cambridge: Cambridge University Press.

Keenan, E. (1976). Towards a universal definition of "subject." In C. Li, ed., *Subject and topic.* New York: Academic Press.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A. & Gerken, L.A. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development, 18*, 111–116.

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

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics, 3,* 129–140.

Lappin, Shalom & Shieber, Stuart (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics, 43*, 1-34.

Lebeaux, D. & Pinker, S. (1981). The acquisition of the passive. Paper presented at the Boston University Conference on Language Development, Boston, October.

Legate, Julie & Yang, Charles (2002) Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review,* 19, 151-162.

Lewis, John & Elman, Jeffrey (2002). Learnability and the Statistical Structure of Language: Poverty of Stimulus Arguments Revisited. In Barbora Skarabela, Sarah Fish & Anna H.-J. Do (Eds.), *Proceedings of the 26th Boston University Conference on Language Development*, 359-370. Somerville, MA: Cascadilla Press.

Lust, Barbara. (2006). *Child Language.* Cambridge: Cambridge University Press.

Macnamara, J. (1982). *Names for things: a study of child language.* Cambridge, Mass.: Bradford Books/MIT Press.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121.

Maye, J. & Gerken, L.A. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. Fish & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Conference on Language Development*, 522–533. Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. & Gerken, L.A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*, 101–111.

Miller, George & Chomsky, Noam (1963). Finitary Models of Language Users. In R. D. Luce, R. R. Bush & E. Galanter (Eds.) *Handbook of mathematical psychology*, *Vol. II.* New York: Wiley.

Mints, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*, 91–117.

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

Morgan, J. (1986). *From simple input to complex grammar.* Cambridge, MA: MIT Press.

Morgan, J. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language, 35,* 666–668.

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

Morgan, J. L., Meier, R. P., & Newport, E. L. (1989). Facilitating the acquisition of syntax with cross-sentential cues to phrase structure. *Journal of Memory and Language*, *28,* 360–374.

Morgan, J. & Newport, E. L. (1981). The role of a constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *20*, 67–85.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66,* 911–936.

Nazzi, T., Kemler Nelson, D. & Jusczyk, P. (2000) Six-month-olds' detection of clauses embedded in continuous speech: effects of prosodic well-formedness. *Infancy 1,* 123-147.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development* 38.

Nespor, Marina & Vogel, Irene (1986). *Prosodic phonology*. Dordrecht: Foris.

Pasdeloup, V. (1990). *Modéle de régles rythmiques du franc¸ais appliqué a` la synthése de la parole.* Doctoral dissertation. Université d'Aix-en-Provence, Aix-Marseille.

Pearl, Lisa (2007). *Necessary Bias in Natural Language Learning.* Doctoral dissertation. University of Maryland, College Park.

Pereira, F. (2000). Formal grammar and information theory: Together again? In *Philosophical transactions of the royal society*, 1239-1253. London: Royal Society.

Peters, A. (1983). *The units of language acquisition.* New York: Cambridge University Press.

Pinker, S. (1979). Formal models of language learning. *Cognition*, 7, 217-283.

Pinker, S. (1982). A theory of the acquisition of lexical interpretive grammars. In J. Bresnan, ed., *The mental representation of grammatical relations.* Cambridge, Mass.: MIT Press.

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

Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics, 20*, 331–350.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.

Rietveld, A. C. M. (1980). Word boundaries in the French language. *Language and Speech*, *23*, 289–296.

Rohde, Douglas (1999). LENS: The light, efficient network simulator. Technical Report CMU-CS-99-164, Carnegie Mellon University, Department of Computer Science, Pittsburgh, PA.

Rohde, Douglas & David Plaut (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72, 67-109.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language, 44*, 493–515.

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

Saffran, Jenny, Hauser, Marc, Seibel, Rebecca, Kapfhamer, Joshua, Tsao, Fritz & Cushman, Fiery. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, 107: 479–500.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word Segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-621.

Santelmann, Lynn M. and Peter W. Jusczyk (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition,* 69, 105 – 134.

Scott, D. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America, 71,* 996–1007.

Seidl, Amanda (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57*, 24–48.

Selkirk, Elisabeth. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shady, M. E. (1996). Infants' sensitivity to function morphemes. Unpublished doctoral dissertation, State University of New York at Buffalo, NY.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*, 193–247.

Slobin, D. (1973). Cognitive prerequisites for the development of grammar. In C. Ferguson and D. I. Slobin, eds., *Studies of child language development.* New York: Holt, Rinehart and Winston.

Soderstrom, M., Kemler Nelson, D. & Jusczyk, P. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development, 28*, 87-94.

Soderstrom, M., Seidl, A., Kemler Nelson, D. & Jusczyk, P. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49,* 249-267.

Swingley, Daniel (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*, 86–132.

Thiessen, Erik & Saffran, Jenny (2003). When Cues Collide: Use of Stress and Statistical Cues to Word Boundaries by 7- to 9-Month-Old Infants. *Developmental Psychology*, *39*, 706–716.

Thiessen, Erik & Saffran, Jenny (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*, 73-100.

Thompson, S. & Newport, E. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development, 3*, 1-42.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74, 209-253.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic boundaries. *Journal of the Acoustical Society of America, 91,* 1707–1717.

Yang, Charles (2006). *The infinite gift: How children learn and unlearn the languages of the world*. New York: Scribner.