UNIVERSITY OF CALIFORNIA,
IRVINE

# A Quantitative Framework for Specifying Underlying Representations in Child Language Acquisition

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Galia Kaas Bar-Sever

Dissertation Committee:
Professor Lisa Pearl, Chair
Professor Barbara Sarnecka
Assistant Professor Gregory Scontras

2019

# TABLE OF CONTENTS

# LIST OF FIGURES

Page

# LIST OF TABLES

# ACKNOWLEDGMENTS

I would like to thank the Department of Cognitive Sciences and the Department of Language Sciences for their support. I am unendingly appreciative of both the instruction and the administrative efforts on my behalf.

I would like to thank my mother, father, and brother, whose support has been invaluable.

I would like to thank all of the members of the QuantLang lab, with special thanks to undergraduates who have assisted in my research.

I want to give special thanks to Alandi Bates, whose guidance was indispensable to this work.

I would like to thank my committee, Dr. Barbara Sarnecka and Dr. Gregory Scontras, for their encouragement, insight, and support.

I especially want to thank my advisor, Dr. Lisa Pearl. This work would not have been possible without her advice, guidance, and wisdom. I am truly grateful for her many years of mentorship.

# CURRICULUM VITAE

## Galia Kaas Bar-Sever

### EDUCATION

**Doctor of Philosophy in Cognitive Sciences** **2019**
University of California, Irvine *Irvine, CA*

**Master of Arts in Psychology** **2018**
University of California, Irvine *Irvine, CA*

**Bachelor of Science in Biological Sciences** **2013**
University of California, Irvine *Irvine, CA*

### RESEARCH EXPERIENCE

**Graduate Research Assistant** **2014–2019**
University of California, Irvine *Irvine, California*

### TEACHING EXPERIENCE

**Instructor of Record** **2018–2019**
University of California, Irvine *Irvine, California*

**Teaching Assistant** **2014–2019**
University of California, Irvine *Irvine, California*

## REFEREED CONFERENCE PUBLICATIONS

**Syntactic Categories Derived from Frequent Frames Benefit Early Language Processing in English and ASL.** **Jun 2018**
BUCLD 42.

**Syntactic Categories Derived from Frequent Frames Benefit Early Language Processing in English and ASL.** **Aug 2015**
BUCLD 40.

# ABSTRACT OF THE DISSERTATION

A Quantitative Framework for Specifying Underlying Representations
in Child Language Acquisition

By

Galia Kaas Bar-Sever

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2019

Professor Lisa Pearl, Chair

My research broadly demonstrates how quantitative approaches can be effectively leveraged for developmental research. In this dissertation, I show one quantitatively precise way to identify the nature of developing mental representations in a variety of domains; my approach utilizes the connection between a learners input, creation of a potential mental representation from that input, and evaluation with respect to the learners output. More specifically, the quantitative approach I use leverages both realistic input data and realistic output data as part of the model design and evaluation. Using modeling, we have the opportunity to concretely evaluate representational options that we would not otherwise be able to disambiguate. I demonstrate this quantitative approach with three case studies in language development: (I) the development of adjective ordering preferences, where I find that the representations that adults use to talk to children are different than the ones used to talk to other, adults, (II) immature individual syntactic category representations, where I identify precisely which immature category representation young children are likely to be using, and (III) the development of adult productive syntactic category representations, where I identify when adult category knowledge emerges in typically and atypically developing populations.

# Chapter 1

# Introduction

## 1.1 The puzzle of language development

As anyone who has tried to learn another language knows, adults can struggle with this task. For most adults, it can take years to develop anything close to fluency, and they rarely reach true native ability. Babies, on the other hand, can't solve equations, compose sonatas, or perform any of the complex tasks that adults can[1]; in learning language, though, they shine. Babies will achieve a high level of proficiency in their native language within five years, while an adult may never reach the same level of proficiency in a second language. Moreover, babies do this naturally, without much explicit correction or instruction (Sakai [2005], Saffran et al. [2001]).

The ease with which babies process, organize, and use their languages rules belies the incredible complexity of language learning. Babies must independently sort out all kinds of language and non-language data, and determine how to construct the rich systems of linguistic representation that underlie language proficiency. This includes knowledge of what

[1]The author would like to reassure any babies reading this that, while an adult, she cannot compose sonatas either.

sounds fit together (phonology), what parts make up words (morphology), individual word meaning (lexical semantics), and how words work together (syntax), not to mention the additional layers built upon these fundamental skills, such as understanding sentence-level semantics and pragmatics.

Given that more sophisticated knowledge is built upon more fundamental knowledge, this means that the development of this knowledge necessarily happens in stages, with more basic steps in the learning process preceding more complex ones. Children create linguistic building blocks, or mental representations, in a variety of domains (like knowledge of phonology, morphology, lexical semantics, syntax, sentence-level semantics, and pragmatics) based on their input. The mental representations that children form during this staged process are critical, since they scaffold future learning. Understanding the development of these mental representations is thus crucial for researchers looking for a complete picture into how language learning works. Relatedly, precise knowledge about when certain linguistics structures typically develop allows for diagnosis of atypical development when we detect deviations from typical development.

While these representations exist in the mind of the child, and are therefore unobservable, we can theorize about them based on behavior that is observable. In particular, because children rely on these representations both to understand and produce language, researchers have traditionally theorized about the representations based on the children's behavior. However, this kind of theorizing is difficult to do, because the link between the representations and the behavior is complex. Even in carefully controlled experiments, it is difficult to draw a causal link between behavior and a specific mental representation. This is especially true in experiments involving very young children. For example, say we observe a child describe an object as a "*small grey kitten*". What is causing this particular order of adjectives (*small* before *grey*) in describing this kitten? It could be that this ordering of adjectives depends strictly on how often the child heard those particular adjectives in those particular locations

in the string. It could also be that the adjective order is dependent on an adjective's lexical class. In fact, a number of mental representations of adjective order could produce the *small grey kitten* utterance we observe. Which could it be? It's impossible to ask adults directly what representation of adjectives they were using to generate their observed adjective productions, much less a child.

Even though we can't observe mental representations directly, we often have a good idea of what the adult mental representations could look like for a particular language phenomena. For example, we have a pretty good idea that words are represented within parts of speech like NOUN, VERB, etc.) and these categories interact with each other in predictable ways. For instance, an adult might represent a noun phrase (like *the kitten*) as a combination of categories, like DETERMINER (*the*) and NOUN (*kitten*). Adults seem to be using these categories productively, and the way these categories are organized is unlikely to change much. The question then becomes how do children develop these categories out of the words they encounter?

Importantly, as children develop, their representations in turn are developing. In particular, children may be considering different immature representations along the way to developing the mature, adult representation. The approach I take allows us to consider different potential representations a child could be using, both immature and mature. My approach allows us to predict what output the child would generate from a particular representation, given the child's input; then, we can compare the predicted outputs against the child's true output to see which one matches best.

More specifically, we can use the child's input to mathematically specify the exact form of a candidate representation. For example, we can disambiguate between a child representing *the kitten* as simply being amalgam based on their input (i.e., how many times they heard *the kitten* in their input) vs. a child representing *the kitten* as a combination based on their own internal representation of a DETERMINER (*the*) combined with a NOUN (*kitten*).

We can then evaluate which potential representation best informs how the child actually used *the kitten* in their output. This connects the child's input, which is observable, to the mental representation, which isn't observable. Then, we can mathematically specify the output that the candidate representation would cause the child to produce. So, the unseen mental representation is again connected to something observable: the child's output. This approach thus allows us to determine which representation the child is most likely to have by connecting that unobservable representation to language data we can observe, namely a child's language input and output.

## 1.2 Underrepresented populations in computational language research

An important thing to consider when investigating child language development is that not every child's development proceeds in a typical fashion. There are many populations whose language learning trajectories diverge from typical development. In particular, there has been a lot of research in the developmental linguistics community focusing on typically-developing children who are learning spoken languages. However, there is much less work in developmental linguistics outside of this population, in both typically-developing children learning a non-spoken language (e.g. American Sign Language) or children from clinical populations (like Autism Spectrum Disorder, Specific Language Impairment, or Down Syndrome).

A large amount of data is necessary for robust analysis of any phenomenon in language acquisition. However, there are nontrivial issues in undertaking computational modeling for these populations that are underrepresented in computational language research. For one, there is limited information about how mental representations and behavior are linked in children who either have atypical input or have atypical cognitive abilities. In cognitively atypical

populations, not only is the population generally smaller than in typically-developing populations, but in some parts of the atypically-developing population (such as children who suffer from particular disorders such as Specific Language Impairment and Down Syndrome), production is necessarily affected by this disorder; this lack of productions in turn results in less observable language data to collect. When we look at language development in other linguistically-diverse, but not necessarily cognitively diverse populations, the modality of the linguistic data adds an extra wrinkle. Such is the case with American Sign Language. It is difficult in itself to code auditory language data, but coding signed languages presents a particular challenge. This is not only because there are different annotation conventions between annotators, but also because the nature of signed languages means features can be simultaneously articulated, making transcribing these elements tricky. For all underrepresented populations, this lack in quantity of data is compounded by the fact that there are fewer able coders of such data, as well as there being varied methods and annotation conventions between able coders, resulting in much less available data.

However, even given these difficulties, it is crucial to take a broad look at data from cognitively and linguistically diverse populations in order to make general claims about how language development works. Lack of information in both arenas further limit our understanding of language development in general. Looking at linguistically diverse populations allows us to disambiguate what aspects of language learning are specific to a particular language and what is true about language learning more generally. This can be even further expanded by looking at signed languages. By examining signed languages, we can gain insight into how language develops irrespective of modality, and what is modality dependent. Looking at underrepresented populations with different developmental profiles also gives us insight in language development. In particular, we can better understand (i) what cognitive faculties are required to achieve certain stages of linguistic development that involve particular mental representations, and (ii) what is actually different (representations, strategies, etc.) in language development in typically vs. atypically-developing populations. That is,

are children with different language learning profiles, whether they be from clinical populations or differing modalities) constructing the same linguistic representations as typically developing children? Do the strategies they use differ?

In the following chapters, I use my novel quantitative approach on three different case studies of language development in order to identify which mental representations children are using at a specific time when we know both their input and output. I first look at the development of adjective ordering preferences in typically-developing children; then I turn to the development of immature syntactic categories in typically-developing children, and finally to the emergence of adult-level productive syntactic category knowledge in typically-developing and atypically-developing children.

# Chapter 2

# The Development of Adjective Ordering Preferences

## 2.1 Introduction

Adults have robust ordering preferences that determine the relative order of adjectives in multi-adjective strings: this is why *small grey kitten* and *nice round penny* are preferable to *grey small kitten* and *round nice penny*. Adults are reliably and robustly uncomfortable with the latter options, yet are typically unable to pinpoint why they have this reaction. Notably, these preferences surface for any multi-adjective string, even ones never before encountered: English adults would probably prefer *tiny green magical mouse-riding gnomes* to *mouse-riding magical green tiny gnomes*, even though it is unlikely they have encountered these particular adjectives strung together before. Even more remarkable than the robustness and productivity of these preferences in English is the fact that these ordering preferences surface in a variety of unrelated languages, both those with pre-nominal adjectives (like English, Dutch, or Mandarin Chinese) and those with post-nominal adjectives (like Selepet

or Mokilese) that follow the modified noun (for discussion, see Dixon 1982, Sproat and Shih 1991).

When it comes to the source of these preferences, there have been a number of hypotheses. The null hypothesis for adults would hold that they simply repeat back what they hear when forming multi-adjective strings, reflecting the statistics of the particular multi-adjective strings in their input. However, this kind of input frequency strategy is limited in its productivity (if you haven't heard it, you don't have a preference about it) and adults are not limited this way. Importantly, because of their productivity, these preferences appear to be based on abstract representations, rather than simply reflecting the positioning of specific adjectives in the input. If you aren't considering both adjectives in a string as a combined unit (for example, assuming the adjectives in *nice round seals* are an atomic unit as *nice+round*), keeping track of the positional frequency of the adjectives is the next best null hypothesis (e.g., keeping track that *nice* appeared before *round*, which appeared before the noun). This will work well if you are working with lexical items you have heard before. For example, I might have a preference for *green mouse-riding gnomes* over *mouse-riding green gnomes* simply because I have previously heard *green* 2 "slots" away from the noun (i.e. like in *green crystalline chinchillas* or *green fire-eating pixies*). We can refer to adjectives like *nice* in this case as being in the "2-away" position.

But how exactly do adults represent these ordering preferences? Prevailing approaches in linguistics advance the idea that adult adjective ordering is determined by abstract syntax, with adjectives grouped into lexical semantic classes that are hierarchically ordered [Dixon, 1982, Cinque, 1994]. These lexical classes and their hierarchical ordering would then serve as primitives in the representation of the preferences. For example, because *green* is a COLOR adjective, it would necessarily be placed before *metal* because a COLOR adjective would always come before a PHYSICAL adjective hierarchically – this is why I say *green metal stars* instead of *metal green stars*. Yet, why should these classes be ordered the way they are, and

how do we handle adjectives that do not fit neatly into a single clear lexical class? Words like *bright*, for example, could refer to the physical notion of emitting light, or apply to sentient creatures as being smart.

Recently, Scontras et al. [2017] identified adjective subjectivity as a robust predictor of adult ordering preferences, with less subjective adjectives preferred closer to the modified noun; they advanced the hypothesis that ordering preferences—and the lexical class ordering observed cross-linguistically—derive from the perceived subjectivity of the adjectives. Thus, perceived subjectivity would serve as a primitive of the adult representation of adjective ordering preferences. This would also mean that it wouldn't matter if an adjective didn't have a clear lexical class it would fit into–all that matters would be its relative subjectivity.

Still, little is known about the development of these adjective-ordering preferences in children, other than that these preferences do in fact develop [Bever, 1970, Martin and Molfese, 1972, Hare and Otto, 1978]. What we do know is incomplete and messy, discussed in more detail later on in section 2.2.3.

Bar-Sever et al. [2018] assessed when more abstract knowledge about adjective ordering emerges, how that knowledge gets represented, and whether the knowledge representation matches what we believe to be active in adults. To perform this same assessment here, I build off of previous work by Bar-Sever et al. [2018] and use corpus analysis and quantitative metrics to connect children's linguistic input, potential underlying representations regarding adjective ordering, and linguistic output, thereby arriving at a clearer picture of children's knowledge in this domain.

But what mental representations are adults using when talking to children? Across many different linguistic domains, it's known that adult-directed and child-directed speech can differ in fundamental ways [Ferguson, 1964, Fernald et al., 1989, Grieser and Kuhl, 1988, Snow, 1977]. Adults are known to adjust the complexity of their child-directed speech based

on the child's age (e.g., Kunert et al. 2011), and so it may be that the representations underlying child-directed adjective orderings vary depending on the age of the child being addressed. For instance, hyperarticulation is thought to be a common feature of child-directed speech, most often considered in highlighting phonetic categories, specifically vowels [Kuhl et al., 1997]. If adults are providing children with input of a fundamentally different character from what they are providing other adults—for example, by hyperarticulating positional differences between adjectives—we ought to understand the pressures that lead to that divergence.

Since adults are probably at the target state, looking at what underlies their productions towards each other might signify the target representation for these preferences. Thus, child-directed speech may differ from adult-directed speech precisely because of these systematic differences. However, there may be differences in child-directed speech in some domains and not others (Pearl and Sprouse [2013], Bates and Pearl [2019]). I will investigate if the mental representations of adjective-ordering preferences are something that changes in child-directed speech. Looking at these two language interaction types will tell us about the one that presumably serves at the target state for learning (adult-directed speech), while the other is what underlies children's input (child-directed speech).

## 2.2 Previous accounts of adjective order

I start by reviewing relevant background for the competing hypotheses surrounding adult knowledge of adjective ordering. I then review behavioral studies aimed at understanding children's preferences, given that there is little known about children's development of these preferences.

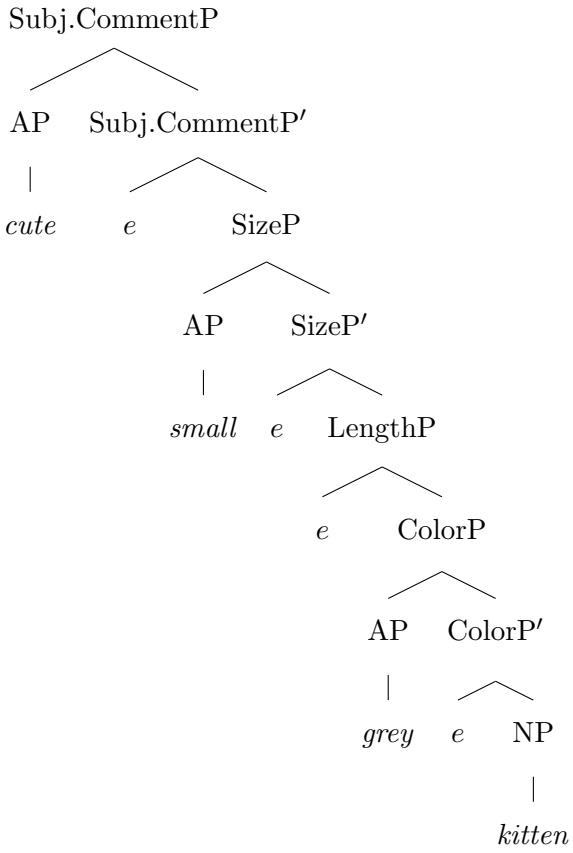
### 2.2.1 The lexical class hypothesis

The lexical class hypothesis begins with the assumption that adjectives come pre-sorted into classes according to their semantic properties: COLOR adjectives group together, SIZE adjectives group together, etc. To account for adjective ordering, these classes correspond to a deterministic hierarchy that maps adjective strings to their linear order, as in (1); higher positioning in the hierarchy leads to greater distance from the modified noun.

(1) *Lexical semantic class hierarchy from Dixon [1982]*:
VALUE > DIMENSION > PHYSICAL PROPERTY > SPEED
> HUMAN PROPENSITY > AGE > COLOR

As proposed by Dixon [1982], these hierarchical lexical semantic classes form part of a speaker's internal grammar and the lexical classes themselves are universal, existing regardless of differences in the morpho-syntactic expression of these semantic types. In an attempt to formalize the linear ordering of these lexical semantic class hierarchies, Cinque [1994] built on these classes and proposed a fully syntactic account of ordering preferences whereby the individual classes project their own phrasal structure, with one phrase hierarchically dominating another. Under a syntactic account, in *small grey kitten*, the COLOR adjective appears closer to the noun than the SIZE adjective because the adjective phrase projected by *small* hierarchically dominates the adjective phrase projected by *grey*. This hierarchical ordering gets expressed as the linear order of adjectives modifying a noun. The proposal has been elaborated on since its initial formulation, with recent authors proposing even richer structure, as in [Scott, 2002] (see also Laenzlinger 2005). In this example, certain phrases (like Color Phrases – ColorP) are constituents of successive hierarchical phrases, with nodes left empty to fit potential adjectives in.

(2) *Phrase structure proposed by Scott [2002] for cute small grey kitten*




Throughout this work on lexical classes, authors have disagreed about the precise specification of the classes themselves. Dixon's classes in (1) gave way to Cinque's (POSSESSIVE, SPEAKER-ORIENTED, SUBJECT-ORIENTED, MANNER/THEMATIC), which depart from the classes proposed by Sproat and Shih [1991] (QUALITY, SIZE, SHAPE, COLOR, PROVENANCE). Table 2.1 shows each of the authors' proposed classes. While there is some overlap (COLOR for Dixon [1982] and Sproat and Shih [1991]), there is little agreement.

Still, despite the fact that it is hard to settle on *the* universal adjective classes, it has been shown that a certain ordering of adjective classes goes some way in matching the patterns observed in adults. What this collection of research does not address is where the hierarchy comes from in the first place: supposing SIZE adjectives do syntactically dominate COLOR adjectives, why should this be the case and not the reverse? This approach also relies on an

| Dixon [1982] | Cinque [1994] | Sproat and Shih [1991] |
|---|---|---|
| VALUE | POSSESSIVE | QUALITY |
| DIMENSION | SPEAKER-ORIENTED | SIZE |
| PHYSICAL PROPERTY | SUBJECT-ORIENTED | SHAPE |
| COLOR | MANNER/ THEMATIC | COLOR |
| SPEED | | PROVENANCE |
| HUMAN PROPENSITY | | |
| AGE | | |

Table 2.1: Comparison of lexical semantic classes proposed by Dixon [1982], Cinque [1994], and Sproat and Shih [1991]

ability to identify the appropriate lexical semantic class for any given adjective, enforcing a sorting into discrete bins based on a static meaning. What about adjectives that fail to fall into an existing semantic class, like *medical* or *multiple*? If an adjective doesn't fall neatly into a class, how can someone make lexical-class-based decisions about how to order these adjectives? Would these words be placed randomly, or by some other strategy?

### 2.2.2 The subjectivity hypothesis

In an attempt to address the concerns facing the lexical class hypothesis head-on, recent work by Scontras et al. [2017] advances the hypothesis that aspects of an adjective's meaning determine its relative position in a multi-adjective string. In particular, Scontras et al. propose that the perceived subjectivity of the property an adjective names influences its ordering. This *subjectivity hypothesis* states that less subjective adjectives are preferred closer to the modified noun than adjectives that are more subjective (see also Hetzron 1978, Hill 2012). This would mean that a turtle that was described as both Italian and gentle would be labeled as a *gentle Italian turtle*, because *gentle* is more subjective than *Italian*, and therefore placed farther from the noun.

Scontras et al. operationalized subjectivity as the potential for faultless disagreement between two speakers about whether an adjective applies to some object [Barker, 2013, Kennedy, 2013, Kölbel, 2004]. In a test of faultless disagreement, two speakers are presented with an

object (say, a kitten); the speakers then disagree about whether the object holds some property (say, being small). To the extent that both speakers can be right while they disagree, the property (and the adjective that names it) admits that degree of faultless disagreement, which stands proxy for its subjectivity. This makes sense, because what makes something subjective or not is whether two people can disagree without either being wrong – i.e., the disagreement being faultless. So, an adjective's subjectivity is defined by how much disagreement speakers can have about that adjective without one of the speakers necessarily being wrong. An adjective like *small* admits a relatively high degree of faultless disagreement (two people can disagree about whether they consider an object small), and is therefore relatively subjective. In contrast, an adjective like *grey* is relatively objective: when two people disagree about whether something is grey, one of those people is likely to be wrong. One might think it would be difficult to rate subjectivity of a given adjective, but in fact, Scontras et al. found that participants' estimates of faultless disagreement matched their ratings for adjective "subjectivity" ($r^2 = .91, 95\%$ CI $[.86, .94]$). So, simply asking adults how "subjective" they believe an adjective to be (a metalinguistic task) can serve as a proxy for the potentially more ecologically valid faultless disagreement task.

To get a clearer picture of the English ordering preferences that need to be accounted for, Scontras et al. measured ordering preferences in a behavioral experiment; participants indicated the preferred ordering for multi-adjective strings (e.g., *small grey kitten* vs. *grey small kitten*). To ensure that the behavioral measure captured the implicit knowledge that speakers use when forming multi-adjective strings, Scontras et al. compared their measure against naturalistic multi-adjective strings from corpora. Finding a high correlation between the behavioral measure and corpus statistics ($r^2 = .83, 95\%$ CI $[.63, .90]$), Scontras et al. concluded that the preferences that were measured accurately capture speaker knowledge.

To test the subjectivity hypothesis, Scontras et al. used their estimates of adjective subjectivity to predict the preferred adjective orderings. They found that adjective subjectivity

accounts for between 51% and 88% of the variance in the ordering preferences. In other words, subjectivity does predict adjective ordering, thus offering a cognitive explanation for the linguistic universal of adjective ordering preferences. Given its promise in accounting for adult knowledge of adjective ordering, one might reasonably wonder about how this subjectivity-based knowledge might develop.

### 2.2.3 The development of adjective ordering preferences

The cross-linguistic robustness of ordering preferences has led many researchers to conclude that the knowledge underlying these preferences is innate, pre-specified as part of the Universal Grammar that shapes human language [Dixon, 1982, Sproat and Shih, 1991, Cinque, 1994]. Part of the appeal of the subjectivity hypothesis is that it allows us to move away from claims of innateness (and the puzzle of genetically specifying linguistic structure) [gkb: see greg notes]. Instead, the subjectivity hypothesis favors an account where subjectivity awareness develops as we use language to communicate; after all, the potential for faultless disagreement is a problem all speakers must attend to. To better understand the role of subjectivity in ordering preferences and the pressures that lead to it, we must therefore ask whether this knowledge is present from the start, or whether it develops—perhaps in stages—into what we observe as the adult state.

There have been several studies examining adjective ordering preferences in children that tested for emergence of preferences according to lexical class, but the results have been unclear. Still, the existing evidence at least suggests that the preferences do in fact develop in the sense that there is a change over time from less adult-like preferences to more adult-like preferences.

Bever [1970] found that with children between two and five years of age, the younger children were more likely to repeat unnatural adjective orderings such as *the plastic large pencil*; older

children corrected the phrase to *the large plastic pencil*. We might therefore conclude that the younger children fail to demonstrate stable adjective ordering preferences. However, Martin and Molfese [1972] attempted to recreate Bever's experiment but were unable to replicate his findings. This replication failure led Martin and Molfese to suggest that the original repetition task is not a reliable measure of adjective ordering preferences. In its place, they used a production task, finding that three- and four-year-olds produced phrases with adjectives denoting CLEANLINESS closer to the noun than COLOR adjectives (e.g., *yellow clean house*), while the adult preference is for COLOR adjectives to appear closer (i.e., *clean yellow house*). This result provides evidence that children's preferences differ from adult preferences, but only with respect to adjectives of CLEANLINESS and COLOR. A later study by Hare and Otto [1978] had children in grades one through five arrange three adjectives of SIZE, COLOR, and MATERIAL to create natural adjective phrases; children in each succeeding grade level chose the adult-preferred order of SIZE–COLOR–MATERIAL (e.g., *little yellow rubber duck*) more often than children in the preceding grade level.

These developmental studies leave much unsettled, but they do suggest that adjective ordering preferences develop or strengthen over time. However, there is disagreement among these studies on the age of acquisition, and what the developmental trajectory looks like. Moreover, none of these studies attempt to tie children's knowledge to adjective subjectivity. If in fact the perceived subjectivity of adjectives is what adults are using to inform their adjective ordering preferences, we ought to wonder when children begin to deploy this strategy.

Notably, this question becomes more complicated in light of recent work showing that children may not have reliable estimates of subjectivity until around the age of seven or eight [Foushee and Srinivasan, 2017]. If subjectivity is not available but children still demonstrate clear ordering preferences, how are these preferences acquired from the input children receive and represented with their available cognitive resources? It may be possible (indeed, likely)

that children evolve through various stages of knowledge representation for their adjective ordering preferences. To investigate this knowledge and its stages of development, I examine children's production of multi-adjective strings in light of the input they are receiving at different ages as well as how adults are forming multi-adjective strings when speaking to each other and to children at different stages of development.

If the current best idea is that subjectivity best accounts for adults' adjective ordering preferences, there are a few questions that emerge which I can try to answer. First, is subjectivity the target state for representing these underlying preferences? Using my approach of connecting input to output via the underlying representation, I can evaluate head-to-head these potential underlying representations (the sophisticated lexical class and subjectivity-based representations as well as the less taxing positional frequency representation, discussed in Section 2.1) as a reasonable baseline strategy that relies on individual lexical items rather than more abstract representations. Second, I can look at which of these representations children seem to be using at different ages and thus concretely identify their developmental trajectory. Third, I can look at how adults talk to children versus each other, because adults are known to adjust linguistic properties of their language when directing it towards children.

## 2.3 Quantitatively assessing representational hypotheses: The approach

### 2.3.1 The representational hypotheses

I consider three representational hypotheses that could underlie speakers' adjective ordering preferences. The first two correspond to the two potential adult representations discussed in Section 2.2: representations based on (i) hierarchically-ordered adjective **lexical classes** or

(ii) perceived **subjectivity** of adjectives. Both hypotheses require speakers to create some abstraction across individual adjective lexical items (i.e., in terms of lexical class or perceived subjectivity), and then order adjectives with respect to this abstraction. In contrast, the third representational hypothesis I consider is a simpler lexical-item-based approach, and does not require additional abstraction. This hypothesis states that speakers track the **positional frequency** of adjectives appearing in certain positions in multi-adjective strings, and their productions mirror the frequencies observed in the input. In particular, for each adjective, speakers would pay attention to how often it appears in the ***1-away*** position closest to the noun vs. the ***2-away*** position farther from the noun (e.g., $small_{2-away}$ $grey_{1-away}$ *kitten*). This positional frequency approach corresponds to the null hypothesis discussed in Section 2.1, and serves as one of the simplest approaches to adjective ordering preferences once you go beyond repeating whole chunks of strings (i.e., *small+grey* as a unit). Tracking this kind of positional frequency information would require some kind of statistical learning ability, which we already have evidence that children have and apply to other learning tasks where they can track distributions in their input (e.g., Saffran et al. 1996, Maye et al. 2002, Gerken 2006, Mintz 2006, Xu and Tenenbaum 2007, Maye et al. 2008, Smith and Yu 2008, Dewar and Xu 2010, Feldman et al. 2013, Gerken and Knight 2015, Gerken and Quam 2017, among others).

Not only are children likely to be able to track statistical distributions in their input, but it may be a less costly strategy for children, meaning it uses less cognitive resources, than a strategy that requires abstraction. In particular, it may be less costly to use a positional-frequency-based representation when a learner isn't explicitly forced to use something more sophisticated. This situation could arise, for instance, if learners were to encounter a novel adjective they need to make a semantic judgement about. For example, if a person heard a sentence *I just heard a lipidub story*, they would need to make a judgement about the meaning of the new adjective *lipidub*, and consider how subjective it is to both the speaker and themselves. Is *lipidub* a word like *fantastic*? If so, you may agree with the speaker or

not, depending on your taste. Or, is it a word like *French*, where they'd either be right or wrong? This process of considering perceived subjectivity certainly would require some cognitive energy, and so be more cognitively taxing than a strategy just based on positional frequency. But, for a novel adjective like *lipidub*, this process would be necessary. In contrast, for familiar words (like *fantastic* or *French*), a less-cognitively-taxing representation could be relied on.

### 2.3.2 Corpus data

I investigate three different types of interactions and three different "speaker" types (where I need to know the input directed towards the speaker and the output produced by the speaker): (I) a child speaker type, where the speaker output is child-produced data and the speaker input is child-directed data; (II) a caretaker speaker type, where the speaker's output is child-directed data and the input is adult-directed data; and (III) an adult speaker type, where the speaker's output is adult-directed data and the input in turn is adult-directed data. For the three speech interactions I'm modeling, I utilize different combinations of the corpus data shown in Table 2.2 below. To identify the representations underlying the development of adjective ordering preferences, I assess naturalistic child input in the form of child-directed speech (CDS), naturalistic child output in the form of child-produced speech (CPS), as well as adult-directed speech (ADS). For the Child learner speaker type, I use child-directed speech (CDS) as the input, and evaluate the underlying representation on CPS. For the Caretaker speaker type, I use ADS as the input and evaluate on CDS. For the Adult speaker type, the input and the output are both ADS.

The child-directed and child-produced data came from the CHILDES database [MacWhinney, 2000b]. I focus on the morphologically-annotated corpora in the North American datasets for children between the ages of two and four, yielding 688,428 child-directed and

1,069,406 child-produced utterances. The strings used for analysis were taken from Bar-Sever et al. [2018]. Adult-directed data came from the Penn Treebank subset of the Switchboard (SWBD) corpus of telephone dialogues with 15744 utterances (Godfrey, Holliman, & McDaniel, 1992), as well as from the spoken and the written portions of the British National Corpus (BNC, see http://www.natcorp.ox.ac.uk/). The BNC-Spoken corpus had 201,261 utterances total and BNC-Written had 89630 utterances total.

Table 2.2: The delineation of input and output data for each speaker type.

| **Speaker type** | **Input** | **Output** |
|---|---|---|
| Child learner | CDS from CHILDES | CPS from CHILDES |
| Caretaker | ADS from SWDB, BNCW, BNCS | CDS from CHILDES |
| Adult speaker | ADS from SWDB, BNCW, BNCS | ADS from SWDB, BNCW, BNCS |

After extracting all instances of adjective-adjective-noun (**AdjAdjN**) strings, like *wonderful calm capybaras*, I arrived at the counts in Table 2.3 and Table 2.4.[1]

| | **Child-directed data** | | | **Child-produced data** | | |
|---|---|---|---|---|---|---|
| **age** | **# AdjAdjN** | **# tokens** | **# types** | **# AdjAdjN** | **# tokens** | **# types** |
| 2 | 1,440 | 2,880 | 131 | 466 | 932 | 79 |
| 3 | 881 | 1,762 | 128 | 274 | 584 | 72 |
| 4 | 745 | 1,490 | 124 | 235 | 470 | 81 |

Table 2.3: Number of AdjAdjN strings and both the adjective tokens and adjective types comprising these strings per age in the morphologically-tagged North American CHILDES corpora.

| **Adult-directed data** | | | |
|---|---|---|---|
| **corpus** | **# AdjAdjN** | **# tokens** | **# types** |
| SWDB | 559 | 1,252 | 412 |
| BNCW | 9,027 | 19,948 | 2,603 |
| BNCS | 5,346 | 10,692 | 1,200 |
| **Total** | 14,932 | 31,892 | 4,215 |

Table 2.4: Number of AdjAdjN strings and both the adjective tokens and adjective types comprising these strings per age in the morphologically-tagged Switchboard Treebank Corpus, the British National Corpus-Written, and the British National Corpus-Spoken).

[1]Because there is no natural split in the adult-directed data between input and output, I performed the evaluation discussied later on on a 90/10 split of the corpora, with 90% of the corpora serving as the input data and 10% as the output data, which I repeated 10,000 times to get a reasonable estimate.

### 2.3.3 Empirical grounding of the representational hypotheses

Each potential representation requires certain information to be known about an adjective: lexical class, perceived subjectivity, or positional frequency. For lexical class, I utilized the assignments from previous work in Bar-Sever et al. [2018], which in turn were based on the 13 lexical classes and adjective assignments reported in Scontras et al. [2017]. These assignments were derived from a synthesis of previous literature [Dixon, 1982, Sproat and Shih, 1991]. I inferred a hierarchical ordering of these classes on the basis of the behavioral data reported by Scontras et al.[2].

| **Sample adjectives and classes** | | | | | |
|---|---|---|---|---|---|
| AGE | COLOR | MATERIAL | VALUE | SHAPE | SPEED |
| young | yellow | wooden | awful | round | quick |
| ripe | brown | plastic | brilliant | oval | instant |
| new | blue | iron | fantastic | square | slow |
| old | golden | wool | awesome | squiggly | fast |
| ancient | pink | silk | crummy | circular | speedy |
| DIMENSION | PHYSICAL | LOCATION | NATIONALITY | HUMAN | TEMPORAL |
| tubby | dry | far | Chinese | sleepy | past |
| fat | sharp | western | French | sorry | next |
| narrow | hard | front | Mexican | brave | late |
| flat | rough | south | Dutch | angry | early |
| wee | damp | upstairs | European | clever | recent |

Table 2.5: All of the lexical classes used in analysis with samples of the words assigned to those classes.

If an adjective had no lexical class entry in Scontras et al. [2017] or Bar-Sever et al. [2018], I attempted to analogize it to an existing entry based on similar meaning (e.g., *teeny* is similar in meaning to *small* and so was assigned to the DIMENSION class). If there was no clear analogy to an existing entry (e.g., *ripe*), I manually assigned it to a lexical class via collective agreement by undergraduate researchers. Some of the adjectives wound up in the X "elsewhere" class as defined in Scontras et al.; these adjectives did not neatly fit into any of the other class categories. Because the elsewhere class is so heterogeneous, its adjectives

[2]A full list of the lexical classes and their assigned adjectives are available on my GitHub `https://github.com/galiabarsever/dissertation_files`

fail to cohere on the basis of meaning. As a result, this collection of adjectives does not stand as a lexical *semantic* class, so is unlikely that there is a clear conclusion about whether it is overall "closer" or "farther" from a noun than another class, nor can a larger semantic meaning be drawn from its contents (some words in this category include *obvious, different,* and *roundabout*). Therefore, I excluded its adjectives from the representational analyses described below. Other exclusions are described in Appendix A.

For positional frequency, I derived both 1-away and 2-away frequencies from the input data's AdjAdjN strings (ex: *nice old dog*, *happy little mice*, etc.). I then calculate how often an adjective appeared in the 1-away vs. 2-away position in the input. A subjectivity score between 0 and 1 (0 being not subjective, 1 being maximally subjective) was assigned to each adjective, based on the mean of participants' judgements on the subjectivity of an adjective [Scontras et al., 2017]. Subjectivity scores were considered the same if they were within +-0.1 of each other.

I took the lexical class assignments and subjectivity scores from previous work from Bar-Sever et al. [2018], and collected additional judgements for common words present in the adult-directed corpus that were not captured in the previous work. To get perceived subjectivity for these 68 additional adjectives, I obtained subjectivity scores from 30 adult participants on Amazon.com's Mechanical Turk crowdsourcing service, replicating the methodology of Scontras et al. [2017] and Bar-Sever et al. [2018]. Participants were presented with 30 adjectives total (one at a time) in a random order and asked to indicate how "subjective" a given adjective was on a sliding scale; endpoints were labeled "completely objective" and "completely subjective." To arrive at the perceived subjectivity score for a given adjective, responses were averaged across participants. These words were additionally grouped into lexical classes in the same way as Bar-Sever et al. [2018][3].

[3]The full list of all the adjectives and associated lexical classes and subjectivity scores are available on my GitHub `https://github.com/galiabarsever/dissertation_files/`

Table 2.6: A sample of the adjectives and their associated lexical class and subjectivity assignments.

| **Adjective** | **Lexical class** | **Subjectivity** |
|---|---|---|
| *new* | AGE | 0.265 |
| *gold* | COLOR | 0.214 |
| *giant* | DIMENSION | 0.622 |
| *brave* | HUMAN | 0.702 |
| *iron* | MATERIAL | 0.1 |

### 2.3.4 Quantitatively linking input to output

Recall that producing an AdjAdjN string requires transforming the input according to the underlying knowledge representation and using that representation to generate the AdjAdjN string. For each representational hypothesis, I can define how this process would occur, thereby linking the AdjAdjN input to AdjAdjN output. I focus on how a given representational hypothesis would generate an adjective in the 2-away vs. the 1-away position when combined with another adjective in an AdjAdjN string.

I consider the collection of AdjAdjN output as a dataset $D$ that is produced according to any of the three representational hypotheses $h_i \in H$, where $H = \{h_{lex}, h_{subj}, h_{pos}\}$. I select the hypothesis that is most likely to have generated the data in $D$ by calculating the likelihood of a given hypothesis $h$ generating the data $D$, $p(D|h)$. The representational hypothesis with the largest probability of generating $D$ (i.e., the highest likelihood) is the hypothesis that best matches the output.

I can conceive of $D$ as the set of AdjAdjN strings involving different combinations of all the adjectives *Adj* observed in the corpus. For example, $D$ might be the set {*grey furry kitten*, *small grey kitten*, *small grey kitten*, *small furry kitten*}, where *Adj* is {*grey*, *furry*, *small*}. To account for the portion of the AdjAdjN strings involving a particular adjective $adj_x \in Adj$, I can calculate the likelihood of the data involving that adjective, $p(D_{adj_x}|h)$. So if we're concerned first with the adjective *small*, our set of AdjAdjN strings that included *small* would be the set {*small grey kitten*, *small grey kitten*, *small furry kitten*}; in this example,

*small* occurs in the 2-away position with probability 1.0. The *grey* strings would form the set $D_{grey}$: {*grey furry kitten*, *small grey kitten*, *small grey kitten*}; here, *grey* occurs in the 2-away position with probability 0.33. I then multiply these individual adjective likelihoods to yield the likelihood for the whole dataset $D$ under that hypothesis, as shown in equation (2.3).

$$p(D|h_i) = \prod_{adj_x \in Adj} p(D_{adj_x}|h_i) \tag{2.3}$$

I define the likelihood for an individual adjective $adj_x$ for a given hypothesis $h_i$ as in equation (2.4), which considers the number of times $N$ that $adj_x$ appeared in an AdjAdjN string in the output, the number of times $t$ that $adj_x$ appeared in the 2-away position, and the probability that $adj_x$ would appear in the 2-away position given the representational hypothesis $h_i$, $p_2 exp(adj_x|h_i)$.

$$p(D_{adj_x}|h_i) = \binom{N}{t}(p_2 exp(adj_x|h_i))^t(1 - p_2 exp(adj_x|h_i))^{N-t} \tag{2.4}$$

To see how this equation works, consider $D_{grey}$ from above: {*grey furry kitten*, *small grey kitten*, *small grey kitten*}. Suppose a given representational hypothesis $h_i$ predicts that *grey* should appear in the 2-away position with a certain probability $p_2 exp(adj_x|h_i)$. The intuition is that I compare this expected probability with the actual frequency of *grey* occurring in the 2-away position to calculate the likelihood of $D_{grey}$ under $h_i$, $p(D_{adj_x}|h_i)$; if the expected probability matches the actual frequency, the hypothesis does an excellent job of accounting for the child output.

To calculate the likelihood, I need to determine the number of ways of generating the pattern in $D_{grey}$ (i.e., *grey* in the 2-away position twice and in the 1-away position once). This corresponds to $\binom{N}{t}$, the number of ways of generating $N$ AdjAdjN strings with *grey* in the 2-away position $t$ times. So, there are $\binom{3}{2} = 3$ ways of generating three AdjAdjN

strings with this pattern (i.e., *grey* in the 2-away position twice and in the 1-away position once). Having determined the number of ways to generate the observed pattern, I then calculate the probability of generating the observed pattern given a specific representational hypothesis $h_i$. I first need to calculate the probability that *grey* would appear in the 2-away position two times, $(p_2 exp(adj_x|h_i))^t$. So, *grey* would be in the 2-away position two of three times ($(p_2 exp(adj_x|h_i))^t = 0.75^2 = 0.5625$) and *grey* in the 1-away position one of three times ($(1 - p_2 exp(adj_x|h_i))^{N-t} = (1 - 0.75)^{3-2} = 0.25$); the probability of this pattern is $0.5625 * 0.25 = 0.14$. To capture the full pattern, I also need to calculate the probability that *grey* would appear in the 1-away position once, $(p_2 exp(adj_x|h_i))^{N-t}$. By multiplying the probability of generating the observed pattern together with the number of ways I could have generated it, I arrive at the likelihood in equation (2.4). So, I multiply the probability of this pattern with the number of ways of generating it to yield the likelihood, $p(D_{grey}|h_i) = 3 * 0.14 = 0.42$.

The calculation of $p_2 exp(adj_x|h_i)$, the probability that a particular adjective $adj_x$ will appear in the 2-away position, depends on the hypothesis $h_i$ under consideration, as well as the input the listener has encountered in their input. For both the lexical class hypothesis $h_{lex}$ and the subjectivity hypothesis $h_{subj}$, the probability that $adj_x$ surfaces in the 2-away position in an AdjAdjN string depends on the kind of adjective it appears with. For $h_{lex}$, if $adj_x$ is combined with an adjective in a hierarchically-closer lexical semantic class, it should surface in the 2-away position 100% of the time ($p = 1.0$); if $adj_x$ is combined with an adjective in the same lexical class, it should surface in the 2-away position with chance probability ($p = 0.5$). For $h_{subj}$, if $adj_x$ is combined with an adjective perceived as less subjective, it should surface in the 2-away position 100% of the time ($p = 1.0$); if $adj_x$ is combined with an adjective perceived as equally subjective, it should surface in the 2-away position with chance probability ($p = 0.5$).[4] These considerations represent the numerator in equation

[4] I considered two adjectives to be perceived as equally subjective if their subjectivity scores were within 0.1 of each other; scores ranged from 0 to 1.

(2.5).

$$p_2exp(adj_x|h_i \in \{h_{lex}, h_{subj}\}) = \frac{f_{input}(< adj_x|h_i) + 0.5 * f_{input}(= adj_x|h_i) + \alpha}{N_{input}(Adj) + \alpha * |Adj|} \quad (2.5)$$

In particular, $f_{input}(< adj_x|h_i)$ represents the number of adjective tokens in the input that are either from a lexically-closer class than $adj_x$ (given $h_{lex}$) or are less subjective than $adj_x$ (given $h_{subj}$); the larger this number, the more I would expect the speaker to produce $adj_x$ in the 2-away position under the relevant hypothesis. Similarly, $f_{input}(= adj_x|h_i)$ represents the number of adjectives that are from the same lexical class as $adj_x$ ($h_{lex}$) or are equally subjective as $adj_x$ ($h_{subj}$); this number gets multiplied by 0.5 to represent the chance probability that $adj_x$ would appear 2-away with adjectives of the same kind. I arrive at the probability of $adj_x$ appearing in 2-away position once I divide these counts by the total number of adjective tokens appearing in AdjAdjN strings in the input, $N_{input}(Adj)$. Both the numerator and the denominator of equation (2.5) contain the smoothing factor $\alpha = 0.5$, which is added to handle adjectives for which there are no observations; in the denominator, $\alpha$ is multiplied by the number of adjective types $|Adj|$. To implement the idea that the target adjective $adj_x$ cannot combine with tokens of itself (e.g., *small small kitten*), the number of $adj_x$ tokens is subtracted from the counts of how many adjectives either are in the same lexical class or have the same subjectivity score in the numerator; this number is also subtracted from the total adjective token count in the denominator. If we take the number of adjectives tokens for a particular adjective under consideration as $n_{adj}$, then the full calculation for $p_2exp$ corresponds to Equation 2.6.

$$p_2exp(adj_x|h_i \in \{h_{lex}, h_{subj}\}) = \frac{f_{input}(< adj_x|h_i) + 0.5 * (f_{input}(= adj_x|h_i) - n_{adj}) + \alpha}{(N_{input}(Adj) - n_{adj}) + \alpha * |Adj|} \quad (2.6)$$

A different calculation is used for $p_2exp$ for the positional frequency representational hypothesis $h_{pos}$, as shown in equation (2.7). The probability of $adj_x$ appearing in the 2-away position given $h_{pos}$ is a simple reflection of how often it appeared in the 2-away position in the input ($f_{2input}(adj_x)$) divided by the total number AdjAdjN strings in which $adj_x$ appeared at all ($N_{input}(adj_x)$). Again, I add the smoothing factor $\alpha$ to avoid assigning zero probability for adjectives not observed; in the denominator, $\alpha$ gets multiplied by 2, corresponding to the two positional options for $adj_x$: 2-away vs. 1-away.

$$p_2exp(adj_x|h_i = h_{pos}) = \frac{f_{2input}(adj_x) + \alpha}{N_{input}(adj_x) + 2 * \alpha} \quad (2.7)$$

Using equations (2.3)-(2.7), I can evaluate how probable it is that speakers would have produced the AdjAdjN strings in their output given the input they heard and a particular representational hypothesis: lexical class, subjectivity, and positional frequency.[5]

## 2.4 Results & Discussion

### 2.4.1 Child speaker type

Log-likelihood is shown below for the three representational hypotheses at 2, 3, and four-years-old using child-directed input and child-produced output (see Figures 2.1, 2.2, 2.3).

[5] I only included AdjAdjN strings in both the input and output sets where both adjectives in the string had been assigned a lexical class and a subjectivity score.

The log of the likelihood was taken to avoid dealing with very small numbers with multiplying very small probabilities together. More probable log-likelihood scores are closer to zero (less negative) than less probable scores (more negative).

Confidence intervals (CIs) allow us to make more reasoned judgements about the results of our analyses, and give a sense of a result's variance. For this and subsequent analyses, I reported 95% confidence intervals. To get these intervals, I drew the individual AdjAdjN strings (like *small grey kitten*) from the input or output with replacement. This resampling and the likelihood analysis was done 10,000 times.

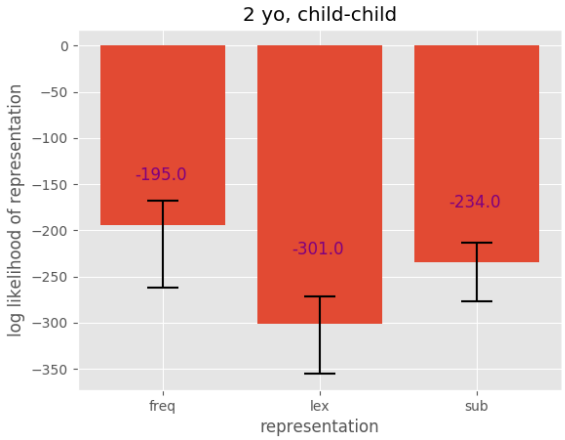



Figure 2.1: Two-year old child speaker type, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

Looking at the means in Figure 2.1 (positional frequency: -195.0, lexical: -301.0, and subjectivity: -234.0), it would appear that two-year-olds are utilizing a positional frequency representation over the other more abstract representations. However, the CIs for the positional frequency hypothesis and the subjectivity hypothesis appear to overlap. What we can certainly tell at two-years-old is that children are unlikely to be using a lexical representation.

At three and four, there seems to be a messier picture. The means for each of the hypotheses at three-years-old are positional frequency: -119.0, lexical: -143.0, and subjectivity: -132.0. The means at four-years-old are positional frequency: -175.0, lexical: -151.0, and subjectivity: -178.0. It seems suggestive that a more abstract representation (lexical or

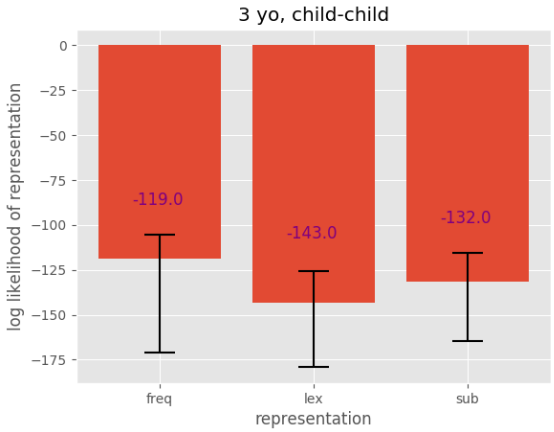



Figure 2.2: Three-year old child-speaker type, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

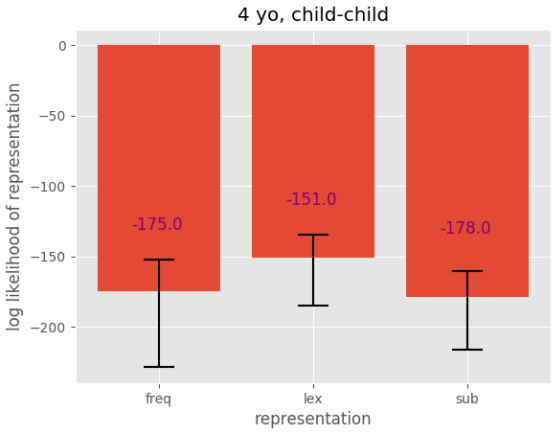



Figure 2.3: Four-year old child-speaker type, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

subjectivity-based) may be used at four.

From Bar-Sever et al. [2018], the quantitative assessment done without CIs suggested that the development of adjective ordering preferences demonstrates that abstract knowledge is likely to underlie children's preferences at age four (but not earlier); moreover, these means were interpreted to indicate that this abstract knowledge is lexical-class-based rather than subjectivity-based, with children initially tracking the word-level statistics of their input when determining adjective ordering. By age four they would shift to a more abstract (and compact) representation based on lexical semantic class.

Looking at the means here, we do see the same trajectory (means at two: **positional frequency = -195.0** , lexical class = -301.0, subjectivity = -234.0, means at three: **positional frequency = -119.0**, lexical class = -143.0, subjectivity -132.0, means at four: positional frequency = -175.0 , **lexical class = -151.0**, subjectivity = -178.0). However, taking a more nuanced look at these findings using 95% CIs makes this interpretation less obviously the only one that's possible. While we can reasonably assume that children are probably utilizing the positional frequency strategy mental representation at two-years-old, at three and four-years-old, any of the three representations is possible, possibly due to competing

representations that are developing at three- and four-years old. It appears we may know something about the mental representations at two (whether children may be using positional frequency or subjectivity, but not lexical class), but it is unclear what mental representation children are using at three and four.

## 2.5 Caretaker speaker type

Here I present the results for the representations adults are using to form the input children receive at ages two, three, and four. The likelihoods in this case are calculated using the adult-directed data as the adult input (from SWDB, BNCW, and BNCS) and the child-directed data from CHILDES at ages two, three, and four is used as the adult output data. I report 95% confidence intervals and utilized the same resampling process as in the child speaker type interaction.

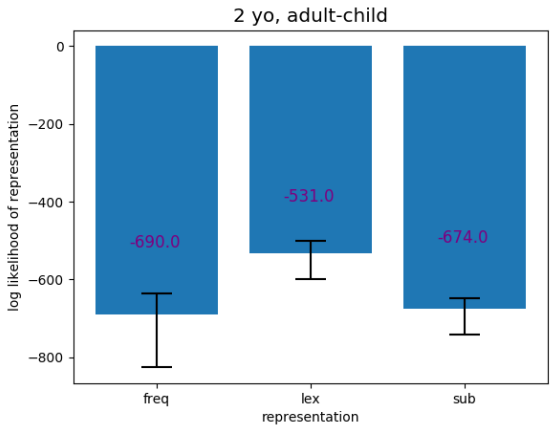



Figure 2.4: Adults talking to two-year old children, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

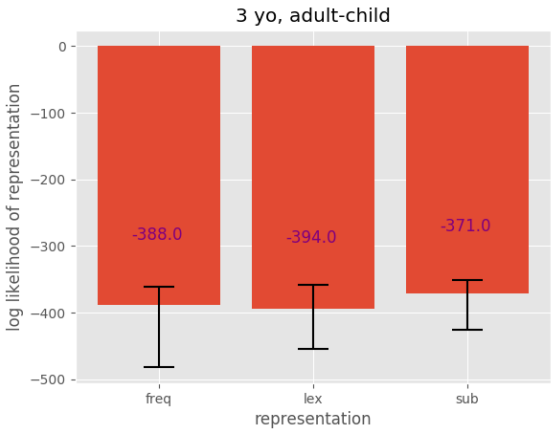



Figure 2.5: Adults talking to three-year old children, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

I find that just looking at the means, when talking to children two-years-old, adults appear to be utilizing a lexical class representation (means at two: positional frequency = -690.0 ,

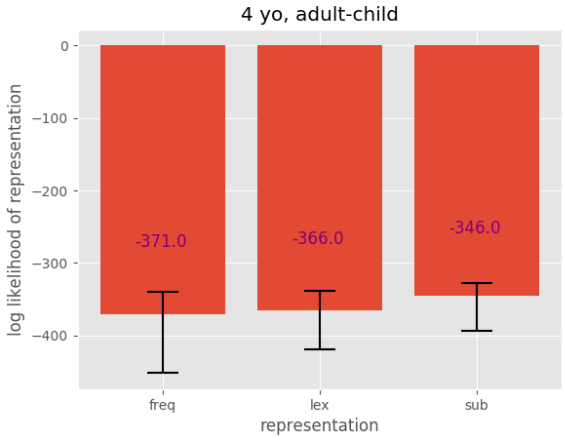



Figure 2.6: Adults talking to four-year old children, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

**lexical class = -531.0**, subjectivity = -674.0). For talking to children who are three-years-old, adults switch to a subjectivity-based hypothesis (means at three: positional frequency = -388.0, lexical class = -394.0, **subjectivity -371.0**) which they continue to use at four (means at four: positional frequency = -371.0 , lexical class = -366.0, **subjectivity = -346.0**).

However, again the CIs suggest that at three and four (as with the child speaker type), it is unclear exactly what mental representation is being used; all representations appear to be equally compatible (see Figures 2.4, 2.5, 2.6).

So, it appears that adults are adjusting their child-directed speech when speaking to children at age two from what we might expect the representation is that they use for adult-directed speech (subjectivity). However, when speaking to children age three and four, it is unclear which representation adults are using. Still, it's interesting that the most likely mental representation that two-year-olds might be using, positional frequency, does not seem to be the mental representation that adults are using to talk to two-year-olds (lexical class). Neither is subjectivity, which is the proposed adult representation.

### 2.5.1 Adult speaker type

Both this analysis and that of Scontras et al. [2017] are trying to get at the same question: what underlying representation is responsible for adults' adjective ordering preferences? What makes this analysis fundamentally different is the way I compare the different representations. Scontras et al. pitted subjectivity against other abstract groupings of adjectives (not just lexical class), and I compare both subjectivity and lexical class additionally against the positional frequency representation, which was not a representation considered by Scontras et al. [2017].

The input and the output in this case comes from the adult-directed corpora of the SWDB, BNCW, and BNCS, with the means and variance calculated from resampling the input (90% of the adult-directed data) and resampling the output (10% of the adult-directed data) 10,000 times.

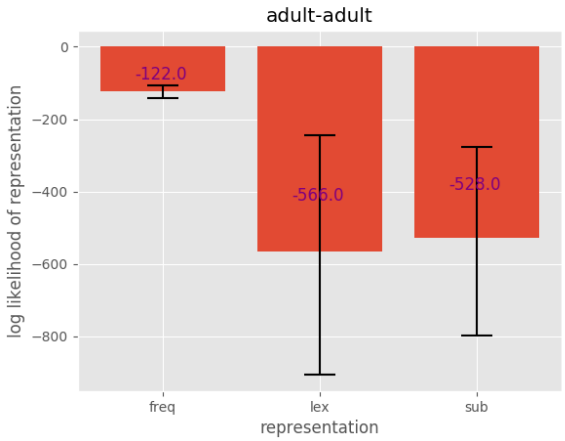



Figure 2.7: Adults talking to adults, where the Y-axis is log-probability and the X-axis is each representation (positional frequency, lexical class, and subjectivity).

I find that adults appear to be overwhelmingly utilizing the positional frequency hypothesis over both the lexical and subjectivity representations, (considering both means and CIs) – see Figure 2.7. In particular, for this speaker type, the means are **positional frequency = -122.0**, lexical class = -565.0, subjectivity = -528.0.

What's striking is how close subjectivity and lexical class are when it comes to data likelihood, while positional frequency is clearly a better fit. Why should this be? It could be that the positional frequency strategy performs so strongly because it demonstrates a better fit for words that occur only in one position. For example, if a particular adjective is always either in the 1-away or 2-away position according to the input (i.e., in a multi-adjective string like *little dog*, the word *little* always appears 2-away from a noun, never 1-away), then a soft score (like a subjectivity score of 0.7 for *little* instead of 0 or 1) will not be as strong of an explanation, and therefore that representational hypothesis will not be as strongly favored. This may also be happening when we consider lexical class as well, although perhaps less often because we have multiple words belonging to a given lexical class. But, it could be the case that an adjective (say *grey*) is the only COLOR adjective a child encounters. If someone only ever hears *cute grey kitten*, *nice grey kitten*, and *happy grey kitten*, where *grey* is only ever in one position, this could bias them away from that lexical class hypothesis because we would again see an extreme positional frequency in the data, but the lexical class hypothesis would soften the score, depending on what other lexical classes the adjective appears in combination with[6].

Given that subjectivity appears to so robustly explain the corpus data Scontras et al. [2017], it seems surprising that the positional frequency performs so much better. Importantly though, Scontras et al. (2017) didn't compare the positional frequency hypothesis to subjectivity and lexical class. So perhaps if they did, it would explain their data even more robustly.

If positional frequency is indeed a better fit for adult speakers, one idea why this should be is due to the function of human processing. It could be that adults who would have already developed robust preferences, perhaps based on subjectivity, are no longer calculating relative

[6]Full tables with each evaluated output adjective and it's associated positional frequency, lexical class, subjectivity, and associated scores as well as input and output strings are available on my GitHub `https://github.com/galiabarsever/dissertation_files/`

subjectivity online while producing multi-adjective strings, and instead are using a strategy that allows them to offload a relatively computationally expensive process in favor of just tracking statistics of the relative positions of the adjectives in the multi-adjective string itself. Calculating subjectivity online requires sophisticated reasoning about theory of mind and other people's beliefs (i.e., as mentioned previously, it's cognitively taxing). So, adults might not rely on it if they don't have to.

Interestingly, humans have been shown to utilize these kind of "cheap" cognitive solutions in other domains of language comprehension (Ferreira et al. [2002]). Ferreira et al. [2002]'s idea is that while using language, adults may not be accessing the semantic content or the "true meaning" of language, instead relying on computationally inexpensive heuristics to process language. This kind of mental off-loading could be an explanation as to why positional frequency could out-explain the other abstract representations for adults talking to adults. In particular, positional frequency might be cheaper to access in the moment than perceived subjectivity.

## 2.6 General discussion

What is most striking about the three different explorations of mental representations, both the development of them in children and their use in both child-directed speech and adult-directed speech, is the story that emerges when I look beyond the means to a more nuanced analysis that includes confidence intervals.

For the three speaker types, I found that in general the representations that children are using at certain ages to generate their AdjAdjN strings do not match the representations that adults are using to direct speech towards children of that age. While it is unclear what mental representation children or adults are using at three and four, there is a marked

difference in the mental representation used by adults to form AdjAdjN strings directed to two-year-olds (lexical class) and the representation that two-year-olds seem to be using to form their own AdjAdjN strings (positional frequency or subjectivity). The child speaker types, which have more viable potential representations, are also markedly different from the adult speaker type, where positional frequency is so strongly favored. Further, the representations that adults seem to be using to talk to two-year-olds are more abstract representations than they use to talk to other adults, where they appear to be utilizing a positional frequency representation. However, I do note that the large CIs we see in the child speaker and caretaker speaker results may be the result of a potential data sparsity issue, which could be remedied by access to more available data.

**Child speaker type.** For the child speaker type, just looking at the means would show a clear progression to a more abstract representation, developing from positional frequency at two-years-old to a lexical class representation at four. However, with the addition of confidence intervals, we can now see much less certainty about how these abstract representations develop. We do know that at two, children seem to be using either a positional frequency representation or a subjectivity-based one, but not a lexical-class-based one.

**Caretaker speaker type.** When looking at the mental representations of the caretaker speaker type, we see an interesting split between the representations that children appear to be considering, given those that adults are using at certain ages. While for three- and four-years-old, there seems to be a similar pattern of uncertainty about the mental representation being used, at two-years-old there is a definite difference between the utilized representation. Two-year-olds appeared to be utilizing a positional frequency representation when producing multi-adjective strings, and adults appear to be utilizing a lexical class representation when producing child-directed strings. It is unclear why this should be different both from the child-produced representation, and from the adult-to-adult representation. It may be that

lexical classes are abstract (and so allow productivity for new adjectives) but aren't as computationally expensive to calculate as subjectivity. Another possibility is that adults are scaffolding development of more abstract representations that are just beyond the child's capabilities (see Chaiklin [2003]). In particular, because lexical class is more abstract than positional frequency, but not as abstract as subjectivity, caretakers are demonstrating an "easier" kind of abstraction. It is possible that for older children who may then be using a more abstract representation like lexical class, the caretakers would further model the next step on the path to the adult state (e.g., subjectivity-based abstraction).

**Adult speaker type.** The other striking element to these findings is in how adults appear to be accessing and forming their multi-adjective strings when talking to each other. I assume that adults have access to a robust abstract mental representation, as demonstrated in experiments on adult notions of adjective subjectivity [Hahn et al., 2017, Foushee and Srinivasan, 2017, 2018], but for this speaker type they appear not to be using it. Adults who would have already developed robust preferences, perhaps based on subjectivity, might no longer be calculating relative subjectivity online while producing multi-adjective strings, and instead are using a strategy that allows them to offload a relatively computationally expensive process in favor of just tracking statistics of the relative positions of the adjectives in the multi-adjective string itself. Another idea relates to the origin of adjective ordering preferences: it is possible that there is an evolutionary benefit to have adjectives ordered according to subjectivity, as it provided a communicative benefit [Franke et al., to appear]. However, while subjectivity may be the results of an evolutionary pressure, it might not be the best explanation for how adults now mentally represent their adjective ordering preferences.

## 2.7 Future work

**Child speaker type.** It remains unclear when (or whether) subjectivity replaces lexical classes as the underlying representation for adjective ordering preferences—this timing no doubt depends on children's development of the conceptual underpinnings of subjectivity, which occurs remarkably late [Foushee and Srinivasan, 2017]. Future work would look at slightly older children, who perhaps have developed a sense of subjectivity, to see when children are firmly landing on an abstract representation for their adjective ordering preferences or if there are certain situations (like novel adjectives) that will force them to rely on more abstract representations.

Future behavioral work can assess children's perceived subjectivity of adjectives at different ages. The subjectivity scores used in our assessment were derived from adult judgments, but children's estimates are likely to differ, given the sophisticated theory of mind involved in evaluating subjectivity. Whether this potential difference in how children consider what is "subjective" affects our understanding of children's productive knowledge of adjective ordering remains an open question. Moreover, it could turn out that adult-like subjectivity awareness develops later than stable adjective ordering preferences. This could mean that adjective ordering could provide clues to children and bootstrap their subsequent subjectivity-based preferences from the adjective ordering they encounter, rather than using subjectivity to figure out the adjective orderings in the first place.

**Caretaker speaker type.** Future work here would involve taking a closer look at the differences between the child-directed speech and adult-directed speech. What makes the speech directed at two-year-olds different than the speech directed to three- and four-year-olds? It could be that there are different kinds of adjectives being used at these ages, perhaps in different contexts. Certainly at older ages children are producing more multi-word

utterances, and it is possible that this rise in sophistication is matched by an adjustment in the speech directed to the child at a particular age. It would be useful to see what conditions promote different behavior in adults that change the representation they appear to be using to produce multi-adjective strings. It may be that adults are using a representation that looks more like how they speak to adults as they are directing their speech to older children. We could see when this happens by looking at analyses of the caretaker speaker type with child-directed speech to older children.

**Adult speaker type.** It may be that whatever forces children to utilize an abstract representation may also force adults to use one. It would be useful to see what conditions might force an adult to access an abstract representation (like subjectivity) over a positional frequency heuristic. This might involve behavioral studies like Hahn et al. [2017]'s, which forced subjectivity judgements for novel adjectives that adults would not already have positional frequency information for.

# Chapter 3

# Frequent Frames: The utility of early syntactic categories in child language

## 3.1 Introduction

Language acquisition happens in stages. This means that early language acquisition strategies probably don't yield adult knowledge right away. Instead, they're more likely to provide transitory representations that scaffold the acquisition of later knowledge [Frank et al., 2009, Connor et al., 2010, 2013, Gutman et al., 2014, Phillips and Pearl, 2015]. For example, syntactic categories that twelve-month-olds have may not look like adult NOUN, ADJECTIVE, and VERB categories. However, if children's developing knowledge representations aren't adult-like, what *do* they look like?

Being able to answer this question depends on the relationship between what's going on in the child's mind and what we see in the world. We can see the reflections of the underlying representations in the observable linguistic behavior (both comprehension and production). This means that given a set of candidate hypotheses about potential underlying representa-

tions, and a child's input, we can assess which potential representation best predicts a child's output, which is the same approach I took in the previous chapter looking at adjective ordering preferences.

In this chapter, I consider different candidates for developing knowledge of syntactic categories that might be used by children to produce their own utterances. First, I discuss the potential hypotheses and underlying representations, as well as around what age representations are likely to be either immature or adult-like. Second, I will talk about Frequent Frames (**FFs**) as a potential source of immature category representations, and discuss the relevant child input and output data. Third, I discuss how to link these potential immature representations to the child's input and output data, which will involve constructing these potential immature representations from the child's input and then assessing the probability of the child's output given the particular representation. Finally, I discuss the results, which indicate which representations best match the output. I find that certain immature representations better predict the children's output than mature categories. I conclude with possible future directions.

## 3.2 Emerging syntactic category knowledge

### 3.2.1 Syntactic categories are useful

In essence, syntactic categories are clusters of individual lexical items that function similarly syntactically. For example, the finer-grained adult NOUN$_{count}$ category includes lexical items like *kitty, penguin*, and *idea*, and each of these can be preceded by a DETERMINER like *the* or *a(n)* and used to create a NOUN PHRASE that can serve as the subject of a sentence. So, one purpose of syntactic categories is to more compactly represent the syntactic patterns of the language (i.e., a single rule NP $\rightarrow$ DETERMINER NOUN$_{count}$ will suffice, instead of multiple

rules like NP → *the kitty*, NP → *a penguin*, etc.).

If language users recognize that individual words are instances of a larger coherent category, it becomes easier to predict the underlying structure of the language input encountered, as implemented by the language's syntax. This is because the structural commonality across different utterances is more readily apparent (e.g., *the kitty is cute* and *a penguin is adorable* are both examples of DETERMINER $\text{NOUN}_{count}$ COPULA ADJECTIVE). Given this, syntactic categories seem like a useful abstract representation to learn.

### 3.2.2 Children's syntactic category knowledge

There hasn't been a clear consensus for when children develop syntactic categories, whether open-class (e.g., NOUN, VERB, ADJECTIVE), or closed-class (e.g., DETERMINER, AUXILIARY). Some studies suggest that knowledge of certain categories – either rudimentary or adult-like – may be in place as early as age two [Pinker, 1984, Valian, 1986, Capdevila i Batet and Llinàs i Grau, 1995, Booth and Waxman, 2003, Rowland and Theakston, 2009, Theakston and Rowland, 2009, Yang, 2010, 2011, Shin, 2012, Meylan et al., 2017, Bates et al., 2018], while others argue that such knowledge only emerges much later [Pine and Lieven, 1997, Tomasello, 2004, Kemp et al., 2005, Tomasello and Brandt, 2009, Theakston et al., 2015]. For example, Booth and Waxman [2003] show that children as early as 14 months old may have rudimentary open-class categories like NOUN and ADJECTIVE that encompass subsets of nouns and adjectives, respectively. Bates et al. [2018] showed evidence for at least one closed-class adult-like category as early as two years (AUXILIARY and possibly NEGATION). More generally, there seems to be some agreement that children may have rudimentary knowledge of open-class categories (e.g., NOUN, ADJECTIVE) fairly early, but don't refine these into adult-like open-class categories until later. However, for closed-class categories (e.g., DETERMINER, NEGATION, AUXILIARY), there isn't yet consensus on when

either rudimentary or adult-like versions of these categories develop. Given this, if we're interested in the syntactic category representations that young children have (e.g., around two-years-old), it seems likely that these representations are of immature syntactic categories; later on, these early representations will be refined into adult syntactic categories.

### 3.2.3 A possible strategy for immature representations: Frequent Frames

What do these immature syntactic categories look like? For example, young toddlers might not recognize all the nouns adults would identify as $\text{NOUN}_{count}$ – instead, toddlers might realize that *kitty* and *penguin* are the same kind of thing, without recognizing that *idea* is, too.

Frequent Frames (FFs) form the basis of an early categorization strategy that is both computationally inexpensive and linguistically-based. This strategy has yielded promising results for many spoken languages with different linguistic properties (e.g, English: Mintz 2003a; Wang and Mintz 2008; French: Chemla et al. 2009b; Spanish: Weisleder and Waxman 2010; German: Stumper et al. 2011, Wang et al. 2011; Dutch: Erkelens 2009; Turkish: Wang et al. 2011; and Mandarin Chinese: Xiao et al. 2006). The basic intuition is that young toddlers pay attention to frequently occurring frames, which identify linguistic units that behave similarly in utterances (i.e., appear in the same linguistic context, as implemented by the frame). For example, in the sentences *I am petting nice kitties* and *I am hugging nice penguins*, the word-level frame *am__ nice* identifies that *petting* and *hugging* have the same linguistic context and so are the same kind of word. The morpheme-level frame *am__-ing* identifies that *pet* and *hug* are the same kind of word.

Experimental evidence for toddlers being able to construct frames like these comes from Mintz [2006]. This suggests that twelve-month-olds are sensitive to word-level frames. More

generally, twelve-month-olds can recognize the non-adjacent dependencies that frames rely on if they already know that adjacent dependencies exist between linguistic elements [Lany and Gómez, 2008].

The "frequent" part of the FFs strategy is meant to capture the intuition that young toddlers have limited attention. More specifically, something that occurs frequently is likely to be salient to toddlers, and so the FFs strategy assumes that toddlers rely on a set of frames that are frequent enough to be noticed. In particular, the intake for early categorization is a set of frequent frames and the output are clusters of linguistic elements captured by each frequent frame. In the cross-linguistic computational investigations mentioned above, these clusters have been compared against adult syntactic categories and generally found to be very precise. For example, a FF might cluster together many VERB items and exclude non-VERB items, and so be very precise with respect to the adult VERB category. This leads to categories that have very high precision, but generally low recall – this is because FFs would produce many clusters of VERBs, instead of just one. However, these small homogeneous categories (where there may be a few categories comprised only of verbs) may be more useful to early learners than larger messier categories (i.e., categories that include all of the adult verbs, but also a bunch of adverbs and prepositions) in developing intuitions about groups of words that function in the same way.

### 3.2.4 Instantiation and evaluation of FFs

There are several considerations for instantiating the FFs strategy and evaluating the FF-based categories that result. I discuss each in turn, including my modeling decisions for each one.

I tested three different implementations of FFs that have been used by other researchers working with FFs in a variety of languages (Chemla et al. [2009a], Mintz [2003b], Wang and

Mintz [2008], Weisleder and Waxman [2010], Wang et al. [2011],) where there were three different approaches to how to define "frequent". The first was utilized by Mintz [2003b]'s first experiment, where they used a strict "most frequent 45 frames" (**Top45**) based on a pilot experiment that deemed this cutoff provided "good enough" categorization. Accuracy at capturing adult-level syntactic categories with these frames in English was very high (between 90% and 93% type accuracy). Mintz [2003b] then in a second experiment selected FFs based on a relative rather than absolute threshold, where the FFs were selected relative to the total number of the frames in the corpus (specifically if they were in the top 0.13% of frames in the corpus (**Top.13**), which roughly corresponded to the same 45 frames from the first experiment). These frames were also highly accurate (between 92% and 94% type accuracy). Many cross-linguistic FFs studies have utilized one of these two frequency thresholds (Wang and Mintz [2008], Weisleder and Waxman [2010], Wang et al. [2011]). Chemla et al. [2009a] choose a frame-frequency-based metric similar to Mintz [2003b], but what they deemed to be more conservative (a frame must contain 0.5% of word types and 0.1% of word tokens to be considered frequent (**T.5T.1**)); this cutoff resulted in comparatively fewer frames (only 6 FFs) – however, these selected FFs were very precise (100% accuracy). I look at each of these major frequency thresholds when implementing FFs here, summarized below.

1. (**Top45**) The top 45 most frequent frames were counted as frequent enough [Mintz, 2003b, Wang and Mintz, 2008, Wang et al., 2011, Weisleder and Waxman, 2010].

2. (**Top.13**) The top 0.13% most frequent frames were counted as frequent [Mintz, 2003b].

3. **(T.5T.1)** A frame must contain 0.5% of of the total word types and 0.1% of the word tokens in the corpus to be considered frequent [Chemla et al., 2009a].

Each of these versions of FFs will also be considered with and without utterance boundaries as potential frame units. For example, if we don't use utterance boundaries in the sentence

*# I love cosmic snowcones #*, a frame would not include a start-of-utterance or end-of-utterance marker. So, *#_love* and *cosmic_#* would not be considered frames in this case; in contrast, if we did include utterance boundaries, those frames would be included. The original implementation of FFs did not include utterance boundaries in framing units (Mintz [2003b]). However, it is likely that children are in fact paying attention to utterance boundaries, and that the words that appear next to utterance boundaries are highly salient (Seidl and Johnson 2006, Shukla et al. 2007, Johnson et al. 2014, Longobardi et al. 2015). Because of this, Weisleder and Waxman [2010] utilized end-frames (FFs using the end of an utterance as a boundary) as well as frames generated without utterance boundaries. Weisleder and Waxman [2010] found that the 45 most frequent frames created with utterance boundaries were messier (i.e., less precise) than the frames created without utterance boundaries. However, it might still be true that, although messy, these frames could still be highly salient to young children, and perhaps still helpful in developing immature category representations. Because of these considerations, I include versions of FFs that include utterance boundaries, as well as versions that do not utilize utterance boundaries.

I identified frames that fit the above frequency criteria for a total of 6 different framing methods: T.5T.1+utt, T.5T.1-utt, Top45+utt, Top45-utt, Top.13+utt, Top.13-utt.

## 3.3 Language data

To determine which syntactic category representation best matches a child's behavioral data, we need input and output data from children at relevant developmental stages, in this case around 2 and just under two-years-old. I looked at typically-developing English-speaking children from the CHILDES database [MacWhinney, 2000b].

The first child's data I look at is the same selection of the Peter corpus that Mintz [2003b]

used in his implementation. [Bloom et al., 1974, 1975] from the CHILDES database (shown in Table 3.1), one of the corpora which has syntactic categories annotated.

Table 3.1: Corpus data for Peter

| Corpus | Age Range | Speaker | # Utts | # Tokens | # Types | MLU |
|---|---|---|---|---|---|---|
| Peter | 1;9 - 2;4 | child-directed | 18317 | 77212 | 2291 | 4.2 words |
| | | child-produced | 14790 | 37588 | 1349 | 2.5 words |

I also wanted to use a corpus from a child around the same age that was not used in the original FFs implementations. I selected Adam from the widely-used Brown corpus from the CHILDES database (Brown [1973], MacWhinney [2000b]) as another typically developing child around two-years-old (shown in Table 3.2).

Table 3.2: Corpus data for Adam

| Corpus | Age Range | Speaker | # Utts | # Tokens | # Types | MLU |
|---|---|---|---|---|---|---|
| Adam | 2;0 - 2;11 | child-directed | 9300 | 37834 | 2065 | 4.1 words |
| | | child-produced | 16289 | 39474 | 1755 | 2.4 words |

## 3.4 Quantitatively connecting underlying representations to output

One practical reason previous studies compared the categories created from FFs to adult categories is that this is a "gold standard" that's both available and fairly easy to agree on (at least, as implemented by syntactic category annotation in many corpora like those in CHILDES: MacWhinney 2000a). However, as mentioned above, the problem is that two-year-olds' syntactic categories may not match adult categories: toddlers might not (i) recognize all instances of a given category as belonging to that category (like NOUN$_{count}$), or (ii) realize certain conceptually subtle categories even exist (like PRONOUNS and DETERMINERS). So, a traditional approach comparing the FF-generated categories against adult categories may not be the best way of assessing if FFs generate the categories toddlers do.

As an alternative approach, just as in the previous chapter, I'm going to be using realistic child input to construct the potential underlying representation, in this case a particular set of potential syntactic categories and the words they contain, and then use those categories to assess how well the potential representation predicts the child's output. In contrast to the previous approach with adjective ordering, here I instead quantify probability of a particular categorization strategy via the output's perplexity [Brown et al., 1992b]. Perplexity is based on surprisal theory, which has been previously used to assess the utility of early syntactic categories in language processing [Bar-Sever and Pearl, 2016] and also used as a standard metric in Natural Language Processing for assessing different language models (Brown et al. [1992a]) given language data the model's meant to capture. In particular, perplexity is inversely related to probability (as shown in Equation 3.1), with the intuition that low probability utterances are unpredictable and therefore highly perplexing. In contrast, high probability utterances are more predictable and so less perplexing. The benefit to using perplexity over log likelihood like I did in the previous chapter is that perplexity allows me to control for length of utterance. This was not a consideration when dealing with 2-adjective-long strings, as all the strings were the same length. Because we are dealing with whole utterances, we can calculate general perplexity over an entire utterance by normalizing with respect to utterance length.

In Equation 3.1, the perplexity of utterance $U$, comprised of words $w_1...w_n$, is the geometric mean of the inverse probability of $U$. So, when the probability of $U$ is low (e.g., a garbled utterance like *penguins I nice like*), the inverse probability is high; in this case, $U$ has a high perplexity. In contrast, when the probability of $U$ is high (e.g., *I like nice penguins*), the inverse probability is lower and so $U$ has a low perplexity. Because probability ranges between 1 and 0, the inverse probability (and so perplexity) ranges between 1 and positive infinity.

$$Perplexity(U = w_1...w_n) = \sqrt[n]{\frac{1}{P(U = w_1...w_n)}} \tag{3.1}$$

Clearly, how we determine the probability of a sequence of words ($P(w_1...w_n)$) matters, since this is the heart of the perplexity calculation. To do this, I follow Bar-Sever and Pearl [2016] and use two potentially plausible assumptions for how toddlers view language generation. First, I assume words belong to underlying (i.e., latent) syntactic categories. This presumably motivates categorizing words in the first place. Second, I assume toddler hypotheses about how language is structured are still developing. So, while toddlers have yet to learn how their native language is truly structured, they likely recognize some local dependencies between syntactic categories (similar to how they recognize local dependencies more generally: e.g., Gómez and Lakusta 2004, Lany and Gómez 2008). One instantiation of this idea is that the current syntactic category depends on the previous category, i.e., a bigram generative mode, a very simple approximation of a child's early syntactic knowledge (Figure 3.1).

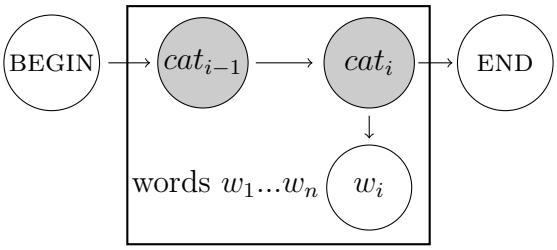



Figure 3.1: A bigram generative model for words $w_1...w_n$. Words are observed, as are the utterance boundaries indicated by BEGIN and END. Categories are latent (shaded).

In the bigram generative model in Figure 3.1, each word $w_i$ is generated based on its latent category $cat_i$, which is conditioned on the previous word's latent category $cat_{i-1}$.[1] To calculate the probability of any sequence $w_1...w_n$, I use Equation 3.2, which is the product of the probability of generating each word $w_i$ in the utterance $U$. This involves two probabilities: the emission probability and the transition probability. The emission probability is the

[1]Note that the first latent category is conditioned on BEGIN, i.e., the utterance beginning.

probability of generating (or "emitting") $w_i$ based on its latent category $cat_i$ ($P(w_i|cat_i)$). The emission probability is multiplied by the transition probability, which is the probability of generating the latent category $cat_i$ given the previous category $cat_{i-1}$ ($P(cat_i|cat_{i-1})$) – in other words, the probability of "transitioning" to the current latent category given the previous context, represented in this model as the previous latent category. The previous category for the first category is the utterance-initial boundary (BEGIN). Additionally, the probability of generating the utterance-final boundary (END) after the last category $cat_n$ is included. I demonstrate this calculation for the utterance *I like nice penguins* in (3.1), assuming the utterance is represented by the syntactic category sequence PRONOUN VERB ADJ $\text{NOUN}_{count}$.

$$P(U = w_1...w_n) = \left( \prod_{w_i \in U} P(w_i|cat_i)P(cat_i|cat_{i-1}) \right) P(\text{END}|cat_n) \tag{3.2}$$

**Example 3.1.** *Words:* *I* *like* *nice* *penguins*

*Categories:* BEGIN PRONOUN VERB ADJ $\text{NOUN}_{count}$ END

$P(U = w_1...w_4) = \left(\prod_{w_i \in U}(P(w_i|cat_i)P(cat_i|cat_{i-1})\right) P(\text{END}|cat_n)$

$= P(I|\text{PRONOUN}) * P(\text{PRONOUN}|\text{BEGIN})$

$* P(like|\text{VERB}) * P(\text{VERB}|\text{PRONOUN})$

$* P(nice|\text{ADJ}) * P(\text{ADJ}|\text{VERB})$

$* P(penguins|\text{NOUN}_{count}) * P(\text{NOUN}_{count}|\text{ADJ})$

$* P(\text{END}|\text{NOUN}_{count})$

Using perplexity and an implementation of $P(U)$ like the one above, we can compare different category representations because each will yield a perplexity score for an evaluation dataset (here, an individual child's productions). This contrasts with the traditional met-

rics that require comparison against adult-level knowledge, which implicitly assumes that early syntactic categorization should yield adult categories. This approach allows me to compare the perplexity of different proposed underlying representations, including the adult categories typically used as the gold standard, as well as the FF-based immature categories I'll be investigating. This comparative evaluation is a marked methodological improvement, precisely because it doesn't assume a child at this stage of development has adult categories, the way a gold standard evaluation does.

I also note that this is a comparative metric only (just as the log likelihoods were for adjective ordering representations in the previous chapter), because a perplexity score is based on the predictability of a particular dataset. For example, a perplexity score of 608 isn't meaningful on its own; instead, it's only meaningful with respect to the dataset used to generate the perplexity score. So, if two category representations are used to generate a perplexity score for a specific dataset, these scores can be compared against each other, with a lower score indicating the a particular representation gives the best fit to the data.

### 3.4.1 Implementing perplexity with potential representations: Evaluation considerations

One evaluation consideration concerns the words that are uncategorized by FFs. This actually wasn't a concern for previous studies that evaluated the accuracy of FF-based categories against adult categories because only words that were captured by FFs were evaluated, and the rest were ignored. That is, words that were not inside any of the FFs were irrelevant for evaluation because evaluation focused only on the accuracy of the words in the FF-based categories. However, the proposed perplexity measure requires us to know the classification of every word, not just the words captured within the FFs. This surfaces in the calculation of $P(U)$, the probability of an utterance, in the perplexity evaluation metric described in

section 3.4. More specifically, every word has both an emission probability and a transition probability, both of which depend on the latent category of the word. So, in order to calculate the probability of an utterance, we need to know what category *every* word in that utterance belongs to – whether the word is inside a FF or not.

Two simple options for determining the category of uncategorized words are (i) assuming each uncategorized word is its own individual category (i.e., as individual as a snowflake), so I'll refer to this as a *snowflake strategy*, or (ii) collapsing all uncategorized words into a single category (i.e., all uncategorized words belong to $cat_{notFF}$), which I'll refer to as a *snowball strategy*. I tested both options as viable potential grouping strategies for words that didn't fall in any FF. The different FF-based categorization options, including how to treat uncategorized words, are shown in Tables 3.3-3.4 for each corpus.

Table 3.3: Number of categories for each version of categorization for both Peter. The total number of categories (Total) is the sum of the FF-categories (FFs) and the uncategorized word types (non-FFs). Framing elements may exclude (-utt) or include (+utt) utterance boundaries (Utt B).

| | | **Categories** | | |
|---|---|---|---|---|
| **Version** | **Utt B** | **FFs** | **Non-FFs** | **Total** |
| Peter gold-std categories | N/A | N/A | 34 | 34 |
| Peter flake, T.5T.1 | +utt | 85 | 2015 | 2100 |
| Peter flake, Top45 | +utt | 45 | 2087 | 2132 |
| Peter flake, Top.13 | +utt | 72 | 2050 | 2122 |
| Peter ball, T.5T.1 | +utt | 85 | 1 | 86 |
| Peter ball, Top45 | +utt | 45 | 1 | 46 |
| Peter ball, Top.13 | +utt | 72 | 1 | 73 |
| Peter flake, T.5T.1 | -utt | 35 | 2222 | 2257 |
| Peter flake, Top45 | -utt | 45 | 2224 | 2269 |
| Peter flake, Top.13 | -utt | 41 | 2224 | 2265 |
| Peter ball, T.5T.1 | -utt | 35 | 1 | 36 |
| Peter ball, Top45 | -utt | 45 | 1 | 46 |
| Peter ball, Top.13 | -utt | 41 | 1 | 42 |

Another evaluation consideration is the number of true syntactic categories. Previous work by Bar-Sever and Pearl [2016] looked at three categorization schemes. First were the two utilized by Mintz [2003b], who collapsed the finer-grained %mor annotation in CHILDES

Table 3.4: Number of categories for each version of categorization for Adam. The total number of categories (Total) is the sum of the FF-categories (FFs) and the uncategorized word types (non-FFs). Framing elements may exclude (-utt) or include (+utt) utterance boundaries (Utt B).

| | | Categories | | |
|---|---|---|---|---|
| **Version** | **Utt B** | **FFs** | **Non-FFs** | **Total** |
| Adam gold-std categories | N/A | N/A | 30 | 30 |
| Adam flake, T.5T.1 | +utt | 80 | 1724 | 1804 |
| Adam flake, Top45 | +utt | 45 | 1800 | 1845 |
| Adam flake, Top.13 | +utt | 80 | 1741 | 1821 |
| Adam ball, T.5T.1 | +utt | 80 | 1 | 81 |
| Adam ball, Top45 | +utt | 45 | 1 | 46 |
| Adam ball, Top.13 | +utt | 80 | 1 | 81 |
| Adam flake, T.5T.1 | -utt | 43 | 1957 | 2000 |
| Adam flake, Top45 | -utt | 45 | 1977 | 2022 |
| Adam flake, Top.13 | -utt | 45 | 1971 | 2016 |
| Adam ball, T.5T.1 | -utt | 43 | 1 | 44 |
| Adam ball, Top45 | -utt | 45 | 1 | 46 |
| Adam ball, Top.13 | -utt | 45 | 1 | 46 |

into categories corresponding roughly to "basic" linguistic categories like NOUN and VERB, using what he called the Standard labeling option with 10 categories: (NOUN (NOUNS and PRONOUNS), VERB (VERBS, AUXILIARIES, and COPULA forms), ADJECTIVE, PREPOSITION, ADVERB, DETERMINER, WH-WORD, NEGATION, CONJUNCTION, or INTERJECTION); the second option was what he called an Expanded Labeling option with 13 categories (NOUNS and PRONOUNS were labeled as distinct categories, as were VERBS, AUXILIARIES, and the COPULA). The other option is simply to use the %mor line annotations from CHILDES as is.

I chose to use the %mor annotations with minimal modifications for the English corpus, which ended up being 34 categories. I decided to use the categories provided by the CHILDES %mor line because I was unable to replicate the Mintz [2003a] results, likely due to differing assignment of words to categories. For example, part of the difficulty in assigning words to larger categories of part of speech is what to do with words that do not obviously fall into one category or the other. An example of this is a contraction like "*we're*". As *we* is a PRONOUN

and *'re*, the contracted form of *are*, is an COPULA, it is not obvious where to put "we're" in the categories laid out by Mintz [2003b]. That is, *we're* isn't obviously PRONOUN or COPULA. The %mor categories help avoid confusion about where to place categorically ambiguous words like contractions (in the %mor line, "we're" has its own particular classification as a combination of a PRONOUN+COPULA). The %mor line categories also have a larger number of smaller categories, which may be more like the categories that FFs generate.

I do note that the number of true categories impacts evaluation both for the traditional approach and the proposed perplexity metric, though in different ways. In the traditional approach where we compare the FF-based categories directly to the gold standard categories, more categories means there are likely to be fewer words in each category. So, FF-based categories may suffer in comparison if they don't make fine-grained enough category distinctions. For example, let's say there are indeed 13 gold standard categories. If FFs collect 7 categories with a big snowball category for the un-captured words, those frames are likely to suffer in their precision, as they would necessarily be clumping together words from different gold standard categories. That is, if there are 13 true categories but only 8 (7 + snowball) FF-based categories, by necessity, at least one of the FF-based categories must have more than one true category in it, leading to lower precision for that category.

In contrast, for perplexity, the number of true categories impacts the probability of the utterance for the true category representation, where more categories means the transitional probability $P(cat_{i-1}|cat_i)$ is likely to be lower on average. However, due to there being fewer words on average per category, the emission probability $P(w_i|cat_i)$ is likely to be higher. The benefit of using perplexity over the traditional approach is that perplexity balances the effect of the number of categories because of the relative effects of the emission and transition probabilities; so, the actual number of categories becomes less important for a perplexity-based evaluation than it would in the traditional approach comparing against a gold standard set of categories.

A third evaluation consideration concerns the implementation of the perplexity metric. Recall from section 3.4 that the perplexity calculation relies on two kinds of probabilities: (i) the emission probabilities of words being generated by a specific category ($P(w_i|cat_i)$), and (ii) the transition probabilities between categories ($P(cat_i|cat_{i-1})$). To calculate perplexity on new language data, the modeled learner must already have some idea of these probabilities from the previous data encountered. Emission and transition probabilities are estimated from the training set and used in the perplexity calculation of the test set. To prevent assigning a probability of 0 for emissions or transitions not seen in the training set, I use add-0.5 smoothing. The emission and transition probability calculations are shown in (3.4)-(3.5). All counts are derived from the training set.

The emission calculation includes the count of word $w_i$ instances in category $cat_i$ ($count_{cat_{w_i}}$), the smoothing factor 0.5, the count of total word instances in category $cat_i$ ($count_{cat_i}$), and the total word types in category $cat_i$ ($W_{c_i}$).

$$Emission = P(w_i|cat_i) = \frac{count_{cat_{w_i}} + 0.5}{count_{cat_i} + (W_{c_i} + 1) * 0.5} \tag{3.4}$$

For example, suppose I encountered a word *puppy* that came from a category NOUN that contained *puppy* 2 times, *kitty*, and *gosling*; then, the emission probability of *puppy* from the category NOUN would be the number of instances of the particular word in that category (2) plus the smoothing factor (0.5), divided by the number of total words in the category (4) plus the number of word types in NOUN (3) plus one (for the smoothing possibility of a new word type), multiplied by the smoothing factor: (2+0.5)/(4+(3+1)*0.5) = 2.5/6.

The transition calculation in (3.5) includes the count of instances where the category sequence $cat_{i-1}$-$cat_i$ occurs ($count_{cat_{i-1}-cat_i}$), the smoothing factor 0.5, the count of total instances of

category $cat_{i-1}$ no matter what follows it ($count_{cat_{i-1}}$), and the total count of category types that can follow any given category $cat_{i-1}$.

$$Transition = P(cat_i|cat_{i-1}) = \frac{count_{cat_{i-1}-cat_i} + 0.5}{count_{cat_{i-1}} + (Cat + 1) * 0.5} \tag{3.5}$$

This last variable includes all the possible category types that could potentially follow $cat_{i-1}$ ($Cat$) and the utterance marker END, yielding $Cat+1$. For a concrete example, let's say I see *the puppy yawns*, where *puppy* is a NOUN and *yawns* is a VERB; suppose I previously saw *the kitty plays* (DETERMINER-NOUN-VERB), *the porg sneezes* (DETERMINER-NOUN-VERB), and *the hero within* (DETERMINER-NOUN-PREPOSITION). To calculate the transition between *puppy* and *yawns*, we calculate the number of times we previously saw a NOUN-VERB sequence (2) plus the smoothing factor (0.5). This is divided by the number of times we see a NOUN followed by any category (3) plus all the possible transition types we've previously seen (NOUN-VERB, NOUN-PREPOSITION = 2) plus 1 times 0.5. Our transition for *puppy yawns* is then (2+0.5)/(3+(2+1)*0.5) = 2.5/4.5.

Importantly, if a child encounters a word she does not know in the test set, she has no reason to put it into any particular category. If the word encountered hasn't been seen in the training corpus, the word will be treated as if it is a newly encountered element, and therefore either belonging to a brand new individual category (i.e. a new snowflake category) or to the larger category containing all uncategorized words (i.e. the snowball category). For example, if an utterance from a test set included *I like nice thestrals*, but *thestrals* had never been seen in training, *thestrals* would be considered a new category and not a NOUN.

Let's calculate the emission of this new word *thestral*, which as far as I the learner know belongs to a new category. If I was considering a *snowflake* strategy, then the number

of times I would have "seen" this new word in this new "category" is 1. This means my numerator in equation 3.4 is (1+0.5). The number of tokens in my new "category" is 1, which means there is 1 type in my new "category" as well, so the denominator of equation 3.4 is (1+(1+1)+0.5). This gives me an emission of 1.5/2.

However, if I am considering a *snowball* strategy instead, my emission calculation is going to look a bit different. All "categories" that I have not seen before are automatically relegated to a giant *snowball*. Therefore I never really encounter a "new" category, just a *snowball* which may or may not have a word I've seen before in it. Let's say *thestral* is assigned to the *snowball*, which is made up of *dog, cat, monkey, porg, horse, cat*. If I have not seen *thestral* before in training, the number of times I would have seen that word in my *snowball* is 0, so the numerator is (0+0.5). The number of word tokens is simply the size of the snowball, or the number of tokens in the snowball, and likewise the number of word types in my snowball. The denominator becomes (5 + (4+1)*0.5). For the *snowball* strategy, my emission is then 0.5/7.5. Therefore, the emission from a *snowflake* category will be higher than the emission from a *snowball* category, which could be very large (so the emission from the *snowball* could be very small).

Let's say instead of seeing *the puppy yawns*, we see *the silly puppy* (DETERMINER-ADJECTIVE-NOUN). In our training data we saw *the kitty plays* (DETERMINER-NOUN-VERB), *the porg sneezes* (DETERMINER-NOUN-VERB), *the hero within* (DETERMINER-NOUN-PREPOSITION), and *the hippogriff is naughty* (DETERMINER-NOUN-VERB-ADJECTIVE). If we are looking at *silly puppy* in our test string, and we did not see did not see a ADJECTIVE-NOUN transition in our training data, this means we have a "new" transition, and we use (3.6) (indicated as

$Transition_{new}$), where $count_{cat_{i-1}-cat_i}$ is 0, because we have never seen this transition before.

$$\begin{aligned} Transition_{new} &= P(cat_i = new\_trans_i | cat_{i-1}) \\ &= \frac{count_{cat_{i-1}-cat_i} + 0.5}{count_{cat_{i-1}} + (Cat + 1) * 0.5} \\ &= \frac{0 + 0.5}{count_{cat_{i-1}} + (Cat + 1) * 0.5} \\ &= \frac{0.5}{count_{cat_{i-1}} + (Cat + 1) * 0.5} \end{aligned} \tag{3.6}$$

If instead of *the silly puppy* we see *the silly thestral*, where *threstral* has not been seen before, and we're using a snowflake strategy so *thestral* does not belong to an existing category, it is both part of a new transition *and* a new category. The transition calculation will be as in (3.7) (indicated as $Transition_{newcat}$), where the probability of seeing a new category is multiplied by a new transition to the category from the previous category.

$$\begin{aligned} Transition_{newcat} &= P(cat_i = new | cat_{i-1}) \\ &= P_{new\ category} * Transition_{new} \\ &= \frac{0.5}{Cat + (Cat + 1) * 0.5} * \frac{0.5}{count_{cat_{i-1}} + (Cat + 1) * 0.5} \end{aligned} \tag{3.7}$$

So, continuing our example, suppose the training data we saw includes *the kitty plays* (DETERMINER-NOUN-VERB), *the porg sneezes* (DETERMINER-NOUN-VERB), *the hero within* (DETERMINER-NOUN-PREPOSITION), *the silly puppy* (DETERMINER-ADJECTIVE-NOUN), *the bouncy ostrich* (DETERMINER-ADJECTIVE-NOUN), and *the hippogriff is naughty* (DETERMINER-NOUN-VERB-ADJECTIVE). We now see *the silly thestral* (DETERMINER-ADJECTIVE-SNOWFLAKE) in our test set, and the probability of seeing a new category (in position $i$) involves the

smoothed count for the new category in the numerator (=0.5). The denominator includes the count of the existing categories that could potentially transition from the previous category ($Cat$) plus the smoothing for those categories and the new category (($Cat + 1$) * 0.5). In our example, this means that the $Transition_{newcat}$ calculation will be the probability of the new transition, 0.5 over the number of times we saw the previous category (ADJECTIVE as the start of a transition with any category (=2), plus the number of possible category types the previous category could transition to, plus 1 (for the transition to END), resulting in (1+1), times 0.5 for smoothing. Then, the probability I generate a new category based on the previous category is the smoothing (=0.5) over the number of category types that could potentially come after the previous category (=1), plus the same number plus 1 for our new category times the smoothing factor ((1+1)*0.5). This is then the probability of generating a new category entirely. The entire probability of $Transition_{newcat}$ will be the probability I see a new transition (0.5/(2+(1+1)*0.5) times the probability of generating a new category (0.5/(1+(1+1)*0.5))

Note that the probability of a transition to this new category uses $Transition_{new}$, but we now have one more category than before, and the count of this new transition ($count_{cat_{i-1}-cat_i}$) is 0. The upshot of this calculation is that it is less probable to have a transition to a completely new category than to have a new transition to an existing category. This is intuitively satisfying.

There are four different possibilities for having a new transition. Let's say a child saw the transitions *Category1__Category2* and *Category2__Category3* in her input. One option is that (I) it is a transition between two previously seen categories, but not in transition together (ex: *Category1__Category3*). In this case, TransitionNew will be used. Another option (II) is a transition between a new category unseen in the training and a previously seen category (new-seen) (ex: *Category4__Category1*); in this case, TransitionNew will be used but the $count_{cat_{i-1}}$ (here, *Category4*) will be 0 and $Cat$ will be 0, because there were no

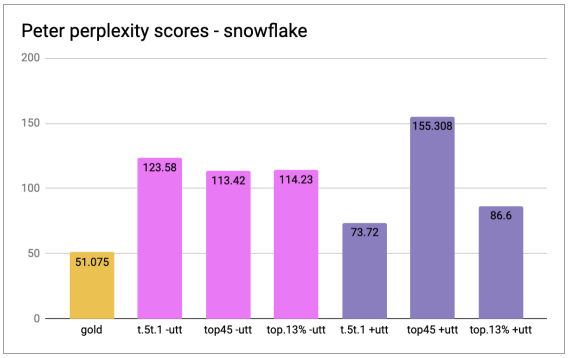
types in training that this "new" category would have transitioned to. A third option (III) would be a previously seen category followed by a new category (ex: *Category1__Category4*). In this case TransitionNewCat will be used. The final option for a new transition would be (IV) when it is a transition between two new categories, new-new (*Category4__Category5*); in this case, TransitionNewCat will be used, but the $count_{cat_{i-1}}$ (here, *Category4*) will be 0 and *Cat* will be 0. For all of the instances where a new category is encountered, the number of existing categories will not be updated because the modeled child is not learning as it analyzes the test data. Rather, the representations are fixed, and are being evaluated on the test data.

## 3.5 Results

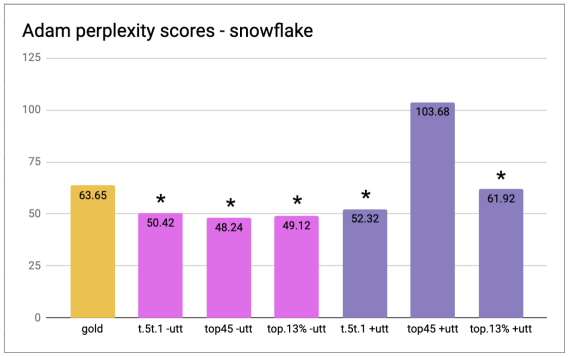
Below I present the average perplexity scores for representations utilizing categories (either gold standard adult or FFs-based) based on the input (training data) and evaluated on the an utterance in the child's output data. The first gold column on the left of Figures 3.2a-3.2d is the perplexity score for the gold standard categories. In Figures 3.2a and 3.2b, the pink columns are the perplexity scores for the *snowflake* FFs instantiations without utilizing utterance boundaries (-utt), and the purple columns are the perplexity scores for the *snowflake* FFs instantiations that do utilize utterance boundaries (+utt). In Figures 3.2c and 3.2d, the green columns are the perplexity scores for the *snowball* FFs instantiations without utilizing utterance boundaries (-utt), and the blue columns are the perplexity scores for the *snowball* FFs instantiations that do utilize utterance boundaries (+utt). 95% confidence intervals (not shown because they're so small, but reported in Tables 3.5-3.6) were generated by calculating perplexity over the output data set based on resampled transitions from the training set. This was done 100 times for each instantiation.

Figure 3.2: The gold standard perplexity scores are given on the left, with the pink and green columns indicate the FFs representations without utterance boundaries (-utt), and the purple and blue columns indicate the FFs representations with utterance boundaries (+utt). An '*' indicates a representation with a perplexity score significantly less than the gold standard category representation

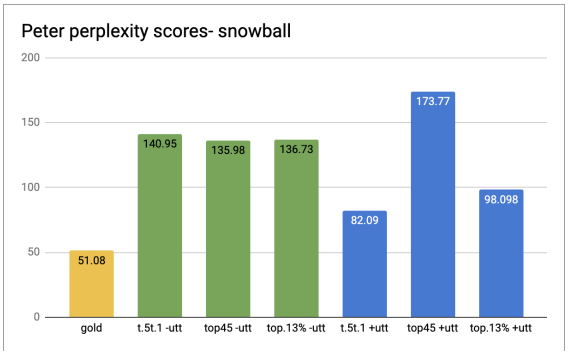
(a)



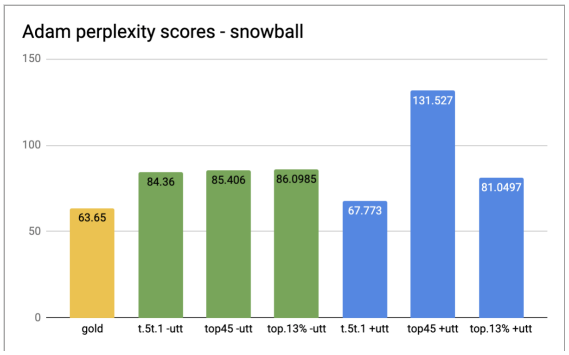

(b)




(c)




(d)




Tables 3.5 and 3.6 shows the perplexity score for each category instantiation as well as the 95% confidence internals associated with them. The gold standard category row is highlighted in yellow, and any FFs instantiation that has a perplexity score that is significantly less than the gold standard perplexity score is highlighted in cyan.

Recall that when a produced utterance is more probable, it is less perplexing (i.e., it has a lower perplexity score). The less perplexing something is (i.e., the closer the perplexity score is to 0), the more probable it is. In this case the lower the perplexity score, the better performing is at explaining the children's output.

To interpret these scores, scores should only be used to compare against each other within the

Table 3.5: Perplexity scores and CIs for Peter

| Version | Utt B | Scores | CIs |
|---|---|---|---|
| Peter gold-std categories | N/A | 51.08 | [50.60, 51.54] |
| Peter flake, T.5T.1 | +utt | 73.72 | [73.41, 74.16] |
| Peter flake, Top45 | +utt | 155.31 | [153.63, 157.69] |
| Peter flake, Top.13 | +utt | 86.60 | [86.19, 87.30] |
| Peter ball, T.5T.1 | +utt | 82.09 | [81.75, 82.60] |
| Peter ball, Top45 | +utt | 173.77 | [171.92, 176.85] |
| Peter ball, Top.13 | +utt | 98.10 | [95.73, 96.96] |
| Peter flake, T.5T.1 | -utt | 123.58 | [123,20, 124.34] |
| Peter flake, Top45 | -utt | 113.42 | [112.50, 115.04] |
| Peter flake, Top.13 | -utt | 114.23 | [113.56, 115.73] |
| Peter ball, T.5T.1 | -utt | 140.95 | [140.38, 142.28] |
| Peter ball, Top45 | -utt | 135.98 | [134.96, 137.91] |
| Peter ball, Top.13 | -utt | 136.73 | [135.51, 138.43] |

Table 3.6: Perplexity scores and CIs for Adam

| Version | Utt B | Scores | CIs |
|---|---|---|---|
| Adam gold-std categories | N/A | 63.65 | [62.90, 64.78] |
| Adam flake, T.5T.1 | +utt | 52.32 | [52.10, 52.67] |
| Adam flake, Top45 | +utt | 103.68 | [103.20, 104.73] |
| Adam flake, Top.13 | +utt | 61.92 | [61.68, 62.30] |
| Adam ball, T.5T.1 | +utt | 67.77 | [67.56, 68.12] |
| Adam ball, Top45 | +utt | 131.53 | [130.82,132.77] |
| Adam ball, Top.13 | +utt | 81.05 | [80.80, 81.53] |
| Adam flake, T.5T.1 | -utt | 50.42 | [50.25, 50.71] |
| Adam flake, Top45 | -utt | 48.24 | [48.12, 48.56] |
| Adam flake, Top.13 | -utt | 49.12 | [48.98, 49.30] |
| Adam ball, T.5T.1 | -utt | 84.36 | [84.00, 85.51] |
| Adam ball, Top45 | -utt | 85.41 | [84.97, 86.42] |
| Adam ball, Top.13 | -utt | 86.10 | [85.75, 87.09] |

same corpus. What we see consistently between Peter and Adam (both typically developing English speaking children around two-years-old) is that there is a similar qualitative pattern between the perplexity scores among the various instantiations of the FFs. In Figures 3.2a, 3.2c, and 3.2d, T.5T.1+utt stands out as being markedly less perplexing than others, including the gold standard categories. For Adam using a *snowflake* strategy, T.5T.1+utt still is relatively less perplexing than the other instantiations using utterance boundaries, though it's more perplexing than the version that doesn't utilize utterance boundaries.

For Peter, there was no particular difference in the qualitative perplexity score pattern between the *snowflake* and *snowball* strategies, except that the *snowflake* scores were slightly lower all around. While the T.5T.1+utt FFs consistently performed the best (73.72 for snowflake and 82.09 for *snowball*), none of the FFs perplexity scores performed better than the gold standard categories (at 51.08). So, in contrast with Adam, Peter seems more likely to be using the gold standard categories than any of the FF-based representations.

For Adam, however, when using a *snowflake* strategy, most of the scores for the FFs-based categories are less perplexing than the gold standard categories (Figure 3.2b); in particular, all of the FFs-based categories that do not use utterance boundaries (-utt) have very similar perplexity scores and perform better at explaining the children's output (are less perplexing) than the gold standard categories. (The gold standard categories are at 63.65 compared to T.5.T.1-utt FFs at 50.42, Top45-utt FFs at 48.24, and Top.13-utt FFs at 49.12. In addition, two of the FF-based categories that do use utterance boundaries (T.5T.1+utt FFs at 53.32 and Top.13+utt FFs at 61.92) have lower perplexity scores than the gold standard categories. Interestingly, the Adam *snowball* strategy does yield a T.5T.1+utt FFs instantiation (at 67.77) that comes very close to the gold standard performance (at 63.65) ((Figure 3.2d).

Overall, these results suggest that gold standard categories tend to be the more likely representation for Peter, while Adam is relying on FF-based categories that ignore utterance boundaries as framing elements and allow each uncategorized word to be its own snowflake category. Also, there seems to be very little difference between the FFs-based categories do not use utterance boundaries, while the Top45+utt FFs-based categories consistently are less a fit to each child's output data than any of the other FFs-based categories.

## 3.6 Discussion

At least for one child (Adam), it seems like one version of immature FF-based categories is more likely to be the underlying representation than a version of the adult categories. Why should this be? It could be that these representations, while immature, are more useful for young children who are also developing other representations and processing abilities.

I also note that these immature representations are also likely to be more useful than having no representation at all, with a child taking every word as its own *snowflake* or all classified as a *snowball* (though that is certainly a useful baseline I could test in the future). Below I discuss the various relevant considerations in interpreting the results of my investigation into developing category representations.

There are certain qualitative patterns in the potential immature representations for the children I investigated. In general, I found that for both children, the *snowflake* strategy (where each uncategorized word belongs to its own category), performs better than a *snowball* strategy (where each uncategorized word belong to a single category). In fact, for Adam, the *snowflake* strategy yielded FFs, both with and without utterance boundaries, that performed better at matching his output than the gold standard categories.

However, in all other implementations of FFs, the gold standard categories performed better than the FFs. The least well-performing FFs implementation was a strict Top45 FFs cutoff, which made the child's output most perplexing, and therefore matched the child's output the least well.

Below I discuss in more detail the implications of different FF implementation decisions (snowball vs. snowflake, the role of utterance boundaries in FFs, the utility of the different FF-frequency thresholds), differences between the two children investigated, and the overall utility of FF-based categories.

### 3.6.1 Snowball vs Snowflake

Overall, the FFs implementations that used a *snowflake* strategy over a *snowball* one performed better at predicting children's output. For Adam, the snowflake strategy yielded FFs that performed better than the gold standard categories. This means that at this stage of development, Adam was likely to consider each new word as belonging to its own special category, rather than assuming all unknown words are grouped together into a giant catch-all category. The superior performance of the *snowflake* strategy across the board also means that the higher probability associated with emitting from a small *snowflake* category matters more than the smaller probability associated with transitions to many different snowflake categories.

### 3.6.2 The role of utterance boundaries

Here, the combination of relative thresholds and the presence of utterance boundaries as potential framing units seems to be the best performing combination in FFs for both *snowflake* and *snowball* versions of FFs for Peter and the *snowball* version for Adam; however, in these cases, the FF-based categories never appear to be the best performing option, with gold standard categories still being the best match for the child's output. In contrast, FFs without utterance boundaries for Adam (using a snowflake strategy) seem to better match Adam's output than gold standard categories, with both T.5T.1+utt and Top.13 also better performing than the gold standard categories. It could be that utterance boundaries may still be a useful cue to young children at certain stages of development. Here, Adam would be in the stage of development while Peter would have already passed through it. It would also be useful to see this kind of analysis applied to a wider age group to determine if there are more general patterns of utterance boundary salience during the transition from immature to mature categories.

### 3.6.3 The utility of frequency thresholds

Overall, the T.5T.1+utt FFs strategy, or FFs where each frame had to capture 0.5% of types and 0.1% of tokens (and where utterance boundaries were included in framing units), performed generally better than the rest of the FFs strategies, even when it didn't perform as well as the gold standard categories at matching the child's output. This version of frequency threshold is thought to be more conservative and potentially more accurate (Chemla et al. [2009a]). It also utilizes a relative frame-frequency-based threshold that is dependent on corpus size instead of a strict numerical cutoff of frame number. Certainly within the versions of frames that use utterance boundaries, T.5T.1 and Top.13% both perform better than the strict Top45 cutoff. While a strict cutoff might be easy to implement for a child (perhaps as a limited memory buffer), it might not be useful enough for a child who is acquiring immature category representations. The combination of a relative threshold, plus the information provided with utterance boundaries, seems like the most useful of the FFs strategies for at least one child. It is unclear if there is anything particularly special about the two relative threshold measures discussed here. It is possible that other thresholds may be useful, and in fact the thresholds might differ between children based on an individual child's cognitive abilities and memory constraints.

### 3.6.4 Differences between children

What does it mean that I only found FFs that performed better than the gold standard categories in one child? There are differences between Peter and Adam, who are both typically developing around two-years-old (Peter: 1;9-2;4 and Adam 2;0-2;11). One is that there are quantitative differences between the corpora for each child, with differences in the amount of input and output data available, but it is unclear what effect that has on the analysis. Besides differences in the amount of data, there are individual differences between

children's language development trajectories (Tomasello and Todd [1983], Bates et al. [1991]). It is possible that at the age I look at, Peter has already developed the finer-grained adult-like categories that were investigated here as the gold standard. It is possible that by looking at younger data for Peter we might see FFs as a better match for generating output (although there is likely to be a data sparsity issue, as multi-word utterance output is less likely under 2). Adam, on the other hand, may still be utilizing more immature representations at 2. It is possible that if I were to look at Adam's data at a little older, I would find that adult-level syntactic categories might generally perform better than any FFs categories, as we see in Peter.

## 3.7 Future work

I'll now briefly discuss possible future directions, including other informative implementations and additional possibilities for how children might deploy FFs in order to create immature categories.

### 3.7.1 Baseline evaluation

Firstly, I'd like to evaluate other baselines, like evaluating FFs against a total snowflake strategy (where every word is its own category), to firmly assess whether any category representation is better than having no representation at all. Having this baseline will give context to our previous results on how well the representations in question best match data that we observe.

### 3.7.2 Further FFs implementations

Secondly, I'd like to explore what other FF implementations seem plausible besides those that I've investigated here. Some of these implementations include "flexible" frames, where categories generated by FFs are merged based on how many overlapping words they have, attempting to create more adult-like categories [Mintz, 2003b, Chemla et al., 2009b]. Here I have only looked at a few FFs implementations, but that does not mean there are not other definitions of frequency that are valid (and possibly more effective) as well. Another question is whether these results depend on using the finer-grained %mor classifications for the adult categories. That is, could these results change with a broader gold standard part of speech classification system more akin to the ones utilized by Mintz [2003b]? Recall that there are implementation difficulties with sorting words into broader categories as Mintz [2003b] and others have done, in that all words need to be categories, and many do not fit neatly into the broader categories they outline in their studies (for example, what do we do with contractions like *we're*?). However, if these issues can be surmounted, it's possible that other potential adult category representations could be evaluated against child data.

### 3.7.3 Investigating FFs deployment

Another consideration involves children whose developing processing abilities don't allow them to deploy FFs as accurately as here. That is, a FFs-based strategy might not be useful for every child. What would a "broken" FFs strategy look like in children who might lack the cognitive resources or skills to fully utilize this strategy? Remember that being able to track frames is just being able to remember what words/morphemes are on either side – for example, remembering that in *At dawn we ride*, *dawn* is in between *At* and *we*. Tracking these non-adjacent dependencies and tabulating their frequencies on the rest of the words in subsequent utterances falls under the general umbrella of "statistical learning"

(Gomez and Gerken [1999], Saffran et al. [1996], Romberg and Saffran [2010]). Interestingly, children with Specific Language Impairment (SLI) tend to have deficiencies in statistical learning [Evans et al., 2009, Haebig et al., 2017], a skill that is crucial for tracking the non-adjacent dependencies that frequent frames rely on. So, SLI children trying to use a FFs-based strategy to form categories may end up with much noisier categories than what I investigated here.

It would also be interesting to see how the different frequency implementations that I utilized in this chapter perform for children who have statistical learning deficiencies. We could do this by introducing different types of noisy FF-based categories. It could also be that if there are memory and processing constraints, there is a higher threshold for frequency than a typically developing child might have, and there might therefore be a bias for fewer, more accurate FFs. To investigate this, I could try increasingly stringent limitations on what counts a frequent, and see if these "highly" frequent frames are better or worse at matching the output of an atypically developing child than either gold standard categories, or the FFs thresholds implemented here for typically developing children. This kind of investigation might give us insight into what kind of memory constraints these atypically developing children may be working with when developing immature category representations.

### 3.7.4 The underlying model

What about children who have different developing syntactic systems (i.e., more rudimentary or more advanced that what I assumed here)? These developing systems could range from something as simple as a unigram model to a syntactic model that is more complex and adult-like, instead of using a bigram assumption as we have done in this chapter. Changing the generative model so that it is more appropriate for children of different developmental stages and abilities can help us assess if a particular underlying syntactic representation

might be more appropriate for this developmental stage, or at least how much the results here depend on the syntactically-immature bigram generative model. It may also be useful to see what this analysis suggests for adult data (similar to what I did in the previous chapter); this would allow us to determine if the standard adult-like categories (and an adult-like generative model) perform better than the other possible category representations for the adults.

# Chapter 4

# The emergence of productive categories in typically and atypically developing children

## 4.1 Introduction

The previous chapter looked at possible immature syntactic category representations that children could have on the way to acquiring adult representations. Although linguists may disagree on what those immature representations should be (as well as the adult representations), we generally agree that typically developing children *eventually* achieve adult linguistic knowledge.

But when is "eventually"? How can we tell exactly when a child has successfully acquired an adult linguistic representation? This chapter investigates the emergence of a particular linguistic milestone: developing productive syntactic categories. I do this by again utilizing the framework of the previous chapters, that of assessing the possible options for underlying

representation, in this case the presence or absence of productive adult categories.

To do this, I start by looking at a variety of children, including a child in the previous chapter (Peter), as well as atypically developing children with Specific Language Impairment (SLI) and age-matched controls. I take the input and output data from each child and diagnose the presence of productive categories by evaluating possible underlying adult-like, productive syntactic categories by connecting that child's input and output.

## 4.2 Target knowledge: what does it mean to be "productive"?

As discussed in the previous chapter, categories are effectively clusters of words that behave similarly in linguistic contexts. Measuring when a cluster of similarly behaving words becomes an "adult" category is a tricky task. One way to assess the presence of a category is to measure the behavior against a gold standard adult category. An alternative option that I will explore in this chapter is looking at how words combine together productively in adult speech, and using this as a signal to assess the presence of a particular adult-like category via the multi-word productions in the child's output.

For instance, let's say I am trying to ascertain if an adult has a DETERMINER category and a NOUN category. What I can do is look at the simplest ways words can combine – 2-word strings. We have some idea about how a DETERMINER category might combine with a NOUN category. In the simplest case (e.g., ignoring exceptions based on agreement), a word that is a DETERMINER should be able to combine with a word that is a NOUN. If I have a collection of DETERMINERs (like *the, an, your, that*) and NOUNs (like *serval, ocelot, jaguar, cheetah, lioness*), I should expect to see *the cheetah*, *your ocelot*, *that lioness*, etc., as any DETERMINER from this set could combine with any NOUN from this set.

The target knowledge, then, is that within these 2-word DETERMINER+NOUN combinations, we should see productivity, where words in two different categories that can combine *do* combine at the same rate seen in adult 2-word productions. We can observe the results of this potential productivity in children's output to see if children are generating 2-word phrases with the same level of productivity for each category under consideration that we believe we should see in adults.

### 4.2.1 Atypically developing children: A look at Specific Language Impairment

Looking at a sophisticated linguistic skill like productive syntax, we might naturally find variation in typically developing populations already. However, in addition to this individual variation, there are many factors that can impact the development of language. What happens when we look at children that seem to struggle with some aspects of language?

Delays and deficits in productivity can result in serious linguistic difficulties. One population that consistently experiences issues with productive language is children with Specific Language Impairment (SLI). Children with SLI are typified by linguistic impairment without any obvious non-linguistic cognitive impairment or belonging to any other clinical group. Cross-linguistic studies in SLI have shown that SLI children tend to struggle the most with whatever in their language is hardest to acquire, often including syntax and morphology, depending on the language (Leonard [2014]). In fact, delays or problems in producing 2-word combinations at 24 months is a strongly predictive measure of presenting with SLI, stronger even than delays in producing first words [Diepeveen et al., 2016, Rudolph and Leonard, 2016]. These later difficulties in producing multi-word utterances could be a symptom of incomplete or differing development of these categories themselves.

The amount of language delay in SLI children is variable and the population itself is very

heterogenous. Some accounts claim that their production is delayed about a year behind their peers, and their comprehension skills delayed about 6 months. However, it is fairly common for children to be far more impaired. All SLI children are typically exemplified by a limited vocabulary and, for English speakers, limited grammatical morphology (such as tense markers). Some sample speech from a 16-year-old with SLI shows how morphology can be limited: "He want to play that violin. Those men sleeping...Can I play with violin?" [Weiner, 1974]. SLI children often never catch up to their typically developing peers [Weismer, 2010].

In this chapter, I look at various children with SLI at three- and four-years-old as well as age-matched control children who are typically developing. The possible representation profiles of children with SLI and typically developing children of the same age might show that there are systematic differences in which kind of productive categories seem to be available. This could give insight into what kind of categories are easier for a child with SLI to acquire. If there are not systematic differences, it could be that either SLI emerges later than the ages we look at here, or that SLI interferes with language at a different level than this kind of productive category representation. Beyond this, because of the heterogeneity of the SLI population, we might expect each individual's productive category representation profile to look different from each other.

## 4.3 Assessing category knowledge

At the category level, the two representations I'll consider here are that the mature category is (i) present (i.e, {*the, that, your*, etc.} $\in$ DET), or (ii) absent (i.e., *the, that, your*, etc. are simply individual words that aren't interchangeable syntactically). At the multi-word combination level, I focus on combinations made up of two potential categories (e.g., *the cheetah*, which could involve DET and NOUN). For these combinations, there are three possible representations: NOT, SEMI, and fully PRODuctive. More specifically, a NOT pro-

ductive representation has both categories absent; a SEMI-productive representation has one category present and the other absent; a fully PRODuctive representation (which we assume adults have) has both categories present.

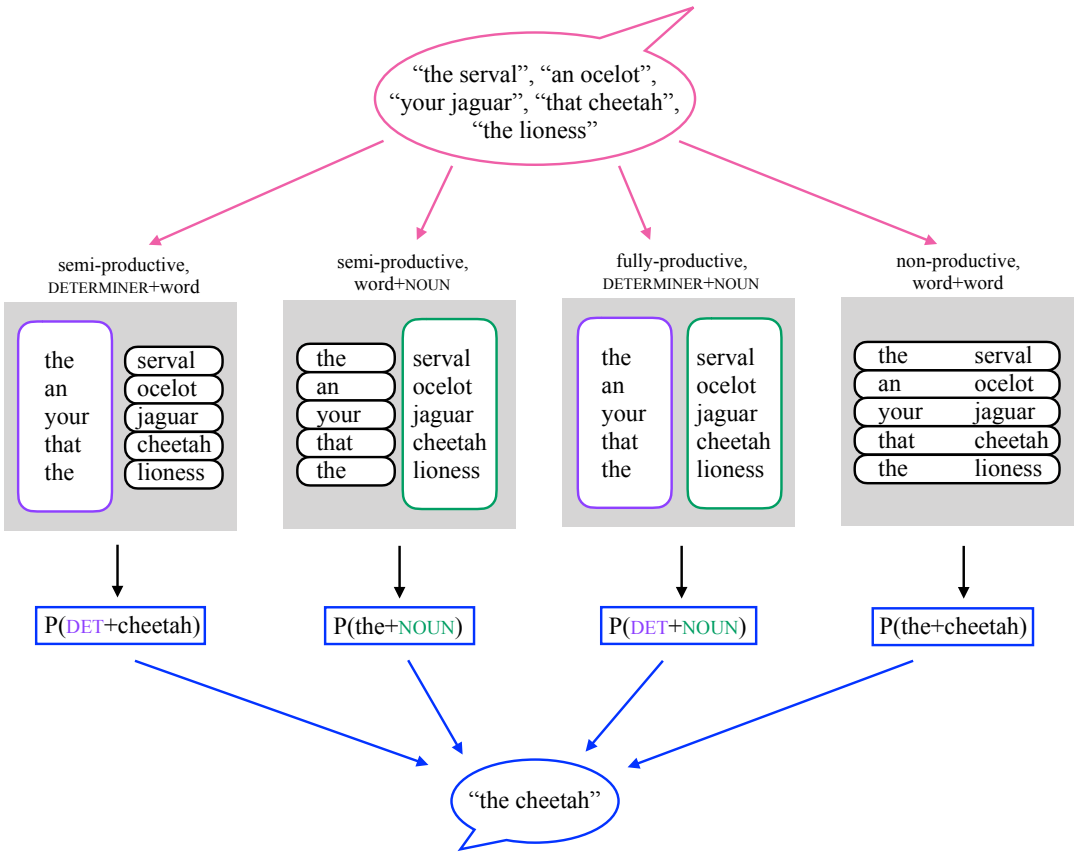



Figure 4.1: A diagram of how potential output 2-word combinations are considered for *the cheetah*, based on the hypothesis under consideration (within the grey box) and the child's input. Combinations involving lexical items in black ovals get their probabilities from the child's input.

### 4.3.1 Possible child category representations for multi-word combinations

I will now consider these three types of syntactic category representation (exemplified in Figure 4.1) that very young children could use to form multi-word combinations (like *the*

*cheetah*). The representation types differ with respect to whether the child produces multi-word combinations according to (i) the distribution of multi-word combinations in her input (NOT productive), (ii) both her input distributions and an internal category representation (SEMI-productive), or (iii) internal category representations alone (fully PRODuctive).

A child using a NOT productive representation can only generate a multi-word combination if she's heard it in her input (e.g., *the cheetah* $\rightarrow$ $(\textit{the}+\textit{cheetah})_{Input}$). So, any multi-word combination she generates is effectively a memorized amalgam; how often she generates a particular amalgam depends on how frequently that amalgam was in her input. (See the fourth column in Figure 4.1.) This contrasts with a child using a SEMI-productive representation, who relies on an internal category for generating one part of the multi-word combination and her input combinations with that category for generating the other part (e.g., *the cheetah* $\rightarrow$ $(\text{DETERMINER}+\textit{cheetah})_{Input}$). (See the first column in Figure 4.1 for a productive DETERMINER category and the second column for a productive NOUN category.) Here, if she's heard *cheetah* used with a DETERMINER – any DETERMINER, not just *the* – she can generate *the cheetah* this way. So, the child can generate some novel expressions, but still relies on input distributions when the expressions involve words that aren't part of a syntactic category. This also works in the other direction (e.g., *the cheetah* $\rightarrow$ $(\textit{the}+\text{NOUN})_{Input}$). Here, if she's heard *the* used with a NOUN – any NOUN, not just *cheetah* – she can generate *the cheetah* this way. However, a child with a fully PRODuctive representation can generate novel combinations by relying on her internal syntactic categories alone, rather than input distributions of multi-word combinations (e.g., *the cheetah* $\rightarrow$ DETERMINER+NOUN). (See the third column in Figure 4.1.) That is, the child draws on her internal category knowledge when generating utterances just the way we believe adults typically do, which allows for productive creation of novel multi-word combinations.

## 4.4 How can we quantitatively measure representational knowledge?

To assess the presence of adult-level category knowledge, we need to measure what particular underlying representation a child has with respect to a certain category. In this section, I outline how to quantitatively analyze children's productions to determine the nature of their underlying representations, following Bates et al. [2018]'s quantitative framework. I review key ideas from this framework below.

### 4.4.1 Lexical overlap as a measure of category knowledge

*Lexical overlap* is often used as a measure for productivity [Yang, 2010, 2011, Pine et al., 2013], and is meant to capture the intuition that words in one category can be freely combined with words from another. That is, category members are effectively interchangeable in those combinations. For example, a DETERMINER category would allow any of its member words (e.g., *the*, *that*, *your*, etc.) to combine with nouns like *cheetah*. So, we would expect to see multiple determiners used with any given noun (e.g., *the cheetah, that lioness, your serval*, etc.) – that is, there would be *overlap* in the use of determiner *lexical* items. So, to assess a category, we need to examine its lexical overlap with respect to words that the category can combine with. For example, when assessing DETERMINER in combination with nouns, we can look at how many nouns have lexical overlap when it comes to determiners.

So, if we want to assess the lexical overlap for DETERMINER, that means we want to know for each individual noun (like *cheetah*) whether it only occurred with *the* or if it also occurred with any other possible DETERMINER, say with *a*. If *cheetah* did occur with more DETERMINERs than just *the*, this would be a case of lexical overlap.

I assess both the *Observed* lexical overlap present in a speaker's productions and the *Expected* lexical overlap if the speaker used a particular representation to generate those productions. While there's only one Observed score per potential category (e.g., a $\text{DET}_{Observed}$ for how determiners combine with nouns), there's an *Expected* score for each potential representation the speaker could be using to generate her productions. If the Expected overlap for a particular representation matches the observed overlap well enough, this indicates that the representation is compatible with the speaker's output.

### 4.4.2 Calculating Observed overlap

I first describe how to calculate the lexical overlap for a potential category with respect to a set of words it combines with. This is the core calculation that will be used for calculating Observed and Expected overlap scores for multi-word combinations. I then describe how to calculate the Observed overlap for multi-word combinations and the Expected overlap for each of the three representation types (NOT, SEMI, and PROD).

For a potential category whose status is $Unknown$ (like DETERMINER), we look at the lexical overlap in words which that category combines with (like nouns, which would be $w_{comb} \in Combine$ in (4.1)). Lexical overlap itself is defined very liberally, following previous studies using it [Yang, 2010, 2011, Pine et al., 2013]: if more than one word $w_{unk} \in Unknown$ (e.g., both *the* and *that*) appears in combination with a word $w_{comb} \in Combine$ (e.g., *cheetah*), then lexical overlap for $w_{comb}$ is 1. Otherwise, if $w_{comb}$ only ever appears in combination with a single word $w_{unk} \in Unknown$ (e.g., *the cheetah* is the only combination of a noun with *cheetah*), lexical overlap is 0. This is $overlap_{w_{comb}}$ in (4.1). The total overlap $overlap_{Combine}$ is the lexical overlap average across all words that the potential category can combine with ($w_{comb} \in Combine$). For example, this would be the lexical overlap average across all nouns when assessing potential category DETERMINER on how it combines with nouns. So, if there

are 50 nouns that combine with determiners in the data sample, then individual overlap scores $overlap_{w_{comb}}$ are calculated for each of these 50 nouns, and the average is taken of all 50 scores.

$$overlap_{w_{comb}} = \begin{cases} 1\text{: } w_{comb}\ occurs\ with\ >1\ word\ w_{unk} \in Unknown \\ 0\text{: } w_{comb}\ occurs\ with\ only\ 1\ word\ w_{unk} \in Unknown \end{cases} \quad (4.1)$$

$$overlap_{Combine} = \frac{\sum_{w_{comb} \in Combine} overlap_{w_{comb}}}{|Combine|}$$

For a multi-word combination involving two potential categories (e.g., DETERMINER+NOUN), Observed overlap can be calculated with respect to each category (e.g., with respect to nouns when assessing DETERMINER and with respect to determiners when assessing NOUN). The observed overlap calculation is just as in (4.1), shown in (4.2) over the set of speaker productions that involve those kind of multi-word combinations $S_{Obs}$ (e.g., all combinations of determiners+nouns for DETERMINER+NOUN).

$$Observed = overlap_{Combine}(S_{Obs}) \quad (4.2)$$

### 4.4.3 Calculation of Expected overlap for the different representation types

Below I provide a walk-through of the calculation of the Expected lexical overlap for the three representational types: NOT, SEMI, and fully PRODuctive.

For the NOT productive representation (e.g., *the cheetah* → *the+cheetah*), the speaker generates her multi-word combinations as memorized amalgams from her input. Using this representation, she will produce a given amalgam with about the same frequency she heard

it in her input. To simulate this process, I generate multi-word combination data samples $S_{Exp_{Not}}$ that are the same size as the observed speaker multi-word combination sample $S_{obs}$; these samples are drawn from the speaker's input. That is, if there are 100 determiner+noun combinations in the speaker's output, I generate 100 determiner+noun combinations, based on the determiner+noun distribution in the speaker's input. This is shown in the top portion of equation (4.3).

The combinations that specific word $w_{unk}$ from the category whose status is $Unknown$ is generated with depend on the combinations from the speaker's input that $w_{unk}$ appeared with. To continue with our determiner example from above, if $det_j$ appeared with noun $noun_k$ (e.g., *the+cheetah*) for 10% of the speaker's input, about 10% of the generated determiner+noun combinations $S_{Exp_{Not}}$ will be $det_j noun_k$ combinations. That is, the probability of sample $s_i$ involving word $w_{unk}$ combined with $w_{comb}$ depends on how often $w_{unk}$+$w_{comb}$ appeared in the speaker's determiner+noun input. Once the sample using the NOT productive representation has been generated, we can calculate the lexical overlap for this sample and use that as the Expected overlap for a child using the NOT productive representation. This is shown in the bottom part of (4.3).

$$
\begin{aligned}
|S_{Exp_{Not}}| &= |S_{obs}| \\
s_i &\in S_{Exp_{Not}} \\
s_i = w_{unk}w_{comb} &\propto p_{w_{unk}w_{comb}{}_{Input}} \\
w_{unk} \in Unknown,\ & w_{comb} \in Combine \\
Expected_{Not} &= overlap_{Combine}(S_{Exp_{Not}})
\end{aligned}
\tag{4.3}
$$

Because we are generating samples of data produced by a child using the NOT productive representation, we repeat this process 1000 times (i.e., generate 1000 expected multi-word combination samples and calculate the Expected overlap). We then average these expected

overlap scores to get the Expected overlap for the NOT productive representation.

We can use a similar approach when calculating the Expected overlap for the SEMI-productive representation (e.g., *the cheetah* → DET+*cheetah* or *the*+NOUN). Using this representation, a speaker generates her multi-word combinations by relying on her internal category representation for one word and her input distributions for combinations with the other word. More specifically, let's consider the case where the word from *Unknown*, $w_{unk}$, comes from a category while the word from *Combine*, $w_{comb}$, doesn't (like DETERMINER+*cheetah*). To generate combination $w_{unk}w_{comb}$, the child relies on her internal category representation to generate word $w_{unk}$ (i.e., DETERMINER ← *the*) and then looks to her input to see how often words from this category combine with word $w_{comb}$ (i.e., *cheetah*). So, she would generate combination $w_{unk}w_{comb}$ (*the+cheetah*) with about the same frequency she heard examples of *Unknown* $w_{comb}$ (DETERMINER+*cheetah*) in her input. To simulate this process, we again can generate multi-word combination data samples $S_{Exp_{Semi}}$ that are the same size as the observed speaker multi-word combination sample $S_{obs}$. Because this representation assumes that all words $w_{unk} \in Unknown$ in the speaker's output were generated from her internal category, they will appear as often as they appeared in her observed output. For example, if DETERMINER is *Unknown* and determiner $det_j \in$ DET (*the*) appears 10 out of 100 times in the speaker's output, the generated sample will include a combination with $det_j$ about $\frac{10}{100} * 100 = 10\%$ of the time. In particular, category *Unknown* (DETERMINER) generates words $w_{unk}$ with some probability, and this is the probability we see these words in the speaker's output. So, the SEMI expected samples involve word $w_{unk}$ proportional to how often they appeared in the speaker's observed productions. This is shown in the top part of (4.4).

The combinations that $w_{unk}$ (*the*) is generated with depend on the combinations from the speaker's input that words of category *Unknown* (DETERMINER) appeared with. Returning to our determiner example from before, if DETERMINER is being assessed in combination with

individual nouns, and determiners appear with noun $noun_j$ (*cheetah*) 5 out of 100 times, the generated sample will include determiners in combination with $noun_j$ 5% of the time. That is, the probability of multi-word sample $s_i \in S_{Exp_{Semi}}$ involving a specific word $w_{unk} \in Unknown$ combined with $w_{comb}$ depends on how often any word in $Unknown$ combines with $w_{comb}$ in the speaker's input. This is equivalent to how often $w_{comb}$ appeared in the multi-word combinations involving words of category $Unknown$ in the speaker's input $w_{comb_{input}}$, as shown in (4.4).

$$
\begin{aligned}
|S_{Exp_{Semi}}| &= |S_{Obs}| \\
s_i &\in S_{Exp_{Semi}} \\
w_{unk} &\in s_i \propto p_{w_{unk_{Obs}}} \\
s_i = w_{unk}w_{comb} &\propto p_{Unknown}\ p_{w_{comb_{Input}}} \\
w_{unk} &\in Unknown,\ w_{comb} \in Combine \\
Expected_{Semi} &= overlap_{Combine}(S_{Exp_{Semi}})
\end{aligned}
\tag{4.4}
$$

A similar process can be used when $Unknown$ isn't a category while $Combine$ is (e.g., *the*+NOUN). To generate combination $w_{unk}w_{comb}$ (*the+cheetah*), the child relies on her internal category representation to generate word $w_{comb}$ (NOUN $\leftarrow$ *cheetah*) and then looks to her input to see how often words from this category combine with word $w_{unk}$ (*the*). So, she would generate combination $w_{unk}w_{comb}$ with about the same frequency she heard examples of $w_{unk}$ $Combine$ (*the*+NOUN) in her input. To simulate this process, we again can generate multi-word combination data samples $S_{Exp_{Semi}}$ that are the same size as the observed speaker multi-word combination sample $S_{obs}$. Because this representation assumes that all words $w_{comb} \in Combine$ in the speaker's output were generated from her internal category, they will appear as often as they appeared in her observed output. For example, if NOUN is $Combine$ and Noun $noun_j \in$ NOUN (*cheetah*) appears 30 out of 100 times in the speaker's output, the generated sample will include a combination with $noun_j$ about

$\frac{30}{100} * 100 = 30\%$ of the time. In particular, category *Combine* generates words $w_{comb}$ with some probability, and this is the probability we see these words in the speaker's output. So, the SEMI expected samples involve word $w_{comb}$ proportional to how often they appeared in the speaker's observed productions. This is shown in the top part of (4.5).

The combinations $w_{comb}$ is generated with depend on the combinations from the speaker's input that words of category *Combine* appeared with. Returning to our noun example from before, if NOUN is being assessed in combination with individual determiners, and nouns appear with determiner $det_j$ (*the*) 2 out of 100 times, the generated sample will include nouns in combination with $det_j$ 2% of the time. That is, the probability of multi-word sample $s_i \in S_{Exp_{Semi}}$ involving a specific word $w_{comb} \in Combined$ combined with $w_{unk}$ depends on how often any word in *Combine* combines with $w_{unk}$ in the speaker's input. This is equivalent to how often $w_{unk}$ appeared in the multi-word combinations involving words of category *Combine* in the speaker's input $w_{unk_{input}}$ (*the*+NOUN), as shown in (4.5).

$$
\begin{aligned}
|S_{Exp_{Semi}}| &= |S_{Obs}| \\
s_i &\in S_{Exp_{Semi}} \\
w_{comb} \in s_i &\propto p_{w_{comb_{Obs}}} \\
s_i = w_{unk} w_{comb} &\propto p_{w_{unk_{Input}}}\ p_{Combine} \\
w_{unk} \in Unknown,\ &w_{comb} \in Combine \\
Expected_{Semi} &= overlap_{Combine}(S_{Exp_{Semi}})
\end{aligned}
\tag{4.5}
$$

As before, once the sample using the SEMI representation has been generated, we can calculate the lexical overlap for this sample and use that for a child using a SEMI representation. This is shown in the bottom part of (4.4) and (4.5). We then do this process 1000 times to get 1000 SEMI samples, compute the lexical overlap for each, and take the average as the Expected SEMI overlap score.

For the fully PRODuctive representation (e.g., *the cheetah* → DETERMINER+NOUN), the speaker generates her multi-word combinations by relying on internal category representations for both words. Yang (2010, 2011) describes an analytical solution for the expected lexical overlap when both categories exist (shown in (4.6)).

$$
\begin{aligned}
overlapprod_{w_{comb}} &= \\
&= 1 - P(no\ w_{comb}) - P(only\ 1\ w_{comb}) \\
&= 1 - (1 - p_{w_{comb}})^{S_{obs}} - \sum_{k=1}^{S_{obs}} \binom{S_{obs}}{k} (p_{w_{comb}} * p_{w_{unk}})^k (1 - p_{w_{comb}})^{S_{obs}-k} \\
&= 1 + (|Unknown| - 1)(1 - p_{w_{comb}})^{S_{obs}} \\
&- \sum_{w_{unk} \in Unknown} (p_{w_{comb}} * p_{w_{unk}} + 1 - p_{w_{comb}})^{S_{obs}} \\
p_{w_{unk}} &= p_{w_{unk_{Obs}}}, p_{w_{comb}} = p_{w_{comb_{Obs}}} \\
Expected_{Prod} &= \frac{\sum_{w_{comb} \in Combine} overlapprod_{w_{comb}}}{|Combine|}
\end{aligned}
\tag{4.6}
$$

We can use this here, rather than generating expected samples and calculating lexical overlap for those samples. The key intuition involves the definition of lexical overlap, where a word $w_{comb}$ shows lexical overlap if more than one word $w_{unk} \in Unknown$ combines with $w_{comb}$. So, we can calculate this analytically as 1 minus the probability that $w_{comb}$ will (i) never appear with any word in $Unknown$, or (ii) only appear with a single word in $Unknown$.

For $w_{comb}$ to never appear with any word in $Unknown$, this means that for all multi-word combination samples $S_{obs}$ involving words from $Unknown$, $w_{comb}$ was never selected. The probability of $w_{comb}$ can be represented as $p_{w_{comb}}$, and so the probability of not choosing $w_{comb}$ to combine with a word from $Unknown$ $S_{obs}$ times is (1-$p_{w_{comb}}$)$^{S_{obs}}$.

For $w_{comb}$ to appear with only a single word $w_{unk}$ in $Unknown$, this means that for every multi-word combination UNKNOWN+COMBINE, either $w_{comb}$ was selected and combined

with $w_{unk}$ (which occurs with probability $p_{w_{comb}} * p_{w_{unk}}$) or some other word – and not $w_{comb}$ – was selected (which occurs with probability 1-$p_{w_{comb}}$). Any given sample with $w_{comb}$ only ever appearing with $w_{unk}$ will have some split between these two options, for all $S_{obs}$ samples (i.e., $k$ samples will have $w_{comb}$ with $w_{unk}$ and $S_{obs} - k$ samples will have some other *Combine* word). So, if we sum up all these possibilities (shown in (4.6)), this is the probability of $w_{comb}$ only ever appearing with a single $w_{unk}$.

Some algebraic rearrangement yields the formula for overlapprod$_{w_{comb}}$ at the bottom of (4.6) for the Expected overlap for word $w_{comb}$ from Yang (2010, 2011). Note that all word probabilities are estimated based on the speaker's productions of $w_{comb}$ and $w_{unk}$ (i.e., $p_{w_{comb}} = p_{w_{comb_{Obs}}}$, $p_{w_{unk}} = p_{w_{unk_{Obs}}}$). This is because all words in these combinations are generated from an underlying internal category, and so don't rely on the speaker's input. As with the original calculation of lexical overlap, these individual word overlaps are averaged to get the Expected overlap.

### 4.4.4 Evaluating possible representations

For each child, the potential representation must include some information about each potential category within that representation. For each potential category, either it is productive (there is a DETERMINER category that determiners are drawn from) or there is not (where all determiners are treated as individual words). Suppose a child, for example Peter, has 4 potential categories (DETERMINER, NOUN, VERB, and PRONOUN). Therefore, there are 16 possible category representations ($2^4$) that Peter could be entertaining, ranging from NOT-productive (determiners, nouns, verbs, and pronouns are not part of categories, and so not productive) to fully-PRODuctive (all potential categories are productive) to somewhere in the middle (some categories may be productive and some may not, yielding SEMI-productive representations).

For each child, I gathered the 2-word combinations that met certain criteria. The combinations had to (i) make up their own phrase (like DETERMINER+NOUN: *the cheetah* and not NOUN+PREP: *cheetah on*) and (ii) have at least 100 tokens of any DETERMINER+NOUN combination). That is, any DETERMINER+NOUN combination would need to have appeared at least 100 times in both the child's input and output. This cutoff is based on the recommended frequency from Bates et al. [2018]. Then, I calculated the Observed and Expected overlap for each child and representational hypothesis, given each potential representation.

Once we get a collection of Observed and Expected scores for a potential representation, how do we decide if it's "good enough" with respect to matching adult-level productivity? To assess whether a potential representation is "good enough", I followed Bates et al. [2018]'s implementation of assessing agreement between Observed and Expected overlap using Lin's Concordance Correlation Coefficient (**LCCC**, represented with $\rho_c$: Lawrence and Lin [1989]) to generate one number that captures how well the Observed and Expected scores match and which can be more easily compared across potential representations.

More specifically, LCCC measures the agreement between two sets of observations on a scale from -1 to 1, with a $\rho_c$ of -1 indicating perfect disagreement, 1 indicating perfect agreement, and 0 indicating no agreement. So, given that there are multiple lexical overlap scores for each category representation (one for each legitimate multi-word combination within a particular category representation), I assess $\rho_c$ for the Observed vs. Expected overlap scores within that category representation.

In the example of Peter, we would have $\rho_c$ scores for each of the 16 possible category representations. Then, we then need to decide which representations have a "good enough" match LCCC-wise between Observed and Expected overlap. Unfortunately, there isn't a current consensus about what the threshold should be for good agreement with the LCCC [Altman, 1990, McBride, 2005]. Given this, as per Bates et al. [2018], I leverage each child's input data, with the idea that the adults producing the children's input had a fully productive

category representation ($\text{Rep}_{\text{PROD}}$). Because of this, the agreement between the Observed overlap in the child's input and the Expected overlap from the $\text{Rep}_{\text{PROD}}$ category representation could serve as a "good enough" threshold of agreement. More specifically, because we believe the $\text{Rep}_{\text{PROD}}$ category representation generated the child's input, the $\rho_c$ obtained for that representation is a reasonable cutoff for when a category representation in general matches sufficiently well with the observed data. The threshold value for each child is taken from each child's input $\rho_c$ when comparing the Expected overlap with a $\text{Rep}_{\text{PROD}}$ category representation against the Observed. This is the measure for when each child's possible category representations are sufficiently compatible with their output.

## 4.5 Child corpus statistics

I now describe the data from the different children, both typically developing and with SLI, that we'll be assessing for productive syntactic categories.

### 4.5.1 Typically developing two-year-old: Peter

The first child, Peter, is a typically developing child under 2;4 [Bloom et al., 1974, 1975] from the CHILDES database [MacWhinney, 2000b], and who we saw in the previous chapter when we were assessing immature FF-based categories against fully-adult productive categories. There, given the available category representation options, the fully-adult categories seemed a better fit. However, here we can more clearly compare different combinations of fully-adult categories and non-adult categories.

Peter's potential categories, whose combinations appeared more than 100 times, consist of Determiner, Noun, Verb, and Pronoun. Table 4.1 shows the frequency of the individual potential categories and 2-word combinations. On the left is the data from the

child-directed input; on the right is the data from the child's productions, or output.

Table 4.1: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Peter's input and output.

| **Peter, Age 2** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| DETERMINER | 16 | 1104 | 11 | 562 |
| NOUN | 323 | 1104 | 188 | 562 |
| VERB | 129 | 761 | 59 | 325 |
| PRONOUN | 36 | 761 | 20 | 325 |
| 2-word combination | Types | Tokens | Types | Tokens |
| DETERMINER+NOUN | 456 | 1104 | 278 | 562 |
| VERB+PRONOUN | 280 | 761 | 113 | 325 |

### 4.5.2 SLI children at three- and four-years-old: Daniel, Nathan, Harry, and Bonnie

I looked at four children identified with SLI: Daniel, Nathan, Harry, and Bonnie, which come from the Conti-Ramsden-3 corpus in the CHILDES database [Joseph et al., 2002, MacWhinney, 2000b]. Two of the children (Harry and Nathan) have data at both 3 and 4 years of age, with a wider spread for age three (Harry 3 (3;05-3;11), Harry 4 (4;0-4;08), & Nathan 3 (3;0-3;11), Nathan 4 (4;0-4;03)). Daniel only has data at age three (3;0-3;11) and Bonnie only has data at age four (4;0-4;11). Tables 4.2-4.5 show the frequency of the individual potential categories and 2-word combinations that appeared over 100 times in each child's data sample. This included 6 potential categories: ADJECTIVE, ADVERB, DETERMINER, NOUN, PREPOSITION, and VERB.

Table 4.2: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Daniel's input and output.

| **Daniel, age three** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 81 | 490 | 31 | 109 |
| Adverb | 59 | 756 | 25 | 246 |
| Determiner | 22 | 3136 | 18 | 891 |
| Noun | 868 | 4601 | 375 | 1308 |
| Preposition | 25 | 737 | 13 | 194 |
| Verb | 161 | 994 | 63 | 362 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 293 | 490 | 76 | 109 |
| Preposition+Noun | 210 | 737 | 65 | 194 |
| Determiner+Noun | 1126 | 3136 | 424 | 891 |
| Verb+Adverb | 267 | 756 | 83 | 246 |
| Verb+Noun | 200 | 238 | 91 | 114 |

Table 4.3: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Harry's input and output.

| **Harry, age three** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 84 | 501 | 40 | 143 |
| Adverb | 57 | 808 | 30 | 239 |
| Determiner | 24 | 2695 | 19 | 590 |
| Noun | 726 | 4114 | 273 | 1047 |
| Preposition | 24 | 681 | 15 | 207 |
| Verb | 149 | 1045 | 60 | 346 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 271 | 501 | 98 | 143 |
| Preposition+Noun | 195 | 681 | 72 | 207 |
| Determiner+Noun | 918 | 2695 | 328 | 590 |
| Verb+Adverb | 267 | 808 | 93 | 239 |
| Verb+Noun | 163 | 237 | 91 | 107 |
| **Harry, age four** | **Input** | | **Output** | |
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 73 | 320 | 38 | 109 |
| Adverb | 48 | 394 | 30 | 130 |
| Determiner | 20 | 1260 | 15 | 493 |
| Noun | 611 | 2019 | 265 | 736 |
| Preposition | 24 | 439 | 20 | 134 |
| Verb | 74 | 395 | 43 | 130 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 197 | 320 | 82 | 109 |
| Preposition+Noun | 192 | 439 | 67 | 134 |
| Determiner+Noun | 625 | 1260 | 293 | 493 |
| Verb+Adverb | 174 | 395 | 87 | 130 |

Table 4.4: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Nathan's input and output.

| **Nathan, age three** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 124 | 1344 | 80 | 680 |
| Adverb | 69 | 1160 | 51 | 711 |
| Determiner | 29 | 5046 | 19 | 1247 |
| Noun | 954 | 7761 | 500 | 2783 |
| Preposition | 30 | 1253 | 23 | 741 |
| Verb | 201 | 1278 | 124 | 826 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 530 | 1151 | 255 | 506 |
| Adverb+Adjective | 95 | 193 | 72 | 174 |
| Preposition+Noun | 260 | 1253 | 220 | 741 |
| Determiner+Noun | 1432 | 5046 | 619 | 1247 |
| Verb+Adverb | 341 | 967 | 207 | 537 |
| Verb+Noun | 239 | 311 | 195 | 289 |
| **Nathan, age four** | **Input** | | **Output** | |
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 50 | 235 | 27 | 105 |
| Adverb | 36 | 255 | 33 | 152 |
| Determiner | 22 | 1051 | 15 | 354 |
| Noun | 458 | 1547 | 241 | 637 |
| Preposition | 21 | 261 | 17 | 178 |
| Verb | 64 | 255 | 50 | 152 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 161 | 235 | 77 | 105 |
| Preposition+Noun | 103 | 261 | 82 | 178 |
| Determiner+Noun | 491 | 1051 | 242 | 354 |
| Verb+Adverb | 118 | 255 | 91 | 152 |

Table 4.5: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Bonnie's input and output.

| **Bonnie, age four** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 52 | 308 | 40 | 300 |
| Adverb | 47 | 430 | 35 | 492 |
| Determiner | 18 | 2330 | 15 | 1738 |
| Noun | 566 | 3225 | 442 | 2818 |
| Preposition | 22 | 466 | 16 | 596 |
| Verb | 103 | 551 | 68 | 676 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 176 | 308 | 129 | 300 |
| Preposition+Noun | 136 | 466 | 192 | 596 |
| Determiner+Noun | 762 | 2330 | 595 | 1738 |
| Verb+Adverb | 168 | 430 | 120 | 492 |
| Verb+Noun | 101 | 121 | 147 | 184 |

### 4.5.3 Typically developing age-controls at three- and four-years-old: Ross & Mark

I selected Mark and Ross from the MacWhinney corpus from the CHILDES dataset as age-matched controls for the SLI children (MacWhinney [1991, 2000b]). Both Mark and Ross have data at age three (3;0-3;11) and at age four (4;0-4;11). Tables 4.6-4.7 show the frequency of the individual potential categories and 2-word combinations that appeared over 100 times in both the child's input and output for each child. This included 6 potential categories for Mark at age three: Adjective, Adverb, Determiner, Noun, Preposition, and Verb, with the addition of Modal for Mark at age four and Ross at age three and four.

Table 4.6: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Ross's input and output.

| **Ross, age three** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 157 | 1345 | 138 | 1163 |
| Adverb | 78 | 745 | 65 | 639 |
| Determiner | 24 | 2582 | 20 | 3176 |
| Modal | 11 | 238 | 10 | 325 |
| Noun | 1122 | 4968 | 1164 | 5298 |
| Preposition | 29 | 754 | 25 | 630 |
| Verb | 233 | 1474 | 270 | 1567 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 613 | 1085 | 527 | 907 |
| Adverb+Adjective | 90 | 158 | 74 | 119 |
| Preposition+Noun | 375 | 754 | 319 | 630 |
| Determiner+Noun | 1148 | 2582 | 1456 | 3176 |
| Modal+Verb | 124 | 238 | 158 | 325 |
| Verb+Adjective | 81 | 102 | 94 | 137 |
| Verb+Adverb | 291 | 587 | 256 | 520 |
| Verb+Noun | 419 | 547 | 476 | 585 |
| **Ross, age four** | **Input** | | **Output** | |
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 119 | 668 | 110 | 421 |
| Adverb | 67 | 430 | 53 | 358 |
| Determiner | 25 | 1856 | 23 | 1773 |
| Modal | 10 | 236 | 11 | 194 |
| Noun | 921 | 3340 | 980 | 2937 |
| Preposition | 30 | 520 | 27 | 433 |
| Verb | 200 | 962 | 227 | 862 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 407 | 668 | 301 | 421 |
| Preposition+Noun | 300 | 520 | 288 | 433 |
| Modal+Verb | 120 | 236 | 127 | 194 |
| Determiner+Noun | 960 | 1856 | 1056 | 1773 |
| Verb+Adverb | 260 | 430 | 235 | 358 |
| Verb+Noun | 256 | 296 | 271 | 310 |

Table 4.7: Types and tokens of potential categories and multi-word combinations involving those categories in the two-word combinations in Mark's input and output.

| **Mark, age three** | **Input** | | **Output** | |
|---|---|---|---|---|
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 160 | 921 | 87 | 335 |
| Adverb | 73 | 550 | 49 | 289 |
| Determiner | 27 | 2311 | 20 | 1347 |
| Noun | 1124 | 4431 | 709 | 2209 |
| Preposition | 27 | 682 | 25 | 288 |
| Verb | 221 | 1067 | 135 | 528 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 572 | 921 | 225 | 335 |
| Preposition+Noun | 367 | 682 | 180 | 288 |
| Determiner+Noun | 1136 | 2311 | 732 | 1347 |
| Verb+Adverb | 319 | 550 | 173 | 289 |
| Verb+Noun | 405 | 517 | 190 | 239 |
| **Mark, age four** | **Input** | | **Output** | |
| Potential category | Types | Tokens | Types | Tokens |
| Adjective | 299 | 2026 | 161 | 824 |
| Adverb | 111 | 1534 | 81 | 736 |
| Determiner | 29 | 4469 | 24 | 2835 |
| Modal | 11 | 482 | 10 | 356 |
| Noun | 1865 | 8457 | 1120 | 4501 |
| Preposition | 33 | 1398 | 29 | 548 |
| Verb | 383 | 2580 | 293 | 1386 |
| 2-word combination | Types | Tokens | Types | Tokens |
| Adjective+Noun | 949 | 1631 | 467 | 705 |
| Adverb+Adjective | 252 | 395 | 94 | 119 |
| Preposition+Noun | 785 | 1398 | 322 | 548 |
| Modal+Verb | 220 | 482 | 170 | 356 |
| Determiner+Noun | 1959 | 4469 | 1358 | 2835 |
| Verb+Adverb | 551 | 1139 | 337 | 617 |
| Verb+Noun | 735 | 959 | 361 | 413 |

## 4.6 Results

The following results give the LCCC scores for the different representational possibilities for each child, given the adult threshold cutoff for that child. Recall that LCCC scores can range between -1 and 1, where 1 is perfect agreement and -1 is perfect disagreement. The adult threshold LCCC cutoff comes from comparing the adult Observed overlap in these same corpora against a fully PRODuctive representation, because we assume that adults are already using this representation.

**Results for Peter.** Table 4.8 shows the only compatible representation that exceeds the adult-derived threshold of 0.805. This representation contains DETERMINER, NOUN, and PRONOUN but not VERB. Peter's results show that there is some development of productive categories in typically developing children under age three, which at least contain DETERMINER, NOUN, and PRONOUN. Like Bates et al. [2018] found in their study with another typically developing child just under two-years-old, Peter's possible representations contains at least one closed-class category (DETERMINER and PRONOUN).

**Results for Daniel, Nathan, Harry, and Bonnie.** For the rest of the children in Tables 4.9-4.15, because there are so many possible permutations, I only show the representations that exceeded the adult-derived threshold for that child [1].

For Daniel, at three-years-old, the common thread between all of the compatible representations that exceed the adult threshold is that there are at least 2 productive categories. However, it is not clear what those two categories are.

For Harry, at three-years-old, a similar pattern to Daniel shows that the compatible rep-

[1] All of the associated category permutations and associated LCCC scores are available on my GitHub `https://github.com/galiabarsever/dissertation_files/`

Table 4.8: LCCC scores for the 16 possible category representations Peter could have, comparing his Observed lexical overlap against the lexical overlap Expected by each possible category representation. Representations with sufficient agreement (>0.805) are indicated in white cells.

| **Representation** | DET | NOUN | VERB | PRO | $\boldsymbol{LCCC}\rho_c$ |
|---|---|---|---|---|---|
| $\text{Rep}_{\text{NOT}}$ | ✗ | ✗ | ✗ | ✗ | 0.303 |
| $\text{Rep}_{\text{PRODUCTIVE}}$ | ✓ | ✓ | ✓ | ✓ | 0.561 |
| $\text{Rep}_{\text{SEMI}_1}$ | ✓ | ✓ | ✓ | ✗ | 0.723 |
| $\text{Rep}_{\text{SEMI}_2}$ | ✓ | ✓ | ✗ | ✓ | 0.821 |
| $\text{Rep}_{\text{SEMI}_3}$ | ✓ | ✗ | ✓ | ✓ | 0.313 |
| $\text{Rep}_{\text{SEMI}_4}$ | ✗ | ✓ | ✓ | ✓ | -0.227 |
| $\text{Rep}_{\text{SEMI}_5}$ | ✓ | ✓ | ✗ | ✗ | 0.501 |
| $\text{Rep}_{\text{SEMI}_6}$ | ✓ | ✗ | ✓ | ✗ | 0.460 |
| $\text{Rep}_{\text{SEMI}_7}$ | ✓ | ✗ | ✗ | ✓ | 0.476 |
| $\text{Rep}_{\text{SEMI}_8}$ | ✗ | ✗ | ✓ | ✓ | 0.214 |
| $\text{Rep}_{\text{SEMI}_9}$ | ✗ | ✓ | ✗ | ✓ | -0.039 |
| $\text{Rep}_{\text{SEMI}_{10}}$ | ✗ | ✓ | ✓ | ✗ | 0.008 |
| $\text{Rep}_{\text{SEMI}_{11}}$ | ✓ | ✗ | ✗ | ✗ | 0.388 |
| $\text{Rep}_{\text{SEMI}_{12}}$ | ✗ | ✓ | ✗ | ✗ | 0.094 |
| $\text{Rep}_{\text{SEMI}_{13}}$ | ✗ | ✗ | ✓ | ✗ | 0.318 |
| $\text{Rep}_{\text{SEMI}_{14}}$ | ✗ | ✗ | ✗ | ✓ | 0.315 |

resentations contain at least 2 productive categories, but again it is not clear what they are. This pattern is maintained at four-years-old, with many compatible representations, but they all contain at least 2 productive categories.

For Nathan, at three-years-old, the only compatible representation is a fully productive one, with the presence of all possible categories, ADJECTIVE, ADVERB, DETERMINER, NOUN, PREPOSITION, and VERB. At four-years-old, there appear to be more compatible representations (including a fully productive one) where every representation contains at least 2 productive categories, but it is not clear which categories those are. However, it is likely that if he developed a fully productive representation at 3, he would have kept it at 4, which is one of the compatible representations at 4.

For Bonnie, at four-years-old, all the compatible representations contain ADJECTIVE, ADVERB, PREPOSITION, and VERB, and possibly also DETERMINER and NOUN.

In contrast to Peter, the SLI child results are more difficult to interpret. Most of the SLI children seem to have many possible representations that they could be entertaining, or that are above the adult-derived cutoff. In all of these potential representations, the children across the board have at least 2 productive categories, with a couple of children certainly having more than that (Nathan at three-years-old with a fully productive representation and likely keeping it at 4, and Bonnie at four-years-old with 4 productive categories). These disparate profiles are in line with what we expect from such a heterogeneous population as SLI.

Table 4.9: Possible category representations for Daniel at 3 yo (adult cutoff: 0.50).Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | 0.54 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.68 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.70 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | 0.53 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.55 |
| ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | 0.54 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.69 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.70 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.53 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.55 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | 0.51 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.52 |
| ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | 0.53 |
| ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.71 |
| ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.74 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | 0.50 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.54 |
| ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | 0.53 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.50 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.50 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.50 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.50 |

**Results for Ross & Mark.** For Ross, at three-years-old, the only compatible representation contains Adjective, Adverb, Modal, Noun, Preposition, and perhaps Verb,

Table 4.10: Possible category representations for Harry at age three (adult cutoff: 0.39). Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.52 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.60 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.42 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.50 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.59 |
| ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.42 |
| ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | 0.41 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | 0.40 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.45 |
| ✓ | ✓ | ✗ | ✗ | ✗ | ✓ | 0.41 |
| ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.65 |
| ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.79 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | 0.44 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.58 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.65 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.80 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.40 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.56 |

Table 4.11: Possible category representations for Nathan at age three (adult cutoff: 0.68). Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.68 |

Table 4.12: Possible category representations for Bonnie at age four (adult cutoff: 0.69). Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.77 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.73 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.80 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.75 |

but not Determiner. At four-years-old, Determiner becomes one of the categories that are certainly in the possible representations, with possibly also Adverb or Preposition, and a fully productive representation also meets the threshold.

For Mark, at three-years-old, he certainly has Adjective, Adverb, Noun, and Verb and perhaps Preposition, but not Determiner. At four-years-old, the same certain

Table 4.13: Possible category representations for Harry at age four (adult cutoff: 0.41). Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | 0.47 |
| ✗ | ✗ | ✗ | ✓ | ✓ | ✗ | 0.42 |
| ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.51 |
| ✗ | ✗ | ✓ | ✗ | ✓ | ✗ | 0.44 |
| ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | 0.54 |
| ✗ | ✗ | ✓ | ✓ | ✗ | ✓ | 0.43 |
| ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | 0.49 |
| ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.60 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.47 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.64 |
| ✗ | ✓ | ✗ | ✓ | ✗ | ✓ | 0.42 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✗ | 0.53 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.74 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.51 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.70 |
| ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | 0.44 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.58 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.83 |
| ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | 0.46 |
| ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.52 |
| ✓ | ✗ | ✓ | ✗ | ✓ | ✗ | 0.42 |
| ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | 0.55 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | 0.49 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.63 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.47 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.43 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.53 |

categories carry over with the possibility of Determiner, the introduction of Modal, and still possibly Preposition.

Both Ross and Mark appear to have at least 3 productive categories for each of the representations at each age and appear to be have consistently more categories that are clearly productive and fewer representations that are above the adult-derived cutoff for "good enough". This issue of the number of matching representations may be due to low agreement between the adult observed overlap and a fully productive representation (0.805 for typically developig two-year-old, 0.68-0.72 for the typically developing three- and four-year-olds, 0.39-0.70

Table 4.14: Possible category representations for Nathan at age four (adult cutoff: 0.48). Each row is a possible candidate representation, with its LCCC in the right hand column.

| Adj | Adv | Det | Noun | Prep | Verb | $\rho_c$ |
|---|---|---|---|---|---|---|
| ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.49 |
| ✗ | ✗ | ✓ | ✗ | ✓ | ✓ | 0.53 |
| ✗ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.57 |
| ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | 0.53 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✗ | 0.51 |
| ✗ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.77 |
| ✗ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.66 |
| ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | 0.51 |
| ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | 0.62 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✗ | 0.61 |
| ✗ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.91 |
| ✗ | ✓ | ✓ | ✓ | ✗ | ✓ | 0.51 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.48 |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.81 |
| ✓ | ✗ | ✗ | ✗ | ✓ | ✓ | 0.50 |
| ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.54 |
| ✓ | ✗ | ✓ | ✗ | ✓ | ✓ | 0.59 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.64 |
| ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | 0.62 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.80 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.62 |

for the SLI three- and four-year-olds).All of the adult thresholds are lower than the one in Bates et al. [2018], but it is unclear whether that is unusual or simply a result of differences in the data. I discuss this in more detail in the next section.

**Results Summary.** My results suggest that each child and each population presents a slightly different profile. At a little over two-years-old, a typically developing child shows the presence of both closed-class and open-class categories. The typically developing children at age three and four consistently have at least 3 productive categories in their possible representations. The SLI children present a slightly messier picture even at three- and four-years-old, with all compatible representations having at least 2 productive categories; however, from the available data it is impossible to tell which categories they are. I'll discuss each of these findings in turn.

Table 4.15: Control kids categories

| Age | Child | ADJ | ADV | DET | MOD | NOUN | PREP | VERB | $\rho_c$ |
|---|---|---|---|---|---|---|---|---|---|
| 3 | Mark (0.70) | ✓ | ✓ | ✗ | N/A | ✓ | ✗ | ✓ | 0.73 |
| | | ✓ | ✓ | ✗ | N/A | ✓ | ✓ | ✓ | 0.73 |
| 3 | Ross (0.72) | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✗ | 0.73 |
| | | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.74 |
| 4 | Mark (0.70) | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✓ | 0.76 |
| | | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | 0.72 |
| | | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✓ | 0.80 |
| | | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.76 |
| | | ✓ | ✓ | ✓ | ✗ | ✓ | ✗ | ✓ | 0.74 |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | 0.78 |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.73 |
| 4 | Ross (0.68) | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | 0.70 |
| | | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.75 |
| | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.82 |

## 4.7 Discussion & Future work

### 4.7.1 Productive categories in a typically developing two-year-old

As I mentioned above, Peter shows the presence of closed-class categories (DETERMINER and PRONOUN) similar to the two-year-old in Bates et al. [2018]. As in Bates et al. [2018], I did not detect the presence of VERB at two years old. Where Peter differs from the younger child in Bates et al. [2018] is that Peter has at least one productive open-class category as well (NOUN). While there is evidence that a rudimentary NOUN category may develop in children as early as 14 months [Booth and Waxman, 2003], there is no guarantee that an adult-level noun category would be present by two years old. Peter, however, is slightly older (at 2;4) than the child in Bates et al. [2018] (who was 1;8-2;0), so it may be that an adult-like NOUN category develops between the ages of 2;0 and 2;4. It is also true that although the adult-derived threshold for Peter (0.805) is lower than what Bates et al. [2018] found in their study with a two-year-old (0.901), we simply don't know what the normal threshold for typically developing two-year-olds is because at this point, there are only two case studies. Further studies looking at typically developing children of the same age will give a clearer

picture of what the expected threshold is in this population.

Moreover, recall that in the previous chapter's analysis, we found that Peter's output was most compatible with a fully adult-like representation, containing (at least) the four categories analyzed here, when compared against immature FF-based categories. However, here we find that a better-fitting representation is one where VERB is actually immature, while the other categories (NOUN, DETERMINER, and PRONOUN) are adult-like. Thus, this analysis allows us more precision than what we saw before.

Given this, it's therefore possible to apply the analysis used in this chapter more broadly to include specific immature categories. That is, we don't have to simply test for the binary presence or absence of adult-like productive categories in young children – we can also assess the presence of productive immature categories in the same way I do here. For example, we could consider the FF-based categories from the previous chapter as a possible immature representation – these are non-adult categories, but they are nonetheless categories, and so can behave productively just as the adult categories I assessed in this chapter. Given this similarity, I would be able to assess which particular FF-based categories might be present by using the fully-productive metric to assess the presence or absence of different immature categories.

### 4.7.2 Productive categories in three-and four-year-olds: Typically and atypically developing populations

While delay of the production of 2-word combinations is a developmental hallmark of SLI children, it is possible that by the age of three, both typically developing and SLI children may have a large number of productive categories. In trying to interpret the potential category representations for each child, there are some things to consider. In general, there are three different explanations for patterns we could observe in the results. First, if some

category knowledge is generally harder to acquire than others, we should see a general pattern between children, with some categories consistently acquired at two, some at three, and some at four. However, the results here suggest that we do not in fact see this kind of general pattern. More specifically, across the SLI and typically developing three- and four-year-olds, there does not seem to be a strong general pattern of which categories may be acquired first.

So, what could be going on? The differences between productive category profiles could be the result of: (I) a child-internal variable – that is, there are some internal cognitive abilities of an individual child that makes them more or less likely to acquire a particular adult-like category before another one, or (II) a child-external variable – that is, a difference between the input that the children are getting that makes one category more easily (or sooner) acquired than another one. Let's now consider the results in light of each of these hypotheses.

First, looking at typically developing children, Mark and Ross seem to have more in common in their productive category development than any of the other children. At four they both consistently have Adjectives, Nouns, Verbs, and likely Adverbs), and both may have Determiners, Modals, and Prepositions. However, this is markedly different from Peter, also typically developing, who seems to have an adult-like Determiner category at two. So, why are Mark and Ross so similar to each other, while showing this marked difference to Peter? This similarity may be due to their child-internal factors (i.e. shared cognitive abilities as typically developing children), or due to child-external factors, (i.e. the similar input they get as siblings with shared caretakers, who are likely giving the same or similar input to each sibling). Their divergence from Peter likewise could also be explained by either child-internal individual differences or systematic child-external differences in input. Further studies could test other typically developing two-, three-, and four-year-olds to analyze what a typically developing child's profile looks across a greater number of children at these ages. To further untangle the role of the child's input, it would take a study where children who

were likely to share child-internal factors (like twins) were separated at birth and raised in different households with different caretaker input. If we see differences between the children's category representation development, we can claim that child-external factors, like input, might have more to do with the category learning trajectory than child-internal ones.

When interpreting the results for the SLI children, it is important to note that the child-internal factors, (i.e. presentation of SLI symptoms and associated linguistic skills) might vary widely. The only consistent pattern between the SLI children is that it is likely that they are working with at least a couple productive categories; however, because of the wide variation both between children and in the large number of compatible representations, it is difficult to say what those categories might be. This large spread of compatible representations is due to the generally low adult-directed LCCC match thresholds for each child. The difference between the threshold ranges is striking (0.68-0.72 for the typically developing three- and four-year-olds and 0.39-0.70 for the SLI three- and four-year-olds). Not only is there a wider variation of thresholds in SLI children, these lower thresholds indicate a lower level of agreement between adult productions and a fully PRODuctive representation, and this results in many more representations that meet this threshold. One way to interpret this difference is as a true difference in input, which would then serve as a child-external factor for why we see the variation we do in the SLI children. In particular, the low level of agreement in the adult speech with using a fully productive representation could be because adults are in fact not using a fully productive representation while talking to SLI children. This is not implausible, as it is well-documented that adults modify their speech when directing their speech at children at particular ages (Weisleder and Fernald [2013], Rowe [2012]). Recall also that in Chapter 2 that I found that adults are changing the representation they use to talk to children versus the one they use when talking to adults.

If the adults are indeed changing how they speak to SLI children at the representational

level, we could no longer assume that the adults are using a fully productive representation in their own speech. To determine which representation the adults are likely to be using, we could calculate the observed and expected overlap for all the potential adult representations, instead of simply the fully productive representation as I do here, and then pick the representation that has the highest LCCC as the adult-derived threshold. If this representation is not the fully-productive representation, adults are likely modifying the input that children are getting at the representational level. This less-than-fully-productive representation might serve as a better indication of the target knowledge for SLI children at a particular stage, and the LCCC threshold associated with it may serve to better disambiguate possible developing SLI child category representations as opposed to using the LCCC threshold score from an adult fully-productive representation.

Another consideration is the uneven spread of data between the SLI children who have data at both three and four (Harry and Nathan). Most of the data for the children at three is between 3;0 and 3;11, with much less data around the early stages of four (4;0-4;08). If we consider that developing adult-like categories can happen very quickly (possibly within a few months), then presenting results from data over the span of a year may disguise some of the development of these adult-like categories. For example, if a child does not have a productive NOUN category at the beginning of the year three, but they develop it at the end of the year, taking all of the data within that year may muddy the actual categories present for that child. In the future, it would be useful to divide the data over the year into smaller sections that may more accurately reflect the developmental stage of the child's category representations at a particular time. This would be useful not just for the SLI children, but for all of the children's data from both typically and atypically developing populations.

What we can take away from all the populations is that some form of a semi-productive representation, with some categories being fully productive and others perhaps not, appears to be normal up to the age of four. This accords with what we know about children's linguistic

development, since at four, in typically developing populations, knowledge of syntax and morphology is still developing [Pine and Lieven, 1997, Tomasello, 2004, Kemp et al., 2005, Tomasello and Brandt, 2009, Theakston et al., 2015]. As children age and as we see children produce more complex utterances and use more complex syntax later on in development, we would expect these semi-productive representations to fade away to more fully productive ones, at least in typically developing populations. It is unclear what the developmental trajectory of atypically developing populations is. It is possible that, as SLI children usually do not catch up with their peers in linguistic skill, they may never reach a typically developing adult-like fully productive representation.

# Chapter 5

# Conclusion

My research broadly demonstrates how quantitative approaches can be effectively leveraged for developmental research. In this dissertation, I've shown one quantitatively precise way to identify the nature of developing mental representations in a variety of domains; my approach utilizes the connection between a learner's input, creation of a potential mental representation from that input, and evaluation with respect to the learner's output. More specifically, the quantitative approach I use leverages both realistic input data and realistic output data as part of the model design and evaluation. Using modeling, we have the opportunity to concretely evaluate representational options that we would not otherwise be able to disambiguate, and uncover the developmental trajectory of a child that might otherwise have been opaque. In the previous chapters, I demonstrated this quantitative approach with three case studies in language development: (I) the development of adjective ordering preferences, where I found that the representations that adults use to talk to children are different than the ones used to talk to other, adults, (II) immature individual syntactic category representations, where I identified precisely which immature category representation young children are likely to be using, and (III) the development of adult productive syntactic category representations, where I identified when adult category knowledge emerges in

typically and atypically developing populations.

Utilizing this framework demonstrates the explanatory power of the framework, as well as highlighting the importance of the assumptions that were made for each of the implementations of this framework. For adjective ordering preferences, assumptions included the judgements that were made about what the null hypothesis should be, how the adjectives should be placed in their appropriate lexical classes, and the probability calculation itself. For immature syntactic category representations, assumptions included choosing which kind of frequent frames to assess as well as the assumption of a bigram generative model for generating observable output. For more mature syntactic category representations assumptions includes the binary presence or absence of an adult-level productive categories. All of these assumptions impact the results of these models, and choices at any point (i.e., anything from how to represent the input data to how to evaluate on the output data) effects what representation is deemed as the best fit from the model as we've designed it.

From a methodological standpoint, this approach allows researchers to better utilize more of the data available in existing corpora. In each of the case studies, by specifying the possible mental representations, whether possible representations of adjective ordering preferences, possible immature syntactic category representations, or collections of productive adult-level categories, we can connect available input data to child output data at specific ages. By using both the input in the creation of the possible representations and the output in assessing the relative probability of the representations, researchers can better harness the data available to them.

Being able to use more of the existing data means that we can use more of the existing data available for populations where the data is less abundant. Most computational modeling work in language acquisition is focused on typically-developing children that receive typical input. The process laid out in these chapters, while having mostly been applied to children from the typically-developing population, could also reasonably be applied to children from

a variety of backgrounds and abilities (as shown with the SLI children for mature syntactic category development). However, what makes running an analysis like this difficult on atypically-developing populations is that it can be difficult to find enough coded data from a relevant population because (i) the nature of the development means that the output itself may be limited (i.e. Late Talkers, SLI), and (ii) that there is a general dearth of available coded data on these populations. To alleviate part of this problem for the field as a whole, we need more linguistically-annotated data from atypically-developing populations. More data, and using frameworks that maximize the usefulness of the data like the one described here, will help bring us closer to understanding developing linguistic knowledge.

# Bibliography


Douglas G Altman. *Practical statistics for medical research*. CRC press, 1990.

Galia Bar-Sever and Lisa Pearl. Syntactic categories derived from frequent frames benefit early language processing in english and asl. In *Proceedings of the 40th Annual Boston University Conference on Child Language Development*, pages 32–46, 2016.

Galia Bar-Sever, Rachael Lee, Gregory Scontras, and Lisa Pearl. Little lexical learners: Quantitatively assessing the development of adjective ordering preferences. In *Proceedings of the 42nd Annual Boston University Conference on Child Language Development*, pages 58–71, 2018.

Chris Barker. Negotiating Taste. *Inquiry*, 56(2-3):240–257, 2013. ISSN 0020-174X.

Alandi Bates and Lisa Pearl. When socioeconomic status differences dont affect input quality: Learning complex syntactic knowledge. In *Proceedings of the 43rd Annual Boston University Conference on Language Development [BUCLD 43]*, 2019.

Alandi Bates, Lisa Pearl, and Sue Braunwald. I can believe it: Quantitative evidence for closed-class category knowledge in an English-speaking 20- to 24-month-old child. In *Proceedings of the Berkeley Linguistics Society*, 2018.

Elizabeth Bates, Inge Bretherton, and Lynn Sebestyen Snyder. *From first words to grammar: Individual differences and dissociable mechanisms*, volume 20. Cambridge University Press, 1991.

Thomas G Bever. The cognitive basis for linguistic structures. *Cognition and the development of language*, 279(362):1–61, 1970.

Lois Bloom, Lois Hood, and Patsy Lightbown. Imitation in language development: If, when, and why. *Cognitive Psychology*, 6(3):380–420, 1974.

Lois Bloom, Patsy Lightbown, Lois Hood, Melissa Bowerman, Michael Maratsos, and Michael P Maratsos. Structure and variation in child language. *Monographs of the society for Research in Child Development*, pages 1–97, 1975.

Amy Booth and Sandra Waxman. Mapping words to the world in infancy: On the evolution of expectations for nouns and adjectives. *Journal of Cognition and Development*, 4(3): 357–381, 2003.

Peter F Brown, Vincent J Della Pietra, Robert L Mercer, Stephen A Della Pietra, and Jennifer C Lai. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1):31–40, 1992a.

Peter F Brown, Vincent J Della Pietra, Robert L Mercer, Stephen A Della Pietra, and Jennifer C Lai. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40, 1992b.

Roger Brown. *A first language: The early stages*. Harvard University Press, Cambridge, MA, 1973.

Montserrat Capdevila i Batet and Mireia Llinàs i Grau. The acquisition of negation in English. *Atlantis*, pages 27–44, 1995.

Seth Chaiklin. The zone of proximal development in vygotskys analysis of learning and instruction. *Vygotskys educational theory in cultural context*, 1:39–64, 2003.

Emmanuel Chemla, Toben H Mintz, Savita Bernal, and Anne Christophe. Categorizing words using frequent frames: what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental science*, 12(3):396–406, 2009a.

Emmanuel Chemla, Toben H Mintz, Savita Bernal, and Anne Christophe. Categorizing words using 'Frequent Frames': What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, 12(3):396–406, 2009b.

Guglielmo Cinque. On the evidence for partial N-movement in the Romance DP. In R S Kayne, G Cinque, J Koster, J.-Y. Pollock, Luigi Rizzi, and R Zanuttini, editors, *Paths Towards Universal Grammar. Studies in Honor of Richard S. Kayne*, pages 85–110. Georgetown University Press, Washington DC, 1994. ISBN 087840287X.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. Starting from scratch in semantic role labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 989–998. Association for Computational Linguistics, 2010.

Michael Connor, Cynthia Fisher, and Dan Roth. Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition*, pages 257–296. Springer, 2013.

Kathryn M Dewar and Fei Xu. Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*, 21(12):1871–1877, 2010.

F Babette Diepeveen, Elise Dusseldorp, Gerard W Bol, Anne Marie Oudesluys-Murphy, and Paul H Verkerk. Failure to meet language milestones at two years of age is predictive of specific language impairment. *Acta Paediatrica*, 105(3):304–310, 2016.

Robert MW Dixon. *Where have all the adjectives gone? and other essays in semantics and syntax*, volume 107. Walter de Gruyter, 1982.

Marian Erkelens. *Learning to categorize verbs and nouns: studies on Dutch*. Netherlands Graduate School of Linguistics, 2009.

Julia L Evans, Jenny R Saffran, and Kathryn Robe-Torres. Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 2009.

Naomi H Feldman, Emily B Myers, Katherine S White, Thomas L Griffiths, and James L Morgan. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3):427–438, 2013.

Charles A Ferguson. Baby talk in six languages. *American anthropologist*, 66(6_PART2): 103–114, 1964.

Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, 16(3):477–501, 1989.

Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15, 2002. doi: 10.1111/1467-8721.00158.

Ruthe Foushee and Mahesh Srinivasan. Could both be right? Children's and adults' sensitivity to subjectivity in language. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, pages 379–3384. Cognitive Science Society, London, UK, 2017.

Ruthe Foushee and Mahesh Srinivasan. Faultless disagreement judgments track adults' estimates of population-level consensus over adjective-referent pairs. In *CogSci*, 2018.

Stella Frank, Sharon Goldwater, and Frank Keller. Evaluating models of syntactic category acquisition without using a gold standard. In *Proceedings of the 31st Annual Conference of the Cognitive Science Sociey*, pages 2576–2581, 2009.

Michael Franke, Gregory Scontras, and Mihael Simonic. Subjectivity-based adjective ordering maximizes communicative success. to appear.

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

LouAnn Gerken and Sara Knight. Infants generalize from just (the right) four words. *Cognition*, 143:187–192, 2015.

LouAnn Gerken and Carolyn Quam. Infant learning is influenced by local spurious generalizations. *Developmental Science*, 20(3), 2017.

Rebecca L Gomez and LouAnn Gerken. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2):109–135, 1999.

Rebecca L Gómez and Laura Lakusta. A first step in form-based category abstraction by 12-month-old infants. *Developmental science*, 7(5):567–580, 2004.

DiAnne L Grieser and Patricia K Kuhl. Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental psychology*, 24(1):14, 1988.

Ariel Gutman, Isabelle Dautriche, Benoît Crabbé, and Anne Christophe. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition*, 22:285–309, 2014.

Eileen Haebig, Jenny R Saffran, and Susan Ellis Weismer. Statistical word learning in children with autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry*, 58(11):1251–1263, 2017.

Michael Hahn, Richard Futrell, and Judith Degen. Exploring adjective ordering preferences via artificial language learning. California Meeting on Psycholinguistics, 2017.

Victoria Chou Hare and Wayne Otto. Development of preferred adjective ordering in children, grades one through five. *The Journal of Educational Research*, pages 190–193, 1978.

Robert Hetzron. On the relative order of adjectives. In H Seiler, editor, *Language Universals*, pages 165–184. Tübingen, 1978.

Felix Hill. Beauty before age?: applying subjectivity to automatic english adjective ordering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 11–16. Association for Computational Linguistics, 2012.

Elizabeth K Johnson, Amanda Seidl, and Michael D Tyler. The edge factor in early word segmentation: utterance-level prosody enables word form extraction by 6-month-olds. *PloS one*, 9(1):e83546, 2014.

Kate L Joseph, Ludovica Serratrice, and Gina Conti-Ramsden. Development of copula and auxiliary be in children with specific language impairment and younger unaffected controls. *First Language*, 22(2):137–172, 2002.

Nenagh Kemp, Elena Lieven, and Michael Tomasello. Young children's knowledge of the determiner and adjective categories. *Journal of Speech, Language, and Hearing Research*, 48(3):592–609, 2005.

Christopher Kennedy. Two Sources of Subjectivity: Qualitative Assessment and Dimensional Uncertainty. *Inquiry*, 56(2-3):258–277, 2013. ISSN 0020-174X.

Max Kölbel. Faultless Disagreement. *Proceedings of the Aristotelian Society*, 104:53–73, 2004.

Patricia K Kuhl, Jean E Andruski, Inna A Chistovich, Ludmilla A Chistovich, Elena V Kozhevnikova, Viktoria L Ryskina, Elvira I Stolyarova, Ulla Sundberg, and Francisco Lacerda. Cross-language analysis of phonetic units in language addressed to infants. *Science*, 277(5326):684–686, 1997.

Richard Kunert, Raquel Fernández, and Willem Zuidema. Adaptation in child directed speech: Evidence from corpora. *Proc. SemDial*, 112119, 2011.

Christopher Laenzlinger. French adjective ordering: perspectives on DP-internal movement types. *Lingua*, 115:645–689, 2005.

Jill Lany and Rebecca L Gómez. Twelve-month-old infants benefit from prior experience in statistical learning. *Psychological Science*, 19(12):1247–1252, 2008.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

Laurence B Leonard. Specific language impairment across languages. *Child development perspectives*, 8(1):1–5, 2014.

Emiddia Longobardi, Clelia Rossi-Arnaud, Pietro Spataro, Diane L Putnick, and Marc H Bornstein. Children's acquisition of nouns and verbs in italian: contrasting the roles of frequency and positional salience in maternal language. *Journal of Child Language*, 42 (01):95–121, 2015.

Brian MacWhinney. The childes language project: Tools for analyzing talk. *Hillsdale, NJ: Lawrence Erlbaum. Miller, J.(1981). Assessing language production in children: Experimental procedures. Austin, TX: ProEd. Oetting, J., & Horohov, J.(1997). Past-tense marking by children with and without specific language impairment. Journal of Speech, Language and Hearing Research*, 40:62–74, 1991.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000a.

Brian MacWhinney. *The CHILDES project: The database*, volume 2. Psychology Press, 2000b.

James E Martin and Dennis L Molfese. Preferred adjective ordering in very young children. *Journal of Verbal Learning and Verbal Behavior*, 11(3):287–292, 1972.

Jessica Maye, Janet F Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, 2002.

Jessica Maye, Daniel J Weiss, and Richard N Aslin. Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, 11(1):122–134, 2008.

GB McBride. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA Client Report: HAM2005-062*, 2005.

Stephan C Meylan, Michael C Frank, Brandon C Roy, and Roger Levy. The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, 28(2): 181–192, 2017.

Toben Mintz. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90:91–117, 2003a.

Toben H Mintz. Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117, 2003b.

Toben H Mintz. Finding the verbs: Distributional cues to categories available to young learners. *Action meets word: How children learn verbs*, pages 31–63, 2006.

Lisa Pearl and Jon Sprouse. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20(1):23–68, 2013.

Lawrence Phillips and Lisa Pearl. Utility-based evaluation metrics for models of language acquisition: A look at speech segmentation. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. NAACL, Denver, Colorado, 2015.

J.M Pine and E.V. Lieven. Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics*, 18(02):123–138, 1997.

Julian M Pine, Daniel Freudenthal, Grzegorz Krajewski, and Fernand Gobet. Do young children have adult-like syntactic categories? Zipf's law and the case of the determiner. *Cognition*, 127(3):345–360, 2013.

Steven Pinker. *Language learnability and language development.* Harvard University Press, Cambridge, MA, 1984.

Alexa R Romberg and Jenny R Saffran. Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914, 2010.

Meredith L Rowe. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, 83(5):1762–1774, 2012.

Caroline F Rowland and Anna L Theakston. The acquisition of auxiliary syntax: A longitudinal elicitation study. Part 2: The modals and auxiliary DO. *Journal of Speech, Language, and Hearing Research*, 52(6):1471–1492, 2009.

Johanna M Rudolph and Laurence B Leonard. Early language milestones and specific language impairment. *Journal of Early Intervention*, 38(1):41–58, 2016.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, pages 1926–1928, 1996.

Jenny R. Saffran, Ann Senghas, and John C. Trueswell. The acquisition of language by children. *Proceedings of the National Academy of Sciences*, 98(23):12874–12875, 2001. ISSN 0027-8424. doi: 10.1073/pnas.231498898. URL `https://www.pnas.org/content/98/23/12874`.

Kuniyoshi L Sakai. Language acquisition and brain development. *Science*, 310(5749):815–819, 2005.

Gregory Scontras, Judith Degen, and Noah D Goodman. Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1:53–65, 2017.

Gary-John. Scott. Stacked adjectival modification and the structure of nominal phrases. In G Cinque, editor, *The cartography of syntactic structures, Volume 1: Functional structure in the DP and IP*, pages 91–120. Oxford University Press, Oxford, 2002.

Amanda Seidl and Elizabeth K Johnson. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9(6):565–573, 2006.

Yu Kyoung Shin. A New Look at Determiners in Early Grammar: Phrasal Quantifiers. *Language Research*, 48(3):573–608, 2012.

Mohinish Shukla, Marina Nespor, and Jacques Mehler. An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, 54(1):1–32, 2007.

Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Catherine E Snow. The development of conversation between mothers and babies. *Journal of child language*, 4(1):1–22, 1977.

Richard Sproat and Chilin Shih. The cross-linguistic distribution of adjective ordering restrictions. In *Interdisciplinary approaches to language*, pages 565–593. Springer, 1991.

Barbara Stumper, Colin Bannard, Elena Lieven, and Michael Tomasello. "frequent Frames" in German Child-Directed Speech: A Limited Cue to Grammatical Categories. *Cognitive science*, 35(6):1190–1205, 2011.

Anna L Theakston and Caroline F Rowland. The acquisition of auxiliary syntax: A longitudinal elicitation study. part 1: Auxiliary be. *Journal of Speech, Language, and Hearing Research*, 52(6):1449–1470, 2009.

Anna L Theakston, Paul Ibbotson, Daniel Freudenthal, Elena VM Lieven, and Michael Tomasello. Productivity of noun slots in verb frames. *Cognitive Science*, 39(6):1369–1395, 2015.

Michael Tomasello. What kind of evidence could refute the ug hypothesis? *Studies in Language*, 28(3):642–645, 2004.

Michael Tomasello and Silke Brandt. Flexibility in the semantics and syntax of children's early verb use. *Monographs of the Society for Research in Child Development*, 74(2): 113–126, 2009.

Michael Tomasello and Jody Todd. Joint attention and lexical acquisition style. *First language*, 4(12):197–211, 1983.

Virginia Valian. Syntactic categories in the speech of young children. *Developmental Psychology*, 22(4):562, 1986.

Hao Wang and Toben Mintz. A dynamic learning model for categorizing words using frames. In Harvey Chan, Heather Jacob, and Enkeleida Kapia, editors, *Proceedings of the 32nd Annual Boston University Conference on Language Development [BUCLD 32]*, pages 525–536, Somerville, MA, 2008. Cascadilla Press.

Hao Wang, Barbara Höhle, NF Ketrez, Aylin C Küntay, Toben H Mintz, N Danis, K Mesh, and H Sung. Cross-linguistic distributional analyses with Frequent Frames: The cases of German and Turkish. In *Proceedings of 35th Annual Boston University Conference on Language Development*, pages 628–640. Cascadilla Press Somerville, MA, 2011.

Paul S Weiner. A language-delayed child at adolescence. *Journal of Speech and Hearing Disorders*, 39(2):202–212, 1974.

Adriana Weisleder and Anne Fernald. Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11):2143–2152, 2013.

Adriana Weisleder and Sandra R Waxman. What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in spanish and english. *Journal of Child Language*, 37(05):1089–1108, 2010.

Susan Ellis Weismer. Typical talkers, late talkers, and children with specific language impairment: A language endowment spectrum? In Rhea Paul, editor, *Language disorders from a developmental perspective*, pages 95–114. Psychology Press, 2010.

Ling Xiao, Xin Cai, and Thomas Lee. The development of the verb category and verb argument structures in Mandarin-speaking children before two years of age. In Yukio Otsu, editor, *Proceedings of the Seventh Tokyo Conference on Psycholinguistics*, pages 299–322, Tokyo, 2006. Hitizi Syobo.

Fei Xu and Joshua B Tenenbaum. Word learning as Bayesian inference. *Psychological Review*, 114(2):245, 2007.

Charles Yang. Who's Afraid of George Kingsley Zipf. Unpublished Manuscript, 2010.

Charles Yang. A statistical test for grammar. In *Proceedings of the 2nd workshop on Cognitive Modeling and Computational Linguistics*, pages 30–38. Association for Computational Linguistics, 2011.

# Appendix A

## A.1 AdjAdj strings excluded from analysis

Any AdjAdj string including adjectives that can also be used as adverbs (in both British and American English): brand (as in *brand new*) real (as in *real good*) jolly (as in *jolly good*) super (as in super fun) awful (as in *awful funny*) dead (as in *dead easy*) massive (as in *massive great*) wicked (as in *wicked nice*) mad (as in *mad crazy*) right (as in *right great*)

All counting adjectives, including: first, second, third, last, next

Other AdjAdj pairs exluded: *eensie weensie, eensy weensy, itsy bitsy, teeny tiny, hot cross, american hard, american short, dark haired, long haired, short haired, blonde haired, haired old, low fat, south central, regional high, gracious good, gracious great, south american, international high, poor sorry*

## A.2 Analysis files: Chapter 2

All lexical class and subjectivity assignments, as well as output results can be found on my GitHub repository: `https://github.com/galiabarsever/dissertation_files`

## A.3 Analysis files: Chapter 4

All possible category permutations and corresponding LCCC scores, regardless of whether they exceed the adult threshold, can by found on my GitHub repository: `https://github.com/galiabarsever/dissertation_files`