Variability and Consistency in Early Language Learning

The Wordbank Project

Michael Frank, Mika Braginsky, Virginia Marchman, and Daniel Yurovsky

Contents

Preface 7
- Overview . . . 7
- Outline . . . 9
- How to read this book . . . 10
- Acknowledgements . . . 10

1 Theoretical Foundations 13
- 1.1 The picture to date . . . 13
- 1.2 Making progress . . . 16
- 1.3 Variability and consistency . . . 17
- 1.4 Process universals . . . 19
- 1.5 Replication and theory-building: conclusions . . . 21

2 Practical Foundations 23
- 2.1 Measuring early vocabulary . . . 23
- 2.2 The logic of parent report . . . 25
- 2.3 Cross-linguistic comparison . . . 26
- 2.4 Wordbank . . . 28

3 Methods and Data 31
- 3.1 Database . . . 31
- 3.2 Datasets . . . 34

4 Measurement Properties of the CDI 47
- 4.1 Strengths and limitations of parent report . . . 47
- 4.2 Longitudinal stability of CDI measurements . . . 51
- 4.3 Psychometric modeling . . . 54
- 4.4 Conclusions . . . 62

5 Vocabulary Size 63
- 5.1 Central tendencies . . . 64
- 5.2 Variability between individuals . . . 73

6 Demographic Effects on Vocabulary Size 85
- 6.1 Sex . . . 85
- 6.2 Birth order . . . 96
- 6.3 Socioeconomic status . . . 103

7 Gesture and Communication 113
7.1 Measurement properties of CDI gestures . . . 114
7.2 The relationship between language and gesture . . . 119
7.3 Conclusions . . . 120

8 Consistency in Early Vocabulary 123
8.1 The first 10 words . . . 123
8.2 Global cross-linguistic similarity . . . 125
8.3 Acquisition similarity and linguistic similarity . . . 127
8.4 Consistency across development . . . 131
8.5 Conclusions . . . 133

9 Demographic Variation in Individual Words 135
9.1 Methods . . . 135
9.2 Results . . . 137
9.3 Conclusions . . . 147

10 Predictive Models of Individual Words 149
10.1 Methods . . . 151
10.2 Results . . . 154
10.3 Discussion . . . 157

11 Vocabulary Composition: Syntactic 161
11.1 Introduction . . . 161
11.2 Methods and data . . . 164
11.3 Results . . . 167
11.4 Discussion . . . 182

12 Vocabulary Composition: Semantic 183
12.1 Introduction and methods . . . 183
12.2 Global results . . . 185
12.3 Individual conceptual items . . . 192
12.4 Discussion . . . 199

13 Morphology, Grammar, and the Lexicon 201
13.1 Introduction . . . 201
13.2 Methods . . . 204
13.3 Results . . . 205
13.4 Discussion . . . 213

14 Individual Variation in Vocabulary 215
14.1 Variation in vocabulary composition . . . 217
14.2 “Spurts” in vocabulary . . . 223
14.3 Variation in production vs. comprehension . . . 228
14.4 Discussion . . . 237

15 Variability and Consistency 239
15.1 Methods . . . 239
15.2 Results and discussion . . . 240

16 Language Development at Scale 243
16.1 Summary 243
16.2 Generalizations 245
16.3 Learning processes 249
16.4 Methodological morals 253
16.5 Conclusions 253

A Individual Datasets 255

B Measures of Variability 263

C Stitching Across Forms 265

D Estimating Age of Acquisition 267

Preface

Overview

The emergence of children's early language is one of the most miraculous parts of human development. The ability to communicate using language arrives with incredible rapidity — most parents judge that their child is producing words with the intent to communicate before his or her first birthday (Schneider et al., 2015) and the onset of comprehension is even earlier (e.g., Bergelson and Swingley, 2012; Tincoff and Jusczyk, 1999).

New words enter children's expressive vocabularies slowly at first, but this process accelerates over the second year such that children reach an average of 300 words by 24 months and more than 60,000 by the time they graduate from high school (Fenson et al., 2007). At the same time, there are significant individual differences in the rate and timing of language acquisition. For example, although some 18-month-olds already produce 50–75 words, others produce no words at all, and will not do so until they are two years or older (e.g., Brown, 1973; Bloom, 2002; Clark, 2003).

How do children learn their first language? To what extent do different children and children learning different languages follow the same path into language? Are these paths similar or idiosyncratic? These questions about the consistency and variability of early language lead directly into the central question of language acquisition: What are the mechanisms that lead to the emergence of human language?

Answering this question is complicated immensely by the fact that there is no one single event that constitutes language acquisition. Early language learning involves the accumulation of thousands of words, grammatical rules, and constructions, and takes place over the course of years of growth and millions of separate interactions. This problem of timescales makes measurement a tremendous challenge. In addition, during the period in which language emerges, language ability varies wildly from child to child and most children are at best reluctant experimental participants. Accurate measurement of language development across individuals is thus a major challenge.

Parent report is one powerful method for addressing these. The MacArthur-Bates Communicative Development Inventory (CDI) is a simple survey instrument for measuring early language outcomes that was designed to address these issues.[1] The CDI is a checklist for parents to fill out to report on their child's progress in language. In different versions of the form, parents mark whether their child "says" or "understands and says" particular words out of a list of several hundred. Separate sections for gestures, word forms, and grammar are also present in some versions. Despite their simplicity, over

[1] For purposes of clarity and ease throughout we refer to CDIs (the family of instruments) rather than the MB-CDI (the particular English forms). We will use this term throughout even though technically some of our data come from "checklists" containing only vocabulary items rather than true Communicative Development Inventory forms.

the past 25 years of use, CDI forms have been shown to be reliable and valid measures of children's early language. In addition, CDI forms have been adapted to more than 100 different languages around the world. Research based on the CDI has contributed tremendously to our understanding of the growth of language in early childhood.

In this book, we examine the question of variability and consistency in early language through the lens of the CDI. Our book is an outgrowth of the Wordbank project (Frank et al., 2016a), which has as its goal to archive CDI data in a structured format so that they can be explored and analyzed in the service of describing early language. The database currently contains data from more than 82055 CDI form administrations across 29 languages. Wordbank is also continuously growing as new researchers contribute data. We believe this database is the largest and most diverse set of data on early language acquisition currently in existence.

Over the course of our work with Wordbank, we have developed a consistent framework for representing and analyzing CDI data. This framework allows us to unify a variety of influential previous analyses of CDI data. Just to take an example, one question of theoretical interest has been whether young children have an over-representation of nouns (names for things) in their vocabulary, and whether this trend is seen across languages. In Chapter 11, we develop an analytic method for measuring the size of this "noun bias" and apply the analysis to the Wordbank data. This measurement can then be compared across languages, and its variability can be estimated and compared to other measurements.

Thus, one first contribution of this book is to synthesize and unify previous work. Research in early language learning often builds off a fragmentary empirical picture, in which many important theoretical conclusions are based on analyses of transcripts from a small number of children, or analyses of experimental or parent-report data from English learners only. We hope that bringing together a large set of analyses of vocabulary data and implementing them consistently, openly, and reproducibly on the same dataset will help to create an empirical starting point for future work.

A second contribution of the book is to develop theory treating the question of consistency and variability in early language. We introduce the notion of "process universals" — that some aspects of the process of word learning may be universal across cultures and may lead to similarities in the dynamics of learning. These universals may arise due to the basic mechanisms of learning, memory, and social cognition that are at play in early vocabulary learning. This general idea has a long history in the field (e.g., Bates et al., 1989), but the Wordbank project provides an opportunity to lend new empirical data and analytic power to these ideas. This notion is contrasted with notions of "content" or "structural" universals in which particular principles regarding the structure of languages are innately given.

This set of contributions reflects an important guiding principle of our work here. Studies which at first glance seem like "mere replication" — in which a particular analysis is replicated on a larger dataset — are in fact important opportunities for theoretical development. Replicating with a larger dataset alone leads to a more precise estimate of the phenomenon, which can be used to confirm reliability, but also for quantitative comparisons and computational models. Further, increased precision allows for the examination of variability and consistency across meaningful units like children, words, instruments, or languages. Such estimates in turn are — as we argue throughout — relevant to foundational theoretical questions. Thus, there is never "mere" replication. More precise measurements sit hand and hand with better theory.

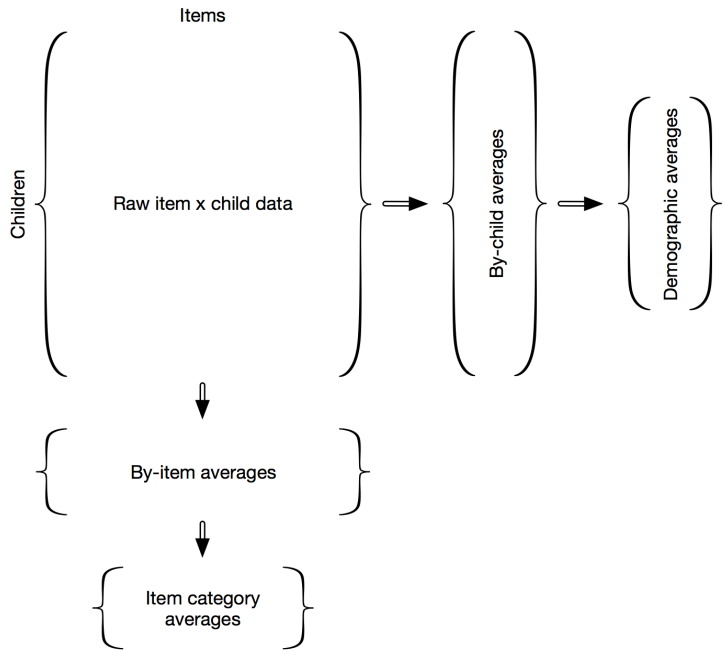


Figure 1: A graphical outline of the book.

Outline

The first chapters of this book provide an overview of the practical and theoretical issues that we cover. Chapter 1 gives a broad theoretical overview of our claims and sets up some of the empirical themes that we return to throughout, especially the notion of using the consistency and variability of phenomena to make generalizations about the process of acquisition. Chapter 2 then discusses the practicalities of the CDI and the Wordbank project. Chapter 9.1 gives a methodological and descriptive overview of the dataset we analyze throughout. And Chapter 4 discusses the psychometric properties of parent report data, addressing questions about the strengths and limitations of this method.

Chapters 5–14 then form the empirical heart of the book. Each applies a particular analysis of interest to our dataset. Although our goal is a full exploration of the phenomena of child language acquisition, the analyses we report are constrained by the structure of the data in Wordbank. At its heart, the individual instrument datasets stored in Wordbank are matrices of item by child data (see Figure P.1).

Considering the data in this way leads to a number of obvious data analytic strategies, many of which correspond directly to previous approaches to CDI data. For example, averaging across items leads to by-child averages, where each child receives a comprehension or production score. Chapter 5 considers this view of the data, examining developmental change and variability in such data. Then in Chapter 6, we report the ways that vocabulary growth varies across gender, birth order, and maternal education (a rough but cross-culturally valid proxy for socioeconomic status).

Averaging across the other margin of the data leads to by-item averages. These can be examined in a number of different ways. Gesture items and their growth trajectory — and relationship to overall vocabulary size — are examined in Chapter 7. Chapter 8 then considers the growth trajectories for

individual words, focusing especially on early vocabulary. Chapter 10 follows this approach and attempts to predict the trajectories of individual words based on both environmental and conceptual features of these words. This last approach calls for the incorporation of other resources, and so we use a variety of English and cross-linguistic resources to supplement Wordbank data in this chapter.

Next, we investigate the grouping of items into categories (both syntactic and semantic). Chapter 11 considers the categorical composition of early vocabulary, giving special consideration to the "noun bias" that is found in many — but not all — of the languages in Wordbank. Chapter 12 adopts the same approach for semantic, rather than syntactic, categories. This approach leads us to consider other aspects of morphosyntax that are reflected in the CDI forms. Chapter 13 explores the relationship between vocabulary growth and the growth of grammar. Finally, Chapter 14 returns to the question of individual variation using tools built up in previous chapters to quantify differences in the style and trajectory of learning across children.

The book ends with two synthesis chapters. Chapter 15 synthesizes observations across languages for the preceding chapters to quantify variability and consistency directly across phenomena. And the concluding chapter, Chapter 16, considers broadly the question of what the process of language acquisition looks like from the birds-eye view afforded by our data.

Finally, a number of appendices provide supplemental analyses, validating particular analytic practices that we adopt.

How to read this book

You can read this book as a narrative monograph. If you intend to do so, we recommend you read Chapters 1, 2, 9.1, and 4 before moving on to substantive chapters of interest.

You can also read this book as a reference. If you take this approach, just dive into any chapter that is of interest, knowing that you may need to use some of the terminology defined in Chapter 9.1 to interpret the constructs and analyses that are used. Further, you may have concerns about the reliability and validity of CDI-type instruments; some of these are addressed in Chapter 4.

A number of the analyses reported here were first described in earlier conference proceedings or publications (e.g., Frank et al., 2016a; Braginsky et al., 2015, 2016). Rather than reprinting these verbatim, this manuscript updates them using the unified analytic approach and larger dataset described in Chapter 9.1. The version of these analyses represented in this manuscript should be considered more definitive than any previously published or presented version.

The dataset in Wordbank is constantly growing and changing as we add new features, new data, and new languages. In addition, as users of the data occasionally identify issues and errors, we make will make corrections to the database. In writing this manuscript, we have attempted to find a middle ground between a completely dynamic document that responds to any change in the database, on the one hand, and a traditional, static book, on the other. A static manuscript would be a shame given the potential for dynamic extension and updating with new data. On the other hand, if the data were completely dynamic, any claim we made in prose would risk being out of date almost as soon as we wrote it.

For this reason, we work using snapshots of the underlying database. Every so often, we will return to the manuscript and recompile the online edition, then check references to the data that might have changed. The current build of the book is from 2019-01-14 11:55:30.

Acknowledgements

Our sincere thanks go to all of the generous researchers — too many to name here, but listed on the Wordbank contributors page and in Appendix A — who contributed their data to the database. Even during our time working on this project, norms of data sharing have shifted; some substantial portion of this shift is due to the generosity of those researchers who shared their data early on in the process. Thanks especially to Rune Nørdgard Jørgensen for sharing the full CLEX-CDI dataset (Jørgensen et al., 2010) and for his kind assistance in the transition from the CLEX-CDI website to the Wordbank website.

Major thanks are also due to the MacArthur-Bates CDI Advisory Board, especially Philip Dale and Larry Fenson, for their continued intellectual, financial, logistical, and personal support of the Wordbank project.

Thanks to Danielle Kellier for substantial work importing data and maintaining the Wordbank website, as well as updating the universal lemma mappings. Initial programming work was done by Ranjay Krishna, and some database imports were performed by Elise Sugarman.

Thanks to NSF Award #1528526, "Wordbank: An Open Repository for Developmental Vocabulary Data" for financial support of the Wordbank project as well as to the Stanford Psychology Department for a small seed funding award that supported the initial development of the site.

Chapter 1

Theoretical Foundations

One of the defining human characteristics is the ability to use language in its lexical and combinatorial richness. The study of language acquisition has been a traditional locus of our search to understand the nature of this ability. What allows human children to acquire a language has been the subject of one of the historical "great debates," in which different proposals about the architecture of the human mind and the nature of human uniqueness have been discussed. Does language arise from domain-specific adaptations for syntactic structure (Chomsky, 1981, 2014)? Or does it arise from a combination of environmental input and sophisticated, general-purpose learning mechanisms (Elman et al., 1996)? Two poles have traditionally emerged in this discussion: from domain-general empiricist proposals to domain-specific nativist proposals, as illustrated in Figure 1.1.

In this chapter, we begin by presenting the perspective from these poles but contend that they are rarely helpful in the practice of understanding the scope and course of early language learning. Instead, we argue for the development of theories that describe the scope and course of language learning as a whole, as well as its quantitative variation across children, languages, and cultures. These concerns lead us to frame our own study in terms of a set of distinct theoretical issues: capturing consistency and variability; drawing connections across timescales of learning; and the notion of process, rather than content, universals. We end by discussing the relation of our theoretical stance to others and the role of replication and larger datasets in building quantitative theories.

1.1 The picture to date

Empiricist proposals emphasize the ability of children to learn structure across domains, the richness of the distributional input that children are exposed to, and their ability to create appropriate abstractions from structured linguistic input. The components of these proposals are children's general statistical learning abilities (Saffran et al., 1996; Fiser and Aslin, 2002). When applied to linguistic input, even general statistical tools can recover some aspects of linguistic structure across a variety of domains (Redington et al., 1998; Frank et al., 2012; Elman et al., 1996). While promising, these proposals have as their primary challenge that children show evidence of abstractions that encode key aspects of linguistic competence, even at an early age and in the absence of substantial input. For example, young children show systematic word orders even in the absence of structured input (Goldin-Meadow and Mylander, 1983). Further, typically developing children interpret the arguments of novel verbs (Gertner et al., 2006) and show evidence of category structure for syntactic

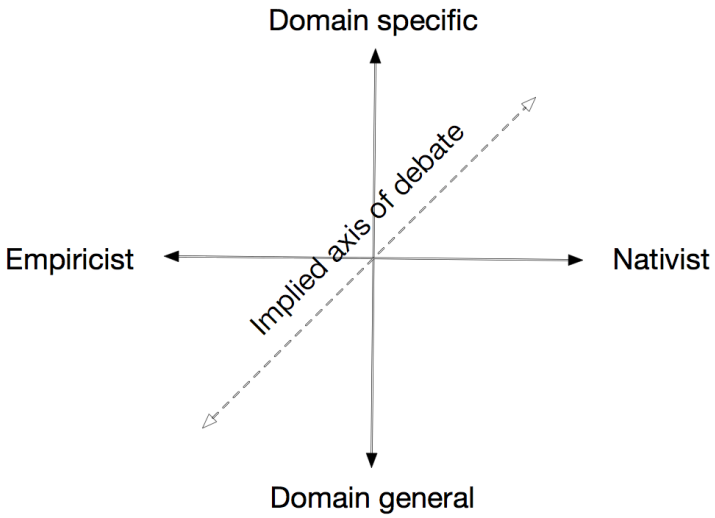


Figure 1.1: Schematic of the space of theoretical debates around language acquisition.

categories such as determiners (Yang, 2013) and structures such as the dative (Conwell and Demuth, 2007), even from an early age.

Nativist proposals, in contrast, tend to focus on the complexity of the grammatical structures that young children appear to possess, relative to their rarity in children's environment. Such arguments emphasize the poverty of the linguistic stimulus (e.g., Legate and Yang, 2002) in contrast to the richness of the structural generalizations children make (e.g., Crain and Thornton, 2000). The type of proposals that are on offer within this space often include "principles and parameters"-type theories, in which languages share a set of syntactic principles that govern syntactic combination but vary on a relatively small set of parameters that control how these structures vary (Baker, 2005; Yang, 2002). These proposals are challenged by the vast cross-linguistic differences in syntactic abstractions (Evans and Levinson, 2009), by the character and scope of children's syntactic generalizations (which are often tied to specific lexical structures; Tomasello, 2000), and by evidence of early input-sensitive learning and generalization both within specific domains (e.g., Meylan et al., 2017) and in artificial language tasks (Gómez and Gerken, 1999). Such proposals also tend to downplay the inherent variability that characterizes language learning across individuals (Bates et al., 1988, 1994), focusing instead on the universals or commonalities that exist across children. While nativist proposals acknowledge that individual variation exists, variation is thought to serve a primarily descriptive function, existing in a "theoretical vacuum" (Bloom, 2002, p. 52), with greater emphasis on how language acquisition works in the general case.

The "great debate" between these viewpoints is philosophically appealing, but has also led to a polarization of the field of language acquisition. Typically researchers work in their own siloed traditions (empiricist or nativist), and focus on individual phenomena that do not make contact with one another — a classic version of Kuhn (1970)'s "paradigms." Research in the nativist tradition often focuses on particular syntactic phenomena that are largely neglected in the empiricist tradition (e.g., the "optional infinitive," Wexler, 1998; but cf. Freudenthal et al., 2010). In contrast, research in the empiricist tradition has often used artificial language learning tasks that are argued not to reflect on the underlying structural properties that are claimed to be innate (e.g., Lany and Saffran, 2010;

cf. Yang, 2004). Empiricist theories are also more likely to recognize the causes and consequences of individual variation in rates of learning, including potential sources of that variation that arise from variation in the circumstances in which children learn and the opportunities they have for engaging with language in a meaningful way (e.g., Huttenlocher et al. (1991); Hoff (2006); Weisleder and Fernald (2013)). By focusing on different paradigms and phenomena and theorizing using distinct vocabularies, these traditions make limited contact. Often, language development conferences feature paired keynote talks from these two different traditions — a clear sign of polarization.

In addition, these theories are frameworks, rather than than actual hypotheses. Few proposals within these frameworks can be said to generate testable and clearly competing predictions, even within a specific domain. And any individual observation typically cannot be said to be inconsistent with any but the absolute strongest nativist or empiricist position. To be tested, proposals must make specific predictions. Computational models have been an important method for allowing proposals to be instantiated to the degree that they can make testable predictions. In practice, however models typically end up less differentiated than framework rhetoric suggests. In order to get off the ground in performing a particular empirical task, theories must often help themselves to generous amounts of both innate structure — in the form of structured inputs from social, cognitive, or perceptual domains — and statistical learning abilities (Roy and Pentland, 2002; Alishahi and Stevenson, 2008; Frank et al., 2009, 2010; Yang, 2004).[1]

Further, when nativist and empiricist viewpoints make differing predictions, they are often in phenomena that — from a bird's eye, or even parents' eye, view — are relatively trivial in the general course of language development. Abstraction debates have played out in the acquisition of the definite determiner "the" (Valian, 1986; Pine and Lieven, 1997; Yang, 2013; Meylan et al., 2017), auxiliary inversion (Pullum and Scholz, 2002; Legate and Yang, 2002), and the use of anaphoric "one" (Akhtar et al., 2004; Regier and Gahl, 2004; Lidz et al., 2003), for example. These phenomena are occasionally observable in the children of linguistically-trained parents, but even the closest observer would be forgiven for being more compelled by watching the increasingly creative and complex ways that children interpret, use, and play with language, rather than the occasional syntactic slip. Further, all of these phenomena concern linguistic behavior at a particular level of abstraction, syntax, reflecting a broader historical argument that syntactic structure is the heart of the uniquely human language faculty (Chomsky, 1957), and that other aspects of language tend to be shared with other species (Hauser et al., 2002) and are therefore somehow less interesting, less amazing and less critical for science to understand.

From an evolutionary perspective, syntax is far from the only unique or notable feature of human communication (Pinker and Jackendoff, 2005; Tomasello, 2010). The nature and range of communicative gestures, the variety of sounds, and the diversity of lexical items all are relatively unprecedented — especially in the primate lineage. And these observable aspects of language — as well as the emergence of syntactic structure more broadly — are some of what makes the broad course of language acquisition striking from the perspective of a clinician or a parent. We notice the first communicative signals, the emergence and rapid growth of of vocabulary, the beginning of the productive combination of words, increases in the length and complexity of utterances, and the patterns of error and overgeneralization that remain in early childhood. Moreover, there is considerable evidence for continuity across these domains (e.g., Bornstein and Putnick, 2012; Tsao

[1] The research on the nature of inflectional morphology — the "past tense debate" — is one place where computational models played a foundational role in instantiating theoretical claims about innateness and representational structure (e.g., Rumelhart et al., 1986; Pinker and Prince, 1988; Plunkett and Marchman, 1993, 1991, 1996; Marcus, 1995; Marchman, 1997).

et al., 2004; Bates et al., 1988; Cristia et al., 2014). Children's earliest gestures and sounds relate to their oral language comprehension and production. And these in turn relate to later skill in using language as a tool for learning, through both the auditory (or visual, in the case of sign language) as well as the written modality.

These broader patterns of language learning are the natural focus of investigations like ours that use parent report to learn about children's language. Parents are attentive and accurate observers of communicative gesture, vocabulary, and word combination. But without linguistic training they may not even notice subtleties like non-productive determine use, auxiliary inversion, or anaphoric "one." Further, these investigations can in many case make productive contact with the rich literatures on early communication, speech perception (e.g., Kuhl, 2004), word learning (e.g., Bloom, 2002; Snedeker, 2009), and grammatical productivity through verb structure (Fisher et al., 2010). While debates over the nature of syntactic knowledge and abstraction have raged, other subfields of language acquisition have prospered.

Research in these subfields makes at most limited contact with broad questions of nativism and empiricism, in part because they deal with phenomena that are language specific — sounds, lexical items, grammatical constructions — and hence that children must learn from their input. The question is then about the mechanisms and constraints that guide this process of learning, rather than about any posited universal or innate content (even at the level of abstractions).

1.2 Making progress

To move beyond great debates, what should a unifying theoretical framework for language learning look like? In spite of the critiques above, we still believe in the importance of the search for core, universal aspects of language learning that elucidate the process by which children acquire this uniquely human ability. Yet the sort of theory that describes such universals will likely look radically different from its historical antecedents. Below and in the remainder of this chapter, we sketch some aspects of what such a theory will look like and how this vision connects to our present investigation.

Any universal is likely to be a statistical or quantitative universal — we refer to these as "consistencies." The variation across the world's languages is such that only the most tautological facts will be truly invariant (Evans and Levinson, 2009). Further, we are unlikely to be able to access the kinds of samples that would allow us to make claims of universality (Piantadosi and Gibson, 2014). Thus, we should talk about the relative consistency and variability of particular phenomena rather than any sorts of absolutes.

In addition, language learning takes place at the timescale of years. CDI forms provide a global snapshot of a child's language at a particular point in time, rather than demonstrating the operation of a particular mechanism or principle. Substantial reconstruction is necessary to understand how processes operating over seconds — for example, online statistical learning or pragmatic inference — would result in particular structures accreting over time in the vocabulary. Thus, consistencies we observe are at best the basis for abductive inferences about underlying mechanisms.

These consistencies are consistencies in the learning of vocabulary and constructions, rather than syntactic rules. These items must be learned from data. Thus any putative universals identified in our investigation must not be "content universals" that specific particular grammatical rules or linkages. They must be "process universals" in the sense that they specify mechanisms or processes that unfold over time and operate over children's interactional input in ways that produce the

observed consistencies. Harkening back to Bates et al. (1988) and Elman et al. (1996), our proposal is a shift from a focus on universal content to universal mechanism.

1.3 Variability and consistency

Observing what "hangs together" in development can provide clues about the architecture of the underlying system (Bates et al., 1988). We think of these correlations as loose targets for theorists: a successful theory can gain support by providing an account for these observations. Crudely put, if a theory posits that some aspect of language acquisition is universal, it should be relatively more consistent in our data.

What are the units over which we compute variability and consistency? We refer to these as "signatures" — loosely, measurements that can vary across populations. In practice, a signature can be the output of any analysis, with the simplest being vocabulary size or variability itself (as in Chapter 5). Signatures are linked to particular theoretical goals by arguments about the validity of an analysis — for example, the argument of Bates et al. (1994) that the over-representation of nouns in early vocabulary is a meaningful dimension of variation between individuals. A signature for our purposes is thus an analysis that yields a set of numbers. In nearly every chapter of the book, we define one or several signatures whose variability we can measure.

Different sources of variance provide different sorts of evidence. One sense of the notion of "universal" that dates to early generative syntax is the notion of typological invariance (Greenberg, 1963). Following this general idea, in the majority of the book we focus on variability in some signature across languages. The implied inference is that consistency across languages points to the idea that a signature results from some mechanism (more on inferences about mechanism below) that is independent of the language being learned and the context in which it is learned.

But when we assess the variability of some signature across datasets, many things vary that are not the target of our inference. Although language and culture typically vary (except in the case of multiple instruments that are assessed on the same data source), many other things vary as well. Different datasets are constructed by different researchers with different goals. They use different instruments with different items — and different length and composition. These instruments are administered to different samples, with different sampling strategies. And the nature of the administration is different as well.

Thus, when a particular response appears to be consistent, we can say a fortiori that none of these sources of variation appear to have affected the consistency of the response. (Or at least that if they have, they have canceled each other out in a highly non-random way). But when variability does occur, we cannot make the opposite inference. Variability has many explanations, but consistency tends to point us towards a single inference.

We focus on cross-dataset variability as the primary source of variability in this book. We refer to this variability throughout as cross-linguistic variability, though in fact there are a number of caveats that must be stated. First, many things vary between datasets far beyond language (as noted above). And second, some datasets represent the same language in different dialects (e.g., Australian and British English). Some even reflect the same language and dialect, measured using two different instruments (e.g., the two Beijing Mandarin datasets). In some cases we will even leverage these parallels to help us rule out alternative explanations. The reasons we focus on cross-dataset variability are three.

First, datasets vary so much that — assuming this variation is somewhat random — claims of consistency are stronger when they emerge from this sort of data. Imagine counterfactually that all of the instruments we used had exactly the same structure and item set, and all were administered identically. Certainly this lack of variation would make our life easier in a number of ways when making quantitative comparisons between datasets! But it also then means that these consistencies would be confounded in our data — particular item sets (plausibly) or administration instructions (somewhat less plausibly) could be the source of an observed consistency in the data. In contrast, while the messiness and inconsistency of the data in the Wordbank dataset make many aspects of our analysis much harder, it actually increases the strength of the inferences we can draw when — despite this — we see some phenomenon emerge with striking consistency.[2]

Second, the genesis of the investigations documented in this book was in part the observation that several phenomena that we examined were strikingly consistent across languages. For example, the gender effects shown in Chapter 6 were much more consistent than any of us thought (being at the time ignorant about the previous literature in this particular area; Eriksson et al., 2012). Empirically, we found a lot to look at that was both surprising and interpretable when we examined well-known signatures as they varied across languages. Thus, our motivation is in part the emergent success of this approach.

The final reason we consider cross-language variability as our primary lever is a negative one. The obvious competitor as a source of variability is variation across individuals. We examine this variability briefly in Chapters 4 and 5 and more extensively in Chapter 14. While we document substantial and stable variability across individuals (echoing Bates et al., 1994), this variability empirically proves to be less of a lever into theoretical issues of interest than we would hope. One aspect of this move is data-related — we have far more cross-sectional than longitudinal data in the Wordbank dataset — and hence we cannot track stability and change over time as easily or powerfully as we would like. Further, we have very few additional measures on most children in the dataset (beyond the occasional demographic feature). In addition, as we show in Chapter 14, though there is some reliable stylistic variation between children, much apparent variability in children's style of language learning can be traced to variation in rate. Thus, and in contrast to the exciting emergent conclusions from cross-linguistic variation, individual variation appears to be a less powerful theoretical lever.

Nevertheless, individual variation across individuals does exist and it is robust. Indeed, in Chapter 5, we show remarkable consistency across languages in the extent to which there is variability across individuals. We remain optimistic that continuing to explore the extent and consistency of this variability will continue to provide a window into the universal processes that guide learning for the mythical "model" child, as well as define the upper- and lower-limits on typical development.

In Chapter 15, we bring together estimates of variability of individual signatures from each of the earlier constituent chapters. We combine these into a single, data-driven continuum from absolute consistency to high variability, and use this continuum to drive speculations about the sorts of mechanisms that would produce a particular set of consistencies.

[2]Of course, we also consider the confounds that do remain at the end of the book. In particular, confounding related to the parent report structure of the CDI is a major risk.

1.4 Process universals

1.4.1 Preconditions

Imagine we were to uncover an aspect of language development that was completely consistent across languages. (Surprisingly, as we'll see in Chapter 15 there are some!). What could we then infer from this observation? Not much, it turns out. The observed regularity could be due to different sources in different datasets or it could be uninteresting from a theoretical perspective.

First, even if the consistency is interesting, any inference from it will always be abductive — an inference backwards from observation to cause These abductive inferences will always be under-constrained and tentative, thus they will always be at best empirically-grounded speculations that should be brought together with other data to make a test. In some sense, this is the fundamental caveat governing our entire enterprise here. The research design is correlational and so causal inferences are not available.

Inferences can go wrong even within this more limited correlational paradigm. For example, we could observe that, across languages, we saw hypothetically that some word was always produced earliest. But it could be the case that the word happened to be learned earliest in some languages because it was short and easy to pronounce, while in other languages it was learned early due to a high frequency of usage in the input. This example illustrates the difficulties of reverse inference from consistency. Similarly, we could observe that a certain distributional form always described children's vocabulary estimates, across languages. This regularity could be due to the operation of the central limit theorem rather than any interesting or substantive mechanism that we might be interested in as psychologists.

These problems mean that we need to have two (somewhat informal) conditions on the consistencies that we posit. First, we need to consider the possibility of multiple routes to the same observed consistency. To the extent that observed regularities are specific and surprising, it will be less likely that there are multiple routes across different languages to observing the same thing. Second, for any potential causal story that we posit, we need to be able to posit a plausible or interesting causal story that does not generate the observed regularity. The tightness of this comparison with a counterfactual governs the strength of the inference.

1.4.2 The nature of the processes

Suppose a consistency we identify meets the conditions we describe above: it is sufficiently surprising that we don't see a parsimonious story for how the data for different languages could have been generated by different processes, and there are close counterfactuals in which this consistency did not emerge. Further, in our example suppose we have a larger and more diverse set of languages and cultures represented in our dataset such that we can justify using the title "universal" rather than the more descriptive and limited "consistency." We can then imagine trying to constrain the nature of the sort of universal that could give rise to this type of consistency.

What can we say about putative universals? By virtue of the learning problem that they arise from, they cannot be universals of content. A child's vocabulary is made up of individual words, each arising from a set of specific interactional circumstances (e.g., the trip to the zoo where a giraffe was seen for the first time). And each of these words is — of course — specific to a particular language. Thus, there is no viable sense in which any possible universals for vocabulary learning can be content

universals: no particular content of this type could be innately given. For this reason, we describe these as "process" universals: they relate to the process by which each individual extracts a lexicon from their own idiosyncratic linguistic experiences.

Further, since these processes are fundamentally learning processes, they operate at the timescale of moment-to-moment interactions (Frank et al., 2009; McMurray et al., 2012). In contrast, using the CDI, we observe the accretion of vocabulary and linguistic competence over the course of millions of these interactions (see e.g., Dupoux, 2018, for estimates). This mismatch makes inferences about the nature of the processes even trickier — in some sense, we are attempting to uncover what forces eroded a hillside when all we see are the uncovered strata.

Yet, from the perspective of the language learning literature, there are still obvious candidates for the sort of universals we are talking about. The general idea of "statistical learning" is one (Saffran et al., 1996; Saffran and Kirkham, 2018). From a very early age, children are sensitive to regularities in their sensory environments and track statistical associations between elements in these environments. Concrete examples of these mechanisms in action include the tracking of syllable-to-syllable conditional probabilities (Saffran et al., 1996) and the tracking of word-referent correspondences (Smith and Yu, 2008), but in principle these mechanisms are likely operating over every level of representation present in early language (Shukla et al., 2011).

Of course, these processes of statistical learning operate in a social context. Statistical learning processes certainly operate over social input that includes information from social partners (Yu and Ballard, 2007). In addition, it is likely that statistical learners take into account the nature of the social context in the inferences that they make (Frank et al., 2009; Shafto et al., 2012; Frank et al., 2013).

Processes of generalization are also strong candidates for process universals. The nature of these generalization mechanisms is of course highly controversial (for all the reasons discussed above), but every account of learning requires some type of generalization from specific lexical items to syntactic constructions or morphological rules (Tomasello, 2003; Yang, 2016).

In addition, a number of processing factors might lead to processes that are universal. At the same time as children's language abilities are growing, a variety of core aspects of cognition are undergoing developmental change as well. Children's general speed of processing is changing (Kail, 1991; Frank et al., 2016b) — including changes in memory (Ross-sheehy et al., 2003; Rovee-Collier, 1997), attention (Colombo, 2001), and executive function (Davidson et al., 2006). Processing factors also influence the speed and efficiency with which children can comprehend words in real time, linking both to developmental change and individual differences that are stable and meaningful and that extend beyond language (Fernald and Marchman, 2012; Marchman and Fernald, 2008; Marchman and Dale, 2017). Although much of the developmental literature on these cognitive constructs focuses either on infancy or the preschool years — because one- and two-year-olds are hard to measure with standard cognitive psychology tasks — the assumption is that these processes are developing continuously throughout the period we focus on here. These developmental changes mean that the processes that we describe are themselves not static but — inasmuch as they draw on these capacities — they are themselves a moving target.

Finally, it is important to note that process universals need not be internal to the child. Instead, they may be universals of interaction between children and their caregivers. Such processes could lead to the emergence of specific signatures just as well as processes internal to the learner. To take a concrete example, the timing of turn-taking in conversation is an example of one such proposed interactional universal (Stivers et al., 2009). In particular, it is entirely possible that some of the

specific consistencies in the content of children's early vocabulary (Chapter 8) that we observe in fact emerge from the nature of children's early environments, including universal features of what children and their caregivers talk about and why.

1.4.3 Alternatives

In Chapter 15, we will examine the empirical support for the claim of consistent signatures in language learning across languages. Yet our further claim is that these consistencies are supported by universal processes. To examine whether there is content to the claim of process universals, it is helpful to consider the alternative hypothesis. One important alternative is that the process of language acquisition is specific and particular, rather than universal. Two prominent particulars pull against universal tendencies.

The first is the vast semantic and syntactic variation across the world's languages. For example, as illustrated by Slobin (1996) and others, languages vary dramatically in the ways that they assign semantic content to verbs. If the semantic partition of verbs led to large-scale differences in the timeline or mechanism of acquisition, we might see systematic differences in the predictors of age of acquisition for verbs in these languages, yet we do not. Further, languages are more and less morphologically complex; while the most morphologically-complex, polysynthetic languages are not represented in Wordbank, we do have data from both Mandarin (less complex) and Russian (more complex). If morphosyntax were relatively more or less important in the acquisition of particular languages, we might expect radical differences in the noun bias, the grammar-lexicon correlation, or the predictors of age of acquisition across languages, yet we do not observe these. Of course, there is always room for finer-grained predictions — with more detailed predictive models and better typological coverage, perhaps we will discover such signatures. Our point here is merely that — to the extent that we observe consistencies, neither morphosyntactic nor semantic variability across languages dominates the process of vocabulary acquisition.

The second set of language-specific particulars that might lead to variance across languages is the vast cultural variability across the communities represented in the Wordbank data. Our data contain both "individualist" and "collectivist" cultures (Markus and Kitayama, 1991; Nisbett et al., 2001) as well as both "loose fit" and "tight fit" cultures (Gelfand et al., 2011). To the extent that parenting differs across these cultures — and there is good evidence that it does (e.g., Bornstein, 2013) — we should see variance in the trajectory of language learning. For example, it would be quite reasonable to predict that the female advantage in vocabulary acquisition might vary as a function of cross-national gender biases (Nosek et al., 2009). Yet it is strikingly consistent overall (presaging the conclusions in Chapter 6), again arguing that — at the broadest level, at least — variable cultural factors do not dominate other processes in the acquisition of vocabulary.

1.5 Replication and theory-building: conclusions

In this chapter, we have sketched a bit of what we see to be the unique theoretical contributions of work with a much larger dataset than is usual in developmental language acquisition research. In a nutshell, doing this work at scale allows for the identification of sources of variability in the "signatures" of language learning. What these signatures are is a matter for further development — each chapter will describe and motivate the particular signatures that it includes. Further, the consistency of these signatures can provide a motivation for positing process universals that underlie

the emergence of these signatures (pending the caveats stated above). We return to the general picture of language learning that emerges from our study in Chapter 16.

One final note about nature of the theory that emerges from the work we do here. One set of concepts that is subsumed in our interest in consistency is that theory be supported by observations that are reproducible, replicable, and robust (Munafò et al., 2017). A theory of consistencies is, again, a fortiori all of these. If a particular characteristic can be shown again and again across individuals, samples, and languages it is replicable. Indeed, one view of our enterprise is that its impact is fundamentally in the consolidation of knowledge through unifying — replicating — previous work.

Crudely put, we have compiled all of the CDI data that we could, and all of the CDI analyses, and executed the cross of analyses and datasets. This project is thus a cross-linguistic replication study. And so, when we state that some phenomenon is consistent across languages, it is by definition replicated — but it is additionally robust to a number of different procedural decisions (such as the design or administration of the CDI form) that end up varying widely in our data.

Finally, this work is also fully computationally reproducible — the analytic conclusions we draw here based on a set of open data and code that can be rerun to create the figures and tables in the manuscript. This characteristic alone does not guarantee their correctness, but their provenance is known.

Chapter 2

Practical Foundations[1]

Almost uniquely in cognitive development, early language learning offers an opportunity to study both consistency and variability in a single phenomenon. Often researchers interested in consistency have measured theoretically-important, carefully-chosen phenomena using small convenience samples that suffice to show a proof-of-concept but do not provide information about variability. In contrast, work on variability between individuals has often focused on larger samples with more reliable tasks, that — perhaps as a consequence of their reliability — are less tightly linked to a particular theoretical construct of interest. And these constructs may depart from the ecological task that is of principal interest (Cronbach and Meehl, 1955).

Early language is a rare case where these problems are minimized. Measures of early language comprehension and production tend to be face valid and tightly linked to the construct of interest in their structure. And yet, in contrast to the measures that usually fulfill these criteria, early language measures are very closely related to the ecological task — linguistic communication — that is the theoretical target for explanation. Thus, early language is the rare case where consistency and variability can both be explored in a single set of measurements (Bates et al., 1994).

Further, the nature and course of early word learning is an important window into children's growing understanding of the world — beyond language. Early words cross-cut a variety of linguistic categories, but generally consist of names for caregivers (e.g., mama), common objects (e.g., bottle, shoe), social expressions (e.g., bye-bye), and actions or routines (e.g., peekaboo, throw) (Nelson, 1973; Tardif et al., 2008). Yet this repertoire expands, and its composition can give insights into the learning mechanisms, biases, and priorities of young children.

2.1 Measuring early vocabulary

Traditional studies of language development typically apply a combination of observational assessment and structured tests, frequently relying on short samples of interactions and small samples of children. Discerning both the universal features and natural variation of early lexical development has been greatly facilitated by the development of parent report instruments like the MacArthur-Bates CDI (Fenson et al., 1994, 2007) and the Language Development Survey (LDS; Rescorla, 1989). The CDIs in particular were developed across a period of more than 40 years. Originally designed for

[1]Some material in this chapter is adapted from Frank et al. (2016a) and Marchman and Dale (2017).

use in a research study (Bates, 1976), the instruments have evolved from a structured face-to-face interview to a paper-and-pencil format and are now increasingly administered online (e.g., the web-cdi project; Kristoffersen et al. (2013) for Norwegian; laboratorium.detskarec.sk for Slovak). While other assessment tools exist for slightly older children, to our knowledge, no other measure allows cost-effective global language assessment for children in the critical age ranges between the emergence of language and the period when children become more able to engage in structured, face-to-face activities (around 30 months).

Naturalistic observations are the other leading candidate for measurement of early language, but such observations are extremely costly and time-consuming to transcribe and annotate. These difficulties lead to a trade-off where most studies either include dense data about a small number of children or smaller amounts of data with a larger sample size. Dense datasets currently provide the best method for in-depth study of the interaction between learning mechanisms and language input in individuals (e.g., Lieven et al., 2009; Roy et al., 2015). The generalizability of these studies is necessarily limited by their small sample sizes, and sample sizes are in turn limited by the costs and practicality of gathering and transcribing such data (see e.g., Bergelson and Aslin, 2017, for the state of the art). At the other end of the spectrum, assessment of many individual language samples can yield information about individual variability (e.g., Dickinson and Tabors, 2001; Cartmill et al., 2013; Weisleder and Fernald, 2013), but at some cost in terms of depth.

Further, standardization and avoidance of confounds in naturalistic observation studies is challenging. Although parent report seems at first glance to be much more subject to the biases of individual parents, in fact many of the same confounds arise in other paradigms. For example, should an observation session be during play with the parent or an experimenter? Given that parents vary in their talkativeness during a play session, play with a parent is bound to measure parents' ability to elicit language as well as children's variation. But for toddlers temperamental variation is extreme so an experimenter play session may simply be impractical for some children (and language use may be limited by shyness rather than a lack of ability). These difficulties can be navigated through careful procedural and statistical control, but the point of this example is that no observational method offers a perfect solution.

Finally, naturalistic observations do not measure children's language comprehension, a variable of interest for many early language researchers. Estimates of production vocabulary from naturalistic observation are highly correlated with the CDI within studies (e.g., Bornstein and Haynes, 1998), but are likely to be affected substantially by length of the session, context, and interlocutor when comparing across studies (see e.g., Hidaka, 2015, for discussion). And although there exist methods to extract insights about global vocabulary from naturalistic observation, these statistical extrapolations are relatively new and have not been validated extensively (Hidaka, 2015).

Experimental testing, in contrast, is an excellent method for measuring individual aspects of children's linguistic competence, for example their knowledge of a handful of words or their speed of processing (e.g., Bergelson and Swingley, 2012; Fernald et al., 2006). These methods are much less subject to the confounding of observational methods. But an infant or toddler can only provide a limited number of trials during a single measurement session, even in implicit tasks using eye-movements. Thus, the ability to measure global language competence is limited. Further, the specific words to measure for children of different ages vary — those words that are appropriate for measuring a 14-month-old's competence are trivially easy for a 24-month-old. And attrition can be quite high for a long measurement session, requiring repeated testing for many participants. Other comprehension vocabulary measures are also available across some range of languages (e.g., the Peabody Picture Vocabulary Test 4, Dunn and Dunn, 2007; the Computerized Comprehension Task (CCT), Friend

and Keplinger, 2008), but most of these assessments are tailored for children older than 2 1/2 years.

In sum, for breadth, depth, and ease of data gathering, parent report is unmatched. In Chapter 4, we provide a more extensive discussion of issues surrounding the reliability and validity of parent report.

2.2 The logic of parent report

Parent-report instruments like the CDI and LDS take advantage of the fact that parents (or other primary caregivers) are expert observers of their child. Parent reports are based on experiences with the child which are not only more extensive than any researcher or clinician can obtain, but are also more representative of the child's ability. Parents have experience with their child at play, at meals, at bath and bedtime, at tantrums — in short, with the full range of the child's life and therefore with the full range of language structures used in these contexts. Parents also have opportunities to hear the child interact with other people: other caregivers, grandparents, siblings, and friends. Because responses on these instruments represent an aggregation over much time and many situations, they are less influenced by factors that can mask a child's true ability in the laboratory or clinic, such as shyness or compliance, or that can impact the validity of naturalistic sampling, such as word frequency. As Bates et al. (1991) point out, "parental report is likely to reflect what a child knows, whereas [a sample of] free speech reflects those forms that she is more likely to use." (p. 57).

Because of its format, parent report enables the collection of data from far larger samples of children than would be possible with standardized tests or naturalistic observation. Information from more adequate samples, especially in the form of norms, can benefit both clinical practice and research. Fenson et al. (1994), for example, used the norming data from English versions of the CDIs - a sample of 2,550 children aged 8 to 30 months - to address questions about variability in communicative development. Large samples are especially needed to provide an accurate statistical description of extreme scores, i.e., what score corresponds to the 10th percentile? What does the most advanced child (e.g., $>$ 90th percentile) look like at a given age? Research on questions such as environmental influences on language development can also benefit from large samples. Correlational research is hampered by the problem of multicollinearity: The predictor variables such as parental education, number of books in the home, family size, use of questions vs. imperatives, are likely to be intercorrelated, making it difficult to separate the effects of each of them individually. Large samples in which there is a substantial amount of non-overlapping variance are essential for addressing these questions.

The main core of the CDIs is the vocabulary checklist. This list is essentially a "bag of words" which represents the set of words that best capture variation in lexical development across the full spectrum of child ages and abilities. Parents choose the words they believe that their child can currently "understand" (comprehension, measured for younger children) or "understand and say" (production, measured for both younger and older children). A child's score on a vocabulary checklist represents their comprehension or production "vocabulary size," indexing that child's relative status against other children assessed with the same list. In their English and Spanish instantiations, the vocabulary checklists come in two versions: Words & Gestures (WG; 8–18 months) which contains about 400 words, and the Words & Sentences (WS; 16–30 months), which contains about 700 words. This structure has often been replicated across cross-linguistic adaptations, though there is some variation in form construction (see below), and some forms include substantially different numbers of words or include/exclude other measures.

The vocabulary checklists contain words from many different semantic (e.g., animal names, household items) and syntactic (e.g., action words, connectives) categories, resulting in broader samples of lexical knowledge than are available from other methods. Importantly, however, these words are not chosen to create a complete list of all words understood or produced by a child. Instead, CDI word lists are constructed to include a set of words that most children will know as well as a sampling of intermediate and more difficult words that will be useful in assessing variability between children.[2]

Thus, an additional advantage of the parent-report method is that parents can report on many different sub-components and correlates of early vocabulary development. In particular, the CDI instruments ask about use of communicative gestures, grammar, and symbolic play, as well as vocabulary comprehension and production. Information about what early vocabulary development correlates with, and what it does not, can yield important theoretical information about the common mechanisms underlying learning. As Bates et al. (1991) note, studies which have the power and scope to examine what "hangs together" across early language development can provide critical clues to how the system is put together in the first place (p. 7).

Of course, parent report has substantial limitations that can lead to both measurement error and bias. These are addressed to some extent by design features of the CDI, and further addressed by evidence for the reliability and validity of the instrument. Because these concerns are so central to our enterprise here, we discuss these issues at length in Chapter 4 both from theoretical and analytic points of view.

2.3 Cross-linguistic comparison

2.3.1 Adaptation, not translation!

Originally designed for English, parallel CDI instruments have now been adapted for more than 100 languages (mb-cdi.stanford.edu/adaptations.html), with data from 29 of those languages currently available in Wordbank (Dale and Penfold, 2011). The ethic behind the development of these instruments is "adaptation, not translation" — in other words, create forms with the same spirit as the English form, but do not simply translate the items (Dale, 2015). Instead, developers have been strongly encouraged to craft instruments that reflect the linguistic and cultural contexts that influence the early acquisition of vocabulary and other aspects of language in that particular language.

The resulting forms vary widely, including differences in length and intended age range. Some forms include hundreds of items more than the original 680 words on the English Words & Sentences form; others are so-called "short forms" and include only a hundred or a few hundred carefully selected words. Some are designed to capture development from the emergence of language through ages 3–4 years, while others are focused on very early development (like the English Words & Gestures form, designed for ages 8–18 months).

While many words on the English-language checklist may easily translate to other languages, others will simply not be relevant within the same developmental time frame for children learning that new language, e.g., cheese in Japanese or snow in Arabic. Conversely, additional words may be needed in the new language which were not included on the English-language vocabulary checklist,

[2]While there have been some efforts to estimate the child's total vocabulary from CDI scores by Mayor and Plunkett (2011), this estimate is calibrated based on a very small sample of diary studies, and cannot easily be extended across forms or languages.

e.g., tortilla in Mexican Spanish. In all languages, though, the vocabulary checklists include words that appear earlier and later in normal development, as well as a similar proportion of words from different lexical classes, for example, nouns, verbs, adjectives, and so on. Taken individually then, each adapted instrument captures key trends in vocabulary development aggregated across all items on the respective checklists.

Further, due to variation in language structure and the interests of the developers of CDI adaptations, the CDI instruments vary in structure across languages. Most adaptations of the WG generally include gestures as well as vocabulary comprehension and production, however, it is not always the case. Further, while adaptations of the WS always include vocabulary production, not all instruments also contain some measures of grammar, for example, early use of closed-class morphology or combinatorial syntax. Note that linguistic differences render the structure and format of many parts of the grammar sections to be very different, and hence, not amenable to comparisons across languages. A few instruments included in our dataset are pure checklists, with no other sections included. In sum, CDIs are a useful tool for many languages, but the forms differ between languages.

2.3.2 Our approach

The wide cross-linguistic adoption of the CDI provides an opportunity for cross-linguistic comparison but it also creates many challenges that are not present in datasets that are designed from the start for such comparisons. Differences in instruments and items as well as differences in samples and administration conditions all make it potentially quite problematic to compare scores and score distributions across forms. We discuss differences in instruments and items here and defer discussion of differences in samples and administration to Chapter 9.1.

Obviously differences in length between CDI forms mean that comparisons of raw scores across instruments are inappropriate. Dividing raw scores by the total number of items on a form results in proportions, which are somewhat more comparable but still potentially misleading. A more comprehensive form with more items on it will yield lower proportions for children with the same vocabulary size. Despite this weakness, we typically use proportions for visualizing differences across forms as it is cumbersome to compare raw scores with different totals. More discussion of absolute and relative vocabulary size differences between instruments can be found in Chapter 5.

Like other psychometric instruments, CDI instruments can also be normed, and many of the most popular forms are. In the standard norming process, the form is administered to a large typically-developing sample so that percentile ranks can be computed. For new administrations, the percentile of a particular raw score can be computed and used in place of the proportions or raw scores. These percentile ranks can be useful for clinical purposes, but they also complicate comparison across instruments because of potential differences in the norming population to begin with. In addition, the Wordbank dataset includes normed and un-normed forms, and for the normed forms, we sometimes have access to both the norming dataset and other data but sometimes only have access to the norming data. For these reasons, we do not employ normative percentile ranks in our analyses.

As the preceding discussion shows, there are serious difficulties that crop up immediately in comparisons across instruments. We will grapple with these difficulties throughout the book, but we generally adopt two approaches that help us navigate this complexity. The first approach, which is used in the majority of chapters, we describe below. The second approach is described in the next subsection.

In general, our approach to cross-linguistic and cross-instrument data is to provide standardized analyses within each instrument and language, without assuming equivalence across words, instruments, or populations. Thus, we will typically investigate a particular phenomenon (say the "noun bias" or the "female advantage") independently and in parallel for each of the instruments available to us.[3] We can then — still with caution — analyze and compare the magnitude of this phenomenon across languages, having abstracted away from the specifics of each particular instrument. We sometimes colloquially refer to this approach as "every form an island," meaning that each instrument is analyzed separately and only the analytic results at the highest level are compared.

Cross-linguistic conceptual comparisons are fraught, both philosophically (e.g., Quine, 1960) and practically. We refer to the practical issue as the tortilla problem: in American English, we have the word bread, which translates to pan in Mexican Spanish. But the Spanish word tortilla takes some of the cultural role of bread in English; bread has two reasonable translations. These translation issues go the other way as well: reloj translates to two distinct words (clock and watch) in English. For this reason, we typically adopt the "every form an island" approach described above. But questions nevertheless arise in the course of analyzing CDI data that can only be answered by comparison across languages (see e.g., Chapter 10).

Thus, in order to facilitate (cautious) cross-linguistic comparison, we developed a set of rough-and-ready translation equivalents. We call these "unilemmas" (short for "universal lemmas"). A "lemma" is a canonical form of a word, typically used for gathering frequency counts across different morphological variants (e.g., walk is the lemma for walks, walked, and walking). Unilemmas are used for mapping distinct lexical forms across languages. Unilemmas enable a number of desirable analyses, but more practically, they also provide consistent glosses that make it easier for researchers to work in languages with which they are not familiar. For convenience, our unilemmas are written in English, but they could of course have been written in any other language as well. Further details on the unilemmas are given in Chapter 9.1.

Even with the care we used here to construct a robust set translation equivalents, individual items are likely to only be roughly equivalent cross-linguistically, and may have significantly different referential scopes for children learning the different languages. That is, if a parent indicates that a child can produce the word dog in English and another parent indicates the translation equivalent in, for example, Spanish (perro), it may nevertheless be the case that these words are heard more or less frequently and in different contexts in the two languages. An important empirical question is the degree to which translation equivalents across languages have consistent developmental profiles.

2.4 Wordbank

To take advantage of the opportunity posed by the broad use of CDI instruments in the child language community, in 2014 we began constructing Wordbank, an open repository for CDI data that allows for interactive analysis and visualization. Our inspiration for Wordbank came from two successful projects for sharing data on children's language acquisition. The first is the Child Language Data Exchange System (CHILDES; MacWhinney, 2000). A database of transcripts of children's speech and speech to children, CHILDES has grown into a robust and important tool for the community, with many contributors and affiliated projects. The second is the Cross-Linguistic Lexical Norms site

[3] An exception to this approach is that we do sometimes interpolate words' trajectories across matched instruments for the same language, e.g. the proportion of children who say the word cat on both Words & Gestures and Words & Sentences forms for American English; see Appendix C.

(CLEX; Jørgensen et al., 2010), which is closer in content to Wordbank, and effectively our precursor. CLEX archived normative data from a range of CDI adaptations across languages, allowing browsing of acquisition trajectories for individual items or age groups.

Wordbank initially built on CLEX, offering the same functionality but allowing flexible and interactive visualization and analysis, as well as direct database access and data download. In addition, Wordbank's goal was always to extend beyond normative data by dynamically incorporating data from many different researchers and projects of varying sizes and scopes. While the resulting datasets in Wordbank are much more heterogeneous than if they were just based on norming samples alone, they are also larger and more representative than the individual norming datasets (in some cases), and available in languages where no norms exist (for others).

From the perspective of the study of child language, there are a number of notable omissions from the datasets represented in Wordbank and the analyses reported in this book. We discuss three of these below: our focus on typical development, monolinguals, and (for the most part) WEIRD populations.

First, our analyses here focus exclusively on typical development. The study of atypical language development is an important part of characterizing the mechanisms of acquisition; further, characterizing language development in these circumstances can have important applied benefits. Studies of language in developmental disorders (Tager-Flusberg et al., 2009; Eigsti et al., 2011), in cases of sensory deficits (Landau et al., 2009), and in cases of abnormal input (Curtiss, 1977), among others. Many of these studies have made use of dense observations of individual children, however, an approach that is fundamentally different than our large-scale, statistical approach here. While CDI-type instruments are increasingly being used with atypical populations (e.g., Charman et al., 2003; Luyster et al., 2007), in practice these datasets still tend to be smaller, concentrated in on English-speaking children, and difficult to access publicly. Thus, Wordbank does not currently archive sufficient data from atypical populations to justify inclusion of these analyses in the current book.

Second, the Wordbank dataset focuses on monolingual acquisition. CDI instruments were initially developed to provide normative measurements of variation within a single language. Since then, however, they have increasingly been used for comparison between monolingual and bilingual groups based on the administration of CDIs in both languages (e.g., Pearson et al., 1993; Hoff et al., 2012). These studies initially focused on specific bilingual populations (e.g., Spanish/English bilinguals). Recent studies have moved beyond this strategy and have begun to examine general trends across multiple bilingual pairings (e.g., Bilson et al., 2015; Floccia et al., 2018). Questions of bilingual acquisition are fascinating and important from both a theoretical and practical perspective. But there are practical obstacles to applying our approach to bilingual data that mean that this book does not consider the bilingual acquisition situation. First, because most of the largest CDI datasets were generated from monolingual norming studies, the vast majority of our data are not bilingual. Second, the combinatorics of bilingualism mean that data on nearly all language pairs will be non-existent. For these reasons, our book focuses on monolingual acquisition, though we recognize this as a limitation that must be addressed by future work.

Finally, the sample of languages we include is limited by our access to data. We have made efforts to include any large CDI datasets whose existence we are aware of — including extensive outreach to CDI authors, professional networking through the CDI board. Despite these efforts, our dataset is limited by both the sample of languages in which such studies have been conducted and international attitudes towards data sharing. Thus, although we do cover many languages around the world (see Chapter 9.1 for a map), these languages are skewed towards Europe and the United States, as well as

towards WEIRD — western, educated, rich, industrialized, and democratic — populations (Henrich et al., 2010).

While there are inherent limitations in comparing different instruments across languages — limitations that we return to again and again throughout the book — our dataset is the first that allows the exploration of both child- and item-level data within- and across such a large and diverse set of languages. As such, the availability of these adaptations remain at the core of the analyses that we offer within Wordbank.

We began the Wordbank project — our first large-scale, data-aggregation project — with a relatively naive attitude. We thought "if you build it, they will come": that contributors would flock to the opportunity to share their data with the world. We were unprepared for the challenges of contacting academics around the world, asking them to volunteer their time and hard-won data to an unknown cause, and then understanding the myriad formats and conventions represented in the data we eventually received. For the first couple of years, our data were largely co-extensive with those gathered by CLEX.

Fortunately, in the years since the Wordbank project began, attitudes towards data sharing have been shifting rapidly (in part as a result of work on replication and reproducibility, e.g. Open Science Collaboration, 2015). In addition, the credibility of the Wordbank project has gradually grown, in part due to the support of the MacArthur-Bates CDI advisory board. And as we received successively more data, our expertise in dealing with heterogeneous datasets has grown. Thus, the dataset has grown quickly in recent years. We hope that in future, authors see contribution to Wordbank as an aspirational endpoint for future studies using CDI instruments.

Chapter 3

Methods and Data[1]

We begin by introducing the structure of our dataset and the database that contains it. In the second section, we give some descriptive information on the datasets included in the database.

3.1 Database

Why use a database to store vocabulary data? Consider the standard format of raw CDI data, illustrated in Figure 3.1 for a small slice of the original CDI norming data (Fenson et al., 1994, 2007).

Each row is a child, each column gives a variable — either a demographic variable or the result of a particular word being administered to a particular child. Although this format is useful for homogeneous administrations of a single instrument, it cannot accommodate multiple instruments, multiple languages, or datasets with different sources or kinds of demographic information. Consolidating data across different instruments is very difficult in this format, and tracking data on children with multiple longitudinal administrations of a single instrument must also be done in an ad-hoc manner.

[1]Some material in this chapter is adapted from Frank et al. (2016a).

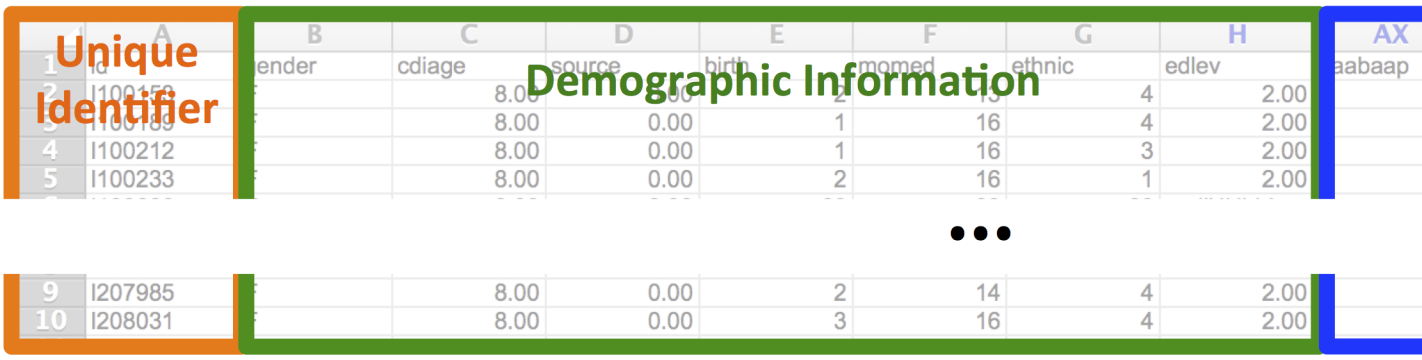


Figure 3.1: Example data from the CDI norming sample (Fenson et al., 2007). Each row has a unique child identifier, demographics, and word-by-word checklist data.

The move to a database format allows far more flexible and programmatic handling of heterogeneous data structures from different sources.

Further, as information about particular entities becomes available — for example, cross-linguistic mappings of lexical items — this information can be added in a way that preserves previous analyses. In a tabular format, such functionality is not guaranteed, and changes to the structure of the dataset will necessarily break previous analyses. A database, especially when supplemented with an appropriate application programming interface (API, see below), can solve this problem elegantly.

3.1.1 Database architecture

A relational database such as Wordbank is at its heart an ontology: a set of entities that are described in a series of tables linked by unique identifiers. The primary entities in the Wordbank database are:

- Instrument: A specific parent-report survey or questionnaire with a particular set of items. For example, the American English Words & Sentences form is an individual instrument.
- Item: A particular question on an instrument. A specific word like dog is our canonical CDI item, but other items include questions about gestures, morphological and syntactic complexity, and other aspects of early language or behavior.
- Administration: A particular instance of an instrument being given to a child, with an associated child age and source (the contributing lab).
- Child: A unique individual, with associated demographics.
- Language: A particular language or language community for which a CDI instrument has been adapted. Note that this definition of language distinguishes e.g. American and British English.

These entities are related by two primary groups of tables in Wordbank. The common tables store data that is shared between CDI instruments, including information about administrations (individual instances of a form being filled out for a child), and items (words and other questions on a form). Then the instrument tables store the item-by-item response data for particular CDI instruments. We currently include all items on CDI instruments, including questions about communication, gesture, morphology, and grammar (though in quite a few of the datasets that we archive these non-vocabulary questions have not been digitized so data on these are sparse at present; see e.g., Chapters 7 and 13).

Wordbank is designed so that it can accommodate data from a wide variety of instruments, both within and across languages. Indeed, at the time of rendering, the site includes data from 82055 administrations of the CDI across 29 different languages and 56 different instruments.

3.1.2 Implementation

Wordbank is constructed using free, open-source tools. The database is a standard MySQL database, managed using Python and Django. All code for Wordbank is hosted in GitHub repositories, with the primary site repository containing data and database code, the R package repository containing code for the API, and the book repository containing the code and text for this manuscript.

All data uploaded to Wordbank are open and freely available for download, both through the site itself and through the GitHub repository. The site includes only de-identified data that cannot be linked to individual parents and children under US Department of Health and Human Services' "Safe Harbor" standard. Because of these features, the Stanford Institutional Review Board has determined that the Wordbank project does not constitute human subjects research.

3.1.3 The wordbankr API

An application programming interface (API) is a set of abstractions that allow applications to interact with a resource (e.g., a set of data like Wordbank) through consistent abstractions. Although in principle it is possible to construct raw SQL queries to Wordbank, in practice all access is through an R API that constructs individual SQL calls. This API is distributed to R users through the wordbankr package, which is available through CRAN.

We developed this package, wordbankr, to provide a simple and flexible API for the Wordbank dataset (Frank et al., 2016a), and our current book depends on it heavily. The package provides a consistent set of function calls for retrieving data from the underlying database, for example get_instruments and get_administrations to retrieve all or subsections of these tables, respectively. We do not describe the package in depth here, since it is described in our previous paper and in its online documentation.

3.1.4 "Unilemmas": cross-linguistic conceptual mappings

As described in Chapter 2, it is sometimes useful to (cautiously) compare the developmental trajectory for a single concept across multiple languages. To facilitate these comparisons, we created "unilemmas," cross-linguistic mappings from lexical items to single (English) forms that stand for a particular conceptual abstraction. Some lexical items are represented on only one or a handful of instruments, but there are many that are common across a large number of instruments, leading to an opportunity for cross-linguistic comparison.

Unilemmas were created for particular instruments by following a two-step procedure. First, using a pool of English unilemmas, we proposed candidate mappings for each lexical item on a form. This first step was often accomplished by a non-native speaker using translation resources and the context of the form (e.g., that an item occurs in the "animal sounds" section). Second, we recruited a linguistically-sophisticated native speaker of the language (often a psychologist or linguist), provided them with the candidate unilemma list, and asked them to review this list item by item and suggest corrections and amendments.[2]

Not every instrument has unilemma mappings, but they are currently available, at least partially, for 46 out of the 56 instruments.

3.1.5 A note on age

Developmental psychologists are very fond of using temporal units like months and years as rough guides. Children tend to begin to crawl between 5 and 8 months, and say their first word around one year. This practice is fine for rules of thumb, but we also use these units for measurement as though they were precise (e.g., "infants with ages between 7;0 and 8;0") when in fact such infants will vary in the number of days since their birth depending on facts like whether their seven months of life encompassed February or not. A similar problem is true of years as a scientific unit — because of

[2]The specific direction they were given was: "We're looking for the best English translation of these words. These are words that are among the first words that children learn, so your translation should be closest to the meaning of the word as it would be used by a young child (say, under 3 years old). For cases when there are two equally good English words, put both. If you don't think there is a good translation into a reasonable English word that a kid might know, you can leave the alternative translation blank."

leap years, years technically include 365.2524 days — though the magnitude of the imprecision is smaller.

Despite these issues, months are the currency of language development research, and we often receive contributed datasets with months as the only measure of age. In Wordbank, we define a standardized month as 365.2524 / 12 = 30.4377 days. When possible, we compute the number of days from birth to testing and then compute the number of standardized months that the child has lived. If this is not possible, we use months as reported in the dataset. We define an eight-month-old (age == 8) as a child who has lived between 8 and 9 standard months: their age is in the range [8–9) standard months. (The alternative definition, from 7;16–8;15, is sometimes used in infancy research but is in our opinion less intuitive.)

3.2 Datasets

This section gives a broad overview of the data we have available. Unlike projects in which data are collected by the organizers, in our work here, we rely on the kindness of others in contributing data that are often years or decades old. Some datasets come via an email containing well-curated tabular data; others were contributed in more idiosyncratic formats or even on paper. One dataset was even retrieved by one of us from a doorstep several hours drive away, in the form of a paper bag full of old paper forms. Thus, the amount and type of meta-data available for some datasets is limited. For example, we have limited demographic information for some datasets and only vocabulary — not complexity or gesture — items for others. In many cases we do not have full details of instructions and administration for a particular dataset. This section gives an overview of data availability and some demographic comparisons of the samples. Specifics of each dataset — to the extent that they are available — are given in Appendix A.

3.2.1 Data provenance

As mentioned above, datasets come from a variety of sources. In all cases, the preferred citation for each dataset and its contributor is given on the Wordbank contributors page. Several of these datasets were transferred second-hand from a pre-existing database (CLEX-CDI; Jørgensen et al., 2010), while many of the others were contributed directly via electronic or paper forms. In the case of paper forms, we re-keyed the forms using double-entry methods (either ourselves or via a commercial contractor).[3]

Each of these datasets is then imported to the database by creating a custom import key that matches individual columns of the dataset to particular database fields (e.g., item types like words or gestures, or standardized demographic fields). These mappings are preserved along with the raw data so that they can be re-checked later.

3.2.2 Overview of the data

Wordbank currently contains data from 29 language communities. Many of these are from instruments in the original Words & Gestures (infant) / Words & Sentences (toddler) format, with around 400

[3]In a check for errors in the re-keying of one Korean dataset, we found that there were 4 incorrect fields in 10 full records for an error rate of ~0.06%.

items in WG and 700 in the WS. Typically, WG forms are intended for children from 8–18 months and WS forms are intended for children 16–30 months, but these ranges are flexible. Some WS forms are used up to 36 months or extended as low as 12 months (in cases where a single form is considered desirable by the researchers constructing the adaptation).

Wordbank also includes some other forms that do not fit into this schema, including short forms and vocabulary questionnaires. Some of these are "short forms" with no internal category structure and fewer items overall, and these are excluded from many item and category analyses. But others have many structural features of WS and WG forms. For example, the Oxford CDI is a WG-style form with comprehension as well as production estimates, but applied to a larger age-range. The Mandarin Infant Checklist (IC) and Toddler Checklist (TC) are checklist forms without grammatical and gesture items but with structured sets of vocabulary items. We include these forms in analyses where WS and WG data are included.

Table 3.2.2 shows an overview of the instruments in Wordbank.

Overview of the available instruments in the Wordbank dataset.

Language	Form	Categories	Items	Words	Age min	Age max
American Sign Language	FormA	18	561	536	8	36
American Sign Language	FormBOne	22	699	674	8	36
American Sign Language	FormBTwo	21	562	537	8	36
American Sign Language	FormC	21	563	538	8	36
British Sign Language	WG	22	569	548	8	36
Cantonese	WS	25	815	804	16	30
Croatian	WG	19	396	396	8	16
Croatian	WS	22	717	717	16	30
Czech	WS	21	553	553	16	30
Danish	WG	20	410	410	8	20

Showing 1 to 10 of 56 entries Previous 1 2 3 4 5 6 Next

The number of administrations available is highly variable across instruments and languages, however. Figure 3.2 shows the distribution of administrations across forms and languages.

These instruments have global reach, although the maximal number cover North America and Western Europe. African, South American, and South/South-East Asian languages are notably under-represented. Figure 3.3 indicates which countries' languages are represented.

3.2.3 Administration details

Data in the dataset were gathered between the beginning of the first CDI norming study in 1990 and the present, with the majority of datasets gathered within the 10–15 years prior to the writing of this book. The details of administration vary widely from dataset to dataset. Though we have different levels of knowledge regarding the exact details of administration, we know that the three most common circumstances of administration (in no particular order) are:

1. On paper in a lab or other space, with instructions given in person by a researcher (e.g., Fenson et al., 1994);

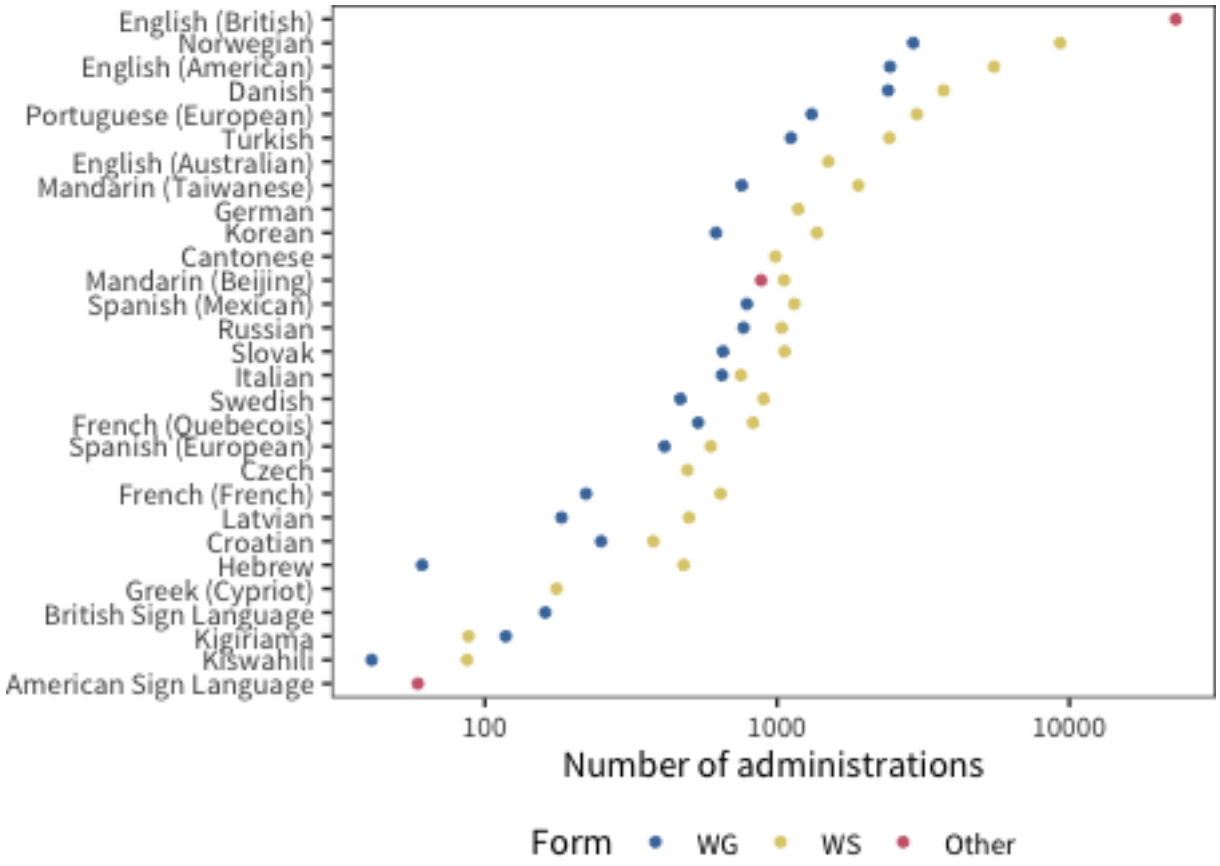


Figure 3.2: Log-scaled number of administrations for each instrument.

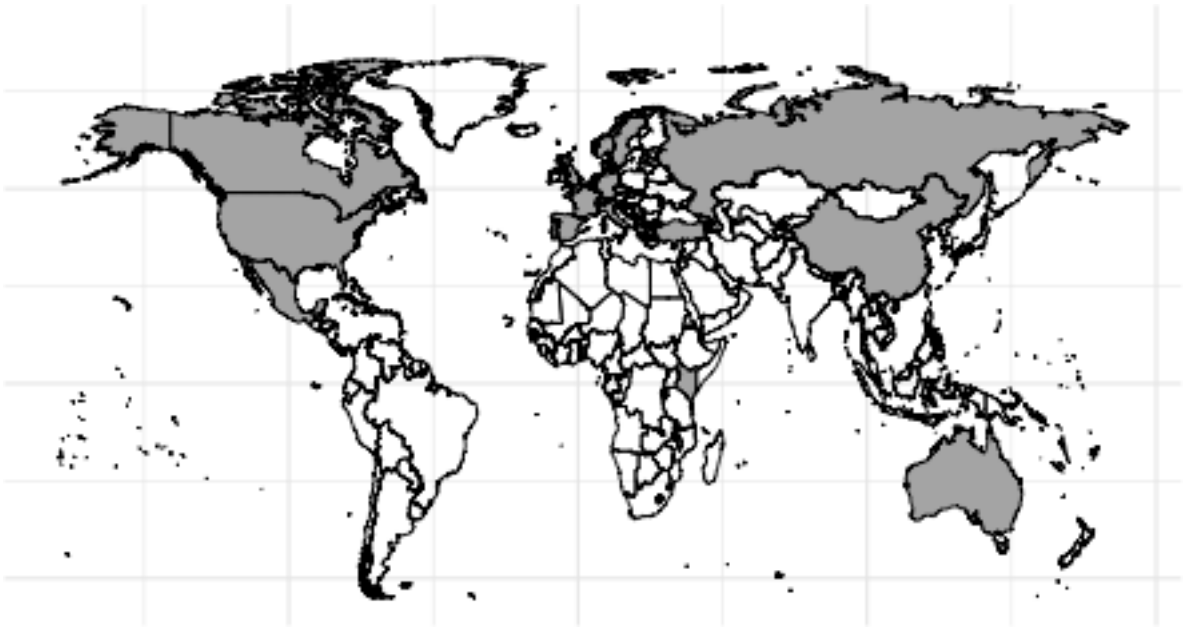


Figure 3.3: World map with countries shaded whose dominant language is represented in the dataset.

2. On paper, with the form sent by mail with written or telephone instructions from a researcher (e.g., the British English Twins Early Development data, which were sent home as part of a packet; Dale et al., 2003); or
3. Electronically, with instructions given either electronically or by phone (e.g., Kristoffersen et al., 2013).

We have limited direct evidence about the effects of particular administration details on the overall results. Such evidence would require random assignment of parents to administration method rather than, e.g., a comparison of administration methods across different populations in which there are obvious sample-related confounds. Nevertheless, the CDI community has amassed a substantial set of anecdotal experiences. For example, improper administration or limited instructions can result in over- or under-reporting, especially with respect to comprehension (see e.g., Feldman et al., 2000).

In one trial we conducted using electronic administration, we found that basic written instructions were misinterpreted by some proportion of parents (as evinced by an atypical number of floor and ceiling responses). This proportion appeared to decrease when we made an attempt to simplify and illustrate the instructions that we gave. Such experiences suggest — congruent with the general warnings above — that caution is warranted in interpreting absolute comparisons between different populations where there are also differences in administration style.

3.2.4 Demographic details

In addition to differences in administration and form, samples from different studies also differ in myriad other ways. The most important of these, especially cultural differences between language communities, are extremely hard to quantify. But we can make a first stab at investigating some similarities and differences between the convenience samples from different studies by comparing demographics where they are available. The demographic makeup of our datasets is shown in Figure 3.4 for sex, Figure 3.5 for maternal education, and Figure 3.6 for birth order.

Sex proportions tend to be quite close to .5, with a few exceptions for small datasets. Several WG datasets (e.g., British Sign Language, Russian, Italian, Quebec French) have more males than would be expected by chance. This pattern is important because (as we will investigate in Chapter 6), there are systematic differences in vocabulary size between boys and girls, and so sample differences in gender will lead to absolute differences in mean vocabulary size.

Although we have maternal education data for far fewer datasets, there are also substantial differences between datasets on this variable (we will also return to this issue again in Chapter 6). Analyses of this variable are complicated by different reporting formats, so for example the German and Mexican Spanish datasets have no separate categorization for graduate education. That said, even for datasets with the most fine-grained maternal education breakdown, we see substantial differences.

Finally, when we examine birth order, we also see differences in the proportion of children who are first- vs. later-born. The majority of the German sample is first-born, while the Czech sample has many more second children, for example.

In summary, our samples differ substantially in their demographic makeup. Presumably these differences are due both to the composition of the societies being sampled as well as the sampling procedure the researchers used.

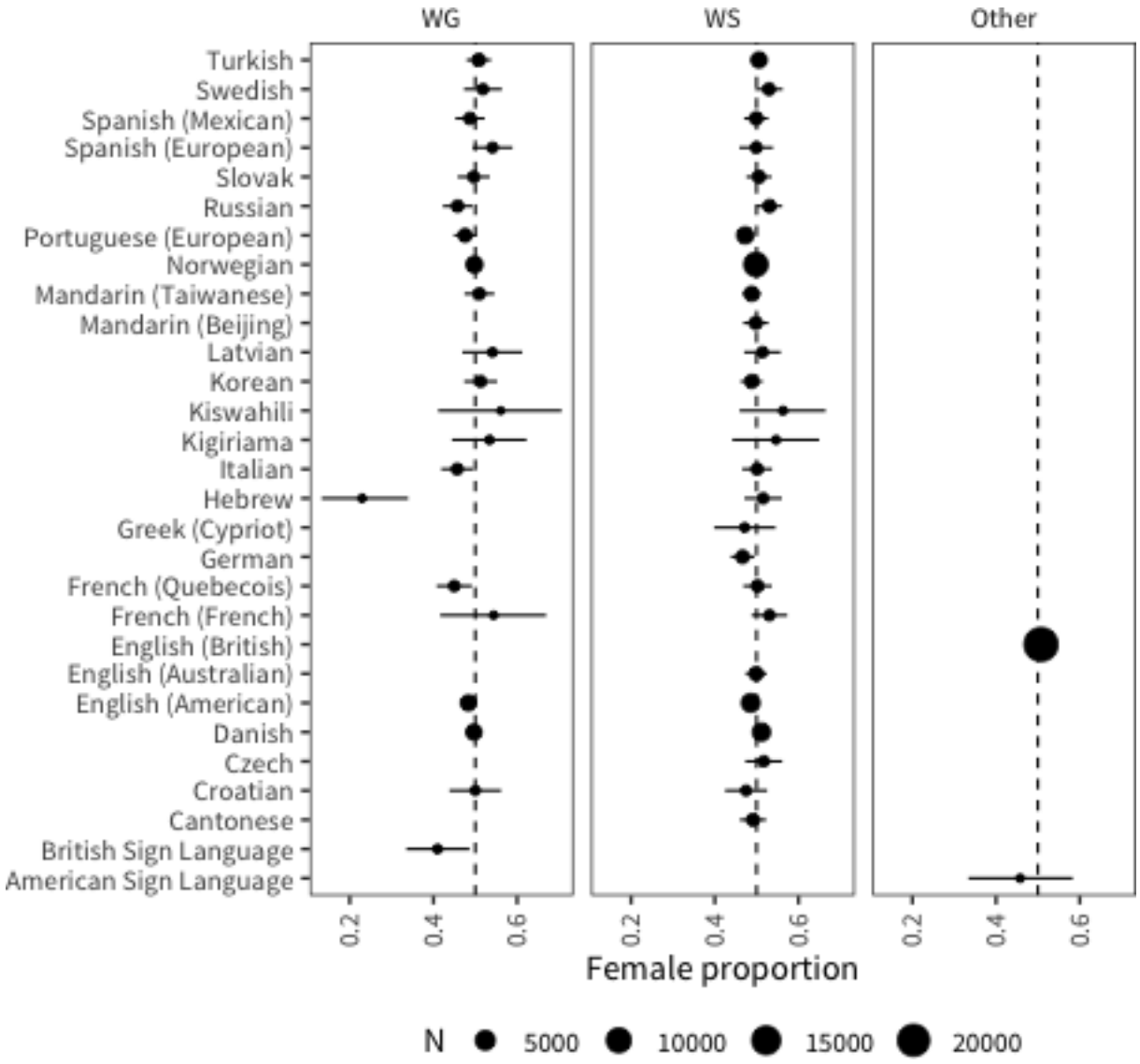


Figure 3.4: Proportion of female-assigned children for each instrument.

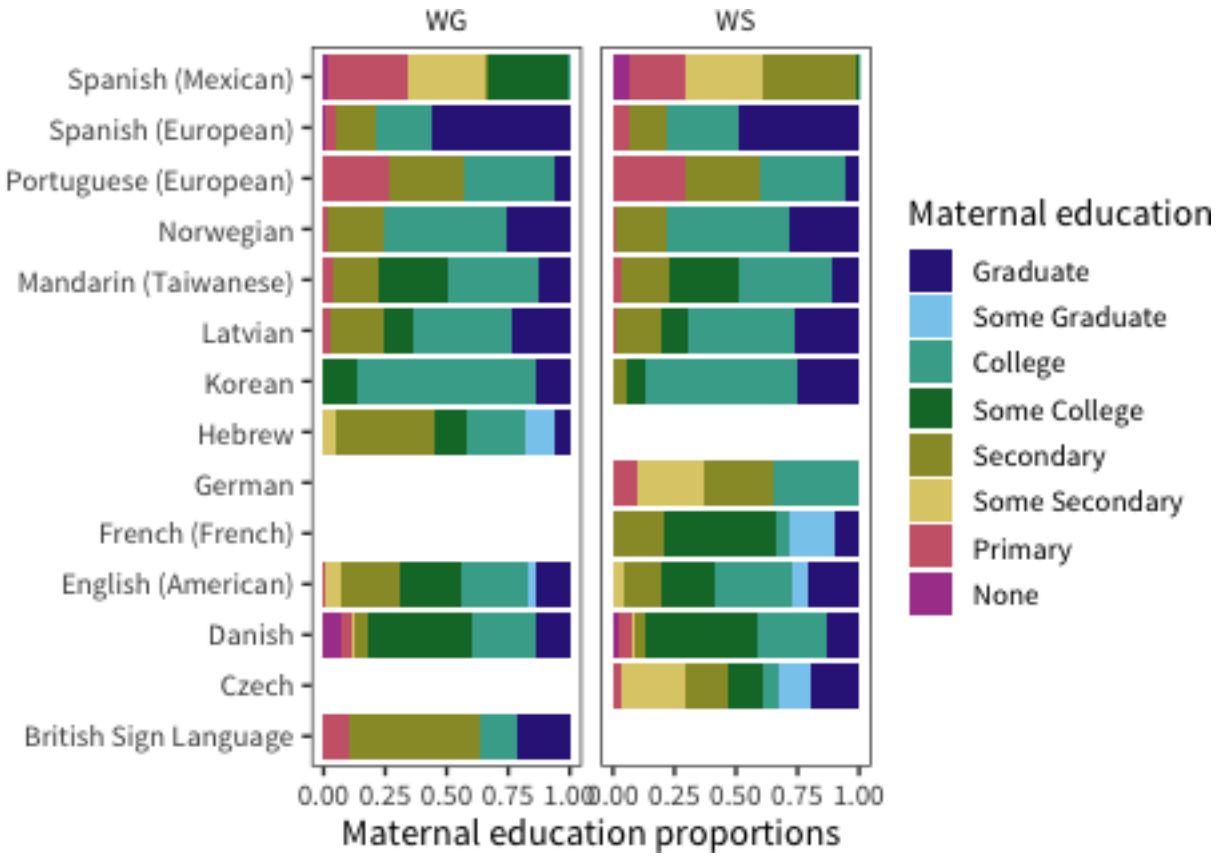


Figure 3.5: Proportions of children with each level of maternal education for each instrument.

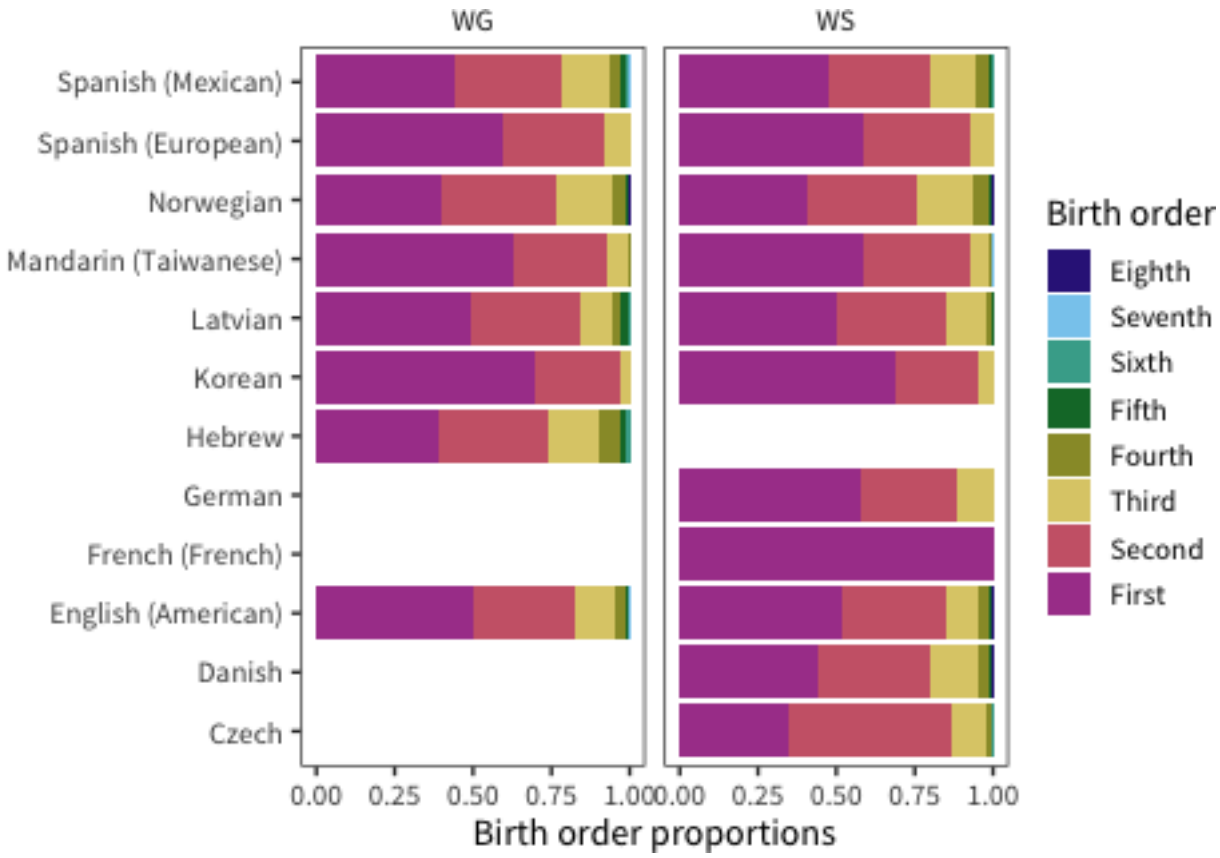


Figure 3.6: Proportions of children with each value of birth order for each instrument.

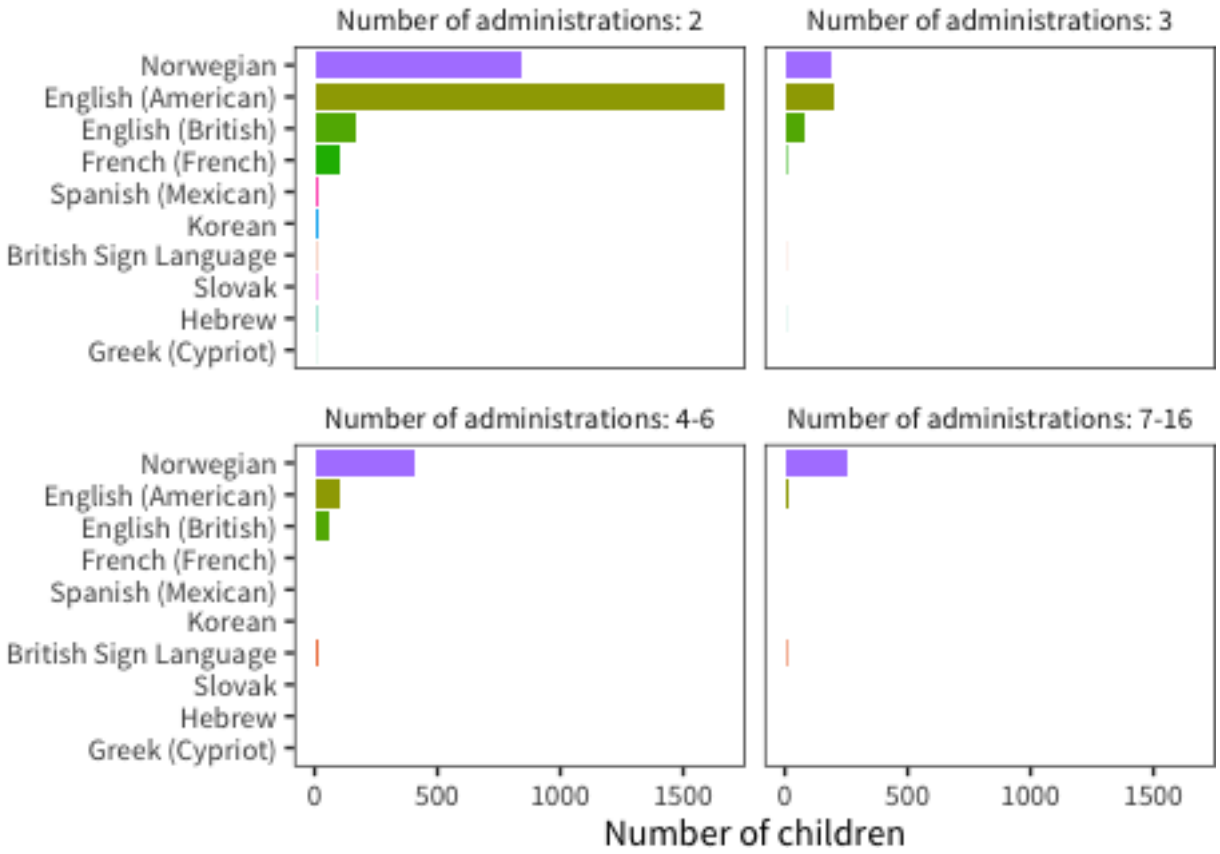


Figure 3.7: Number of children for whom there are multiple administrations for each instrument, split into bins.

3.2.5 Longitudinal vs. cross-sectional data

The strongest developmental inferences can be made by the examination of longitudinal data, in which children's individual development is measured multiple times using the same instrument. Unfortunately, relatively little of our CDI data comes from this type of repeated administration. Figure 3.7 shows the number of administrations for particular languages that come from longitudinal datasets with a particular depth. There is a substantial amount of two-administration longitudinal data for several languages, but only a few have more than two observations for an individual child.

In general, this aspect of our data is a consequence of the fact that, for normative datasets, pure cross-sectional data collection is used to ensure statistical independence between datapoints. Thus, we must typically settle for using the large amount of available cross-sectional data to average out individual variability. We do use the more extensive Norwegian and English longitudinal data in Chapter 14, however.

3.2.6 Difficult datasets

One feature of dealing with data from such disparate sources can't be glossed over. There are "difficult" datasets — data that do not make sense with respect to our other analyses. This short

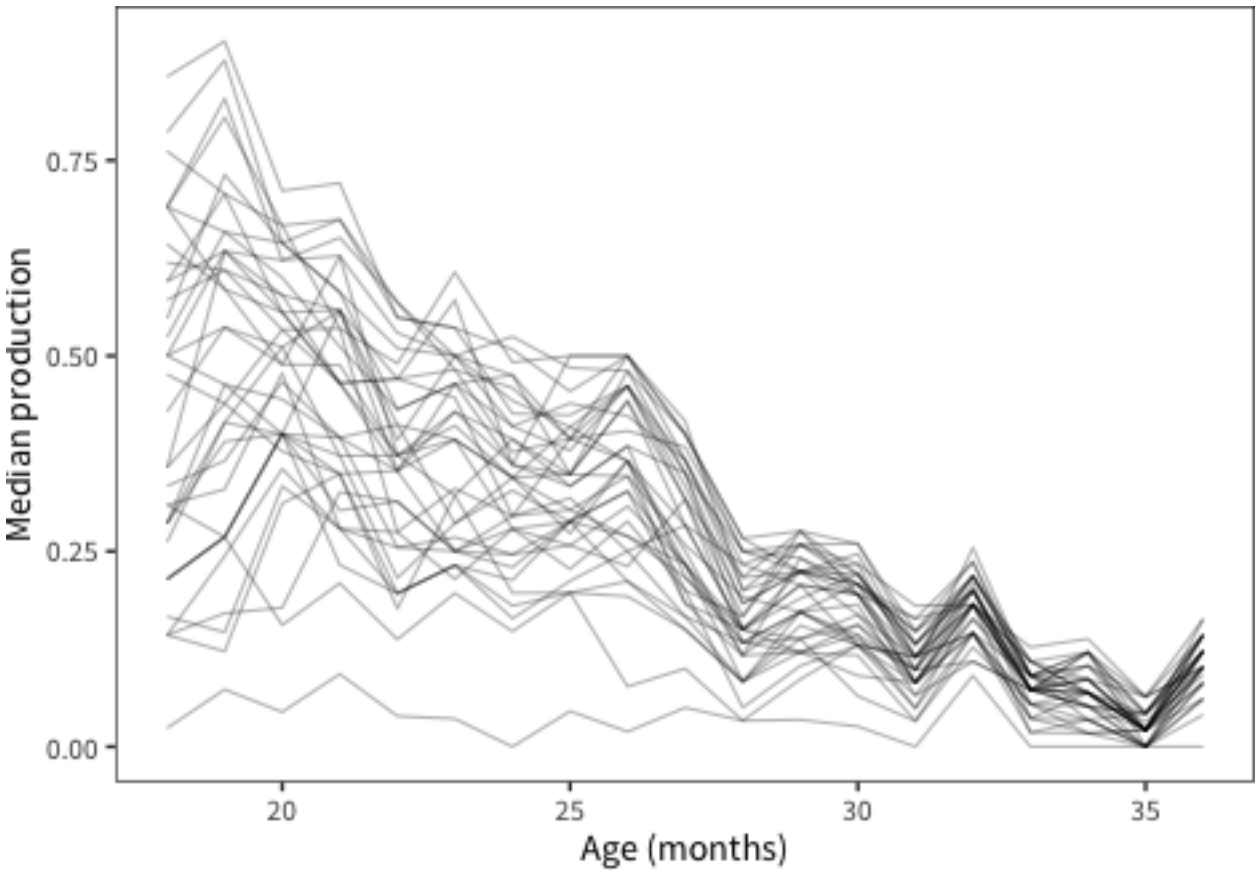


Figure 3.8: Mean proportion children reported to produce each item in the sounds category for Russian Words and Sentences.

section documents some of these issues (helping itself more intuitively to a few visualizations that will be developed in more detail in subsequent chapters).

In general, our approach with respect to these data is to embrace the messiness of the data we have. While it is very tempting to remove specific datasets from consideration when they deviate from our expectations, this practice creates a strong circularity in all of our inferences: they will be estimates of variability or consistency stemming from cases where we ourselves have imposed certain consistency standards on our data. While there are some cases where we have a relatively likely explanation close to hand for the pattern we observe in the data, unless we can confirm this pattern externally, we have chosen not to exclude these data.

One small example of this kind of situation comes from the Russian dataset. Although — as we will explore in depth — nearly every individual item in every dataset shows a positive developmental slope (indicating learning over time), Russian animal sounds are a distinctive exception, as shown in Figure 3.8. Every item in this category decreases developmentally in a very consistent and reliable way. What happened? One possibility is that this set of items was reverse coded (and so it should be asymptoting at three years). Another possibility is that Russian parents treat these as "baby words" that a three-year-old wouldn't or shouldn't produce (e.g., rather than saying oink they should say pig.) We can speculate but we will likely never know.

Extending more broadly (and presaging the discussion in Chapter 5, our analyses have revealed

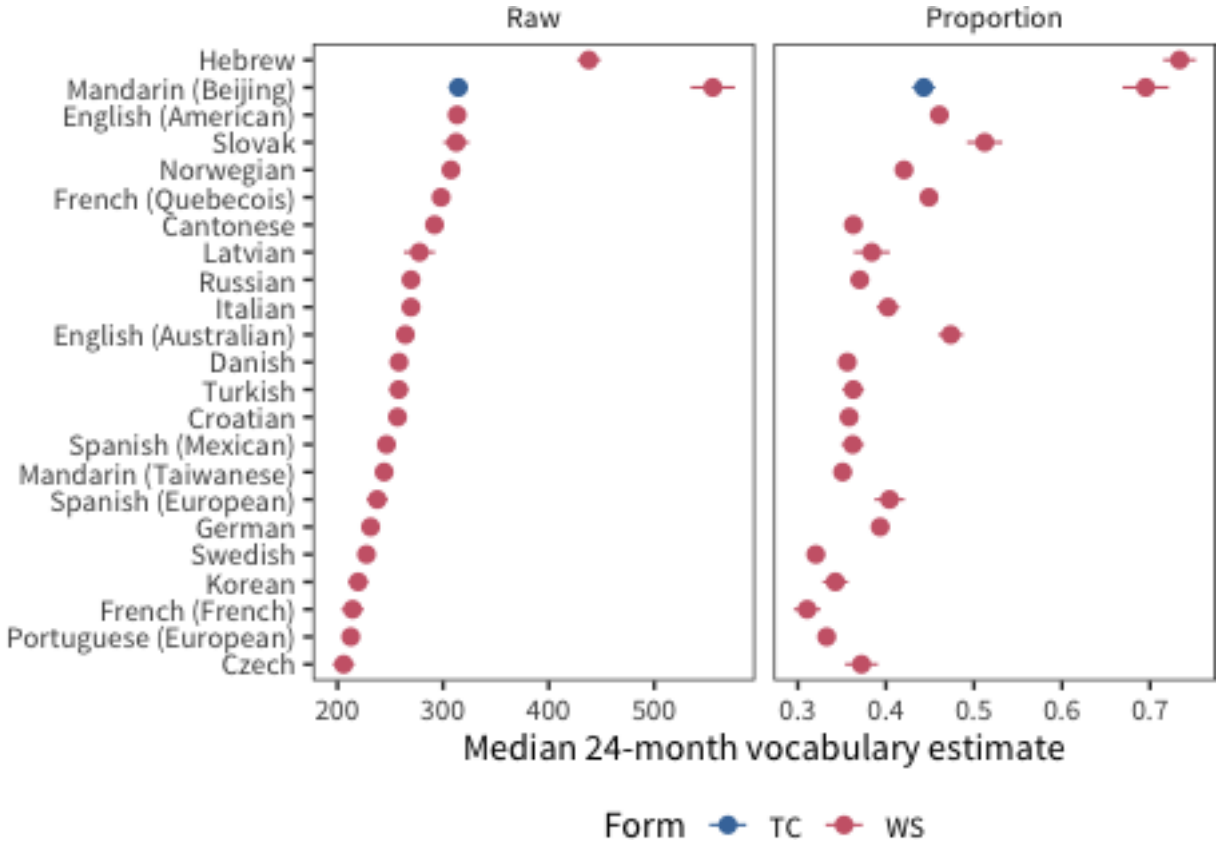


Figure 3.9: Median production vocabulary for 24-month-olds, with total item scores shown in the left panel and proportions on the right. Scores are sorted by total item score. To increase stability, the plotted value is the intercept of a linear model predicting vocabulary as a function of centered age between 18 and 30 months.

two datasets that show large disparities not just in a single category but in the pattern of overall vocabulary sizes: Mandarin (Beijing) Words & Sentences production and Mandarin (Taiwain) Words & Sentences comprehension.

Turning to our first example, Mandarin Words & Sentences data are reported by Tardif et al. (2009) in a study of both Mandarin- and Cantonese-learning children. The data reported there show a pronounced Mandarin advantage. As it turns out, this advantage is almost unprecedented relative to other languages. We plot the median production for 24-month olds in Figure 3.9. This figure reveals both how large the Mandarin advantage and the high level of vocabulary reported for Hebrew speakers as well, which is less unusual in raw scores because of the relatively smaller number of items on the Hebrew form.

To investigate the Mandarin disparities further, Tardif et al. (2009) discussed a number of possible explanations, given that the administration and sampling procedures were similar in these two languages. The children in the Mandarin sample are nearly all monolingual, only (first born) children; but these factors did not account for variation between samples. Tardif et al. (2009) therefore, speculate that structural factors regarding Mandarin (e.g., phonological structure relative to Cantonese) might be accounting for the Mandarin advantage.

These speculations seem unlikely in light of the data presented here. First, and perhaps most

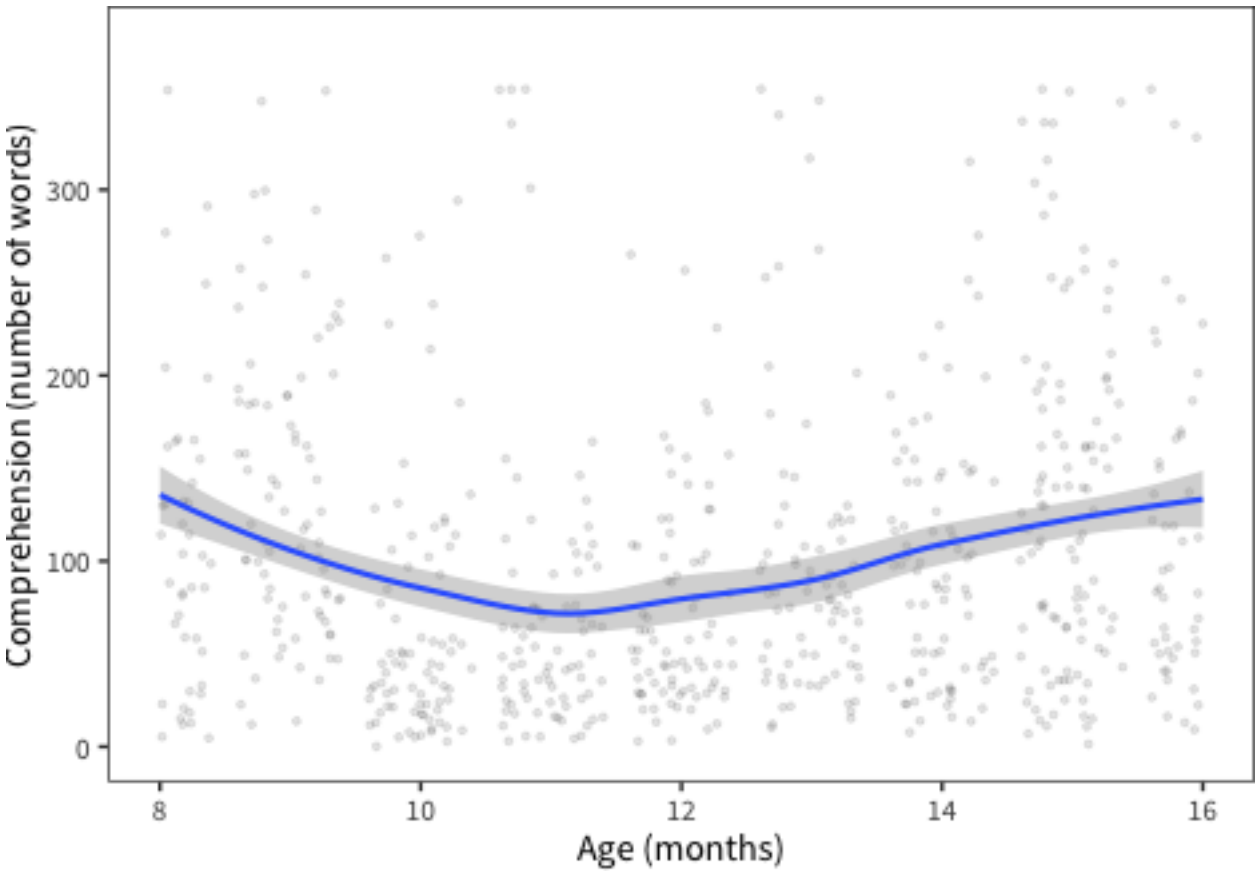


Figure 3.10: Comprehension data from Taiwanese Mandarin.

importantly, the same magnitude response is not shown in the data from the analogous Toddler Checklist questionnaire of Hao et al. (2008) (blue points). Second, this unusual trajectory is not apparent in the production data from the Mandarin Beijing WG form. Finally, given the surprising difference between Mandarin and all other languages in the sample, pure phonological factors seem unlikely to account fully for the differences. These differences thus remain somewhat mysterious; perhaps some quirk of administration instructions led to relative over-reporting, or perhaps the populations being sampled truly were different. Alongside the Hebrew data, these data serve as an important caution against simple cross-linguistic comparison in raw scores or even percentiles.

Turning to our third case study, Mandarin (Taiwanese) comprehension scores, we see that they are relatively flat and show very high medians very early in development. Deeper inspection of the full distributional pattern (Figure 3.10) suggests that there is relatively little developmental change in comprehension scores on this dataset. In contrast, production, seen above, appears to follow a more typical pattern. In our experience, this pattern results from parents who do not understand what is being asked on the comprehension section of a form; sometimes they report whether they think a child has heard a particular word, or whether they respond to language in more general ways. We have observed a population of "over-responders" of this sort in a number of self-report contexts — often they are parents of very young children who appear loathe to return a form having checked essentially no items at all. But such an explanation is only speculation.

These are a few examples, but in fact there are quite a number of "difficult datasets" in one way or

another. While we have offered some tentative explanations of a few features, these are necessarily post hoc and rely on our assumption that they should be relatively similar to other datasets from other cultures and with other forms. Thus, in our further analyses we choose not to omit these data but instead consider them as a caution on making strong inferences from variability rather than from consistency. As we discussed in Chapter 1, variability may be caused by a wide variety of sources; consistency is somewhat more surprising.

3.2.7 Conclusions

The strength of the Wordbank framework is that it allow access to CDI data in a consistent format, such that analyses can be applied uniformly. Yet we must not allow this ease to blind us to the difficulties of comparing across measurements that are gathered using different forms, under different administration conditions, and from convenience samples in different countries and cultures using different sampling schemes. Each of these differences has the potential to complicate cross-linguistic comparisons. We will return to each throughout the book.

Chapter 4

Measurement Properties of the CDI

Many researchers are initially shocked to hear that one of the most important methods for studying child language is parent report. Yet, as we argued in Chapter 2, alternative methods like naturalistic observation or lab experiments can be biased, and are quite costly to revisit at scale. Thus, the goal of this chapter is to revisit the strengths and weaknesses of parent report in depth, since the remainder of our manuscript depends on the use of CDI data.

Broadly speaking, we would like to provide evidence for the reliability and validity of the CDI. Many studies provide evidence for reliability in the form of concurrent and longitudinal correlations between CDI scores and validity in the form of correlations between the CDI and other language measures; some of the most prominent of these studies are cited below and others are reviewed in Fenson et al. (2007). Here we address some issues that have received a little less attention: in the first part, we discuss the limitations of the CDI (and the design features that address these limitations); in the second part, we use longitudinal data to examine the test-retest reliability of the CDI; and in the third part, we present evidence for the measurement properties of the CDI (including comprehension questions) from a psychometric perspective.

4.1 Strengths and limitations of parent report

Although the standardization of parent reports using the CDI contributes to the availability of large amounts of data in a comparable format, there are significant limitations to the parent report methodology that are important to understand (Tomasello and Mervis, 1994; Feldman et al., 2000). To do so, it is useful to reflect on what it means when a parent reports that their child "understands" or "understands and says" a word. In an ideal world, the parent's responses would be an unbiased reflection of their observations of their child's language development. For example, when asked if their child produces the word ball, a parent is likely recalling situations in which their child has used the word ball correctly, and then reporting on the success or failure of this process of recollection. Of course, this judgment clearly depends on the parent's ability to accurately judge that the child intended to say the word ball, that the child's target word form was ball, and that the child has some meaning for the word form ball that at least approximates the expected meaning. There are also a number of other sources of information that the parent might bring to bear on these judgments.

Figure 4.1 shows a sketch of the process of parent report. For each word on the CDI, the parent is asked to report whether their child has produced or comprehended the word. This report could depend

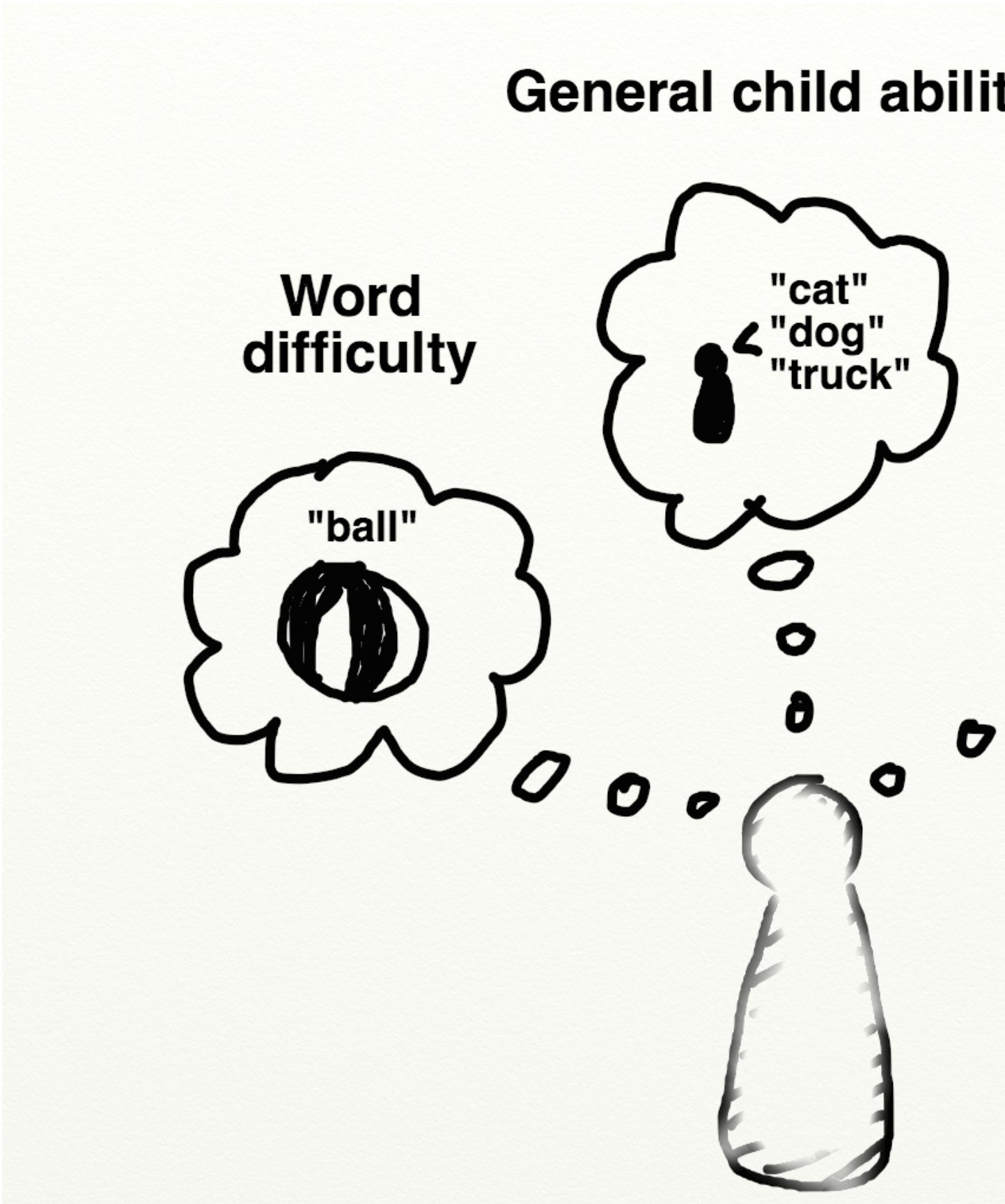


Figure 4.1: The intuitive structure of parent report.

on direct recall of a particular case when their child actually produced or showed comprehension. But in addition to these factors, parents probably draw on their general assessment of the difficulty of the word and on their overall assessment of the child's linguistic abilities. As even this simple sketch shows, parent report judgments are based on a fairly complex set of factors. And hence there are legitimate concerns about the ability of parents to provide detailed and specific knowledge about their children's language. We discuss specific concerns below.

First, parents may be biased observers generally. Most parents do not have specialized training in language development, and may not be sensitive to subtle aspects of language structure and use. Further, a natural pride in the child and a failure to critically test their impressions may cause parents to overestimate the child's ability; conversely, frustration in the case of delayed language may lead to underestimates. Parent report is most likely to be accurate under three general conditions: (1) when assessment is limited to current behaviors, (2) when assessment is focused on emergent behaviors, and (3) when a primarily recognition format is used. Each of these conditions acts to reduce demands on the respondent's memory. For example, parents are better able to choose from a list of items that are likely candidates, rather than requiring that the parents generate the list themselves. In addition, parents are likely to be better able to report on their child's language at the present time than at times past and when their child is actively learning the particular words on the list (e.g., names for animals).

Second, parent reports likely suffer from a number of biases that interact with sub-portions of the forms and the ages of the target children. For example, it is likely that parents may have more difficulty reporting on children's comprehension or production of function words (e.g., so, then, if) than content words (e.g., baby, house) — relying more on their estimates of the words general difficulty. We return to this question below in our psychometric analyses. Moreover, in typically-developing samples, parents can track their child's receptive vocabulary to about 16–18 months, after which it is too large to monitor. Expressive vocabulary can be monitored until about 2.5–3 years, after which the number of words a child can say becomes too large. Different instrument developers make different choices about the ceiling of CDI-type forms but relatively few have considered CDI-type parent report for measuring older children's vocabularies (but cf. Libertus et al., 2015).

The CDI instruments capitalize on the greater ease of recognition, as contrasted with free recall, to help offset these memory limitations. That is, it is better to ask parents to report on their child's vocabulary by selecting words from a list of possible words rather than having them write down all the words they can recall hearing their child use (or, even worse, asking the global question "Does your child know at least 50 words?" that is so commonly used in pediatric assessments).

In addition, asking parents to reflect on their child's language abilities may be particularly difficult for early vocabulary and especially for early comprehension. As Tomasello and Mervis (1994) point out, for the youngest children, especially 8–10 month olds, vocabulary comprehension scores can be surprisingly high, possibly reflecting a lack of clarity in what the term "understands" means for parents of children at this young age. On the other hand, more recent evidence has suggested that children in this age range do plausibly have some comprehension skill even if it is somewhat fragmentary (Tincoff and Jusczyk, 1999, 2012; Bergelson and Swingley, 2012, 2013, 2015). Thus, the degree to which very early comprehension reports are artifactual — or were actually ahead of the research literature — is unknown. (Resolving this question will require detailed studies of the correspondence between parent reports and experimental data for individual children). Below we assess some of the measurement properties of comprehension items, but we are unable to resolve the issue fully.

One study that bears on the earliest production data is Schneider et al. (2015), who compiled a number of sources of data on children's first words. Surprisingly, they found relatively few differences for the age and topic distribution of this very salient milestone across datasets collected via a number of different methods, including concurrent (CDI) and retrospective report. The age at which a first word was reported was also relatively similar between CDI data and the concurrent diary reports of a sample of psycholinguists (though some CDI data appeared to be shifted a little bit earlier such that more parents were reporting first words in the 7–9 month period).

Third, there is some evidence that variability in reporting biases may be moderated by factors such as SES (Feldman et al., 2000; Fenson et al., 2000; Feldman et al., 2005). Some studies suggest that parents from some SES groups may be more likely to underestimate child's abilities (Roberts et al., 1999), while others report that parents from lower-SES groups may over-estimate children's abilities, especially comprehension at younger ages (Goldfield and Reznick, 1990; Feldman et al., 2000). Later studies, however, have shown that for children over 2 years patterns of validity were consistent in lower and higher-SES groups (Feldman et al., 2005; Reese and Read, 2000). Thus, SES-differences could reflect valid delays in children's language development that parallel those obtained with different methods, such as naturalistic observation or standardized tests (e.g., Hammer et al., 2010).

Fourth, as discussed in Chapter 2, the items on the original CDI instruments were chosen to be a representative sample of vocabulary for the appropriate age and language (Fenson et al., 1994). The checklists contain some words that most, including the youngest, children are able to understand or produce, some words that are understood or produced by the "average" child, and some which only children who are relatively more advanced will understand or produce. This structure ensures that the list has the psychometric property of capturing individual differences in vocabulary both across younger and older children and across children of different developmental levels. Validity of the CDIs has been demonstrated in reference to both standardized tests and naturalistic language sampling (see Chapter 4 of Fenson et al., 2007).

But the checklists were not originally constructed with the intention that responses on individual items would be reliable. While item-level responses provide useful information about patterns of words that children are likely to understand or produce, responses on the vocabulary checklist do not necessarily license the conclusion that a child would respond appropriately when asked "can you say _____?" by an experimenter in a confrontation naming task. Nonetheless, if parents' observations at the item level reflect any signal — even in the context of significant influence from other factors — then this signal should be observable by aggregating together data from many children. Thus, the item-level analyses we present in Chapter 10 (for example) are not predicated on an assumption of high item-level reliability for individual children.

Fifth, while the lengths of the vocabulary checklists on the CDIs may give the impression that they yield an estimate of the child's full vocabulary, in fact the vocabulary size estimates only reflect a child's relative standing compared to other children assessed with the same list of words (see Mayor and Plunkett, 2011, for discussion). Such estimates should not be misconstrued as a comprehensive estimate of the child's vocabulary knowledge, as CDI scores likely understate the size of a child's "true" vocabulary substantially, especially for older children.

Sixth, when a parent reports on a word on the vocabulary checklist, there is no information about the actual form of the word used, and hence, these vocabulary estimates can say little about phonological development (e.g. segmental vs. suprasegmental approaches to the analysis of speech). Parents are instructed that they should check that a child can produce a word even if it is pronounced in the

child's "special way," and only approximates the adult form. Thus, throughout this book we refrain from analyzing the phonological forms of words reported on CDI instruments (with the exception of Chapter 10, in which we use word length as a predictor of production).

Finally, we also gain little information about the frequency with which children use a particular word in their spontaneous speech, nor can we know the range of contexts in which individual lexical items are used (e.g., is that word used productively vs. in a memorized chunk of speech). Thus, the vocabulary size that is captured by the CDIs reflects the number of different word types (not tokens) that the child is able to understand or produce, with little information about nuances in meaning that might be reflected in actual usage.

In sum, despite these limitations, when used appropriately, the CDI instruments yield reliable and valid estimates of total vocabulary size. Because the instruments were designed to minimize bias by targeting current behaviors and asking parents about highly salient features of their child's abilities, they have proven to be an important tool in the field. Dozens of studies demonstrate concurrent and predictive relations with naturalistic and observational measures, in both typically-developing and at-risk populations (e.g., Dale and Fenson, 1996; Thal et al., 1999; Marchman and Martínez-Sussmann, 2002). In addition, a variety of recent work has shown that individual item-level responses can yield exciting new insights, for example about the growth patterns of semantic networks when aggregated across children (Hills et al., 2009, 2010). Such analyses have the potential to be even more powerful when applied to larger samples and across languages.

4.2 Longitudinal stability of CDI measurements

The first question that we address here is with regards to the longitudinal stability of CDI reports. A classic test of the reliability of a psychometric instrument is its test-retest correlation. Assessing this correlation for CDIs for a single reporter is a bit impractical however, since — unlike e.g., a math test with objective answers and different question forms — this procedure would involve asking a caregiver to fill out the exact same survey twice in a row, and presumably they would remember many of their answers. An alternative possibility would be to measure the same child via multiple caregivers. This procedure was followed by De Houwer et al. (2005), who found that caregivers varied substantially from one another in their responding; but plausibly this is due not only to parent bias but also to the different contexts in which caregivers interact with children (e.g., one caregiver takes the child to the zoo more often, another plays kitchen at home).

Avoiding the issues of these procedures, we instead examine correlations in CDI measurements across developmental time. There are only a small number of deeply longitudinal corpora in Wordbank, so we will limit our investigation to two languages: Norwegian and English. Furthermore, the largest group of longitudinal data cover the WS form so we restrict to these data for simplicity. Within each of these datasets, the modal number of observations is two, but there are some children with more than 10 CDIs available.

Differences between a particular individual's measurements could vary for two primary reasons: first, measurement error (parent forgetfulness, mistakes, etc.) and second, true developmental change (learning new words). Since all children's vocabulary increases over time, we can look at the relative magnitudes of CDI scores via correlations; this is our first analysis. Our second analysis attempts to normalize these absolute differences by extracting percentile ranks and finds that this procedure in fact increases longitudinal correlations. Because there are two sources of differences between

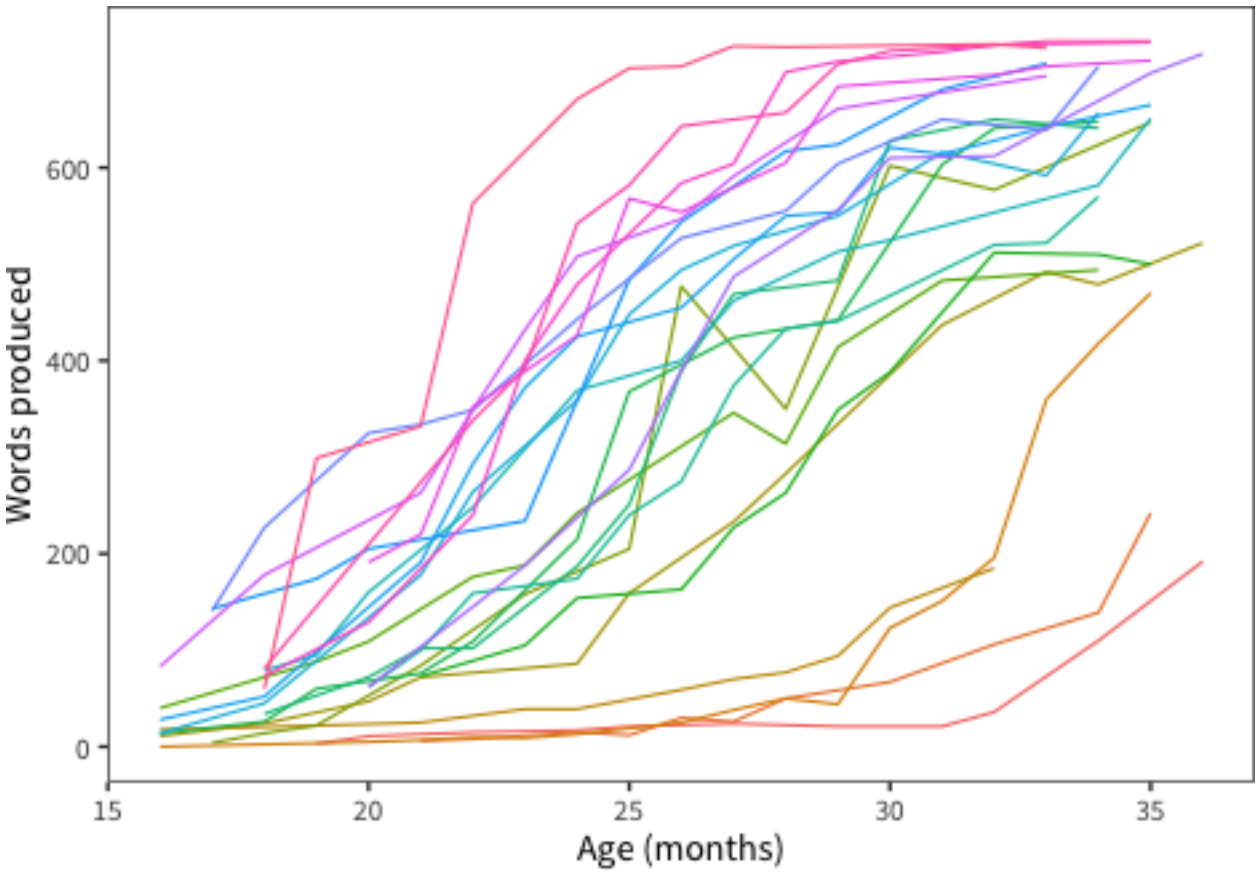


Figure 4.2: Vocabulary size as a function of age for children with more than 10 administrations (color indicates child).

measurements, when correlations are low, we do not have direct evidence for whether 1) children's relative linguistic abilities are shifting with respect to one another or 2) we are observing measurement error. But when correlations are high, we can assume the converse: measurement error is low and developmental stability is relatively high.

It turns out that this is the case: The first locus for individual differences in language acquisition is the rate of growth. As we will discuss in more detail in Chapter 5, there is substantial variability between children in vocabulary size. This variability appears to be quite stable longitudinally.

Figure 4.2 shows the trajectories of children (individual colors) who were measured more than ten times, and includes Norwegian data only due to data sparsity issues in English. These trajectories appear quite stable; the ranking of individuals does not appear to change much over the course of several years. We quantify this trend below. This general conclusion — longitudinal stability of language ability as well as limited measurement error — is ratified by other studies using different datasets, for example Bornstein and Putnick (2012), who found substantial stability (r = .84) between latent constructs inferred from early language at 20 months and later language measured at 48 months.

One way to operationalize the question of stability is how children's percentile ranks tend to change

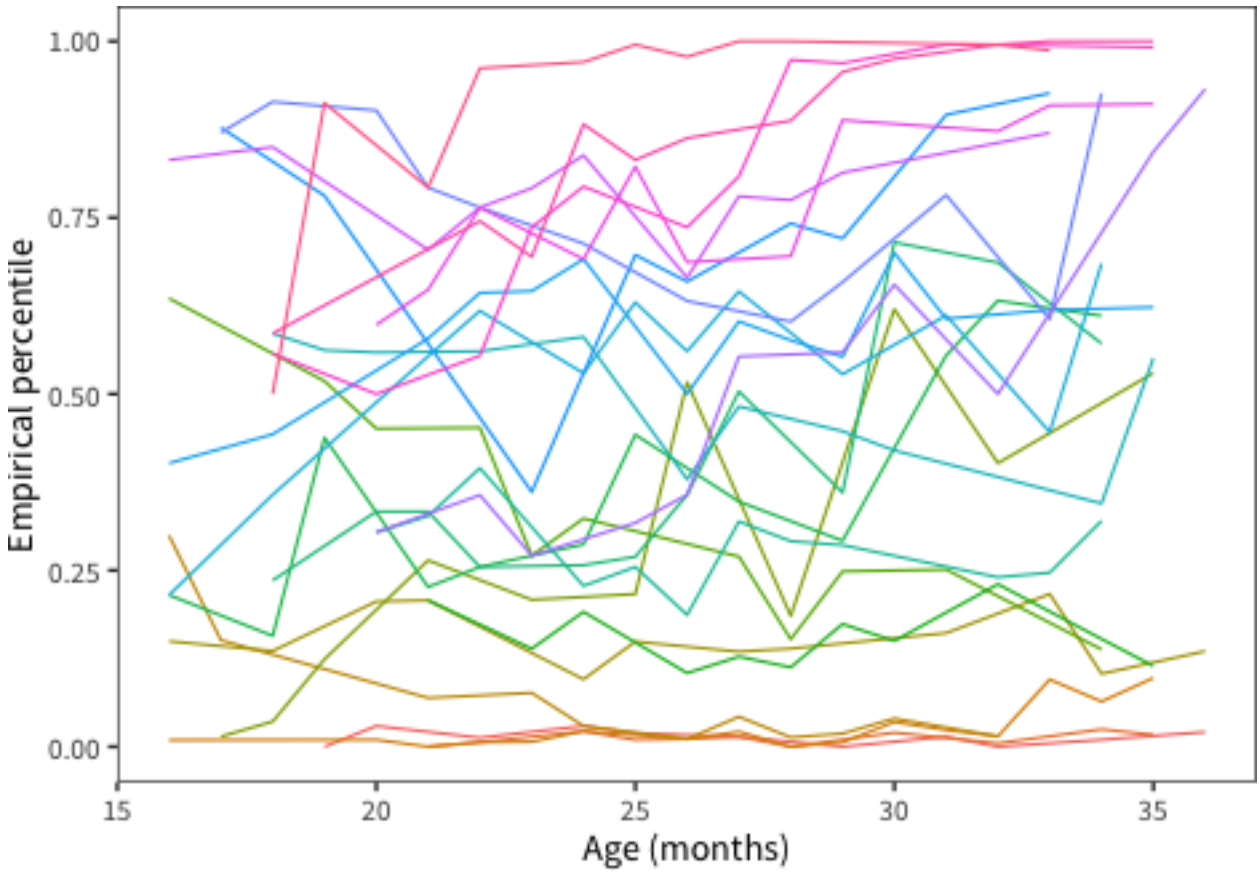


Figure 4.3: Vocabulary percentile as a function of age for children with more than 10 administrations (color indicates child).

over time. We examine this question creating an empirical CDF for each age group.[1] As shown in Figure 4.3, these ranks are visually quite stable.

The transformation to percentile ranks allows us to assess the correlation between a child's percentile rank at time 1 and their rank at time 2, depending on the gap between these two. Because of sparsity, we bin children into two-month age bins and eliminate age bins with fewer than 50 children, then calculate between-bin correlations in percentiles. Figure 4.4 shows this analysis, which reveals that percentile ranks are quite stable; across a 2–4 month age gap they are correlated at better than .8. This stability declines to around .5 at 16 months, but this decline should be taken with a grain of salt. First, this range is a doubling of the child's age, so stability might be expected to be lower. But second, many children who are measured longitudinally across a 16-month gap will be expected to move from the floor of the form to the ceiling, compromising measurement accuracy. To test this last hypothesis, we evaluated the longitudinal stability of correlations using the same analysis as above, but varying whether we used raw scores or percentiles. The percentile method substantially increased correlations.[2]

[1] We could use a model-based method (e.g., the gcrq method used in the Wordbank app and Chapter 5 and 6) but in practice we have enough data in each of these languages that this method should perform well.

[2] We also used latent abilities derived from a 4-parameter IRT model as below. While the IRT-derived ability parameters showed a consistent improvement in longitudinal correlations over the use of raw scores, percentiles realized a further gain over the IRT parameters.

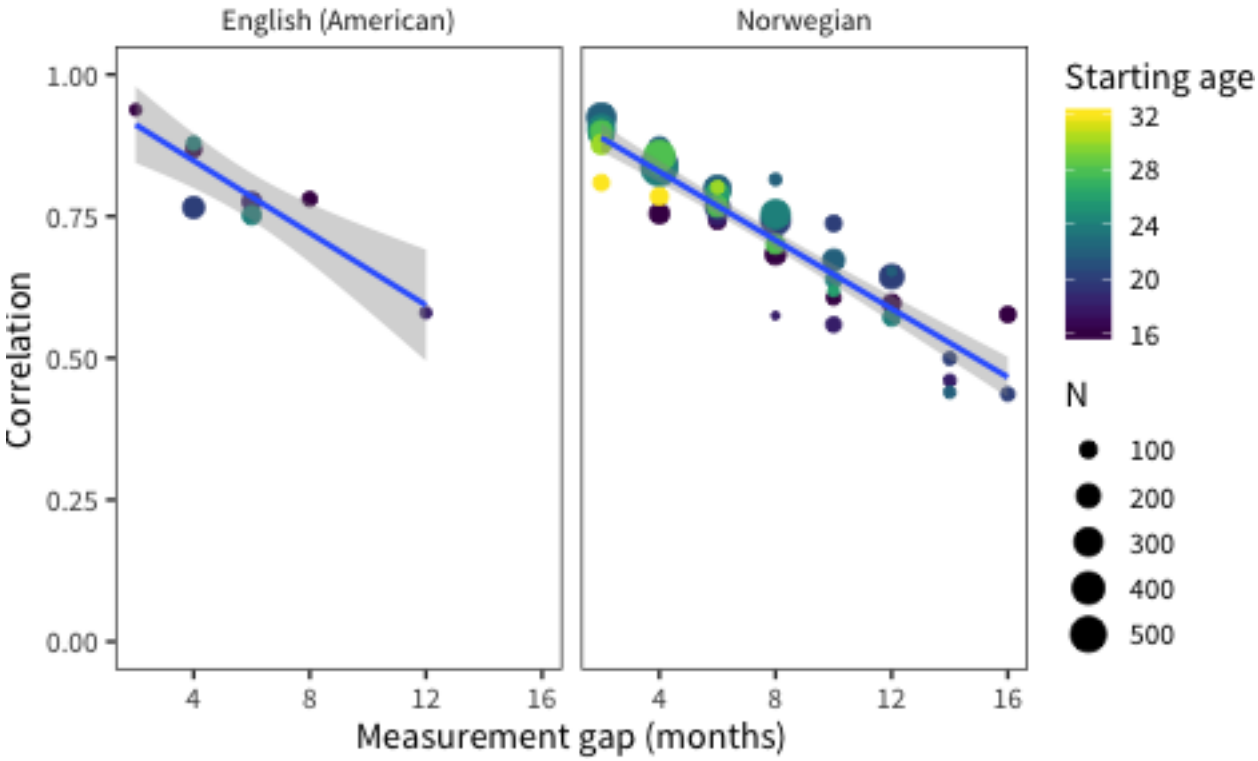


Figure 4.4: Correlations between vocabulary percentile's at multiple age points as a function of the age difference between them.

In sum, the variability between children that we observe in the CDI is quite stable longitudinally. It declines over time, but some of this decline may simply be due to the unavoidable limitations of CDI forms with respect to floor and ceiling effects.

4.3 Psychometric modeling

In this next section, we examine the psychometric properties of the CDI through the lens of Item Response Theory (IRT). In brief, IRT provides a set of models for estimating the measurement properties of tests consisting of multiple items. These models assume that individuals vary on some latent trait, and that each item in a test measures this latent trait (see Baker, 2001, for detailed introduction). IRT models are a useful tool for constructing and evaluating CDI instruments, as they can help to identify items that perform poorly in estimating underlying ability. For example, WEBER et al. (2018) used IRT to identify poorly-performing items in a new CDI instrument for Wolof (a language spoken in Senegal). IRT can also be used in the construction in computer-adaptive testing (Makransky et al., 2016).

IRT models vary in their parameterization. In the simplest (Rasch) IRT model, each item has a difficulty parameter that controls how likely a test-taker with a particular ability will be to get a correct answer. In the more sophisticated two-parameter model, each item also has a discrimination parameter that controls how much response probabilities vary with varying abilities. Good items will tend to have high discrimination parameters across a range of difficulties so as to identify test-takers at a range of abilities.

We examine IRT models as a window into the psychometric properties of the CDI. In the first subsection, we explore latent factor scores using the English WS data. In the second subsection, we examine individual items and find generally positive measurement properties, although with some items at ceiling (included via carry-over from the Words & Gestures form). In the third subsection, we look at differences between comprehension and production in the WG form. In the fourth subsection, we look at the properties of the instrument by word category in both WS and WG.

Overall, the conclusions of our analysis are that:

- Latent factor scores may have some advantages relative to raw scores in capturing individuals' abilities, but for the purposes of the analyses we perform in the main body of the manuscript, they may carry some risks as well; hence we do not adopt them more generally.
- In general, CDI WS items tend to perform well, but from a pure psychometric perspective there are a number of items that could be removed from the English WS form.
- Comprehension items in general tend to have less discrimination than production, suggesting that they are not as clear indicators of children's underlying abilities.
- Function words tend to have lower discrimination than other items but the lexical class differences are not huge and do not interact with whether they are measured using production vs. comprehension.

These analyses generally ratify the conclusion that the measurement properties of the CDI are good, even for function words and for comprehension measures. These questions may carry slightly less signal about the specifics of a child's vocabulary and load more heavily on a parents' general estimation of the child's linguistic ability, but they do carry some signal that relates to other responses. Further, when the English CDI departs from good measurement practice it generally does so for completeness (e.g., including mom and dad words because these are important to parents, even though they do not show good measurement properties).

4.3.1 Measurement properties of individual WS items

A first question that we can ask using a fitted IRT model is how well individual items relate to children's overall latent abilities. Practically speaking, in these analyses, we use the mirt package (Chalmers, 2012; Chalmers et al., 2016) to estimate the parameters of a four-parameter IRT model. As described above, the two-parameter model includes difficulty and discrimination parameters for each item. The four-parameter model supplements the standard two-parameter model with two parameters corresponding to floor and ceiling performance for a particular item. Items with high rates of guessing or universal acceptance across test takers would tend to have abnormal values on these bounds.

Figure 4.5 shows item discrimination and difficulty, with outlying items labeled. Difficulty refers to the latent ability necessary for a child to produce an item, on average. Discrimination refers to how well an item discriminates between children of lower and higher ability (as judged by their performance on other items). Visual inspection shows a long tail of items with limited discrimination and low difficulty (e.g., mommy, ball, bye, etc.). These are clearly those items that are produced by nearly all of the children in the sample — they do not discriminate because they are passed by all children in the sample. If the only goal of the instrument were discrimination of different ability levels, they could likely be removed. But, as discussed above, these items tend to be included for

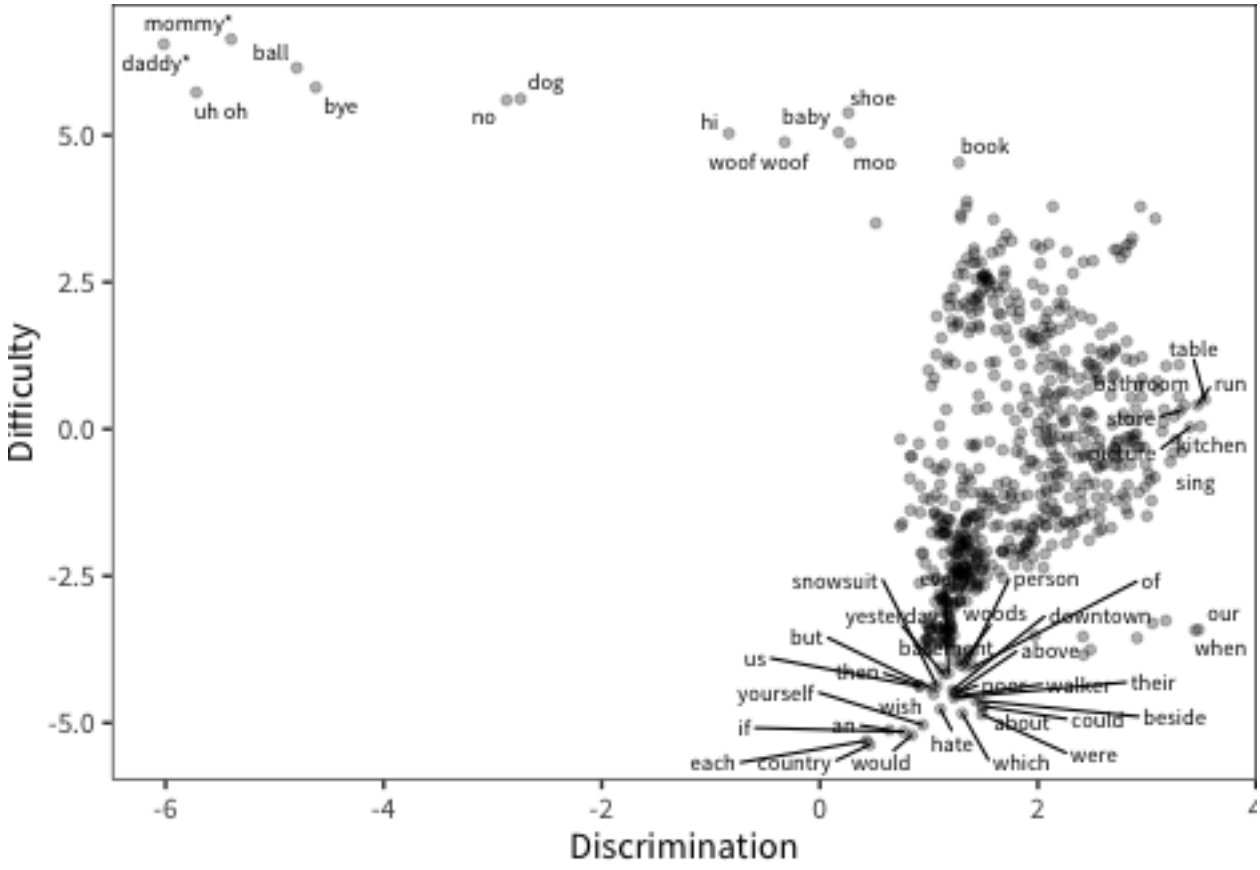


Figure 4.5: Words (points), plotted by their difficulty and discrimination parameters, as recovered by the 4-parameter IRT model (see text). Outliers are labled.

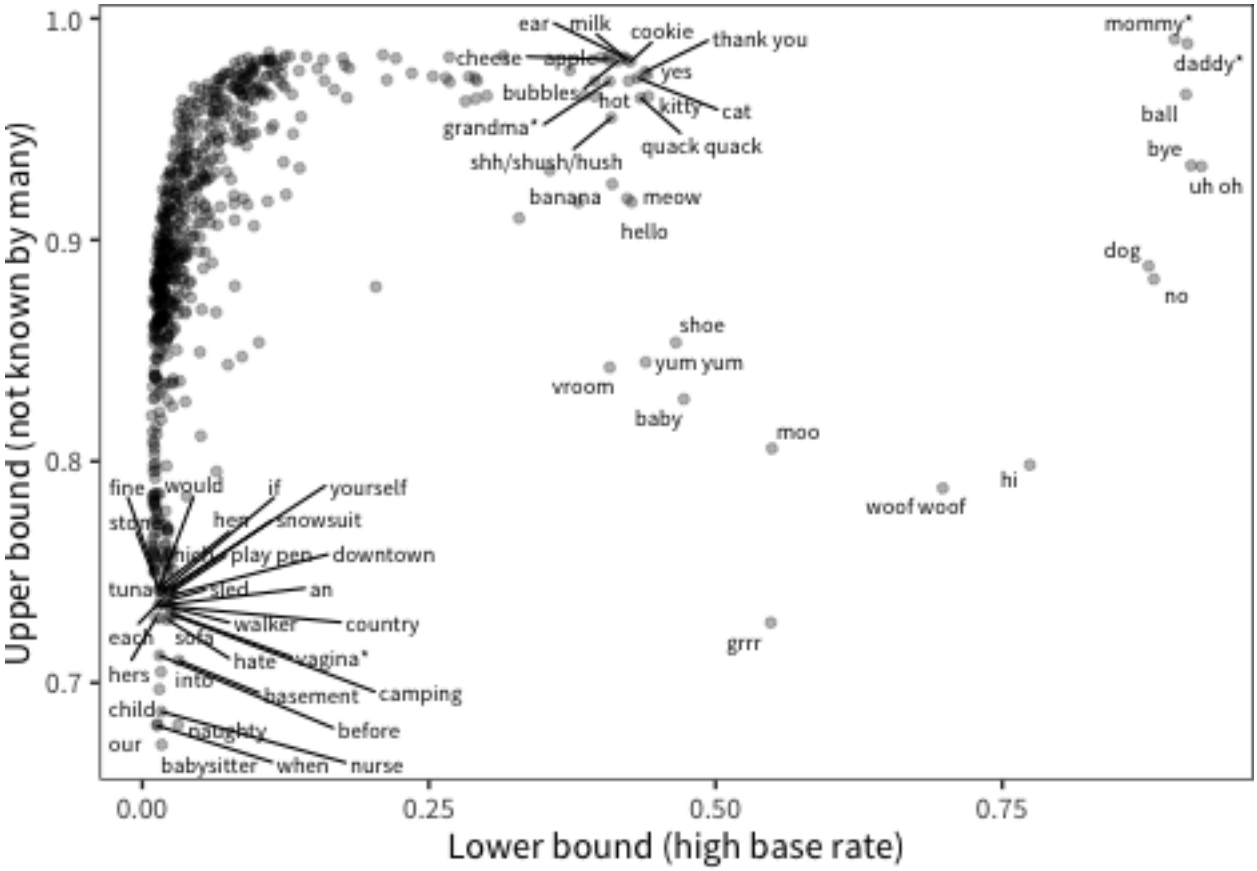


Figure 4.6: Words (points), plotted now by their lower and upper bound parameters from the same 4-parameter IRT model.

completeness (and so the WS instrument is a strict superset of the WG instrument, which is used with younger children).

On the lower right hand side of the plot, the remainder of items are clumped, with discrimination above zero and somewhat higher difficulty. The right-hand tip of this triangle shows the most diagnostic words (e.g., run, kitchen, and table), all of which effectively distinguish between the upper and lower groups of children in the sample. Finally, at the bottom of this triangle is a large cluster of words that are quite difficult. Some of these do not show good discrimination (e.g., country), since it is likely too difficult for nearly all children in the sample.

We can also examine the upper and lower bounds estimated for particular words, as shown in Figure 4.6. These bounds show words that are known by only a small number of children (low ceiling) or are known by almost all children (high floor), respectively. Examining those with a very low ceiling, we see items that are likely to be quite idiosyncratic, for a variety of reasons. For example, babysitter, camping, and basement likely vary by children's home experiences (further mediated by access to resources, parenting practices, and circumstances). In contrast, genital items (e.g. vagina*) vary by gender (see Chapter 9). Examining those items with a very high base rate shows a similar set to those with very low discrimination patterns, suggesting that the four-parameter model may have fit these words as having a high chance level with essentially no discrimination ability.

One way to think about these analyses is that they show that the CDI has not only a large core

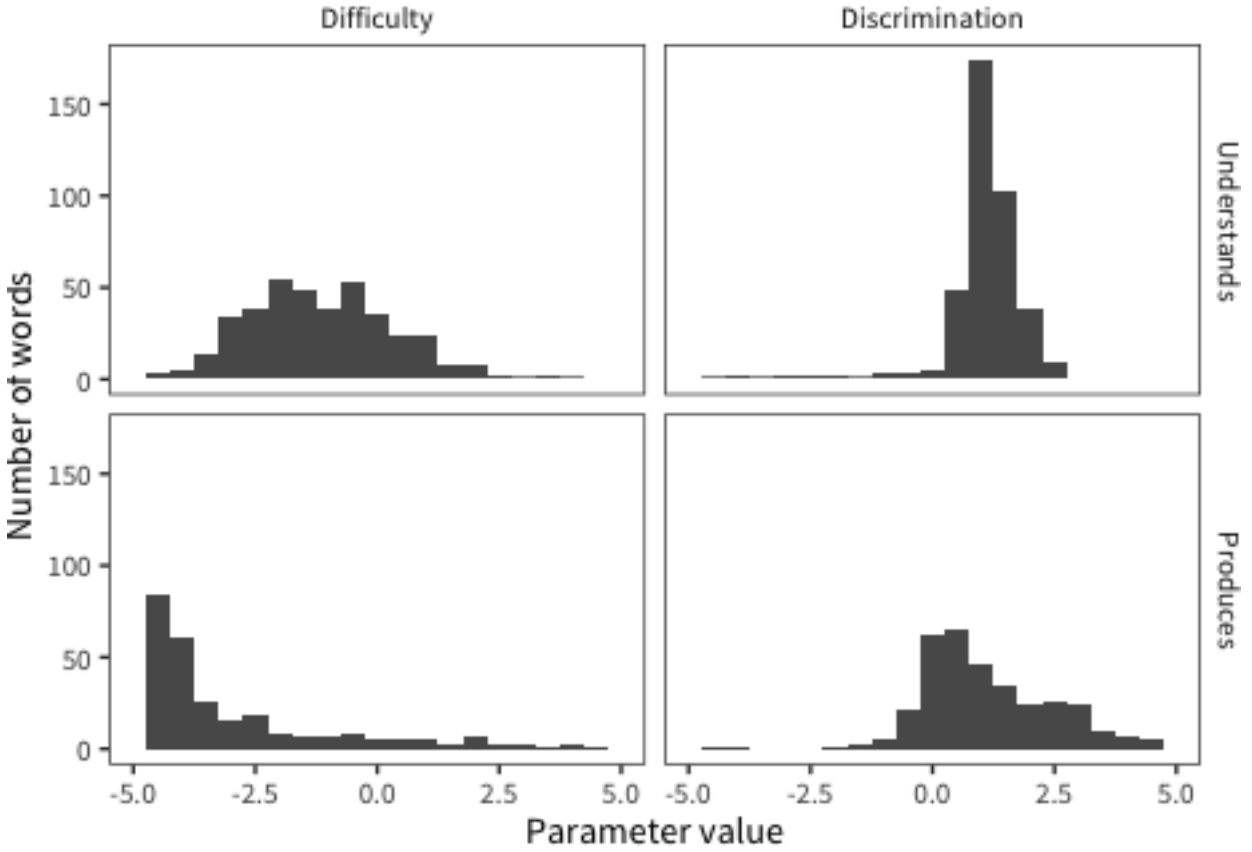


Figure 4.7: Histograms of words' difficulty and discrimination parameters, for comprehension and production.

of words with good measurement properties but also some other words that do not contribute as substantially and add length without adding much signal. In revisions of the CDI, some of these words might be good candidates for deletion.

4.3.2 Production and comprehension

We next use IRT to estimate whether there are differences between production and comprehension, using WG data. Figure 4.7 shows IRT parameter values for discrimination and difficulty for production and comprehension (a few extreme parameter values are truncated in the plot for ease of seeing general trends). There are clear distribution differences on both measures. Difficulty is much higher (negative values) for production relative to comprehension, reflecting the expected asymmetry of production coming "after" (being more difficult than) comprehension.

The second generalization is that comprehension questions largely have positive discrimination parameters. Thus, these questions on the whole carry signal about children's latent linguistic ability. There do appear to be more discrimination values that have negative discrimination parameters, however, indicating more items that are not measuring ability appropriately (perhaps because they are difficult for all children or because they are too hard to assess).

Finally, mean discrimination is substantially lower for comprehension relative to production (1.1

vs. 1.8). This pattern is consistent with the hypothesis that production behavior is a clearer signal of children's underlying knowledge than assumed comprehension. This pattern of findings — lower discrimination values for comprehension — could be due to at least two possibilities. One is that parents are better reporters of production than comprehension, and hence these items are more discriminative of true behavior. The source of error in this case would be parents' mistaken belief that their child understands a word. The second is that comprehension is a fundamentally more variable construct and that, hence, individual word knowledge consistent with understanding could be due to partial knowledge. Here the source of error is variance in how well children know the meanings of words. We cannot distinguish between these two models, but they have different underlying implications for the CDI.

4.3.3 Lexical category effects on item performance

One hypothesis that we have often speculated about is the question of whether there are special psychometric issues with particular word classes. For example, do parents struggle especially to identify whether children produce or understand function words?

Figure 4.8 shows WS item difficulty and discrimination (as above) and the histogram of discrimination, but broken down by lexical class (color). Many of the easy, non-discriminating items are found in the "other" section. In contrast, the hardest items tend to be function words. These items tend to have lower discrimination on average (1.3) compared with nouns (1.8), adjectives (1.9), and especially verbs (2.1). Nevertheless, the situation is not dire: most have a discrimination parameter above one. Thus, although function words are not the most discriminative items on the CDI WS, these items still appear to encode valid signal about children's abilities.

In our last analysis, we turn to the WG data. Figure 4.9 shows the mean (error bars show SD) for discrimination parameter values. In production, the higher discrimination shown on the whole (above) is likely due to the strong performance of nouns, which index distinctive and memorable productions. In contrast, mean discrimination for other words is low. This pattern may be due to the overall sparsity of early production data for non-noun items. In comprehension, in contrast, there is a moderate level of discrimination for all classes except "other" (which includes items like mommy and daddy and a variety of animal sounds and social routines). One hypothesis about this finding is that, especially early on, parents are very generous in their interpretation of whether their child understands these specific words.

In sum, we do not find evidence that function words are particularly low-performing items from a psychometric perspective — even in comprehension assessments! Rather, there are some low-performing items spread across all categories of the CDI form, and many of these likely perform poorly for the reasons described above — especially difficulty in interpretation of very early behavior and variability in home experience.

4.3.4 IRT models: conclusions

One question regarding IRT-model derived parameters for individual children is whether they should be used in place of percentiles or raw scores for some of the measurement problems we encounter throughout the rest of the book. Although these latent ability scores might be overall better reflections of children's vocabulary than other measures, we do not find strong evidence to support that conclusion. For example, in the analysis above, we compared longitudinal correlations derived

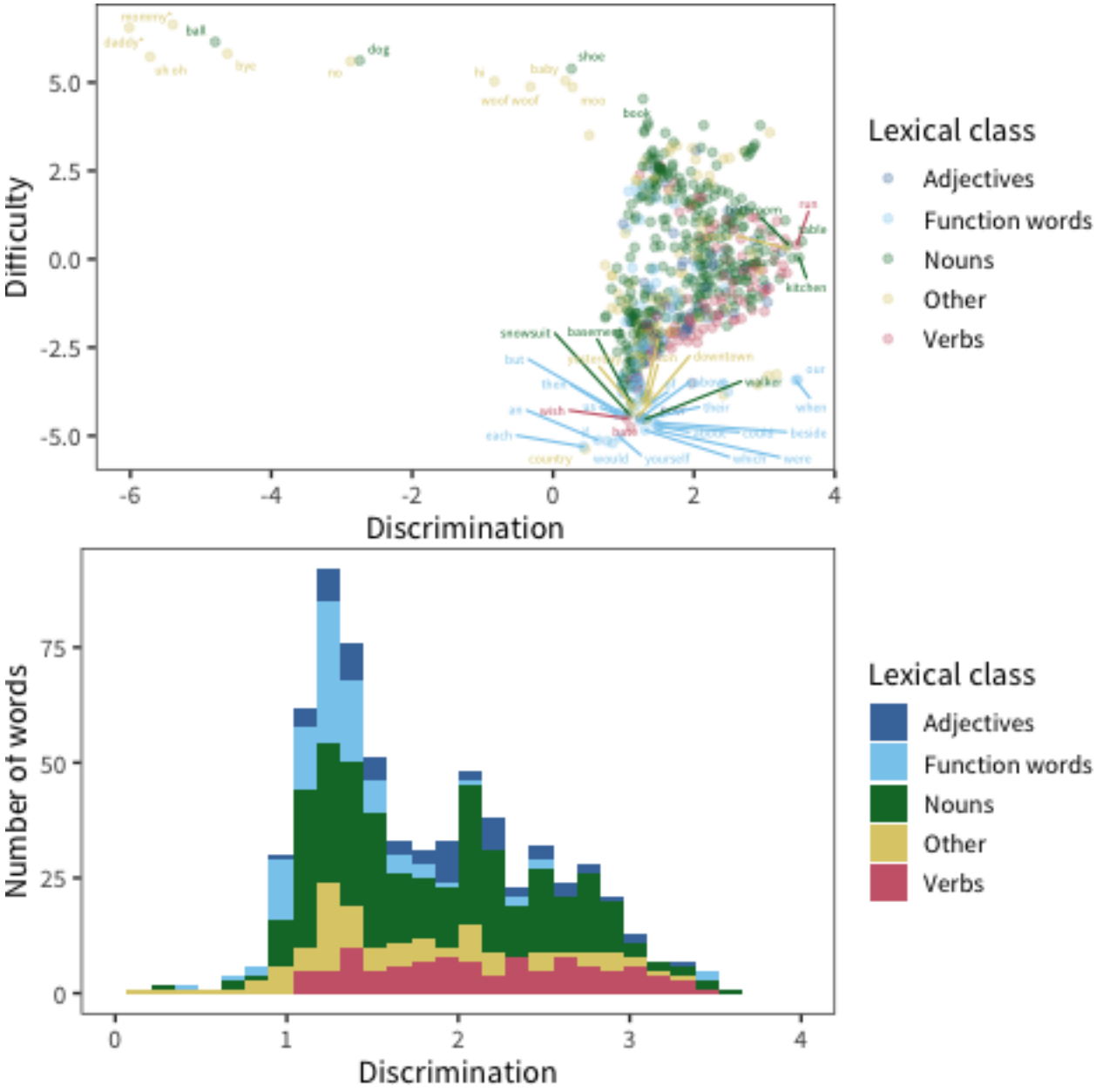


Figure 4.8: Lexical class effects on difficulty and discrimination for Words and Sentences. The top plot shows individual words plotted by their parameter values, with color representing the lexical class of the words. The bottom plot shows discrimination information in the form of a histogram.

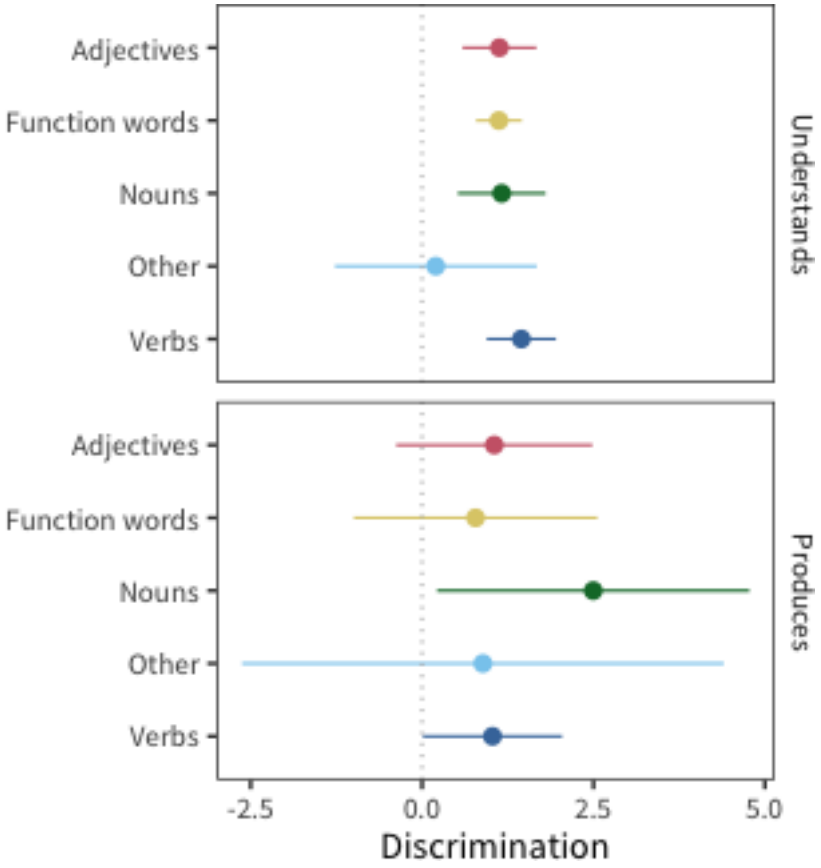


Figure 4.9: Mean discrimination values for individual words' in production and comprehension measures from the Words and Gestures form (error bars show SD).

from raw scores, percentiles, and IRT ability parameters. While IRT parameters yielded higher correlations than raw scores, empirical percentiles performed better still (at least for Norwegian and English, two languages for which we have large amounts of data).

Furthermore, there are other negatives associated with swapping an imperfect but straightforward measure (raw and percentile scores) for a model-derived measure (latent ability). Interpretation clearly suffers if we use the model-derived measure, since readers will not be able to map scores back to actual behavior in terms of the checklist. In addition, model estimation issues across instruments introduce further difficulties in interpretation. Most obviously model estimates with smaller datasets may vary in unpredictable ways; similarly, the presence of poorly-performing items in certain datasets may lead to systematic issues in the latent estimates for those datasets.

4.4 Conclusions

In this chapter, we examined the measurement properties of the CDI from three perspectives. From a theoretical perspective, we reviewed why the design features of the CDI make it a reasonable tool for measuring child language, even if there are opportunities for error and bias throughout. (Of course, one of these design features are dependent on the style of administration for a particular study, so of course a poorly-administered form will yield a dataset with lower reliability and greater bias). Then, we took advantage of the deep longitudinal data available for two languages and showed quite strong longitudinal correlations between CDI administrations. This pattern indicates that early language is a stable construct across development (Bornstein and Putnick, 2012). It also signals that measurement error between CDI administrations appears to be limited, at least when the span of time between administrations is not too great. Finally, we used item-response theory to examine the measurement properties of individual items. While the CDI includes some items with limited measurement value (if all that the user cares about is a single ability score), most items show good psychometric properties. This analysis also revealed that comprehension questions and questions about function words do not appear to be particularly worse than other items, contrary to previous speculations. In sum, the CDI appears to be a reliable instrument for measuring children's early language, with measurement properties that support a range of further analyses.

Chapter 5

Vocabulary Size

This chapter begins our substantive analysis of properties of language learning and their variation across languages and children. We begin with one canonical view of CDI data, in which each child is represented by a single vocabulary score: the proportion of words that child knows, out of the total in the form. We first quantify the median pattern of vocabulary growth observed in our data; we then turn to characterizing variability across individuals in these data. In these analyses as well as subsequent chapters, our inspiration comes from what we think of as the "Batesian" approach to variation.

> Far from simply reflecting noise in our measuring instruments or variability in low-level aspects of physiological maturation, the variations that we will document (in vocabulary development) are substantial, stable, and have their own developmental course. Because this variation is substantial, it is critical for defining the boundary between normal and abnormal development; because it is stable, it provides a window onto the correlates and (by inference) the causes of developmental change; and because it has its own developmental course, it can be used to pinpoint critical developmental transitions that form the basis for theories of learning and change. (Bates et al., 1995)

We are interested in these theoretical uses of variability. But variability is only meaningful in the case that it is stable; that is, that it reflects signal about individuals (or cultures) rather than measurement error. With respect to the CDI, the strong evidence for the reliability and validity of the forms — reviewed in Chapter 4 and in Fenson et al. (2007) — provides support for the contention that observed variability is meaningful. Such evidence has primarily been collected for the English CDI, however. In this chapter we examine variability across the full set of languages, and it is worth noting up front that we project the reliability and validity of the English instrument to its adaptations.

One study nicely illustrates why such an approach might not be misguided. Bornstein and Putnick (2012) collected multiple language measures at 20 and 48 months in a sample of nearly 200 children and used a structural equation model to estimate the stability of a single latent construct, language ability. Essentially all measures related strongly to this latent variable and the coefficient on its stability over time was r = .84, suggesting that early language is quite stable, at least when measured appropriately. Notably, the ELI, a precursor to the CDI, was included in the measures at 20 months and was found to correlate with the 20-month latent construct at r = .87.

This finding — along with the other evidence, mentioned above — justifies the implicit conceptual model of the following analyses. That model is that there is a single quantity, early language ability,

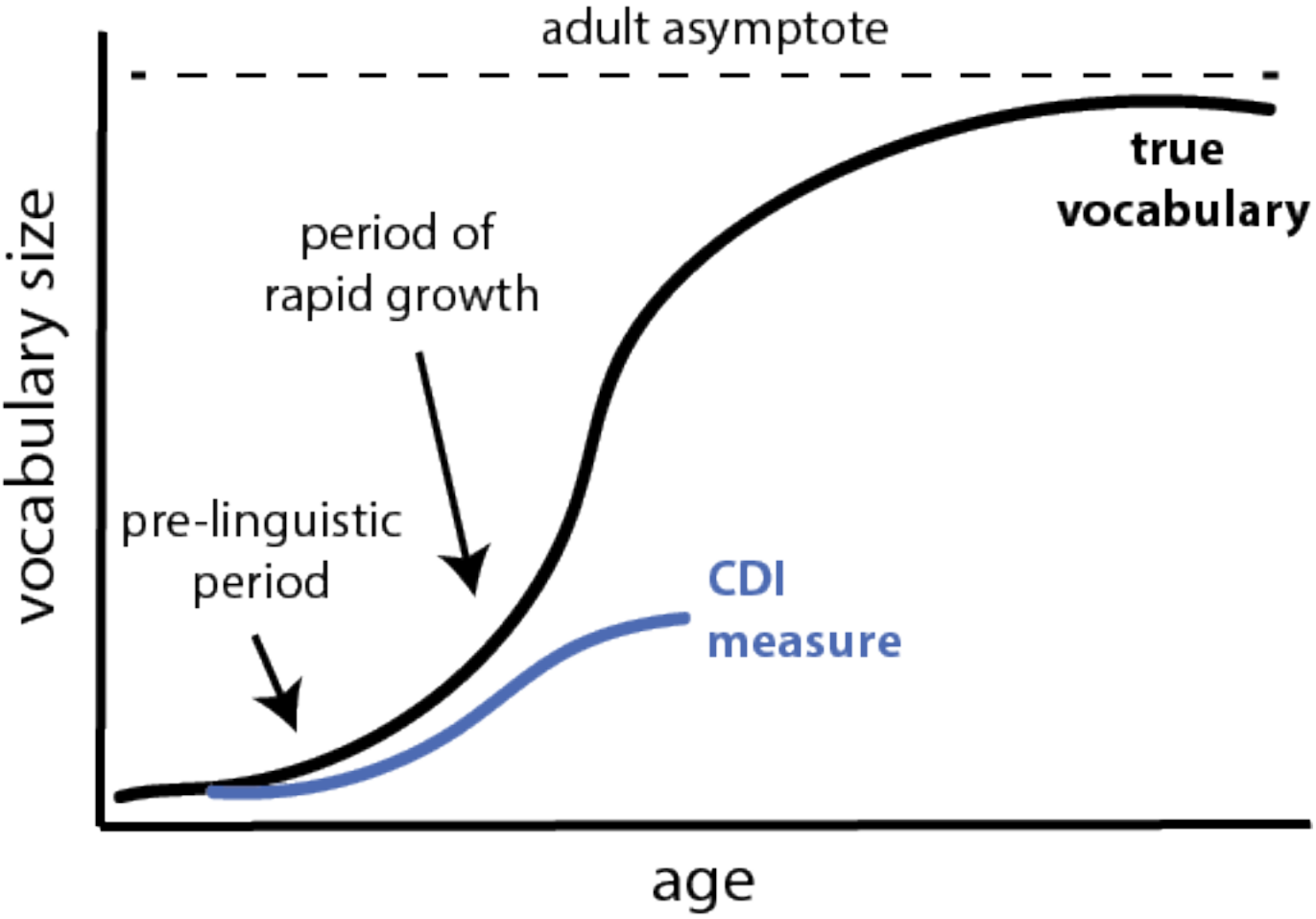


Figure 5.1: Schematic true vocabulary growth and vocbaulary growth as measured by the CDI.

that is stably measured by parent report and that can be approximated as the raw proportion of words a child "understands" or "understands and says" on CDI forms.

5.1 Central tendencies

The first question we can ask about CDI data is about its central tendency — the median pattern of vocabulary growth. Our general expectation is shown in Figure 5.1.

This schematic reveals a number of patterns that are explored in this and subsequent chapters. The CDI necessarily captures a small fraction of any individual's true vocabulary, but even within the measured range there are a number of specific questions that can be addressed by different analyses. The question of the exact slope of children's growth, especially in the period immediately after the emergence of language, is treated in Chapter 14 — this question is sometimes posed as whether children undergo a vocabulary "spurt" (Ganger and Brent, 2004). On the other side of the CDI curve, the question of the divergence between CDI-measured vocabulary and true vocabulary (and whether true vocabulary can be recovered via a statistical correction) is treated in work by Mayor

and Plunkett (2011).[1] In the current chapter, we focus on the middle section of the CDI curve, in which children's vocabulary is neither at the floor or the ceiling of the instrument.

5.1.1 Commonalities across languages

Figure 5.2 shows the median patterns of growth for early production. Rather than showing proportions, as we will do more standardly throughout the book, here individual item totals are plotted.[2] In general, the median child before the first birthday is reported to produce a small number of words. (These data raise a number of questions about the specific reliability of very early parent reports, which we take up below.) Overall, however, these curves accord relatively well with our intuitive sense of early vocabulary development: they reveal that most children tend to speak at most only a few words before their first birthday, but that production accelerates across the second year.

This analysis also motivates the decision made in most our large-scale analyses to omit early production from Words & Gestures forms. Many WG forms end at 16–18 months, meaning that the median production is only around 50 words. For analyses of vocabulary composition or predictive modeling, these numbers are often too small to yield reliable and meaningful estimates, although they can be combined with Words & Sentences data (see Appendix C).

The acceleration in early vocabulary is even clearer when looking at production reports from older children using Words & Sentences. Figure 5.3 shows this pattern. In every language, the median child is reported to produce 50 words between 16–20 months (dotted line). As we will see below, this analysis masks the tremendous variability apparent during this period. In addition, as we discuss below, languages vary considerably in the absolute number of words reported. (As it is a major outlier, we discuss the Beijing Mandarin WS data in Chapter 9.1). Nevertheless, there are still substantial consistencies in the shape and general numerical range across languages.

During the period of 24–30 months, we see children beginning to produce a large enough sample of words that curves are leveling out. Presumably this leveling does not reflect a slowing in the rate of acquisition, which most researchers assume continues unabated for many years (Bloom et al., 2001). Instead, it reflects the limitations of the CDI instrument, in that there are many possible "more advanced" words that they could be learning, of which only a small subset are represented on any form.

We next turn to comprehension medians, shown in Figure 5.4. Comprehension is only queried on the Words & Gestures form. Reported comprehension increases much faster than production; so much so that most parents are reporting that their children understand most words on the form by 18 months (Chapter 14 discusses differences in the balance between comprehension and production between children). As with the production data, we see substantial differences across languages in reported vocabulary, discussed below (see Chapter 9.1 for more discussion of Taiwanese Mandarin comprehension data).

Outside of that particular dataset, one striking aspect of the comprehension data is how early comprehension is reported. For example, from 8 months, we see parents reporting medians of 4.5 (Swedish) and 123.5 (Mandarin (Taiwanese)) words. To many researchers (and some parents) these high numbers feel unlikely. We are largely agnostic on this issue, but the literature does provide

[1] Because of the English-specific nature of this work, we do not take up this issue here.

[2] As Eriksson et al. (2012) write, "Using raw data assumes that each form is exhaustive, while using percentages assumes that each form is equally exhaustive. Neither is correct and the truth lies somewhere in between."

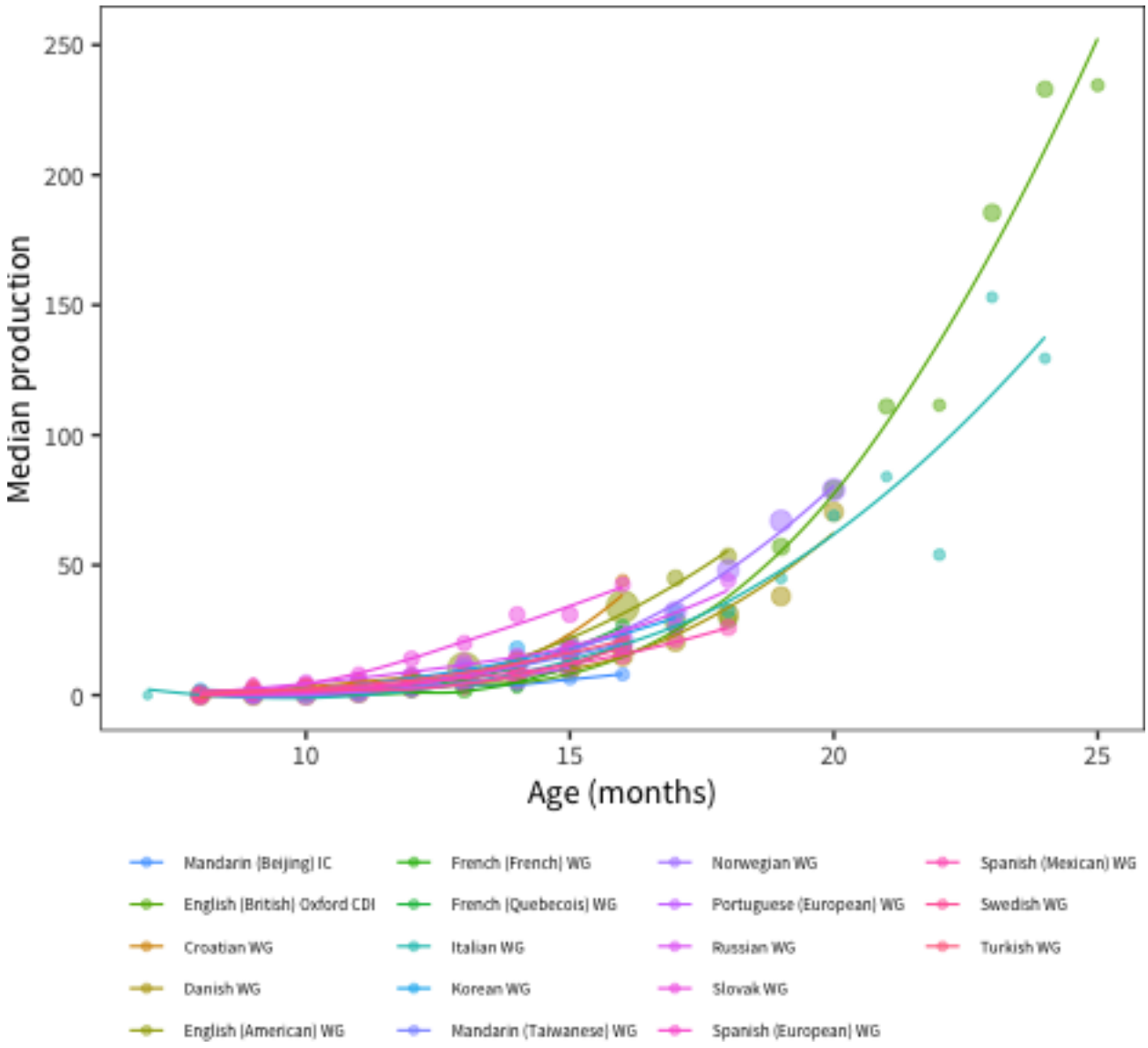


Figure 5.2: Median production using Words and Gestures-type forms. Included are only languages where there are more than 200 administrations total.

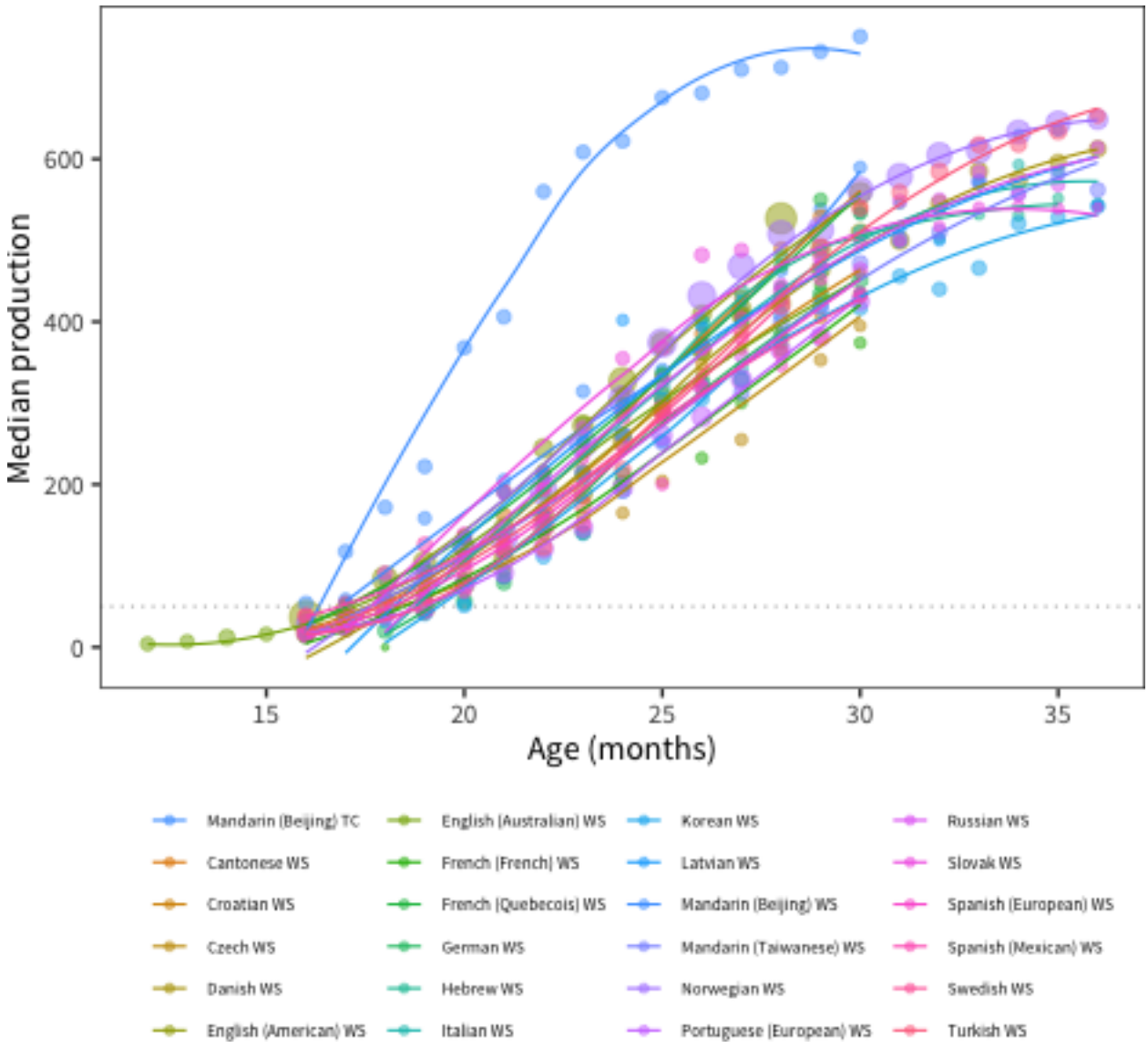


Figure 5.3: Median production using Words and Sentences-type forms. Included are only languages where there are more than 200 administrations total.

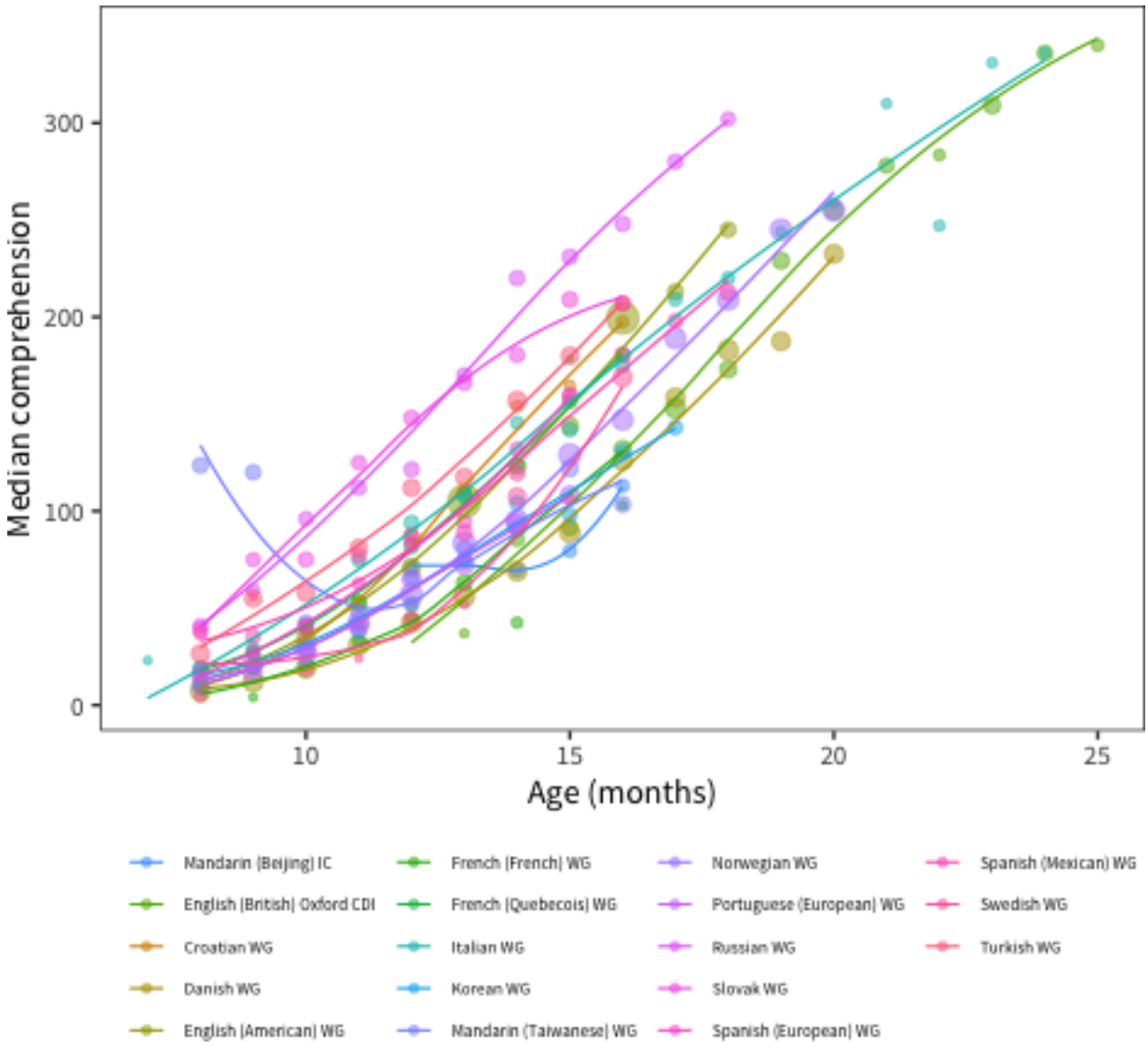


Figure 5.4: Median comprehension using Words and Gestures-type forms. Included are only languages where there are more than 200 administrations total.

some support for early comprehension reports. A spate of recent infancy experiments suggest that in fact, children in the second half of the first year do have some fragmentary representations of many common words available (e.g., Tincoff and Jusczyk, 1999, 2012; Bergelson and Swingley, 2012; Bergelson and Aslin, 2017). The representations revealed in these tasks are quite weak — often amounting to a 2-5% difference in looking to a target on hearing a word uttered — but, depending on the criterion used by parents, may be what is detected in these early reports. Thus, these estimates may not be as far off as we initially suppose.[3]

5.1.2 Cross-language differences

While we have reason to believe that there are some outlier languages in our data, there is still other observed variation that does not seem unusual. What are the sources of this variation? In these analyses, we examine production on the Words & Sentences form, as the data are the densest and most reliable for this instrument. Clearly there are substantial differences in raw scores, even leaving aside Mandarin and Hebrew (which we discussed in Chapter 9.1, in the section on "difficult data"). We consider a range of explanations for this pattern.

Differences could be due to differences in form length. As shown in the plot above, however, medians for production and raw scores are highly correlated ($r(22) = 0.906$, $p = < 0.001$), suggesting that this ordering is not only a function of form length. Further, although raw scores are correlated with form length ($r(22) = 0.336$, $p = 0.109$), this correlation changes direction and is no longer reliable for proportions ($r(22) = -0.0789$, $p = 0.714$). In sum, it appears that there are form-length differences (motivating the use of proportions in general), but that there is still stratification between languages even correcting for this issue. Figure 5.5 shows the relevant proportion trajectories, highlighting remaining differences between English and Danish (two languages for which we have substantial datasets with full demographic information).

Differences could also be due to form construction. For example, the Czech form could contain harder words, leading to fewer words being checked. We cannot directly address questions about the difficulty distribution items without moving to psychometric models (see Chapter 4). These models in turn would need to be equated across forms in order to compare latent ability scores across instruments. While we have experimented with these procedures, there is a circularity to these procedures that makes us leery of proceeding. In particular, in order to equate across tests in standard item response theory models, it is critical to have test items that are share across instruments. But although we have concepts that are shared across instruments (see Chapter 10), we do not believe the words that represent these concepts are equally difficult across languages — in fact, the premise of our later analyses is that they are not. Thus, assessing form difficulty across languages is a complex proposition that we do not address directly here.

Differences could also be due to demographic differences across samples. We can examine sample composition in Chapter 9.1 and see that — to the extent we have access to demographic data — sample composition does vary in features that affect vocabulary (e.g., maternal education, birth order; see Chapter 6 for fuller analysis). We are not yet in a position to conduct a full analysis of these differences, controlling for demographics, as data are sparse and demographic differences also vary across cultures. But we can examine the difference between Danish, and English (American)

[3] An alternative possibility is that both accounts are true, but unconnected: 8-month-olds could in fact know some common words, but parents could be overestimating their vocabulary based on observed behaviors — in essence, parents could be right, but for the wrong reasons.

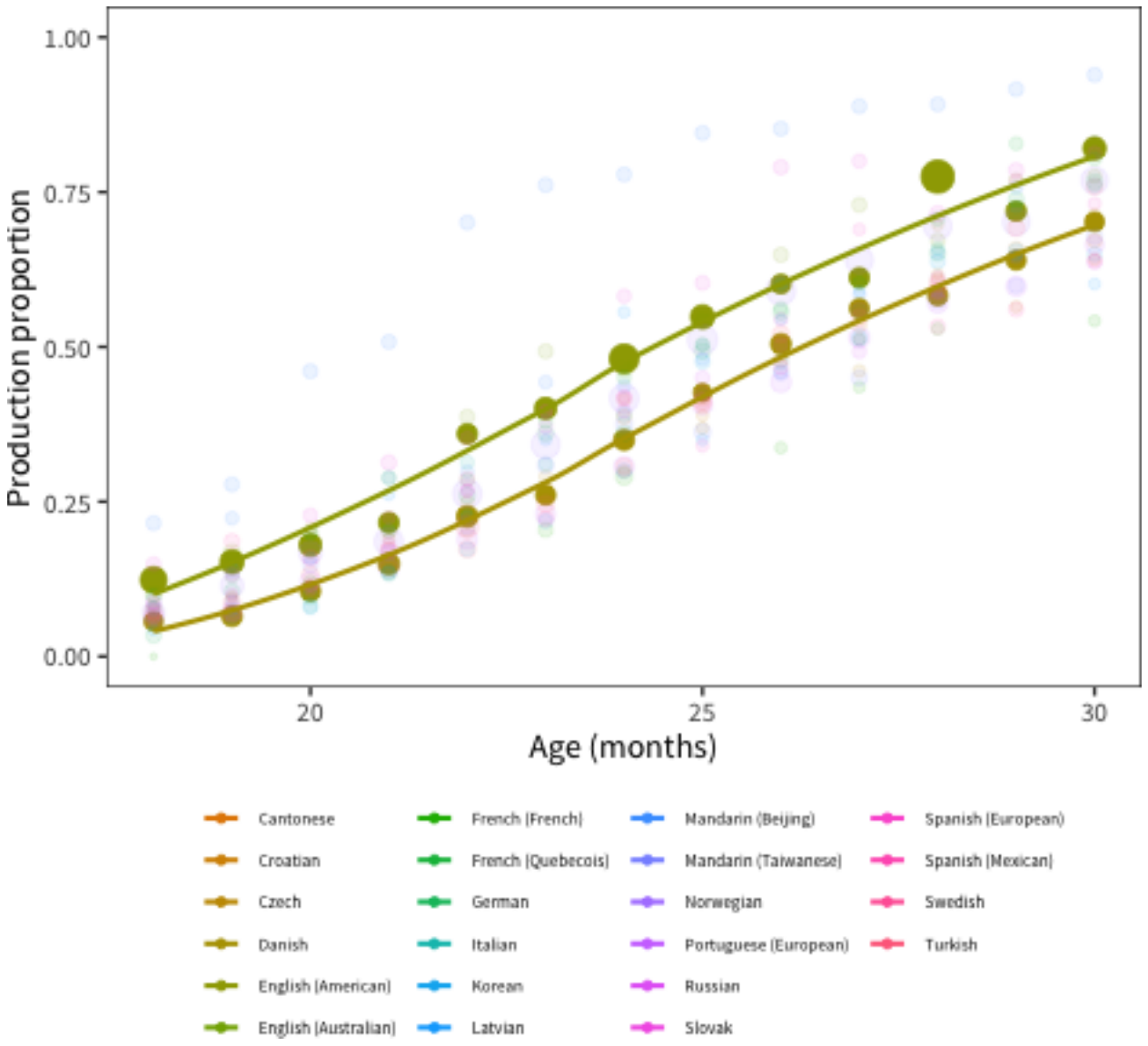


Figure 5.5: Cross-linguistic production data, proportions plotted by age. English (American) and Danish are highlighted.

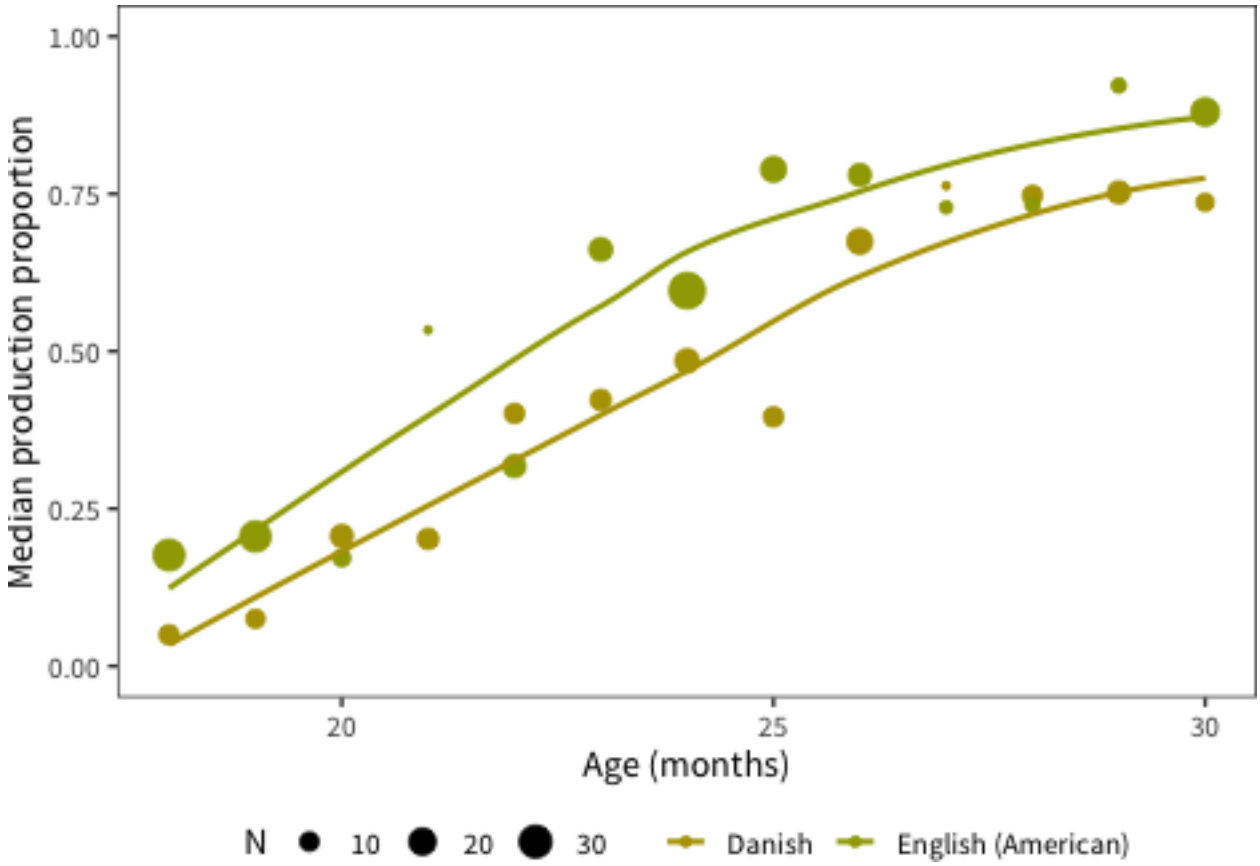


Figure 5.6: Proportion production plotted by age for Danish and English samples, now subsetting to the first-born female children of college-aged mothers.

for example, and note that these differences look quite similar (though noisier) in the female, first-born children of college-educated mothers (Figure 5.6). Thus, we do not believe that demographic differences fully explain the cross-linguistic differences observed.

Differences could be due to cohort effects, in which older sets of data show differences from newer datasets. Most of our data date from the period 2005-2015, but some of our English and Spanish data are older, as they date to the period in which CDI forms were first being designed (the early 1990s; Fenson et al., 1994). Unfortunately we do not have reliable information about the collection data for many datasets, so we cannot do regression analyses straightforwardly. Naively, we would expect later cohorts to show higher vocabularies, consistent with the Flynn effect (Flynn, 1987). Yet, the Danish data, for example are relatively recent and were collected online using standardized instructions. Danish is subject to its own issues, however, but the Norwegian data are also relatively recent and are quite comparable to the English data.

Differences may relate not to demographics of the sample but to the procedure at administration. These differences are not transparent to us in all cases, and so, similar to cohort effects, we cannot control for statistically. For example, instructions at administration — whether written on the form or given by experimenters — might have been more liberal in the case of Slovak or English (American) samples. Such instructions could have emphasized completeness in reporting vocabulary. Or the circumstances of administration could have been different — for example, Danish data were collected online while most English data were collected using paper and pencil forms. English (American) data are contributed by many different labs, so there are likely many different administration styles represented.

Differences could be due to cultural or experimental differences in reporting bias. Slovak parents might have a lower criterion for reporting knowledge of a word. Recalling our discussion of these issues in Chapter 2, their model of children's overall competence might be shifted up. (Such an explanation could be true in principle for the case of the Mandarin and Hebrew data discussed above, as well, though this would be a case of extreme differences!) This explanation is as an extension of the discussion above of administration and instructions — perhaps cultural expectations for what it means to be producing a word or cultural expectations for how verbal children are expected to be.

Finally, differences could be due to true differences in language acquisition. While this explanation is a possibility, we hope we have emphasized that it is only one among the many enumerated above. Many researchers working on Danish believe that, due to its phonological properties, it is truly a difficult language to learn (Bleses et al., 2008, 2011). In particular, Danish is characterized by some highly distinctive phonological reduction processes which greatly reduce the frequency of obstruents, and more generally lead to "an indistinct syllable structure which in turn results in blurred vowel-consonant, syllable and word boundaries [where]... word endings are often indistinctly pronounced" (Bleses et al., 2008, p. 623). The authors of this study also able to provide evidence against the alternative view that Danish parents are simply more reluctant to respond "yes"'– there were no differences on either gestures or word production. They concluded that the phonological structure of Danish produces an initial obstacle to breaking into the stream of speech and is reflected in overall patterns of vocabulary development. While this generalization may in fact be true, it does not answer the question of why children learning other languages are equally slow by both raw and percentile metrics.

In summary, differences between languages in the sheer number of words reported are unlikely to be accounted for purely by differences in form size or demographic differences between samples. In our very speculative view, they likely result from a combination of cultural attitudes towards children's

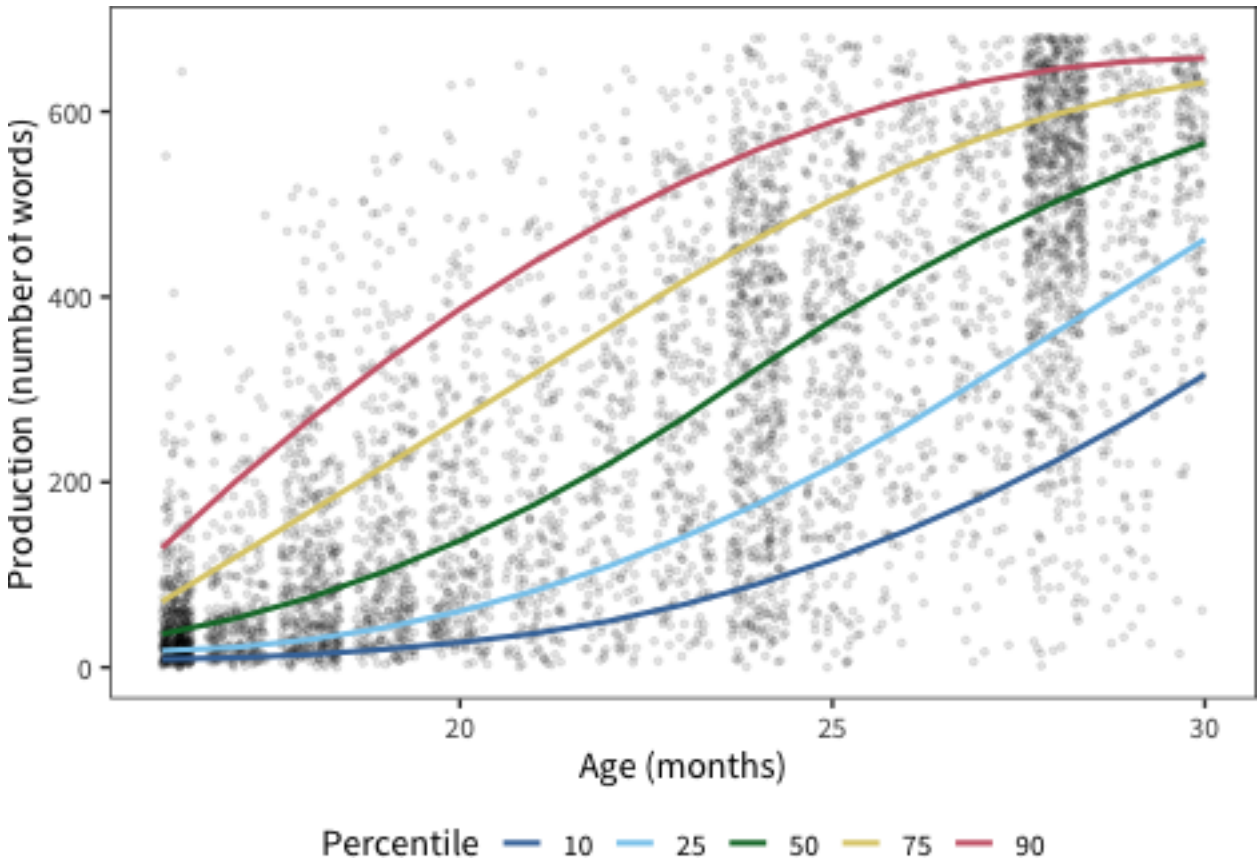


Figure 5.7: Raw production scores for English (American) production data. Dots show individual participants, while lines give standardized percentiles, computed via spline-based quantile-regression.

language, differences in administration instructions, and real differences in learning across languages. Partialling out these differences would likely require better-controlled data that included constant administration and sampling methods. For this reason, in the remainder of our analyses, we attempt to avoid interpreting overall differences in vocabulary size wherever possible and limit ourselves to quantities that can be effectively normalized.

5.2 Variability between individuals

We next turn from the question of central tendencies in early vocabulary to the question of variability. One of the most important features of early vocabulary development is its variability (Fenson et al., 1994). To examine variability, we switch to a view of the data that reveals the full range of variation across the samples in the database. Examining Words & Sentences data, we can see that across every language in the database there is a tremendous range of vocabulary sizes reported. How systematic is this variability? That is the question addressed by our next set of analyses.

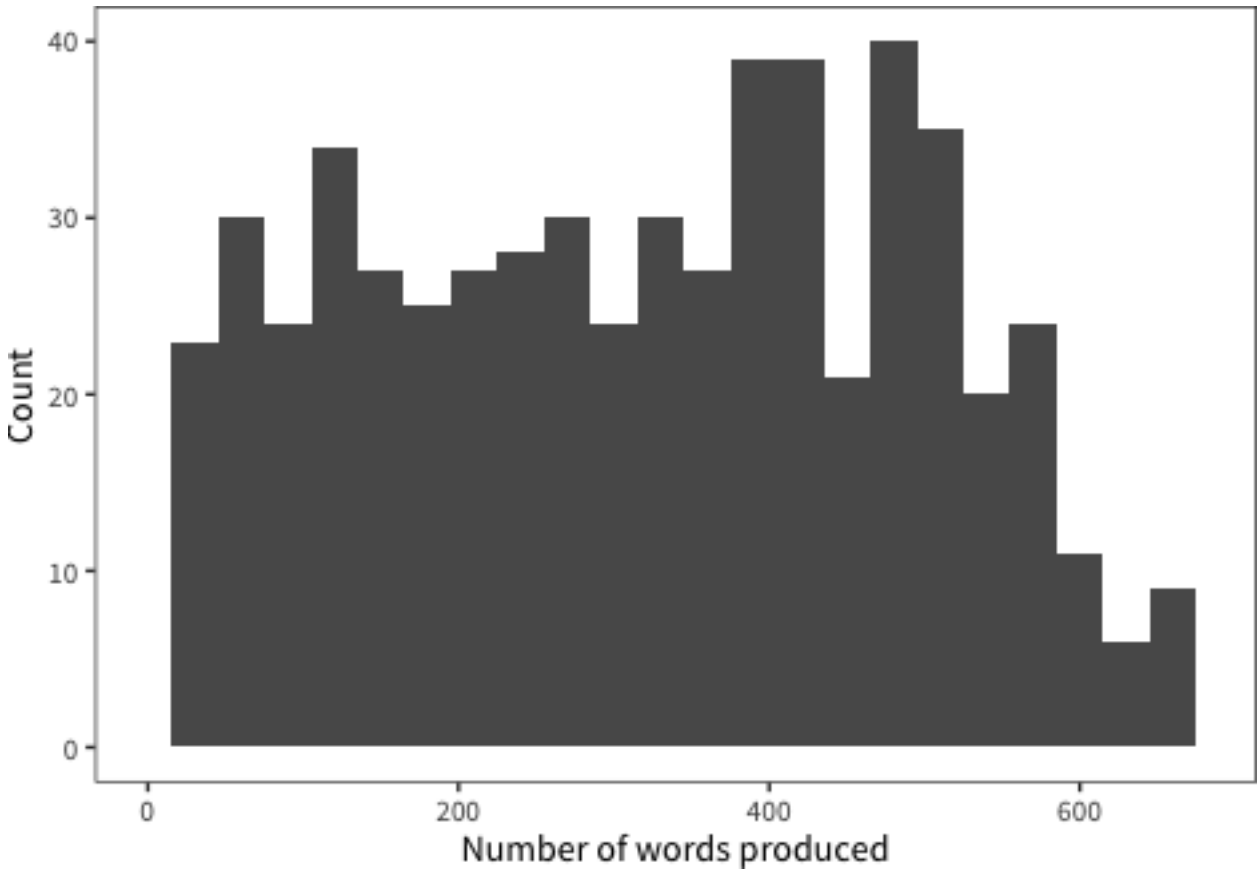


Figure 5.8: Histogram of English (American) production values for 24-month-olds.

5.2.1 Quantifying variability

As an example, we zoom in on the English (American) production data from the Words & Sentences form. The canonical view of these data is given in Figure 5.7. It is very clear that variability is the norm! Children are all over the map at almost all age groups.

Next consider just a single age group, 24-month-olds. A histogram of two-year-old production vocabulary is show in Figure 5.8. The distribution of vocabularies across children is far from normal, with many children at the very bottom of the scale and almost as many at the top. Quite a few two-year-olds on their second birthday are producing only a handful of words (or at least their parents say they are) and others are producing nearly all of the 680 listed on the form (as well as others, in all likelihood). (It will quickly get tiresome to acknowledge ceiling effects and parent report biases in every analyses, so we acknowledge them up front and then mention them only when relevant throughout.)

One way to describe these data is to consider the relationship of the variance to the central tendency. The "coefficient of variation" (CV) is a common measure used for this purpose:

$$CV = \frac{\sigma}{\mu}$$

This statistic allows standardized comparison of variability across measurements with different scales,

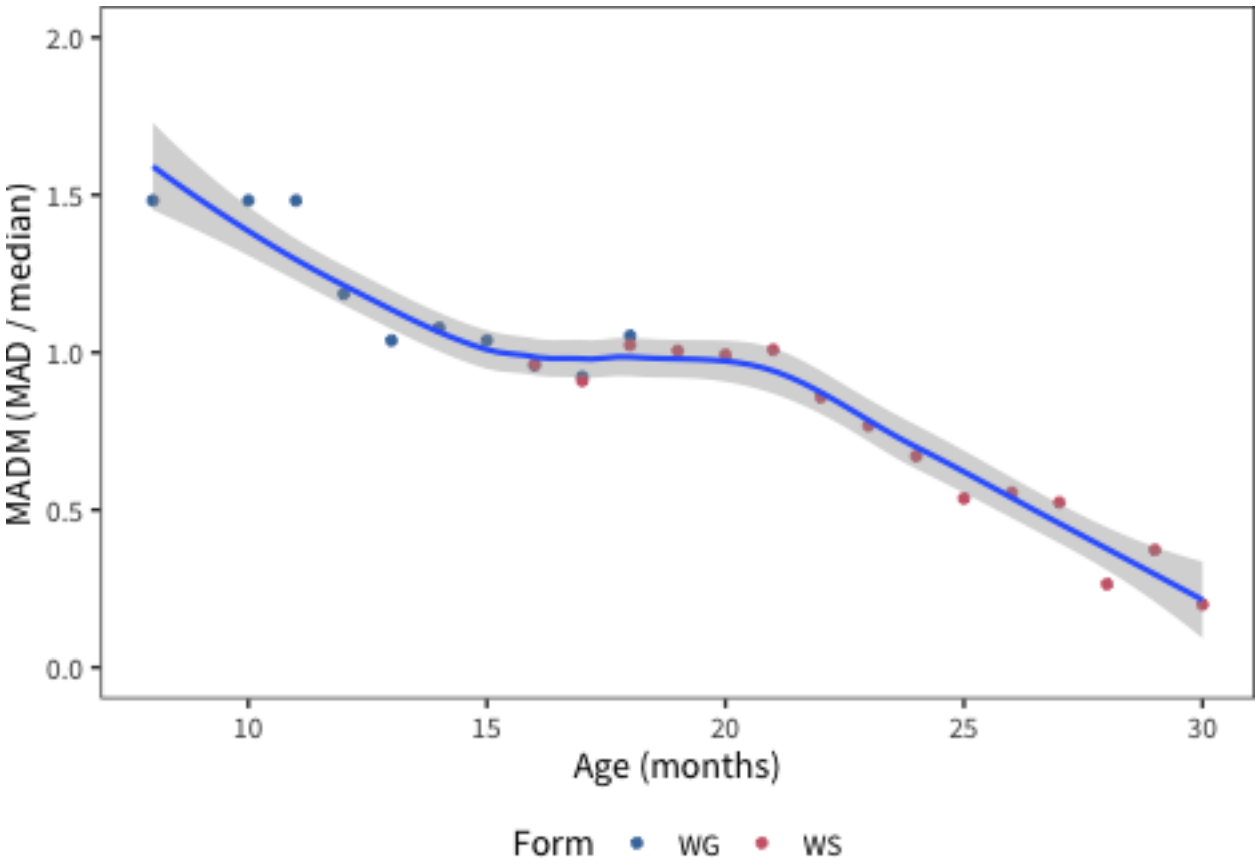


Figure 5.9: MADM values plotted by age for English (American) production data, across forms. The smoothing line is produced by a loess smoothing function.

an important concern when we want to compare forms with very different numbers of vocabulary items. For example, for two-year-olds, the mean productive vocabulary is 319 words, and the standard deviation is 175, words, leading to a CV of 0.55.

But, again as seen in Figure 5.8, the distribution of productive vocabulary scores is far from normal. And distributional forms deviate even more from the standard normal distribution at younger and older ages. Thus, a non-parametric approach is more appropriate. Accordingly, we compute the MADM statistic, the non-parametric equivalent of the CV. In MADM, the mean μ is replaced by the median ($m(x)$), and the standard deviation σ is replaced by the mean absolute deviation (which captures how far away values are from the median):

$$MADM(x) = \frac{\frac{1}{n}\sum_{i=1..n}|x_i - m(x)|}{m(x)}$$

Appendix B demonstrates that, although MADM is more appropriate for our data, CV and MADM are very highly correlated with one another.

Figure 5.9 shows the MADM value for American English production data, plotted by age. In these data, the MADM is actually close to 1 from age one until almost age two, suggesting that the standard difference from the median is actually as big as the median itself!

To get a sense of this variability, it can help to have a smaller dataset to consider. Imagine groups of

three children. A group where one produced 30 words, one produced 100, and another produced 170 would have a MADM of 0.99. In contrast, one where they were more closely grouped — say 70, 100, 130 — would have a MADM of 0.44.

Does the general level of variability observed in American English hold for other languages? Figure 5.10 shows MADM, now plotted for production across languages and instruments. This similarity in variability structure is quite striking, such that between the first and second birthdays, children's early language is remarkably variable. Yet this variability itself is quite consistent.

We can summarize the MADM in the second year of life by taking its mean. This summary is shown in Figure 5.11. This mean is close to 1 for almost every language and form for which we have data. Confirming the analysis above, we see cross-linguistic consistency in the variability of children.

One question that could be raised regarding the analysis above is the extent to which variability is caused by variability across children vs. variability in reporting. The extreme values seen in the English data, for example, could easily be the result of a mixture of lazy parents who stopped answering the form with overly diligent parents who misunderstood and checked every box for a word they thought the child had been exposed to. But to the extent these biases are the source of variability, they are extremely consistent across languages — which, recall, is the exact opposite argument from the one we considered above (where parent diligence was supposed to be variable enough across samples to lead to differences between languages). Although there are certainly some reporting biases represented in the data, we do not believe that the particular results of this analysis are an artifact of reporting bias.

We can complete the same analysis using comprehension data, shown in Figure 5.12. In this measure, we see a gradual decrease in variability throughout development. The intercept for the 12-18 month period appears to be lower than that observed in production, despite (or because of?) the higher scores. This observation matches one made by Mayor and Plunkett (2014), namely that production vocabulary appears more idiosyncratic in distribution of words than comprehension vocabulary. One speculative explanation for this difference would be the tremendous differences in speech-motor development (as well as general differences in loquacity) between toddlers. This variability would then carry over into production. Another possibility, however, is that true variability is masked by the overall lower reliability of comprehension items (see Chapter 4). Our data do not allow us to distinguish between these two explanations.

The final plot in this sequence is shown in Figure 5.13, which shows 12-18-month MADM values. These are slightly lower and slightly more variable than the production values shown above, but still display a quite consistent level of variability.

5.2.2 Is there a ceiling to variability?

The analysis above suggests that variability between individuals decreases. But this conclusion is compromised by ceiling effects: once children begin to reach the ceiling of the CDI form, variability is necessarily truncated. No analysis can completely eliminate these effects, but the use of item response theory-based analyses can partially address the issue by estimating variation in latent ability rather than variation in raw scores themselves.

Chapter 4 provides a summary of our approach to using item response theory (IRT) with CDI data. In brief, IRT provides a framework in which the full test (the CDI) is broken down into a series of items, with each having its own logistic model predicting the response for a particular child on the

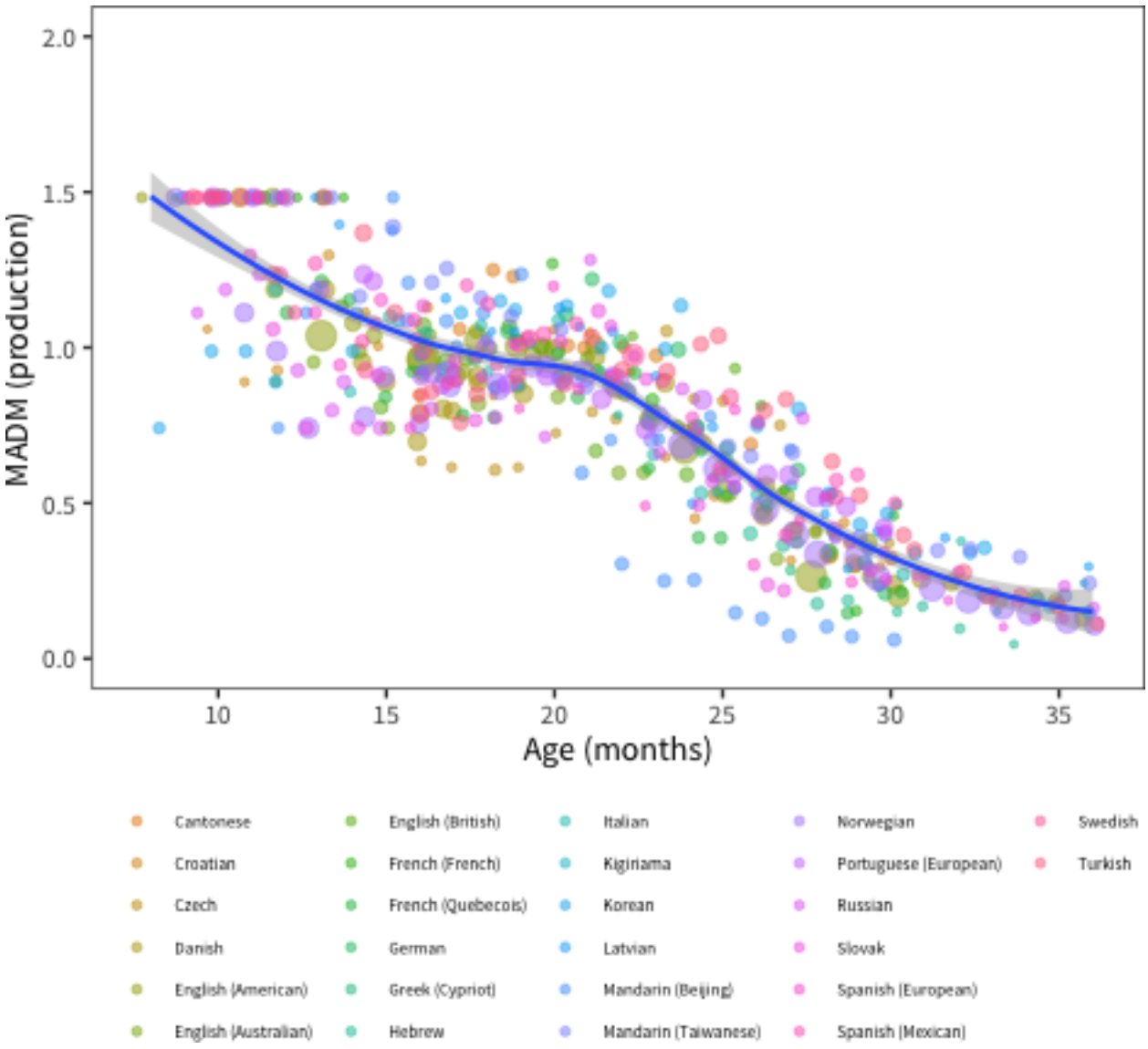


Figure 5.10: MADM for production, plotted by age group, for the full sample of languages in our dataset.

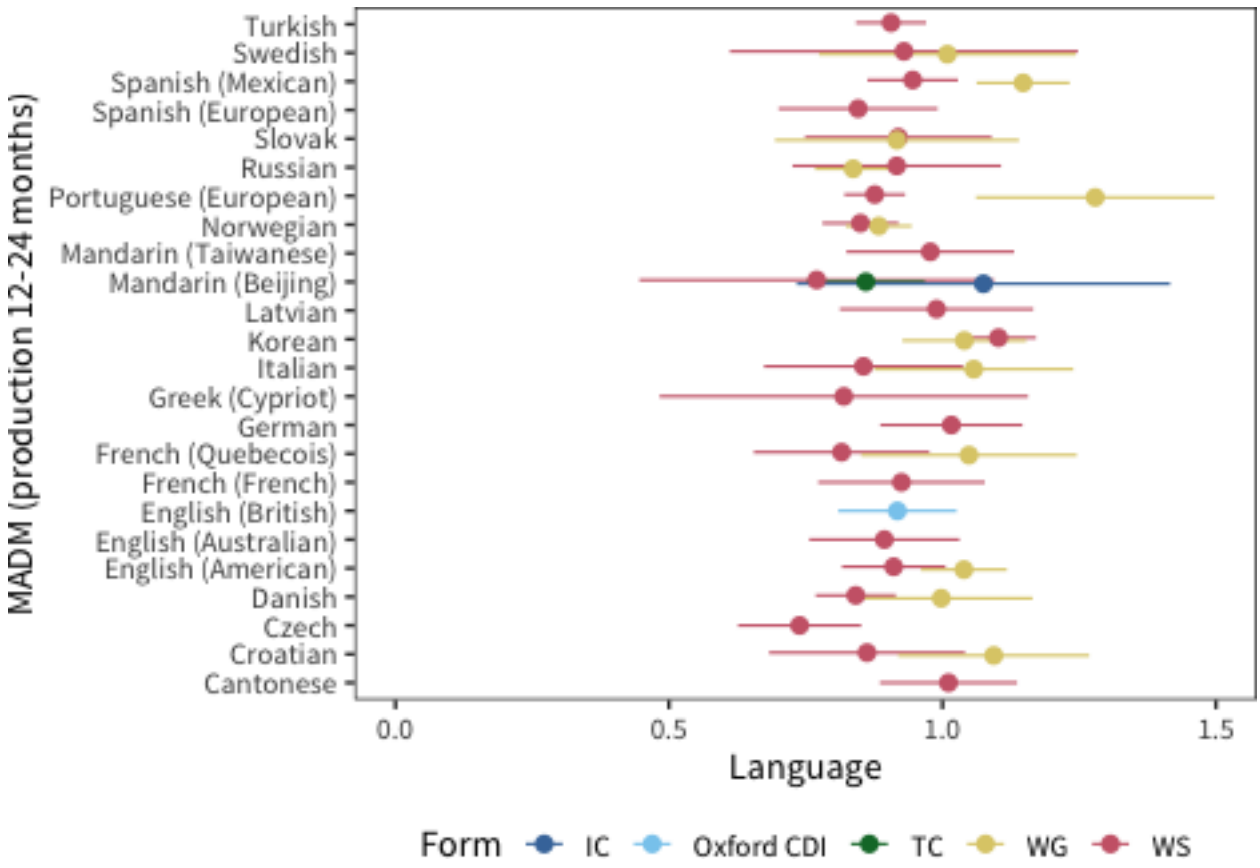


Figure 5.11: MADM values from 12-24 months for all languages and forms.

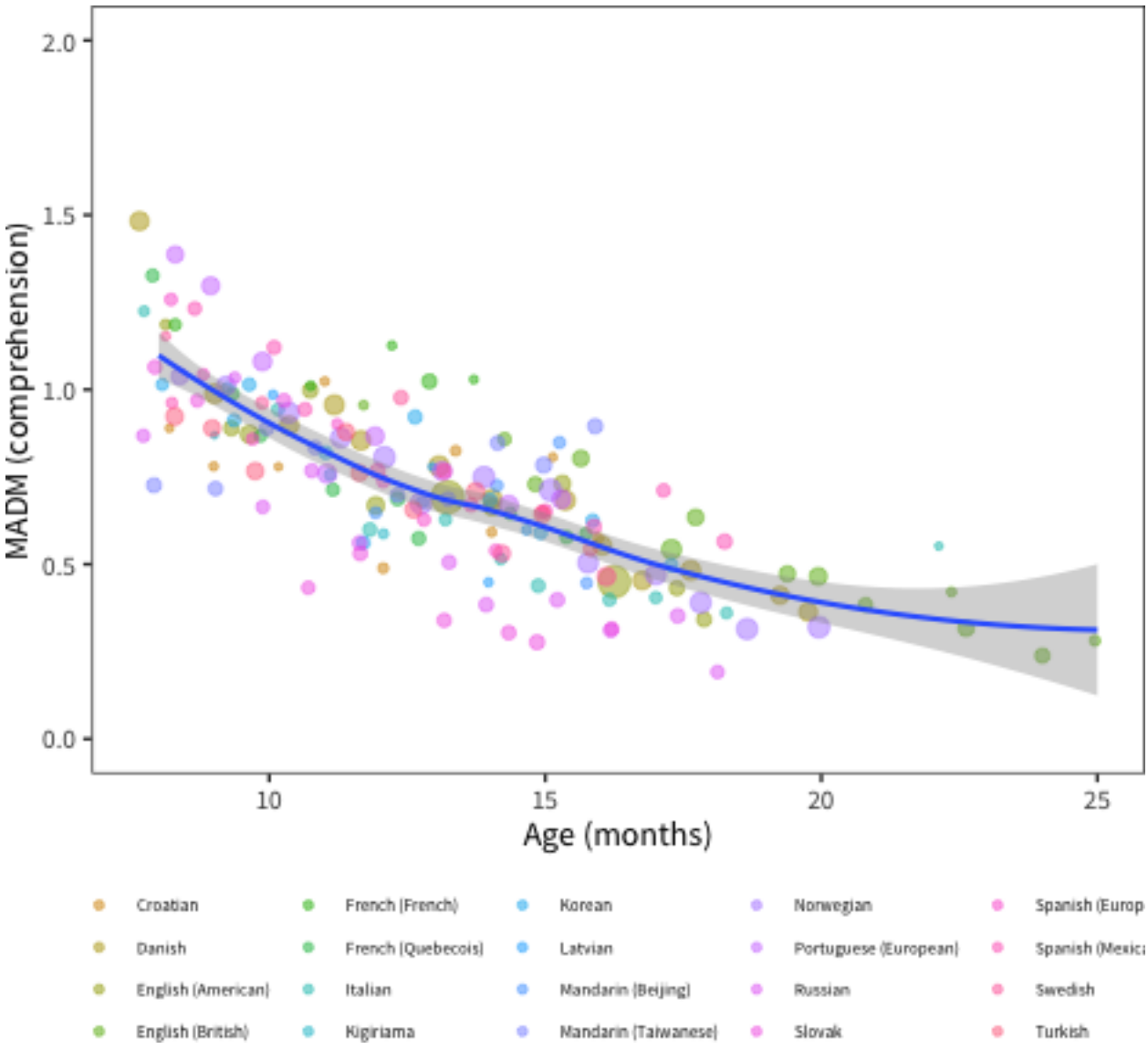


Figure 5.12: MADM for comprehension, plotted by age group, for the full sample of languages in our dataset.

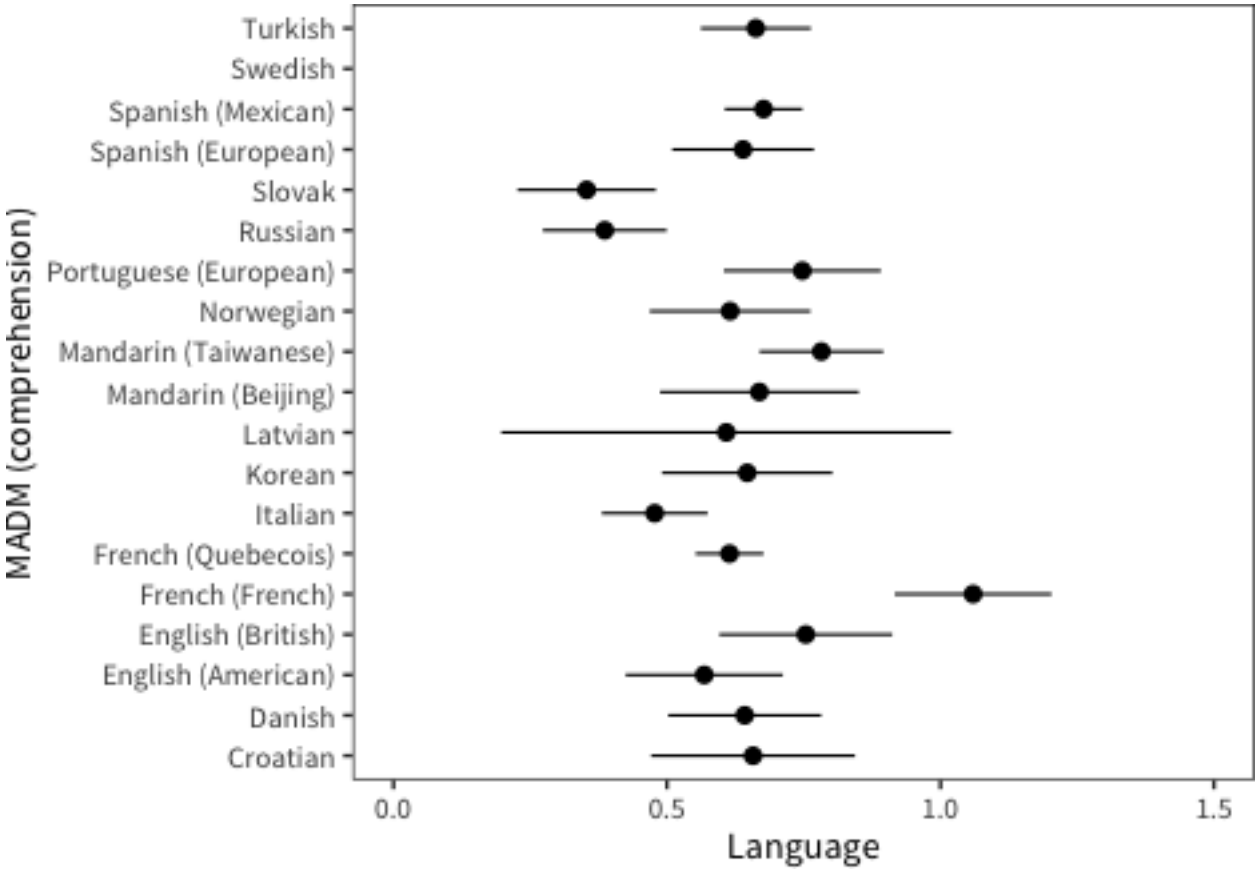


Figure 5.13: MADM values from 12-18 months for all languages and forms.

basis of their latent ability. In Chapter 4, we examine the parameters of individual items with respect to their properties; but fitting an IRT model also implies estimating a set of latent ability parameters for individual participants. These latent ability parameters are logistic regression coefficients and hence are not bounded in the same way that individual responses (and hence raw scores) are. Thus, we can examine their variability as a way of dealing with ceiling effects.

Figure 5.14 shows the normalized standard deviation of latent ability scores, plotted by age. The absolute size of the standard deviations is not easy to interpret — the range of latent ability scores on a form is a function of how consistently difficult or easy the items are (and as we noted above, we do not think it is trivial to equate these scores across tests) Thus, the standard deviation of these scores will be larger or smaller based on this feature as well as the consistency of children's abilities within an age group and is not interpretable. On the other hand, age-related trends in the standard deviation are interpretable and can be used as an index of whether variability in ability stays constant, increases, or decreases.

Interestingly, slopes for the variability of latent ability are fairly flat or increasing (with Mandarin, which has extreme ceiling effects, being the exception). This finding suggests that the decreases in variability observed above are very likely due to ceiling effects. In sum, when we remove ceiling effects, we find that variability is constant — or perhaps even increasing — throughout the full measured range of early language.

5.2.3 Discussion

We observed a striking consistency in the individual variability of children's vocabulary during their second year and perhaps beyond. Across languages and forms, it appears to be the norm that toddlers vary.

What does it mean to have such a high level of variability? For one comparison, we compare age of walking onset (as measured by a Norwegian national survey with parent 47,515 respondents) and age of achieving production and comprehension milestones (also in Norwegian). Walking data are from Størvold et al. (2013).

Comparison of language milestones with walking, in terms of month at which a percentile ranking is achieved.

Response	25th percentile	75th percentile	Range (months)	Range (proportion)
walking	12	14	2	0.15
produces 10 words	13	16	3	0.21
produces 50 words	17	20	3	0.17
produces 100 words	18	23	5	0.25
understands 50 words	10	15	5	0.42
produces 200 words	20	26	6	0.26

Table 5.2.3 shows the 25th and 75th percentiles for a variety of behaviors. The spread of achieving walking (defined as taking a step independently) is quite tight with a mean of 12.9 months and a spread of only a month between 25th and 75th percentile. Very early language comprehension and production are relatively similar with 2 and 3 month spreads. In contrast, production and comprehension at a higher level has quite a large spread in comparison to walking (even as a percentage of age).

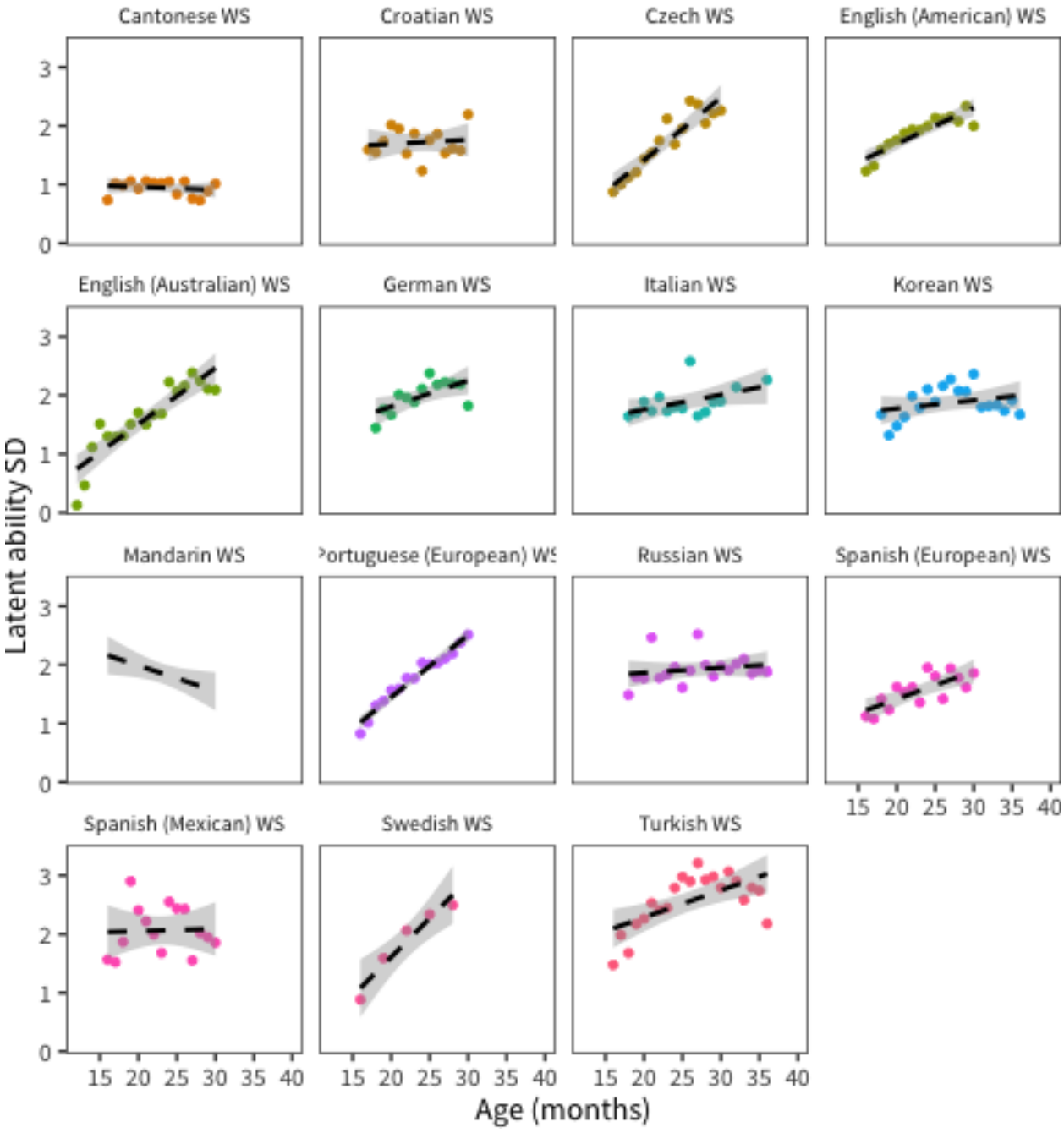


Figure 5.14: Standard deviation of latent ability scores from IRT models fit to each Words and Sentences-type dataset. Panels show individual languages. Smoothing lines are linear model fits.

In sum, these results echo the conclusions of Bornstein and Cote (2005) based their comparative study of Spanish, English, and Italian. They noted that "individual variability is probably a universal feature of early language acquisition" (p. 311).

Chapter 6

Demographic Effects on Vocabulary Size

Chapter 5 examined cross-linguistic consistency and variability in the size of children's reported vocabulary. In this chapter, we follow up these analyses by beginning the process — which will span throughout the book — of attempting to understand the nature and sources of these differences. In particular, we take advantage of the sample diversity described in Chapter 9.1 to explore differences in the median trajectory of vocabulary growth within demographic groups, focusing on three variables that are relatively available in our data: sex, maternal education, and birth order. In particular, we develop a framework for comparing the magnitude of these effects.

One important tension in this chapter reflects many earlier discussions of variability in CDI scores (from Fenson et al., 1994; Feldman et al., 2000; and Eriksson et al., 2012, among others). While some demographic differences in vocabulary are quite consistent — substantially so, as it turns out — the overall proportion of variance in vocabulary that they capture is still relatively limited. We will examine a number of different perspectives on how to quantify this relationship, moving back and forth between emphasizing consistency in differences in the central tendency and emphasizing the limited size of these effects relative to the variability documented in Chapter 5.

Our analyses in this chapter are limited to a subset of languages, as demographic data for many contributed datasets were not available. We begin with sex, the variable for which the most analysis has already been done and for which we have the most data, then move on to birth order, and finally, turn to maternal education. Throughout, we focus on earlier comprehension differences from Words & Gestures-type forms and later production differences from Words & Sentences-type forms. For the sake of length, we omit analysis of production data from WG-type forms, since these data are typically sparse.

6.1 Sex

Our first analysis examines how vocabulary development differs by children's sex.[1] The literature on cognitive differences due to sex is vast, controversial in many places, and difficult to summarize (see Miller and Halpern, 2014, for a useful recent review). Focusing on development, Maccoby and Jacklin (1974) began the enterprise of systematizing and summarizing gender differences. Their

[1] Throughout, we will assume that parents report on children's assigned sex at birth, rather than their gender identity.

conclusions were largely deflationary but did suggest some differences in aggression and verbal ability (which were suggested to emerge in the period of early adolescence). This latter claim is most relevant for our analysis, but has been controversial as well.

Using meta-analytic tools, Hyde and Linn (1988) found that differences in verbal ability were minimal, but more recent studies have suggested consistent verbal ability differences. For example, Stoet and Geary (2013) found differences in reading ability across nations in a massive elementary education dataset (the PISA assessment), with variance in the magnitude of difference, but with girls very consistently showing an advantage. Similarly, Robinson and Lubienski (2011) reported consistent differences in reading ability (favoring girls) at the onset of kindergarten in a nationally-representative US sample. A weak prediction from this literature is thus a modest but consistent female advantage in vocabulary.

Of course, a complication of our analysis is the presence of caregiver reporting bias added to any true sex differences. In contrast to these findings suggesting modest and consistent female advantages, there is substantial cross-linguistic variation in gender stereotypes (Nosek et al., 2009). Thus, a qualitative but plausible speculation is that, if stereotype-based reporting bias plays a major role in gender effects, the cross-national variance should be high.

Despite these predictions, it almost goes without saying that any finding from our analyses here is subject to the full range of possible explanations articulated in the literature. These range from caregiver and academic socialization (e.g., self-fulfilling expectations that girls are more verbal) to "self socialization" in which affiliative differences produce differences in behavior, all the way to biological explanations.[2] While descriptive data of the type cited above (and reported in our analyses below) can be more or less consistent with some of these theories, conclusive evidence will not be forthcoming.

Our analyses below replicate and extend the results of Eriksson et al. (2012), who used an overlapping sample of CDI data from 12 languages to explore sex effects on vocabulary size.[3]

6.1.1 Comprehension (WG)

We begin by examining data from WG-type forms using comprehension measures. Figure 6.1 shows our approach. Each subplot shows median reported comprehension for each age and sex group. Smoothing lines show the predictions of a robust generalized linear model (we selected a robust GLM to avoid some pathological effects from outliers in a small subset of situations).

Visual inspection of the data suggest limited sex differences, but a female advantage is present in some languages (most pronounced in Korean, Latvian, and Hebrew). Note that many authors do not find gender differences in early comprehension. For example, using an overlapping 12 language dataset, Eriksson et al. (2012) concluded that there were no major comprehension differences. And in an earlier study, Feldman et al. (2000) also did not find gender differences in comprehension using a large, relatively representative American dataset, though this study included data only from younger children (10–13 months).

[2] As an illustrative example, some literature has implicated fetal testosterone, though by a Lutchmaya et al. (2001) used CDI measures and recovered an effect somewhat similar to ours with a small sample (d = .64 at 18 months with N = 87), and d = .60 at 24 months for a subsample). They found some relationship with fetal testosterone across sexes, but it did not hold up within sex (perhaps due to small samples). The mechanism by which testosterone translates into vocabulary growth is unclear however.

[3] An earlier version of this analysis was reported in Frank et al. (2016a).

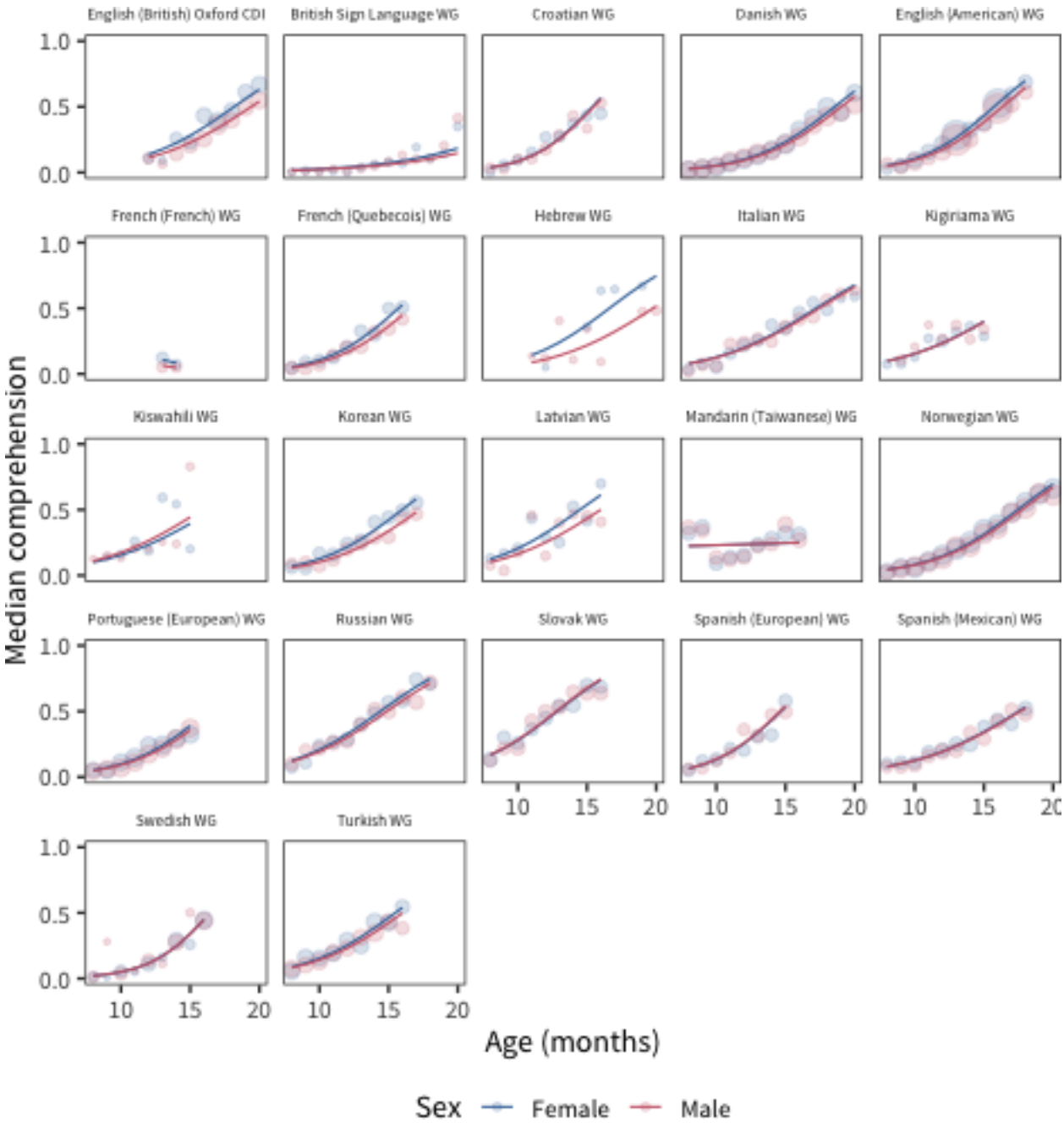


Figure 6.1: Differences in WG comprehension scores by sex, plotted across age by language.

We can examine the statistical models to get a clearer picture. For each language, we the robust generalized linear model predicts vocabulary size (number of words understood out of the total) predicted by age, and the interaction of age and sex We specified this simple model so that the coefficient estimate on the age by sex interaction (as shown in Table 6.1.1) provides a convenient summary of the difference in vocabulary growth across groups. Despite the small magnitudes of the coefficients, 16 of 22 languages had a female advantage. In contrast, 2 showed a male advantage and the remainder did not show a significant sex by age interaction.

Interaction term between age and sex for WG comprehension data in each language.

Language	Form	β	*p*
Hebrew	WG	0.0510	< 0.001
French (French)	WG	0.0438	< 0.001
Latvian	WG	0.0284	< 0.001
Korean	WG	0.0239	< 0.001
French (Quebecois)	WG	0.0197	< 0.001
English (British)	Oxford CDI	0.0189	< 0.001
English (American)	WG	0.0138	< 0.001
British Sign Language	WG	0.0134	< 0.001
Turkish	WG	0.0106	< 0.001
Russian	WG	0.0097	< 0.001

Showing 1 to 10 of 22 entries Previous 1 2 3 Next

Moving away from the model-based method above, we can look for a measure of effect size (similar to that used in the previous chapter). Effect size quantifies the size of the difference between groups in terms of the variability, producing a scale-free measure of difference that is appropriate for comparison across languages. Normally we'd use a measure like Cohen's d here, where

$$d = \frac{\mu_2 - \mu_1}{SD_{pooled}}$$

But as in the previous chapter, we have the problem of non-normal distributions. To circumvent this issue, we use a non-parametric measure derived from the same components: the difference between medians, divided by the MAD. (We call this the MMAD).

Applying this measure to the data on comprehension, we see a quite small average female advantage that appears relatively constant across age (Figure 6.2). For those languages with dense enough data, we can take a weighted average of this pattern across ages, which reveals substantial variability (Figure 6.3). The overall median for these 18 languages is quite small as well, 0.08. In summary, there is some evidence for a modest female advantage in comprehension.

6.1.2 Production (WS)

We next turn to production data on the Words & Sentences instrument, which we expect to be much more informative regarding production (see Figure 6.4. Visual inspection confirms differences in almost every case, and an analysis of the fitted models (see Table 6.1.2) shows that 25 of 26 languages show a statistically significant female advantage!

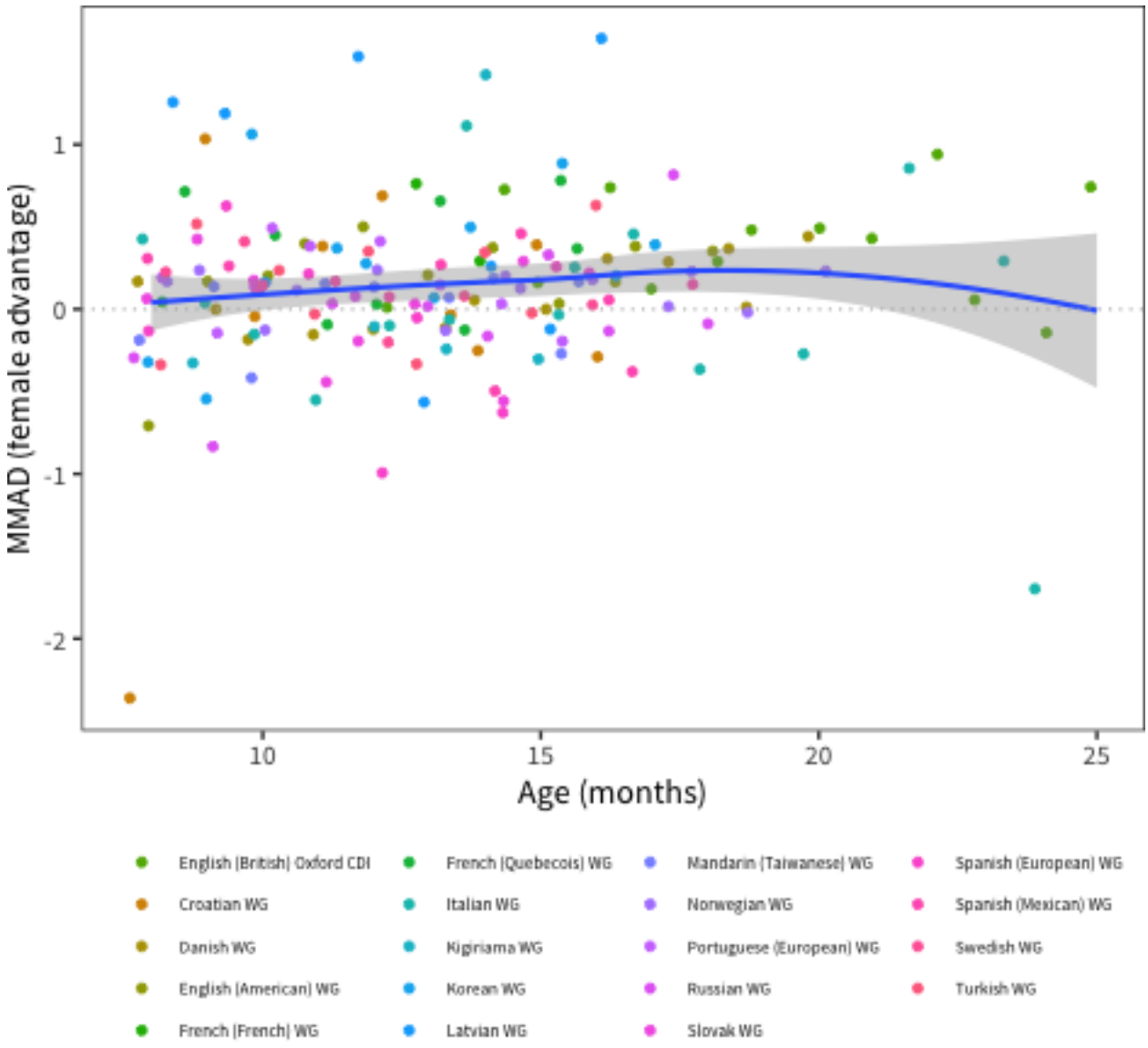


Figure 6.2: MMAD female advantage for WG comprehension data in each language across age.

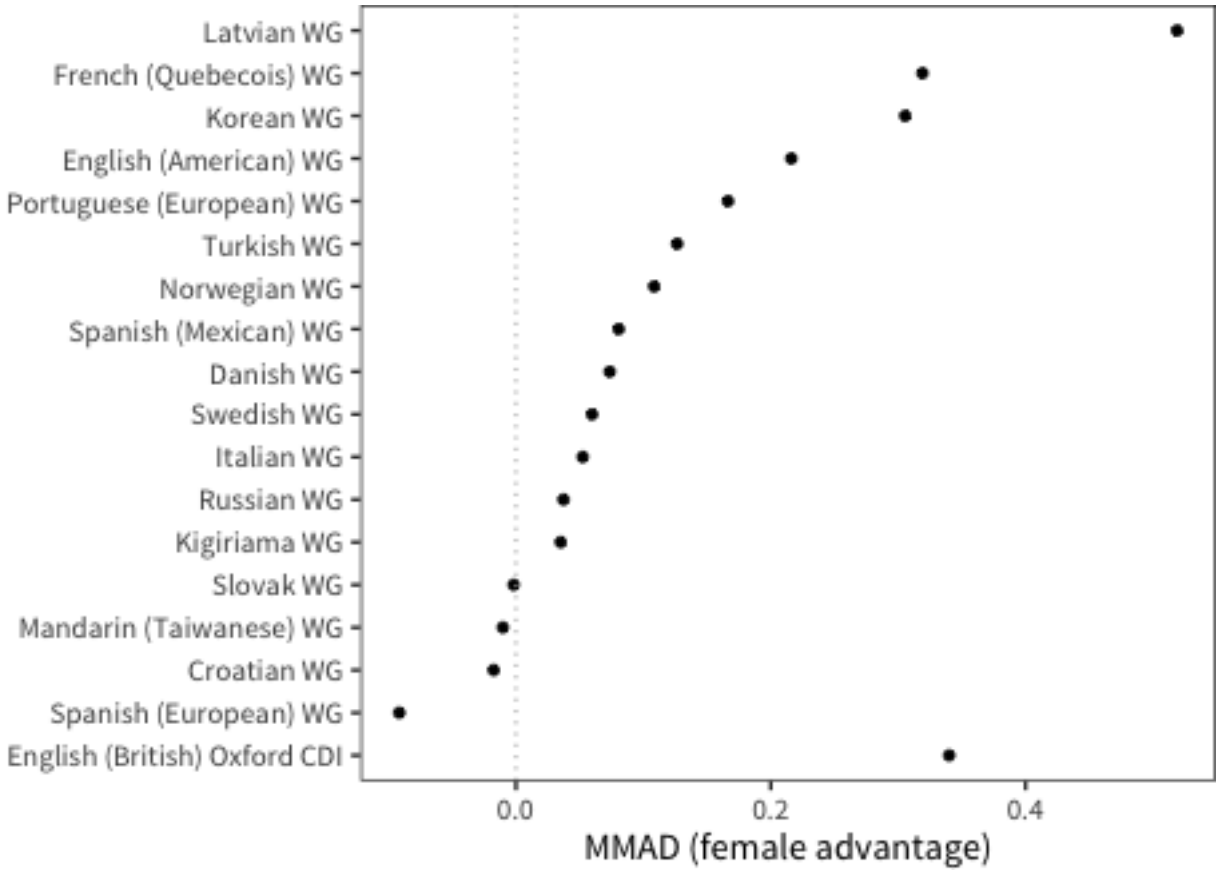


Figure 6.3: MMAD female advantage for WG comprehension data in each language averaged over age.

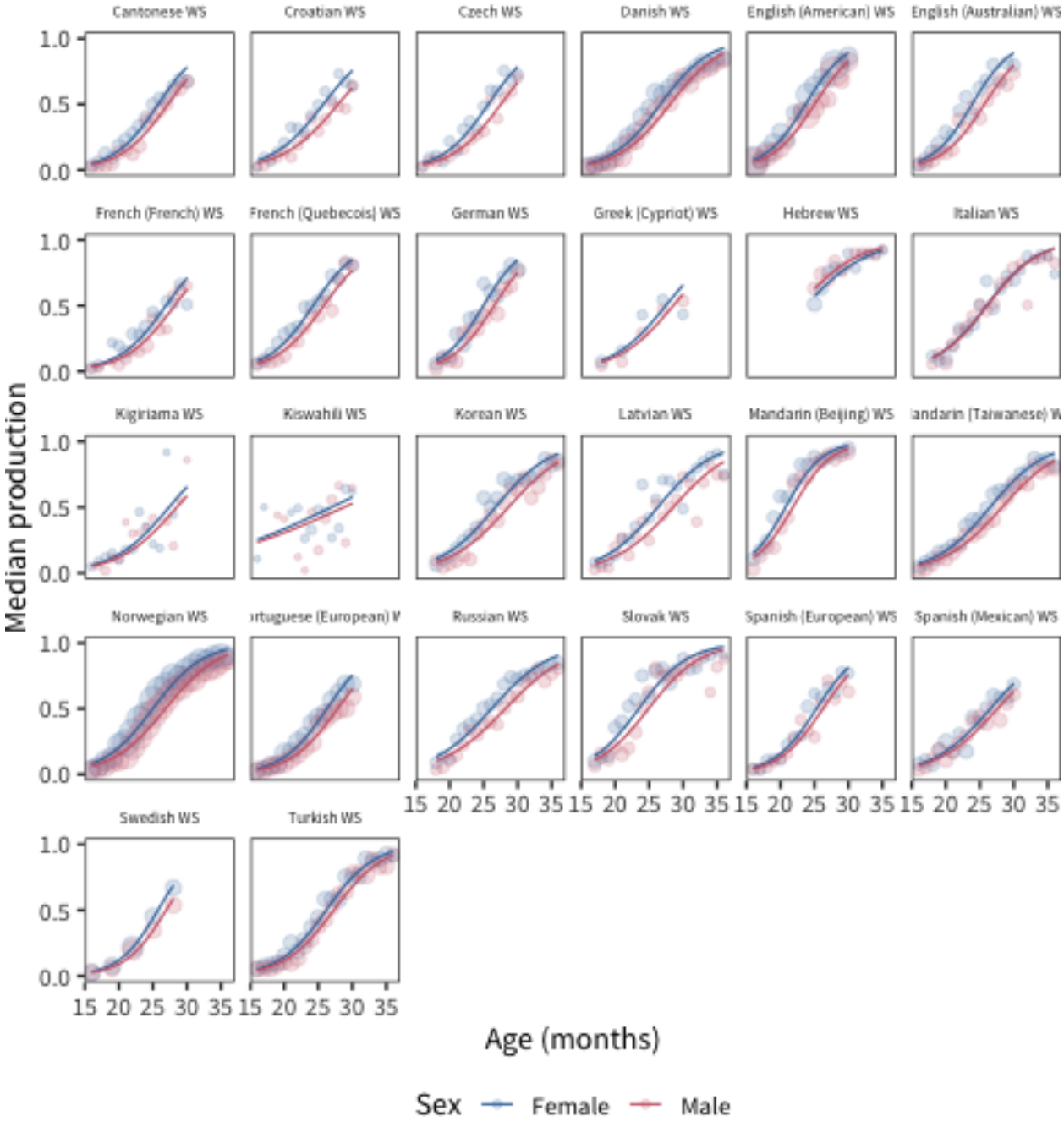


Figure 6.4: Differences in WS production scores by sex, plotted across age by language.

Interaction term between age and sex for WS production data in each language.

Language	Form	β	*p*
English (Australian)	WS	0.0244	< 0.001
Croatian	WS	0.0208	< 0.001
German	WS	0.0204	< 0.001
Czech	WS	0.0201	< 0.001
Latvian	WS	0.0196	< 0.001
Norwegian	WS	0.0188	< 0.001
Mandarin (Beijing)	WS	0.0187	< 0.001
French (Quebecois)	WS	0.0186	< 0.001
English (American)	WS	0.0180	< 0.001
Portuguese (European)	WS	0.0171	< 0.001

Showing 1 to 10 of 26 entries Previous 1 2 3 Next

We next turn to the MMAD effect size measure (see Figure 6.5 and Figure 6.6). Here we see a relatively consistent difference across ages with perhaps a slight downward trend in effect size. This downward trend might be a function of ceiling effects on the form, however, as seen in the model-fit curves above for, e.g., Danish. When variability is limited by the form ceiling, effect size estimates will necessarily be depressed. The median female advantage is 0.4, substantially larger than that seen in early comprehension and somewhat larger than that seen in early production (perhaps due to floor effects early on).

6.1.3 Reporting bias?

Do these differences reflect differences in measurement that are unique to the CDI? One way of addressing this question is to examine other studies of gender differences that have been found in other studies. Unfortunately, many of the studies reporting differences themselves rely on the CDI, likely for the reasons reviewed in Chapter 2, (e.g. Bauer et al., 2002; Fenson et al., 1994; Feldman et al., 2000). For example, Feldman et al. (2000) collected CDIs with a large dataset of low-income American English speakers at 12 and 24 months. In those data, early comprehension showed no significant gender differences, but production at 24 months showed a difference comparable to what we observed here (N = 2156, d = .35, recomputed from provided summary statistics). These data, while providing replication in an independent dataset, do not speak to whether reporting biases contributed to or created the observed sex effects.

For external validation, we turn to two other studies that provide more objective (non parent-report) measurements of early language. First, a seminal study by Huttenlocher et al. (1991) measured gender effects in vocabulary production as estimated from a naturalistic language sample, finding substantial differences in vocabulary growth favoring girls. Although the measures from this study are not comparable to the current data, the effects are quite large (and are relatively unaffected by controlling for maternal language exposure).

Second, Bornstein and Putnick (2012) use a particularly powerful study design to examine stability in early language estimates across different measures. They gathered longitudinal data at 20 and 48 months using a wide range of standardized and parent-report measures, and then use structural

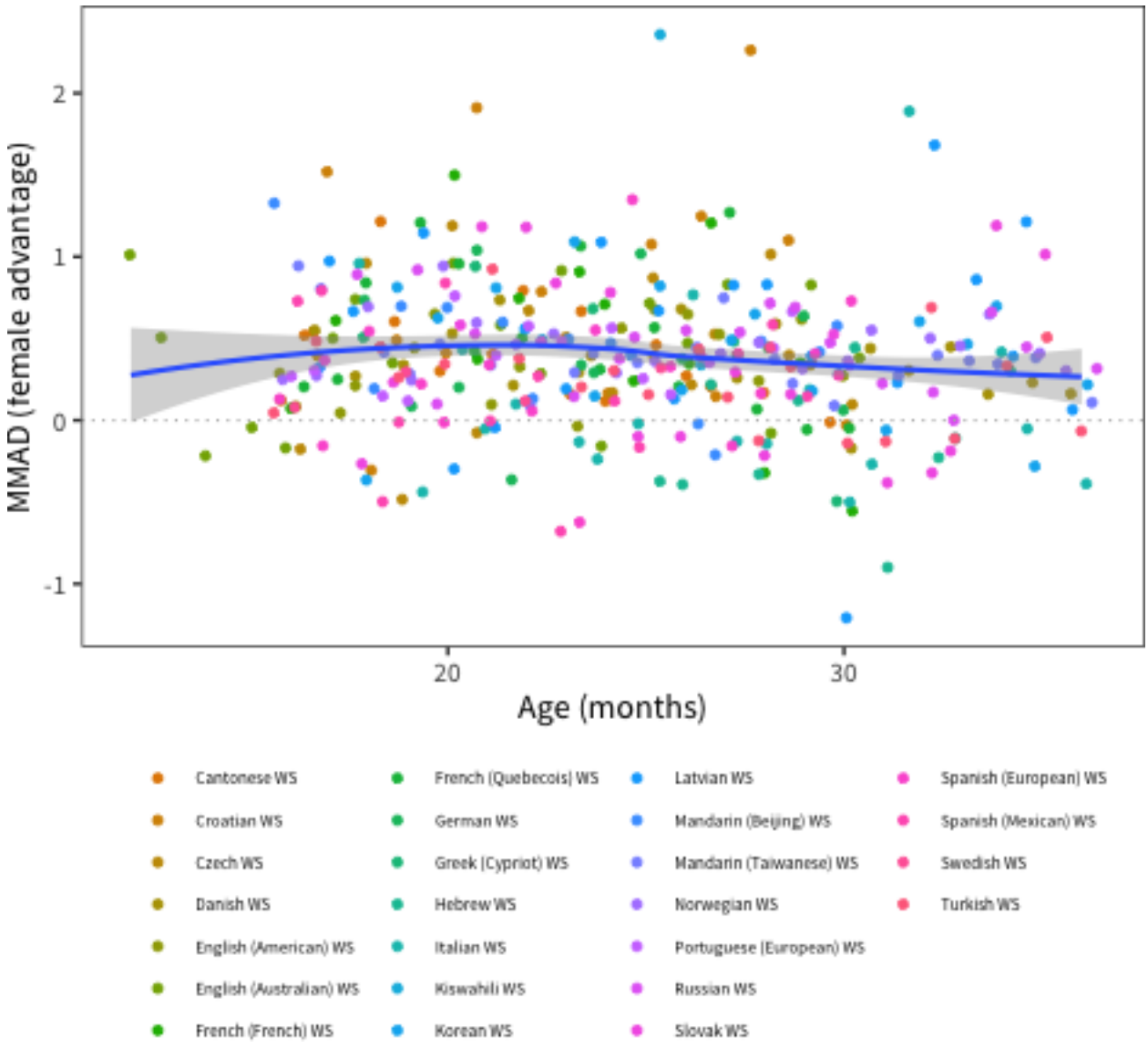


Figure 6.5: MMAD female advantage for WS production data in each language across age.

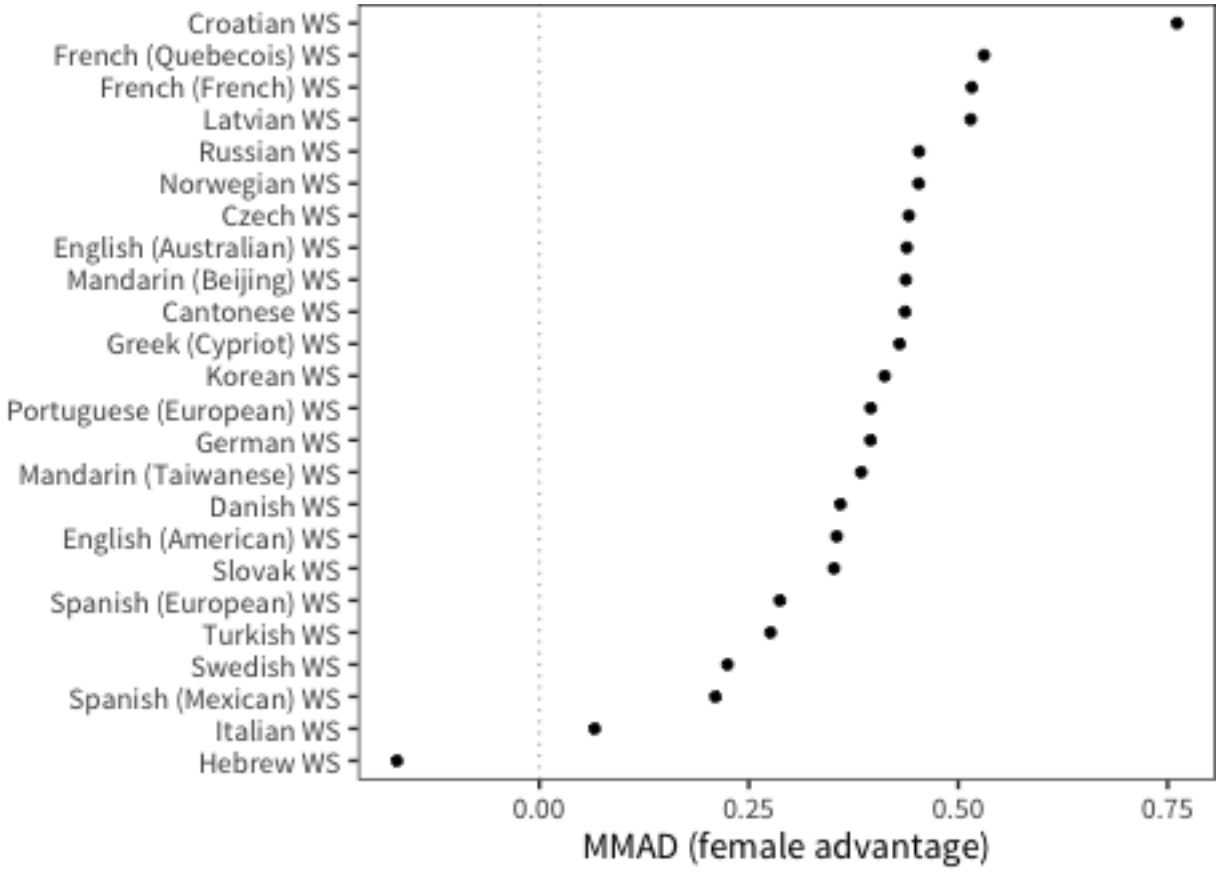


Figure 6.6: MMAD female advantage for WS production data in each language averaged over age.

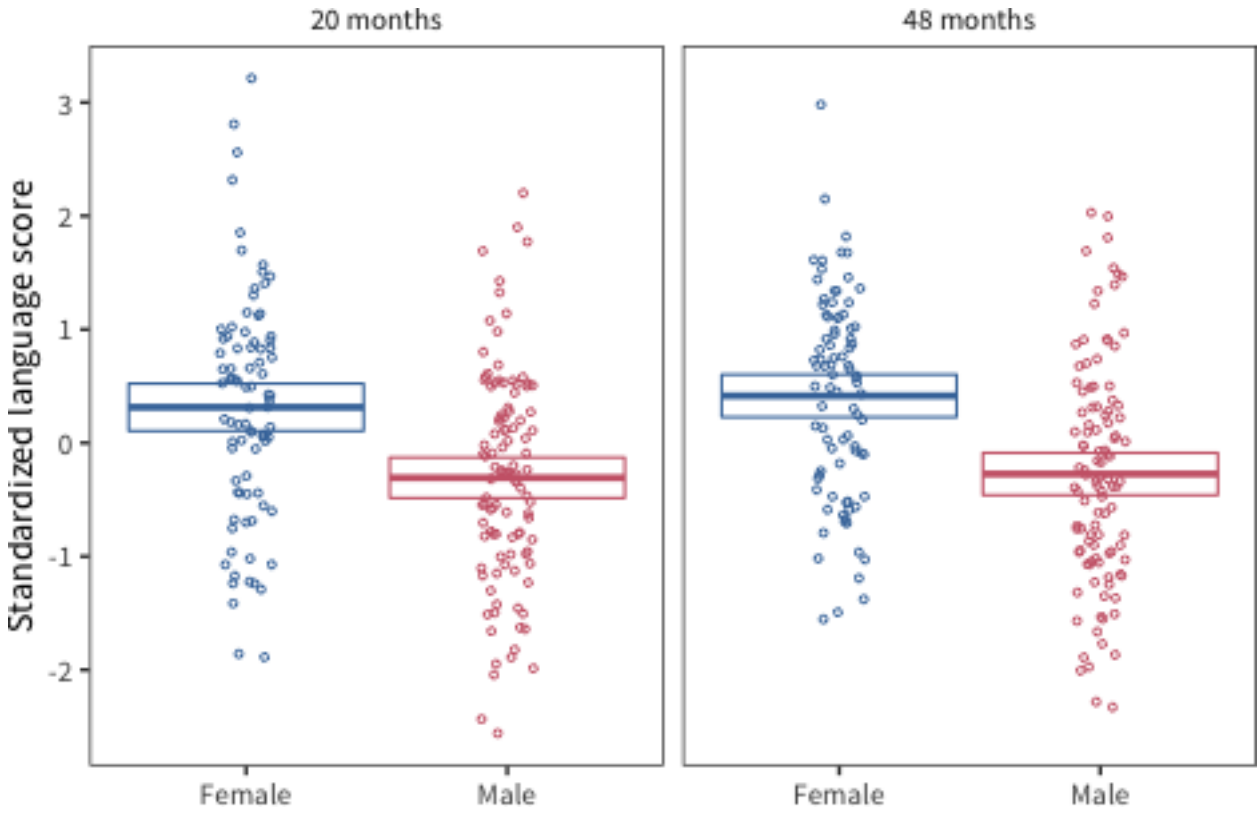


Figure 6.7: Latent vocabulary scores from Bornstein and Putnick (2012) by age and sex (crossbars show means and 95 percent confidence intervals).

equation modeling to model shared variance due to parent report as well as standardized latent language ability at each age. We digitized data from their Figure 2 to examine the size of the gender differences in the latent vocabulary construct that they recovered (see Figure 6.7).[4] As these scores are standardized, we can examine the difference in means and recover the standardized effect size for gender in the data, which is 0.62 standard deviations. Since this measurement is greater than that found using the CDIs (the comparable American English measurement was 0.35), we doubt that our observed effect is due solely to reporting bias.

6.1.4 Discussion

In summary, we found a considerable and strikingly consistent cross-linguistic female advantage in early language production (replicating and extending Eriksson et al., 2012). A much smaller but still relatively consistent female advantage was reported in comprehension. We suspect that neither of these effects are due to parental reporting bias. First, our review of the literature suggests that studies using direct assessments yield similar effects to the extent that we were able to compare. Second, comprehension is very likely to be the measure more affected by reporting biases as it is likely to be more subjective (Feldman et al., 2000; Fenson et al., 2000). In contrast, we find a much smaller gender effect in comprehension. As noted above, we remain agnostic about the causes of these differences, but in Chapter 15 we discuss inferences from consistency across languages in much

[4]These data are not perfect, we've have double-digitized one point.

more detail.

6.2 Birth order

Another factor that may contribute to individual differences in children's vocabulary development is birth order. The literature suggests some evidence for a first-born advantage in early vocabulary development, but these differences are small and tend to be most evident early in development. For example, Bornstein et al. (2004a) found that mothers report larger receptive and expressive vocabularies in their first-borns. Using naturalistic language samples, Berglund et al. (2005) found that first-born children reached the 50-word milestone earlier than later-born children, but that birth order differences diminished later in development. Finally, Hoff-Ginsberg (1998) found that first-born children were more advanced in vocabulary development than later-born children, but that later-born children were more advanced in their conversational skills.

Here, we can examine birth order effects in early vocabulary comprehension and production in a few languages in our sample. Only 12 languages have birth order data, with data available for 11 languages for Words & Sentences, and 8 languages for Words & Gestures.

6.2.1 Comprehension (WG)

We perform the same set of analyses as for sex, shown in Figure 6.8, Figure 6.9, Figure 6.10, and Table 6.2.1.

Interaction term between age and birth order for WG comprehension data in each language.

Language	Form	β	*p*
Hebrew	WG	0.0398	< 0.001
Latvian	WG	0.0187	< 0.001
Spanish (European)	WG	0.0142	< 0.001
Mandarin (Taiwanese)	WG	-0.0002	0.852
Norwegian	WG	-0.0022	< 0.001
English (American)	WG	-0.0035	< 0.001
Spanish (Mexican)	WG	-0.0035	< 0.001
Korean	WG	-0.0305	< 0.001

6.2.2 Production (WS)

The parallel set of analyses for WS Production data are shown in Figure 6.11, Figure 6.12, Figure 6.13, and Table 6.2.2.

Interaction term between age and birth order for WS production data in each language.

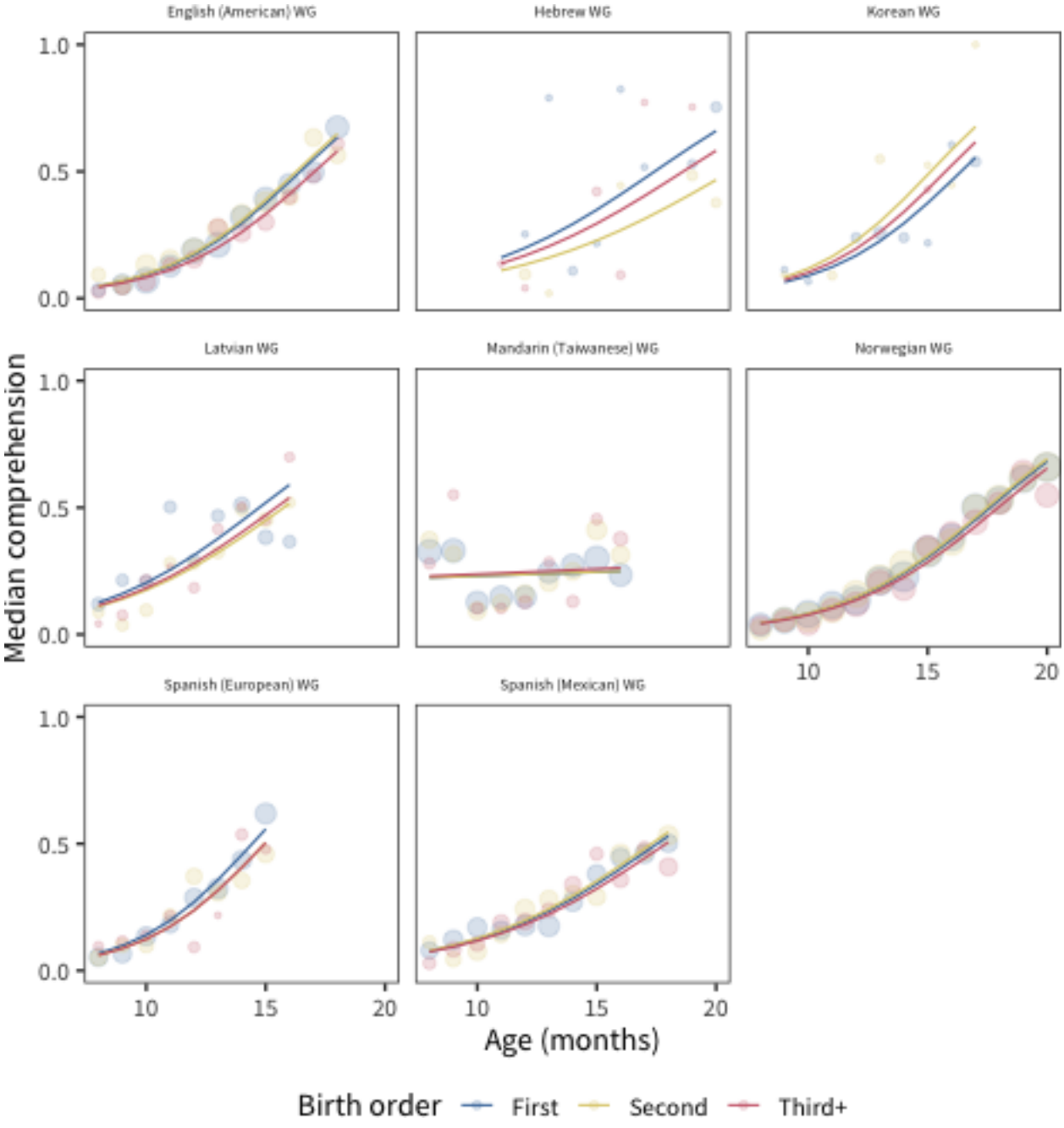


Figure 6.8: Differences in WG comprehension scores by birth order, plotted across age by language.

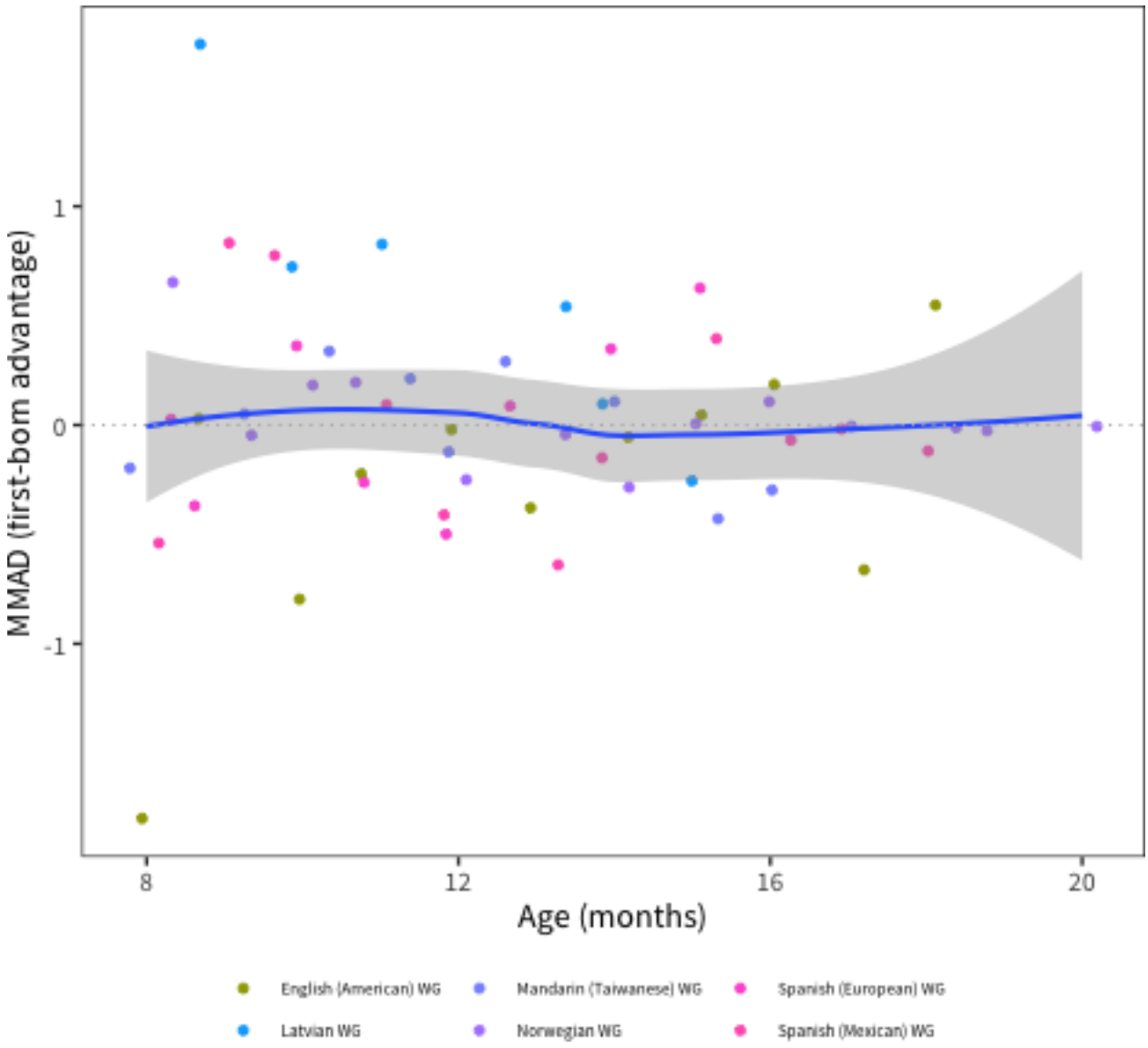


Figure 6.9: MMAD first-born advantage for WG comprehension data in each language across age.

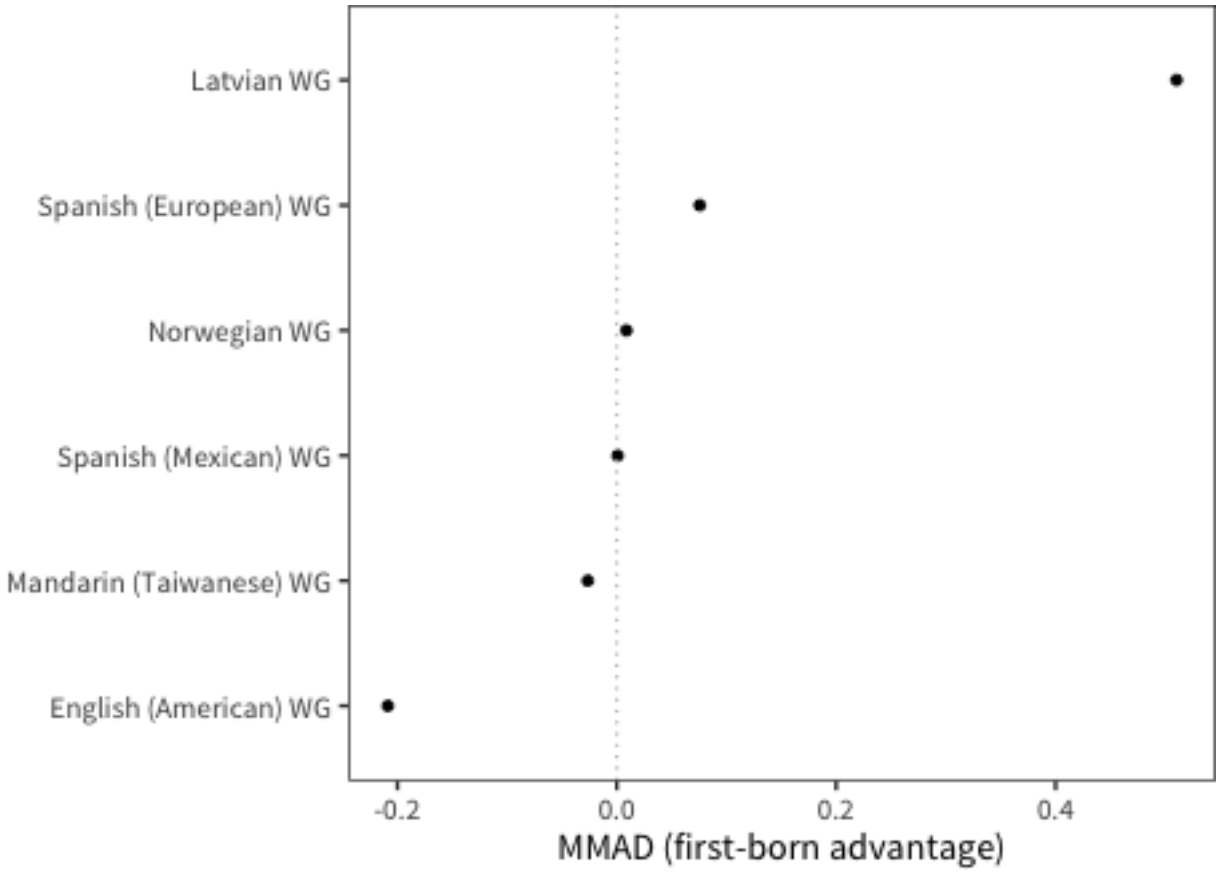


Figure 6.10: MMAD first-born advantage for WG comprehension data in each language averaged over age.

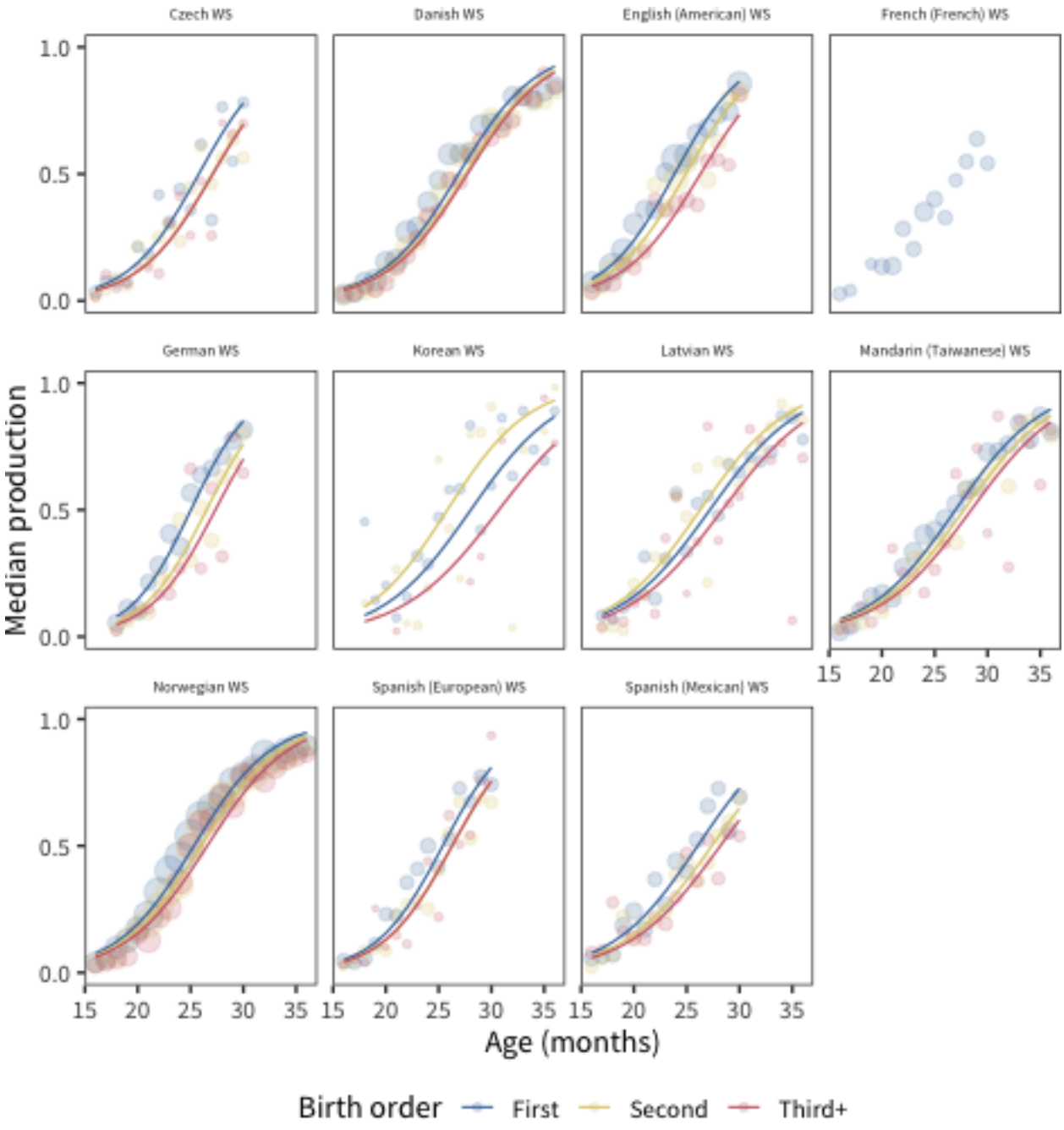


Figure 6.11: Differences in WS production scores by birth order, plotted across age by language.

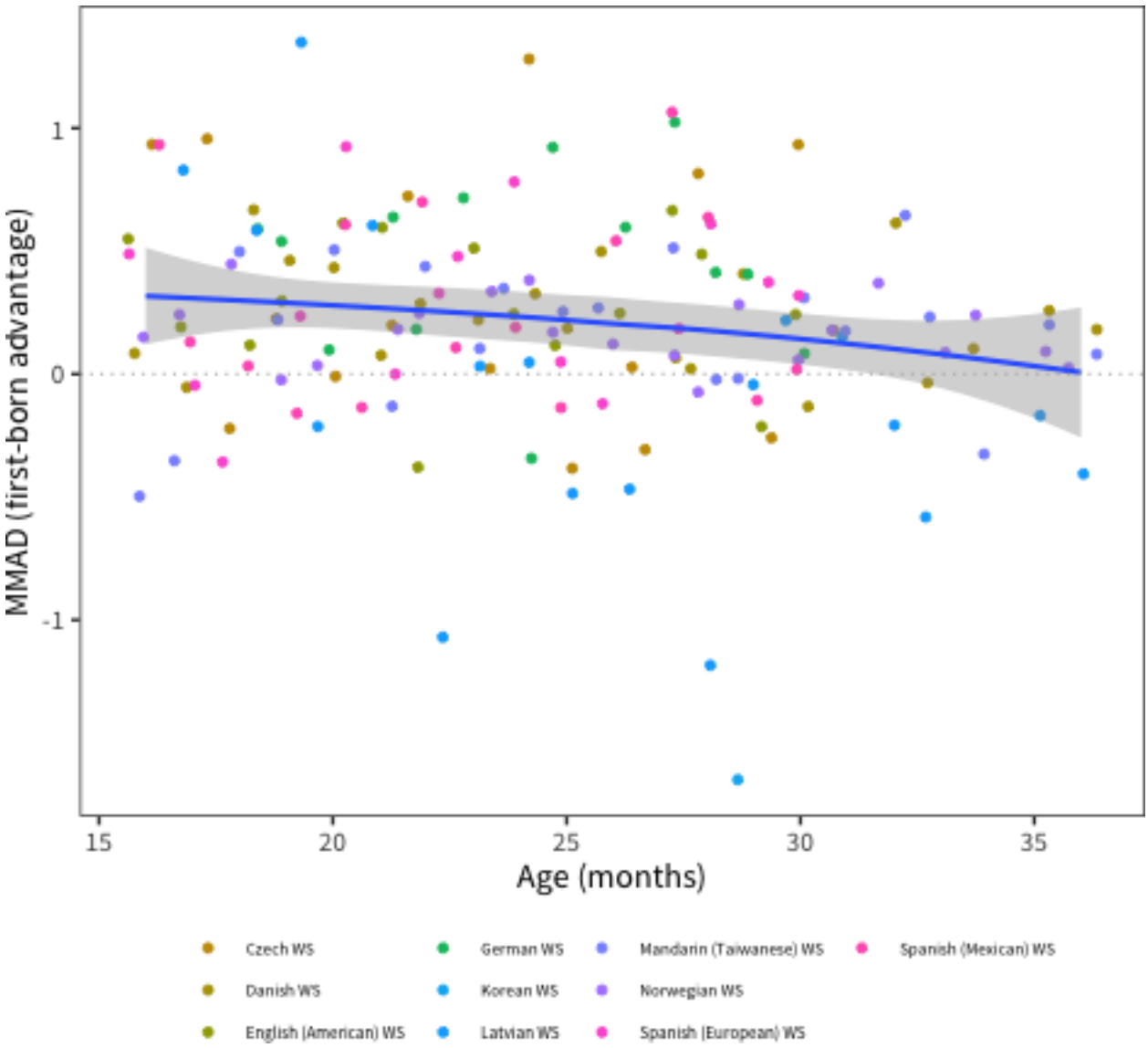


Figure 6.12: MMAD first-born advantage for WS production data in each language across age.

Language	Form	β	*p*
German	WS	0.0194	< 0.001
Czech	WS	0.0138	< 0.001
English (American)	WS	0.0130	< 0.001
Spanish (Mexican)	WS	0.0121	< 0.001
Spanish (European)	WS	0.0110	< 0.001
Norwegian	WS	0.0073	< 0.001
Mandarin (Taiwanese)	WS	0.0070	< 0.001
Danish	WS	0.0062	< 0.001
Latvian	WS	-0.0081	< 0.001
Korean	WS	-0.0204	< 0.001

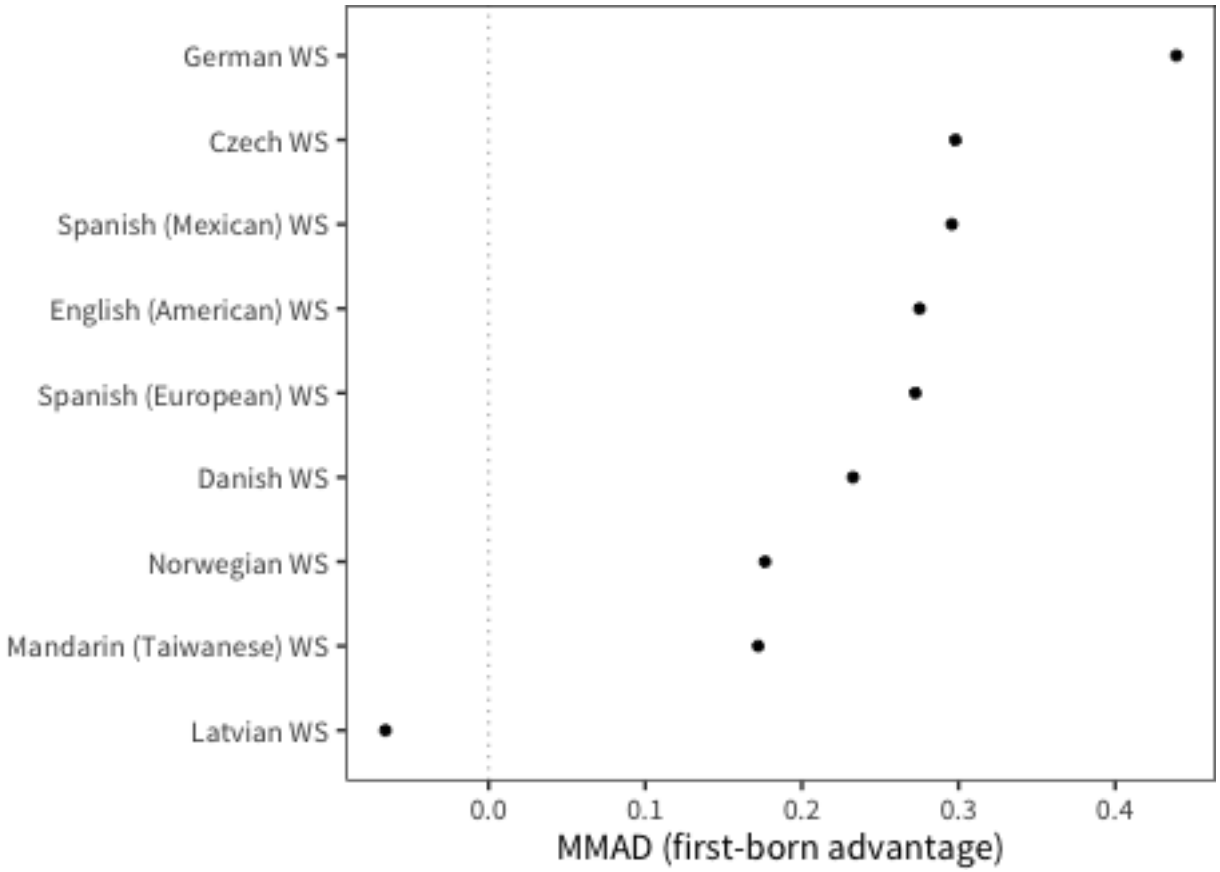


Figure 6.13: MMAD first-born advantage for WS production data in each language averaged over age.

6.2.3 Discussion

In sum, we see a relatively consistent cross-linguistic pattern: earlier-born children show larger vocabularies in production (though not in comprehension for the most part). This general finding is consistent with previous literature reporting a first-born advantage for individual languages. Our results suggest that the same pattern appears in most languages, with only a few showing different magnitudes. While the current dataset cannot rule out reporting-related reasons for these demographic differences, this explanation does seem unlikely for two reasons. First, our results largely mirror non-parent report findings in the literature (Berglund et al., 2005). Second, reporting bias would be relatively more likely to influence comprehension relative to production vocabulary.

We can only speculate as to the cause of birth-order related differences in early language given our current data. That said, it seems very reasonable to assume that parents speak more to first-born children as the addressee, just because of the pure statistical fact of having a second possible addressee for other utterances. And although it is certainly possible to learn from overheard speech under optimal conditions (e.g., Akhtar et al., 2001; Akhtar, 2005), a variety of studies suggest that speech directed to a particular child is the best predictor of that child's learning outcomes (Weisleder and Fernald, 2013; Shneidman and Goldin-Meadow, 2012).

6.3 Socioeconomic status

From health to education, children from lower socioeconomic status (SES) backgrounds tend to be at higher risk for a variety of negative developmental outcomes, compared to their higher-SES peers (Bradley and Corwyn, 2002). A large literature documents specific relations between SES and children's early language abilities, especially oral vocabulary, which is in turn related to outcomes when children begin formal education (e.g., Hart and Risley, 1995; Hoff, 2003; Fernald et al., 2013). The parent report method allows the assessment of the influence of SES on vocabulary outcomes earlier in development than is possible with direct assessments. Using the CDI Words & Sentences form, Arriaga et al. (1998) compared the language skills in 103 very low-income toddlers with a sample of middle-income toddlers from the Fenson et al. (2007) norming sample, matched on age and sex. They found that the vocabulary production scores for the low-income group were consistently about 30% lower than those for the middle-income group. The size of these effects suggest that differences in SES are evident from the earliest phases of language development.

Environmental explanations of these SES effects are often given in terms of indirect factors that affect life opportunities or experiences, such as nutrition and access to health care, as well as more direct factors that impact daily life, such as smoking during pregnancy, access to quality child care, or amount of time caregivers spend in interactions with their young children. Alternatively, even early language shows a significant genetic component, raising the possibility that SES-vocabulary links may instead be genetically mediated (Hayiou-Thomas et al., 2012).

In the current dataset, we have the opportunity to explore the extent of these effects across several language communities. Cross-language comparisons may shed light on the factors that lead to relations between SES and children's vocabulary outcomes. On the one hand, relatively constant relations across language communities that vary widely in indirect and direct factors that shape learning would provide prima facie support for genetic explanations. In contrast, a greater degree of cross-language variability would point to the origins of SES effects in aspects of children's early environments that vary with SES to differing degrees across countries (e.g., Fernald et al., 2012).

We use maternal education as a proxy for SES, following previous work suggesting that maternal education is strongly related to SES variation (Bornstein et al., 2003; Hoff, 2003).

6.3.1 Comprehension (WG)

We again perform the same set of analyses, shown in Figure 6.14, Figure 6.15, Figure 6.16, and Table 6.3.1.

Interaction term between age and maternal education for WG comprehension data in each language.

Language	Form	β	*p*
Spanish (Mexican)	WG	0.1092	< 0.001
British Sign Language	WG	0.0260	< 0.001
Danish	WG	0.0029	< 0.001
Mandarin (Taiwanese)	WG	-0.0019	0.013
English (American)	WG	-0.0024	< 0.001
Latvian	WG	-0.0063	< 0.001
Portuguese (European)	WG	-0.0068	< 0.001
Norwegian	WG	-0.0100	< 0.001
Spanish (European)	WG	-0.0231	< 0.001
Hebrew	WG	-0.0321	< 0.001
Korean	WG	-0.0353	< 0.001

6.3.2 Production (WS)

The final set of analyses are shown in Figure 6.17, Figure 6.18, Figure 6.19, and Table 6.3.2.

Interaction term between age and maternal education for WS production data in each language.

Language	Form	β	*p*
German	WS	0.0171	< 0.001
Latvian	WS	0.0166	< 0.001
English (American)	WS	0.0139	< 0.001
Mandarin (Taiwanese)	WS	0.0122	< 0.001
Czech	WS	0.0063	< 0.001
Danish	WS	0.0049	< 0.001
Portuguese (European)	WS	0.0046	< 0.001
Norwegian	WS	0.0043	< 0.001
Korean	WS	-0.0003	0.733
French (French)	WS	-0.0016	0.018
Spanish (European)	WS	-0.0050	< 0.001

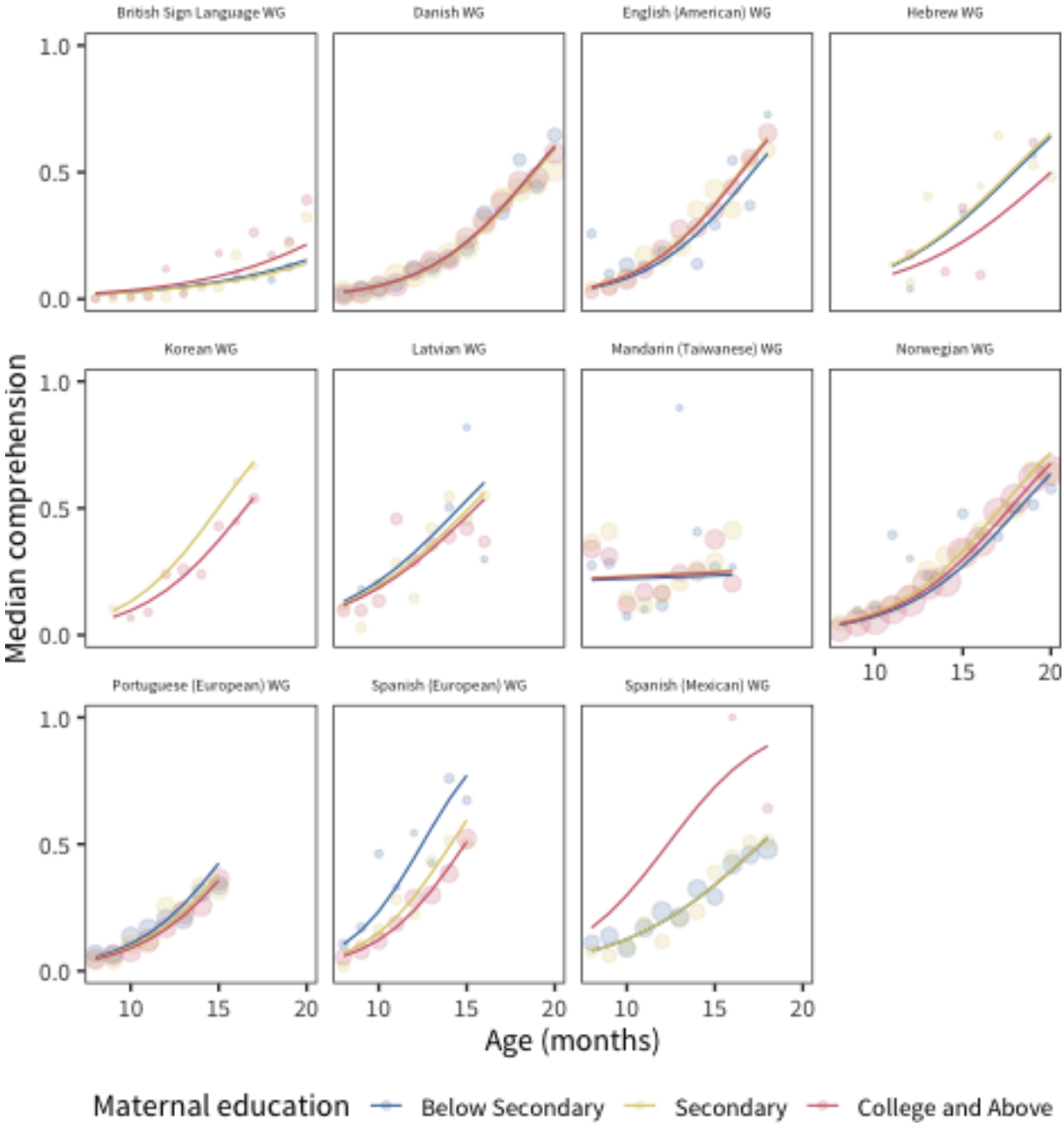


Figure 6.14: Differences in WG comprehension scores by maternal education, plotted across age by language.

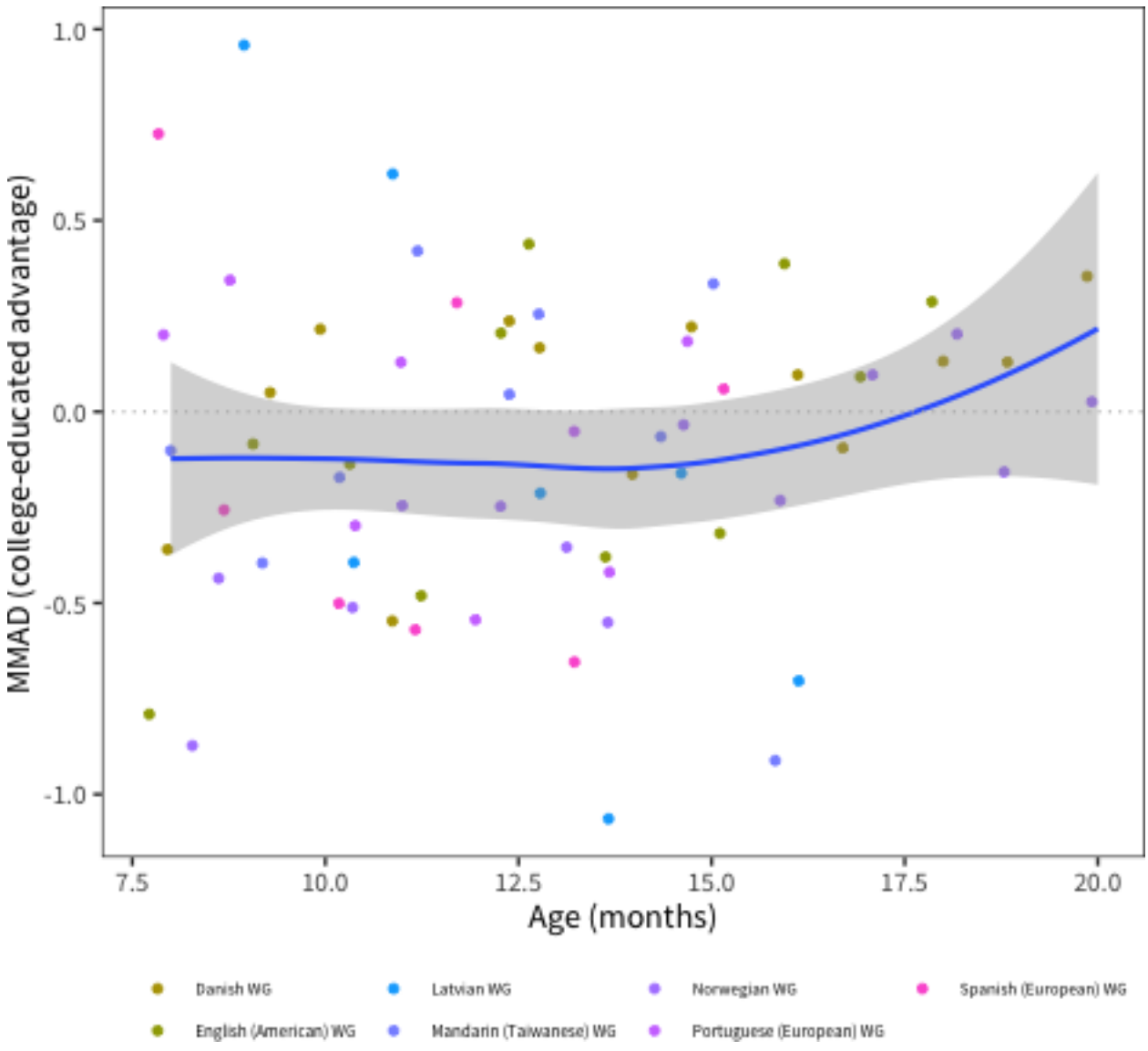


Figure 6.15: MMAD college-educated advantage for WG comprehension data in each language across age.

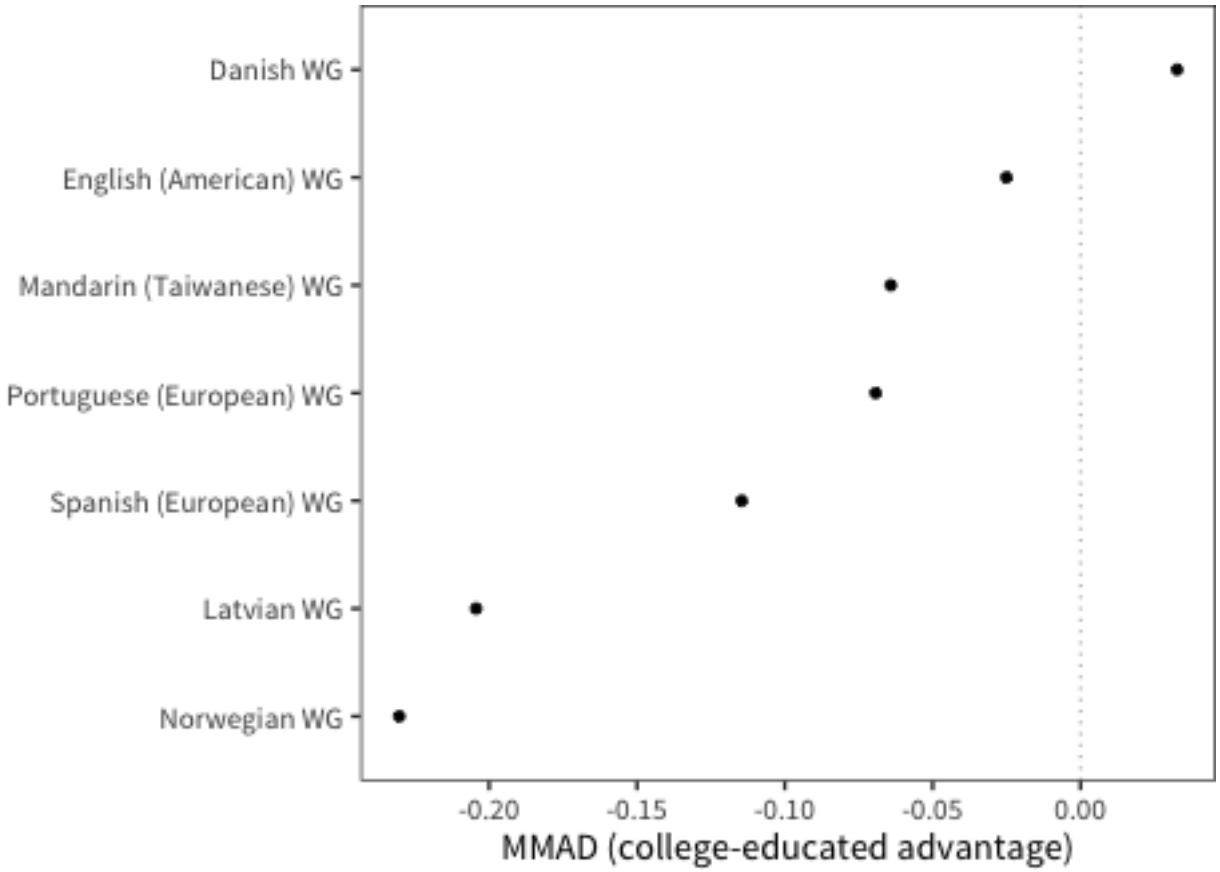


Figure 6.16: MMAD college-educated advantage for WG comprehension data in each language averaged over age.

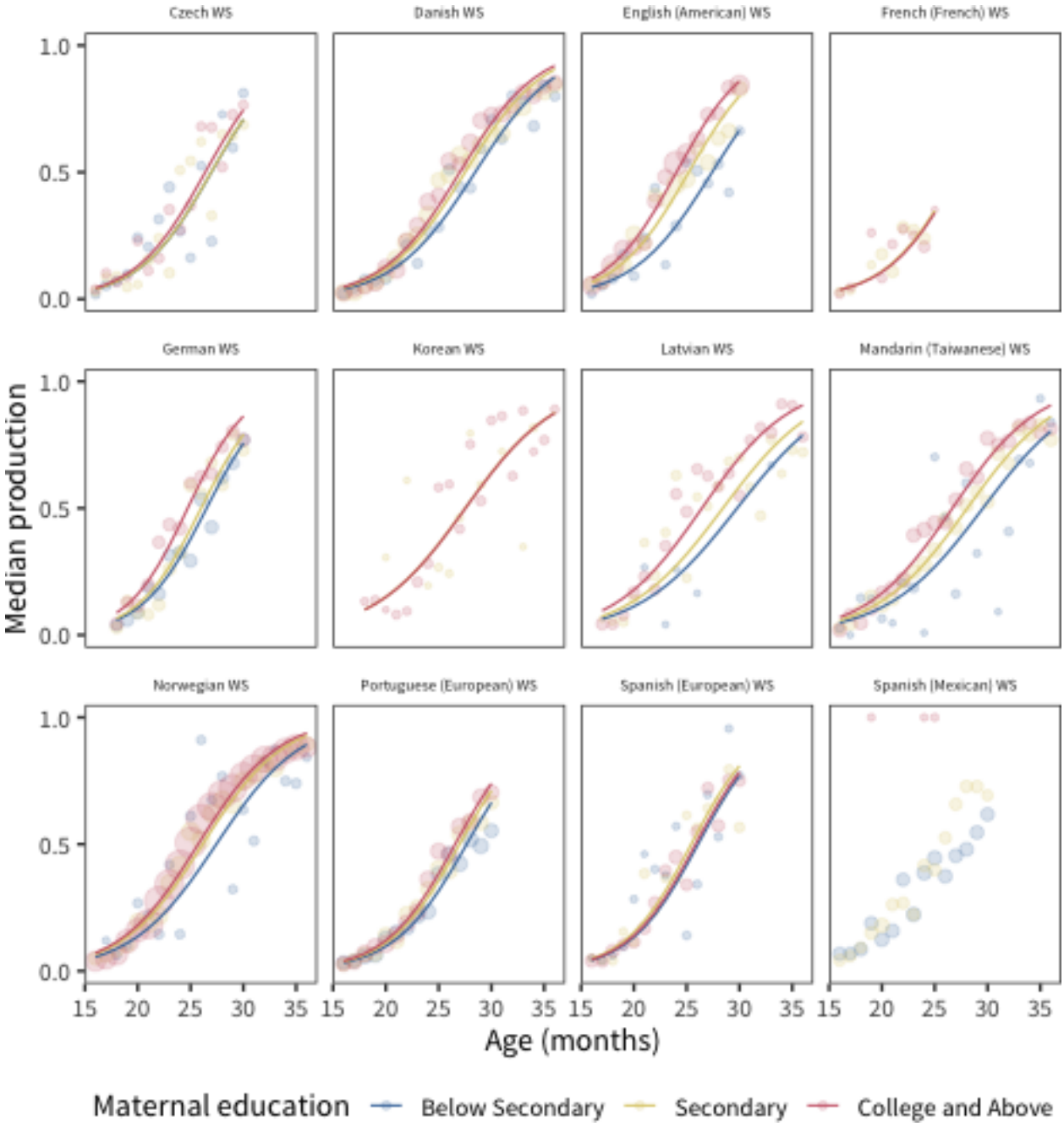


Figure 6.17: Differences in WS production scores by maternal education, plotted across age by language.

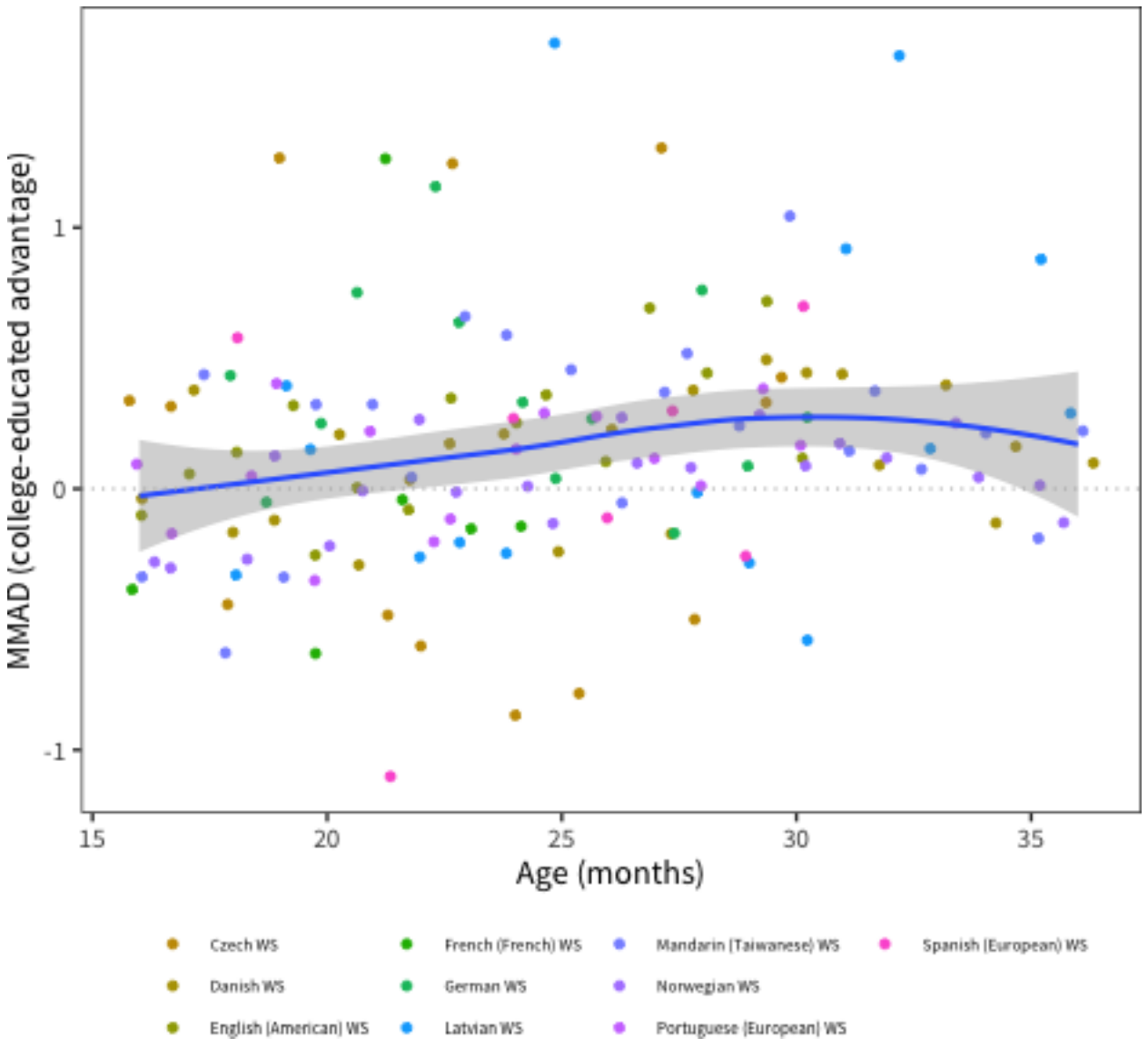


Figure 6.18: MMAD college-educated advantage for WS production data in each language across age.

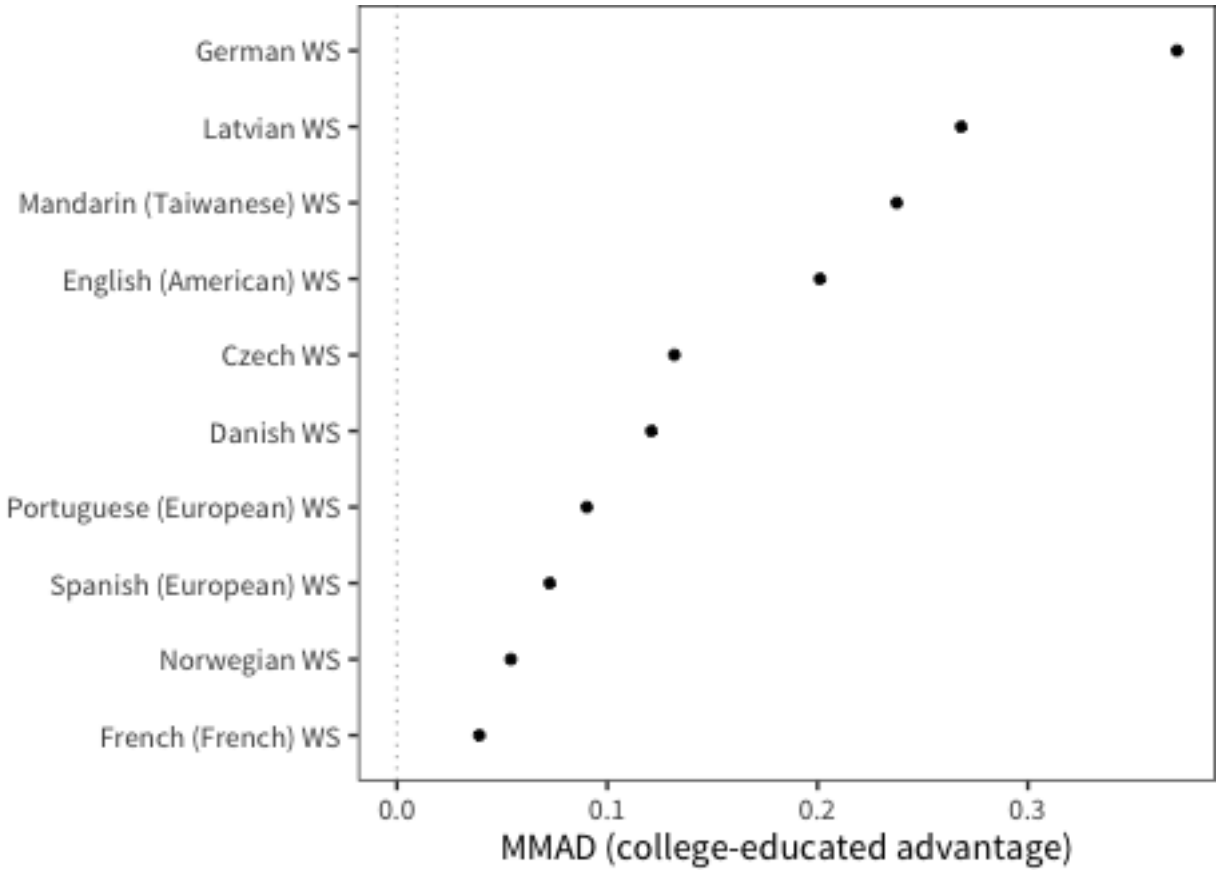


Figure 6.19: MMAD college-educated advantage for WS production data in each language averaged over age.

6.3.3 Explanations

The relationship between maternal education and children's vocabulary is highly variable across countries in our data. The observational nature of our dataset precludes strong inferences about the precise factors that lead to these differences, though speculation is certainly possible. The large magnitude of the differences across countries suggests a powerful role for environmental factors in shaping variation in child outcomes even before the age of 3 years. Findings from the present study support the proposal that inequalities in factors that are linked to child well being should be addressed early in life. Since features of the quantity and quality of caregiver linguistic input has been strongly implicated in early language development, future research should focus on policy interventions that facilitate children's access to high quality talk by caregivers, including care availability and parental leave.

While lower scores on the vocabulary checklists can reflect authentic differences across children from different SES groups, it is also possible that some of these SES effects are the result of differential reporting biases (for discussion, see Fenson et al., 2007). For example, in Feldman et al. (2000)'s study of more than 2000 children, vocabulary comprehension scores on the Words & Gestures form were higher for caregivers with lower education than for caregivers with higher education, whereas, the opposite relation was found for vocabulary production and later grammar skills from the Words & Sentences form. Thus, parents with low educational and income levels seem to overestimate their child's comprehension abilities in comparison with parents with more education and income, especially early in development. Parents, in general, are likely to have difficulty making judgments about early comprehension. Over-reporting may occur for those sub-scales of the CDI that require inference on the part of the parent, for example, when they have to separate evidence of comprehension of a word in isolation from that word taken together with the non-linguistic cues that are likely to support a child's appropriate behavior in a highly constrained context. In addition, some parents may recall times when the words were used in the child's presence and may confuse exposure with understanding. These difficulties may explain why parental reports of verbal comprehension in this age range are higher than results of direct testing (Tomasello and Mervis, 1994) and why correlations between parental reports and comprehension scores from other methods are extremely low (Goldfield and Reznick, 1990).

6.3.4 Discussion

In this chapter, we have observed relationships between vocabulary and sex, birth order, and maternal education. Although all of these have some potential to be influenced by reporting bias, in all cases we see some reasons from prior literature to suspect that similar effects are present (at least in English) for non-parent report measures. More validation is certainly always useful but we see no reason to discount our measurements further.

The relationships we observed were more prominent in every case for production than comprehension. This prominence could be a function of the relatively greater psychometric reliability of production compared with comprehension (see Chapter 4) or it could reflect demographic differences increasing over developmental time, since the production measures we examined are largely from older children than the comprehension. (We did not include early production as we found that the data were noisy and hard to interpret due to children's small production vocabularies for much of the measured range).

We could only speculate about the origins of the relations we found, but the directionality of these relations was similar across languages and they showed reasonable consistency across the full language sample (especially for sex differences). For sex differences, this consistency leads us to speculate about the cognitive origins of verbal ability differences, which are found quite consistently outside of the CDI as well (e.g., Maccoby and Jacklin, 1974). In contrast, birth order and maternal education-related differences were slightly more variable than sex differences and have been argued in previous work to relate to children's input. We return to the interpretation of demographic differences in vocabulary in Chapter 16.

Chapter 7

Gesture and Communication

Children's most recognizable early linguistic accomplishments are surely their first words — a topic we turn to in the Chapter 8. However, even before infants approach this important milestone, they are already communicating with their through another modality: gesture. For example, a child who extends their hands and opens and closes their fist likely wants something. A child who points to a bird up in a tree likely wants to get their caregiver's attention so that they can share in the delight together. Sometimes, children's early vocalizations are accompanied by gestures, for example, a child might raise both of their hands in the air and says "up!" Indeed, the social and communicative routines that these gestures allow children to establish with their caregivers may form the supportive context in which early language learning happens (Bruner, 1985). Gestures thus are an important aspect of children's communicative development.

Early gestures have long been thought to have a common mental status with later-developing linguistic accomplishments because both may reflect the child's understanding of symbols, i.e., that a name or gesture can "stand in" for things in the world. The classic theories of Piaget (1962) and Werner and Kaplan (1963) proposed that all symbols have their origins in actions carried out on objects and moreover, such symbols can be manifested in either the vocal or the gestural domain. These proposals suggest a common underlying mental function that is critical to the development of all symbolic skills, both language and in certain types of gestures.

While the strong representational claims in these theories may be too extreme by modern standards, they do (correctly) predict the developmental continuity between early gesture use and children's later lexical and syntactic development (e.g., Bates et al., 1975; Thal and Bates, 1988). For example, children's ability to point to distant objects is linked to the onset of the production of first words (Fenson et al., 2007; Brooks and Meltzoff, 2008), and children with delayed onset of pointing are likely to also be delayed in first word production (Clark, 1977; Butterworth, 2003). In addition, children's early gesture use is correlated with their later comprehension abilities (Bates et al., 1991), and children's use of gestures in combinations with words is linked to the later production of multi-word combinations (e.g., Goldin-Meadow, 1998; Iverson et al., 1994).

These early correlational findings could simply reflect a common cause: Children who use gestures might also be better better at learning words. More recent studies have demonstrated more specific links between early gesture use and later lexical and syntactic development, however (e.g., Rowe and Goldin-Meadow, 2009). For example, the particular lexical items that enter a child's vocabulary are likely to be names for objects that are labeled using a gesture several months earlier (Iverson

and Goldin-Meadow, 2005). Moreover, early gesture vocabulary is specifically linked to later word vocabulary, whereas early gesture plus word combinations are linked specifically to children's later word combination skills (Rowe and Goldin-Meadow, 2009). Taken together, the pattern of data suggests that children's early gestures provide an important social, communicative, and linguistic foundation for later language development.

Early gestures serve many different functions. Children typically first begin to use "deictic gestures," for example, giving, pointing, or showing (e.g., Volterra and Caselli, 1985). Such deictic gestures are clear precursors to important linguistic and communicative functions, including establishing reference and promoting shared attention (Carpenter et al., 1998). However, these deictic gestures do not necessarily have symbolic content per se (i.e., they do not stand for objects in the world, Bates et al., 1980). Early on, pointing gestures generally may serve an imperative function, e.g., to request something from an adult, whereas, later, pointing is more likely to direct a caregiver's attention to another object or person (Bates et al., 1975; Masur, 1990; Vygotsky, 1980). Children might also use gestures as part of a social activity, for example, waving "bye bye" or signaling "all done."

At first, social gestures might occur simply as imitations, but then later, a child may be able to produce these social gestures spontaneously in certain communicative contexts. Children's social gestures also reflect children's ability to engage in certain activities during pretend play, e.g., talking on a pretend phone or pretending to stir a soup. Such social gestures reflect children's ability to tune into contextual cues, mentally reconstruct activities, and engage in sequences of events. Later, children's gestures might take on a "true" symbolic meaning, as a child might use a conventional gesture to recognize or classify objects as an instance of a category (e.g., pretend to drink from a cup or sniff a flower). Children's ability to use gestures in this symbolic way may reflect a common underlying "vocabulary" in both the verbal and gestural domain (e.g., Acredolo and Goodwyn, 1985; Bates et al., 1980).

This chapter contains analysis of the "early gesture" items from the CDI. Our goals here are to examine (1) the robustness of the measurement properties of these non-verbal parent-report measures, (2) the degree of cross-linguistic consistency and variability of reporting milestones like first pointing, as well as social routines like waving hi and playing peekaboo, (3) the relationships between gestural development and linguistic development, and (4) the relationship between gestural development and two demographic variables: sex and socio-economic status.

7.1 Measurement properties of CDI gestures

7.1.1 Measuring the development of gesture

Unlike the word items on the CDI , which typically ask parents to make a binary decision about whether a word is in their child's vocabulary (although comprehension and production are separate decisions), the First Gestures on CDI forms ask parents to make a 3-way decision, determining if their child produces a given gesture "often", "sometimes", or "not yet." We begin by asking whether parents responses are sensitive to this distinction, as the choice of whether to treat all three levels as meaningful impacts downstream analytic decisions. We perform this sensitivity analysis on the American English CDI as it is the inventory for which we have the best a priori intuitions.

Figure 7.1 shows the proportion of American English learning children who give each of these responses. If each of the three responses is meaningfully different, the developmental trajectory of

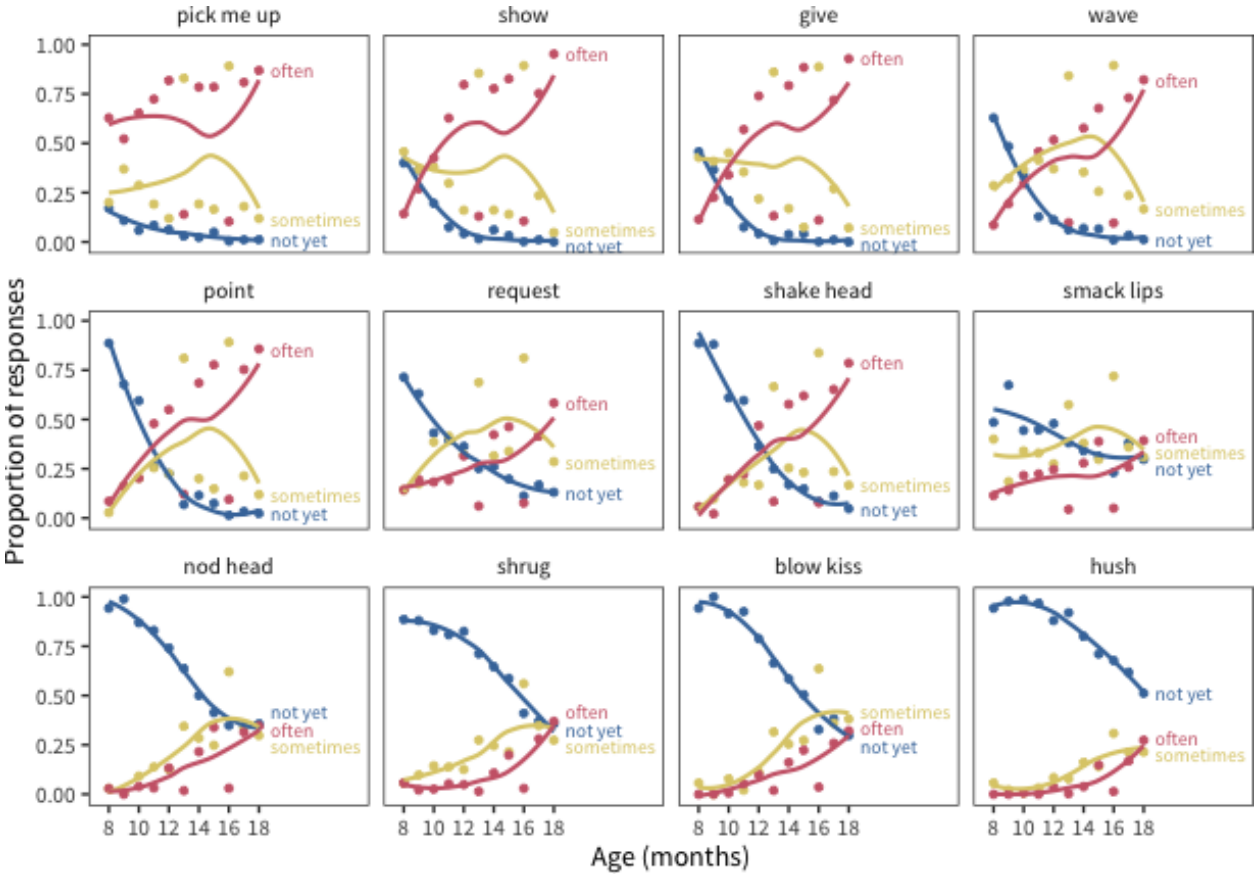


Figure 7.1: Proportion of each response type over age on each First Gestures item for American English.

each should be distinct and predictable. The proportion of children whose parents indicate that they do "not yet" produce each gesture declines predictably over development. However, the other two responses — "sometimes" and "often" — do not appear to have reliably different trajectories. Perhaps they are used differently by different parents or in different samples.

For comparison, we collapse the "sometimes" and "often" into a single value, and plot the proportion of children at each age whose parents report that they produce each gesture (Figure 7.2). The trajectories look generally smooth and prima facie reasonable, with the potential exception of the smack lips gesture for which there is very little developmental change (this gesture, which corresponds to the vocalization yum yum, may be unusual or less stereotyped).

While these gestures are categorized on the CDI as "first gestures," the form also asks parents about a variety of other kinds of gestures that children produce, including those involved in games and pretend play. Do these gestures have similar trajectories? Figure 7.3 plots developmental trajectories for these other categories of gesture.

While some are clearly learned later than the early gestures, a number of these appear to be learned quite early as well — for example, peekaboo and pretend play with cups and spoons. They all also appear to have generally smooth and increasing trajectories with the exception of so big (from Gestures Games) which, like smack lips from the First Gestures section appears to be either less stereotyped, more difficult to identify, or more variable across children.

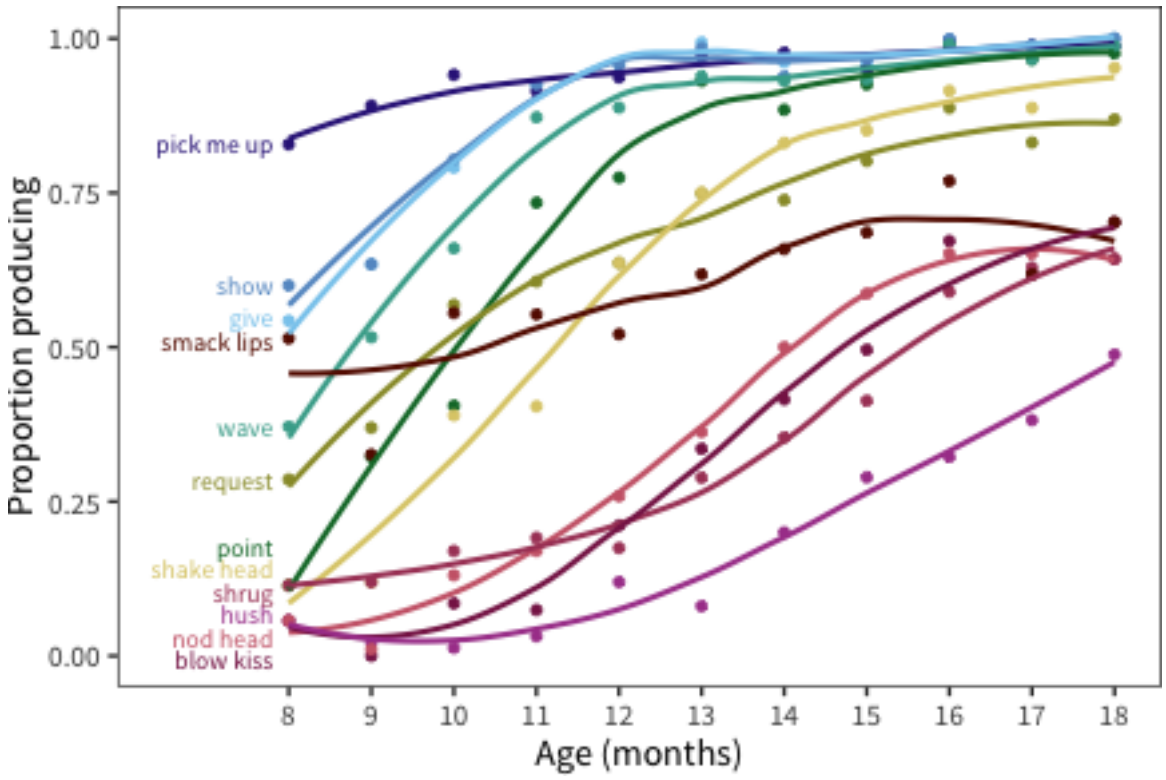


Figure 7.2: Trajectory over age of each First Gestures item for American English.

Taken as a whole, it is clear that almost all of the gesture items have developmental trajectories not unlike word items, and that they thus have the potential for informative analyses. Further, trajectories look qualitatively similar across categories. Consequently, for general cross-linguistic analyses, we will consider all of the gestures together, and compress "often" and "sometimes" into a single affirmative choice.

To estimate the coherence of these categories, we compute age of acquisition estimates for each of the American English gestures by gesture type: First Gestures (e.g. pick me up, point), Game Gestures (e.g. play peekaboo, chase), Object Gestures (e.g. brush teeth, push car), Adult gestures (e.g. type, use pen), and Parent Gestures (e.g. sweep, feed from a spoon). These estimates were produced by fitting a robust linear regression to the proportion of children who produce each gesture and estimating the age at which 50% of children produce the given gesture. The resulting ages of acquisition for each gesture type are shown in Figure 7.4. These categories vary in coherence, but overall First Gestures and Games tend to be produced early, and Adult and Parent gestures — more representative of pretend play — are produced relatively later. The object gestures vary substantially in their ages of acquisition.

7.1.2 Consistency of the first gestures

While the First Gestures are not universally learned before the other gestures measured, they are among the earliest learned. Because of the particular theoretical importance of these early communicative gestures (i.e. deictics like pointing and showing, showing routines like pick me up), we analyze the cross-linguistic consistency of these at the item level.

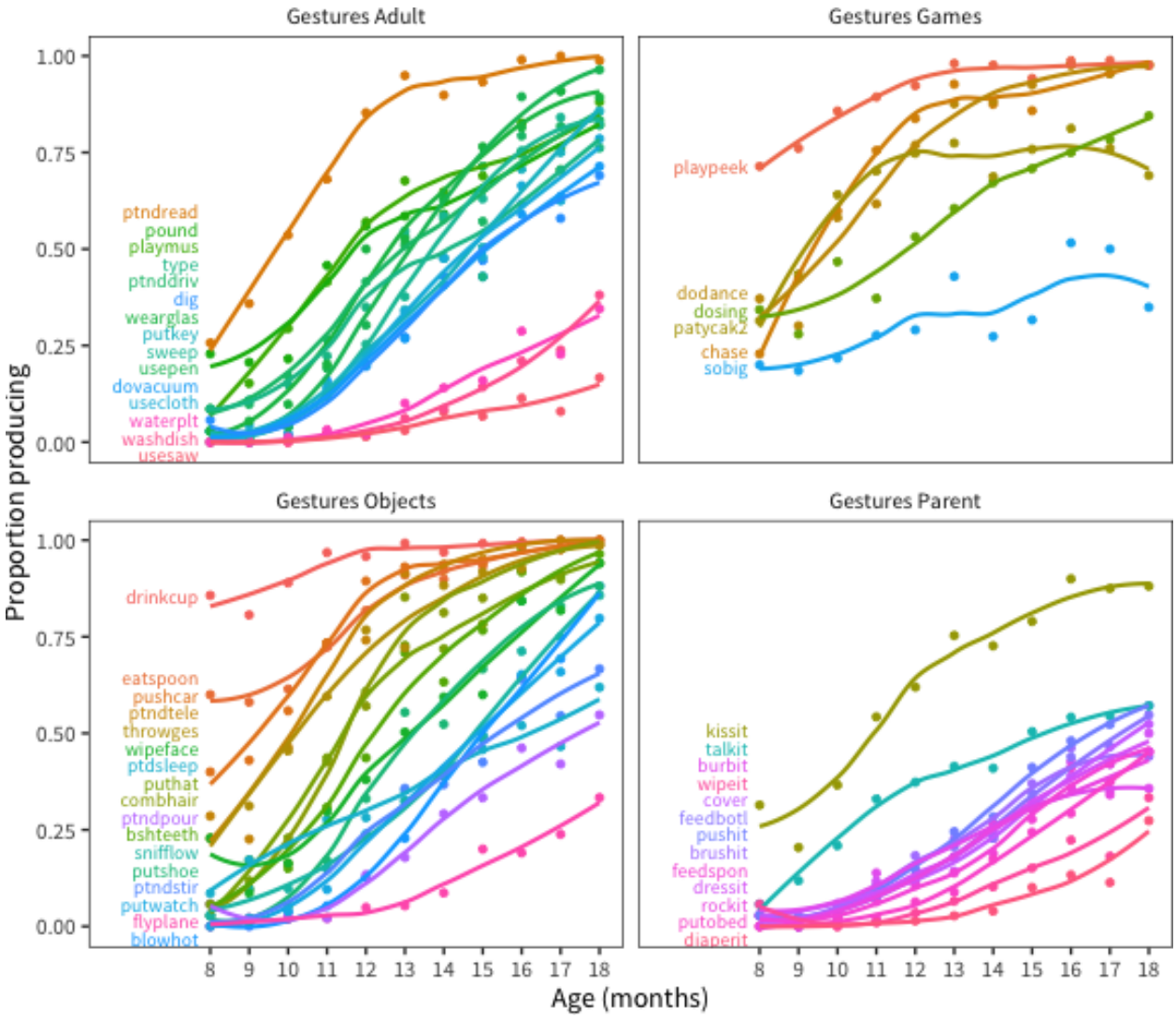


Figure 7.3: Trajectory over age of each gestures item by type for American English.

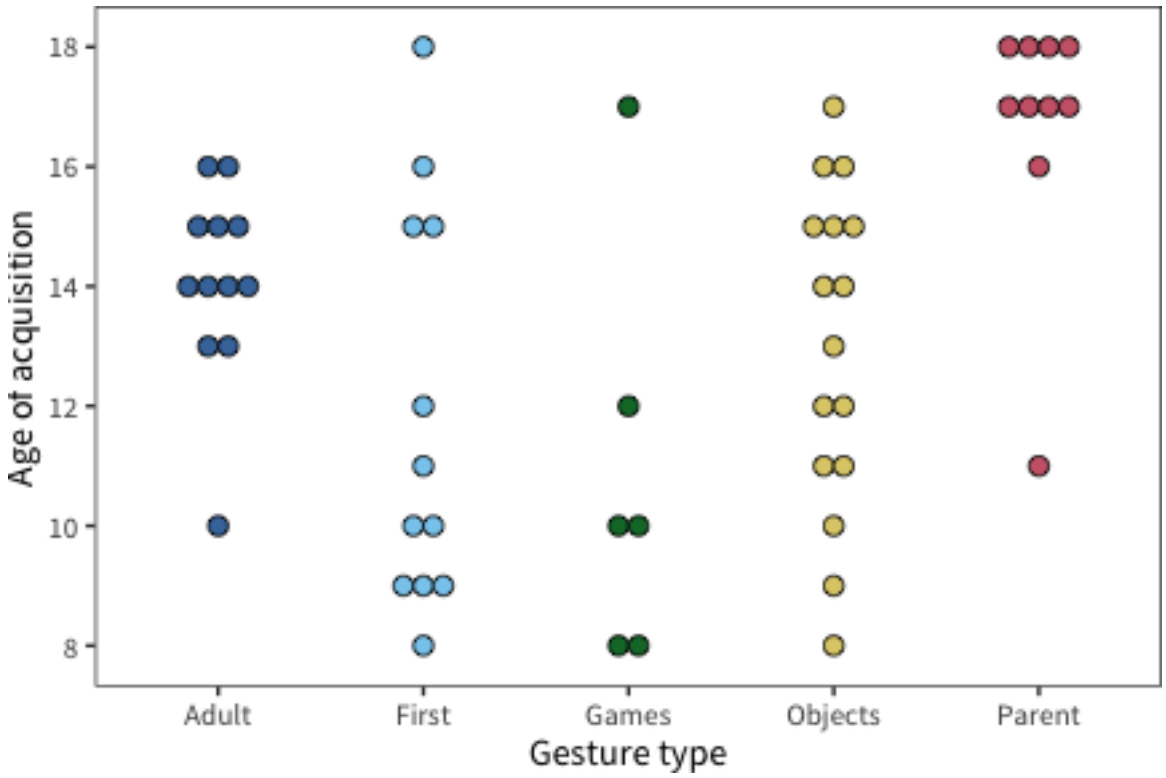


Figure 7.4: Age of acquisition for each gestures item by type for American English.

Table 7.1 and Figure 7.5 show both consistency and variability across items. As in the learning of words, the means and variances of these ages of acquisition were correlated (r = 0.46; Mollica and Piantadosi, 2017). The primary outliers were request, which appears to be produced surprisingly late in American English, and shrug which was produced surprisingly late by French-learning infants. In general, however, most of the cross-linguistic differences appear to be consistent across the gestures (i.e. French-learning infants gestures later). It is difficult to tell from this small sample of mostly European languages whether these differences are driven by linguistic factors or rather by properties of our samples or variability in parents' interpretation of the form. Nonetheless, they provide some evidence for consistency in the process of gestural development cross-linguistically. To get additional leverage on this process, we next consider the full set of gestures.

7.1.3 Intercorrelation among gestures

Given both the similarity and the variability in the developmental trajectories of different gestures, as well as the cross-linguistic variability in first gestures, a natural next step is to quantify the relationship of gestures to each-other. We begin by computing the average intercorrelation between each of these gesture categories. In this analysis, we take gestures in pairs (e.g. Adult Gestures and First Gestures) and ask how the proportion of items that kids know in one predict the proportion of items they know in the other. For American English learning children, the proportion of gestures they know across categories is correlated at r = 0.6000678 — nearly identical to the value of ˜.6 reported by Fenson et al. (1994). For comparison, the same intercorrelation computed across categories of words (e.g. "animals" and "places") yield 0.643456 for production and 0.5657221 for comprehension.

Table 7.1: Summary statistics for ages of acquistion for each First Gestures item across languages.

Item	Mean	SD	N
pick me up	8.0	1.7	7
give	8.9	1.5	7
show	8.9	1.2	9
point	9.1	2.2	8
request	9.8	4.2	8
shake head	9.9	1.9	8
wave	9.9	1.9	8
smack lips	10.9	3.9	7
nod head	14.5	3.8	8
hush	15.2	3.0	7
shrug	15.7	3.8	7
blow kiss	16.0	1.9	6

Table 7.2: Summary statistics for intercorrelations between gesture categories for each language.

Language	Mean	SD
English (American)	0.60	0.113
French (French)	0.56	0.178
Hebrew	0.60	0.141
Italian	0.68	0.104
Korean	0.84	0.129
Norwegian	0.57	0.110
Slovak	0.69	0.080
Spanish (Mexican)	0.67	0.087

This cross-category intercorrelation is quite similar cross-linguistically, ranging from 0.56 in French (French) to 0.84 in Korean. The full set of intercorrelations is shown in Table 7.2.

7.2 The relationship between language and gesture

A critical theoretical question in early communicative development concerns the relationship between language and gesture. As alluded to above, a number of early influential theories (e.g., Piaget, 1962; Werner and Kaplan, 1963) held that gesture and language should be intimately related because of their reliance on a shared system of symbolic reasoning. To the extent that they are underpinned by the same system, words and gestures should have related developmental trajectories–children who gesture early should also speak early and vice versa (Bates et al., 1991). Following in the footsteps of Fenson et al. (2007), we ask this question at larger scale, and cross-linguistically. To assess this relationship, we will look at the correlations between children's gestural and linguistic vocabularies.

To first provide a baseline, however, we compute the correlation between children's language and gesture development and their age. As Table 7.3 below shows, gesture shows as much or more development than comprehension and production over the ages measured by the CDI Words & Gestures forms, and the variability in the correlation with age in all three measures hangs together within-language: Languages where there is more developmental change in linguistic development

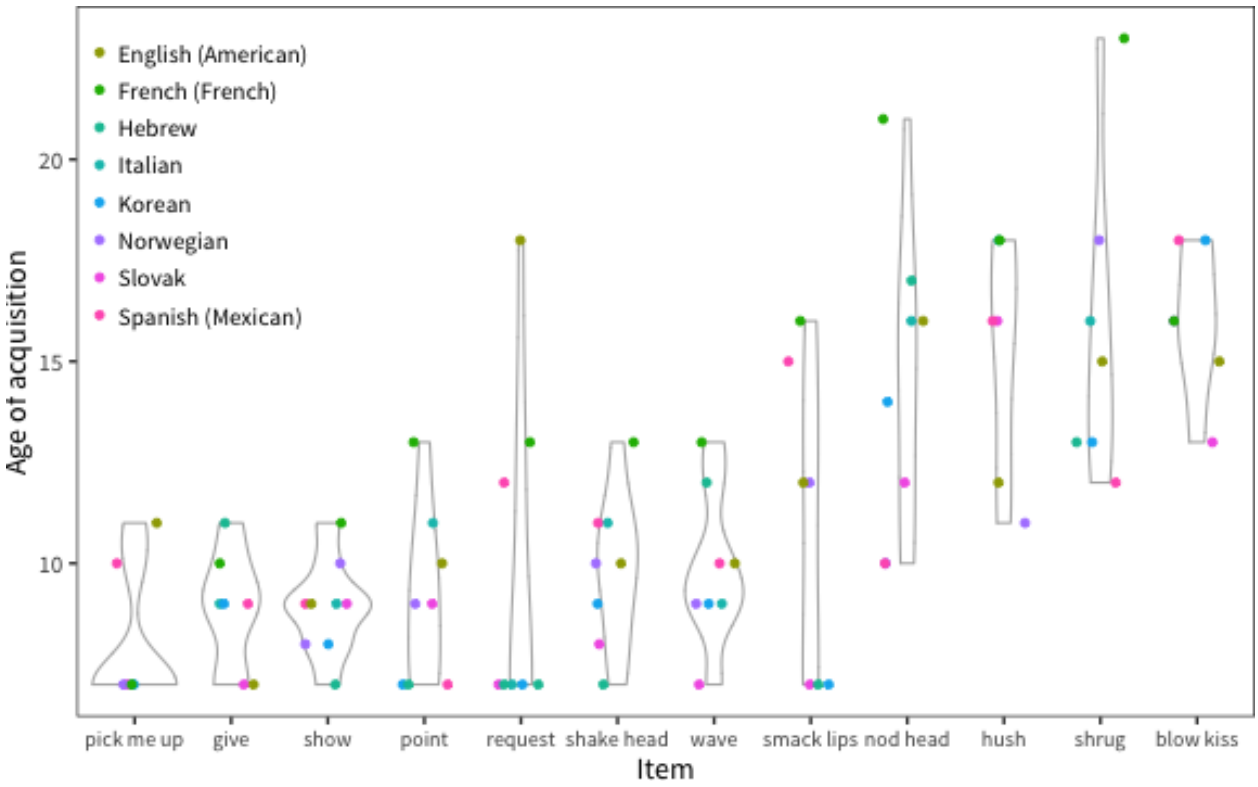


Figure 7.5: Distribution of ages of acquistion for each First Gestures item for each language.

also tend to have more gestural development.

However, as we noted in Chapter 4, comprehension and production do not proceed in lock-step and comprehension generally outpaces production. This is, in part, because production requires additional control over the developing motor systems necessary for speech. To the extent that gesture and language are related by their shared reliance on symbolic understanding, their correlation should be highest when only this shared system is tapped. In this case, we should predict that gesture production and language comprehension are more tightly correlated than gesture production and language production. In contrast, if the correlation is due primarily to a shared desire to communicate and engage socially with caregivers, we should predict a stronger correlation between gesture production and language production. Across these 8 language, children's production of gestures is consistently more highly correlated with their comprehension (Table 7.4).

7.3 Conclusions

Gestures appear early in young children's communicative repertoires, with an understanding of the deictic purpose of pointing being found as early as 12-months of age (Liszkowski et al., 2007). The frequency of gesture in children's input - as well as the precociousness of their own productions - are also reliably correlated with their linguistic development (Rowe and Goldin-Meadow, 2009; Brooks and Meltzoff, 2008). In this Chapter, we extended the analyses developed in Fenson et al. (2007) to look at these relationships cross-linguistically using large-scale CDI data.

Table 7.3: Correlation with age for each subscale in each language.

Language	Gesture	Comprehension	Production
English (American)	0.73	0.60	0.46
French (French)	0.56	0.31	0.29
Hebrew	0.66	0.67	0.65
Italian	0.81	0.76	0.61
Korean	0.54	0.61	0.44
Norwegian	0.70	0.72	0.55
Slovak	0.66	0.66	0.50
Spanish (Mexican)	0.71	0.55	0.36

Table 7.4: Correlation with gesture subscale for each vocabulary subscale in each language.

Language	Comprehension	Production
English (American)	0.75	0.53
French (French)	0.67	0.37
Hebrew	0.78	0.65
Italian	0.81	0.49
Korean	0.51	0.45
Norwegian	0.69	0.50
Slovak	0.63	0.54
Spanish (Mexican)	0.76	0.55

In the first parts of the chapter, we evaluated the measurement properties of the gesture items, as well as the coherence of the groupings of items on the CDI forms. Our analyses confirm that parents are likely able to report reliably on their children's gesture development. We also found a high degree of consistency in development of different gestures within child–children who learn some gestures early are also likely to learn other gestures early, and a high-degree of cross-linguistic consistency in the age of acquisition of different gestures.

Finally, we examine the relationship between gesture and language directly. We found that individual differences in gesture development were highly correlated with language development, and that gesture was more related to comprehension than production. We take these results to mean that precociousness in the gestural domain may be related to children's developing understanding of symbolic communication systems, rather than a strong desire to communicate with their caregivers.

Chapter 8

Consistency in Early Vocabulary

Which words do children learn first? In spite of tremendous individual variation in rate of development (see Chapter 5; Fenson et al., 1994; Hart and Risley, 1995), the first words that children utter are reported to be quite consistent (Tardif et al., 2008). In particular, in English, Mandarin, and Cantonese, babies' first 10 words tended to be about important people in their life (mom, dad), social routines (hi, uh oh), animals (dog, duck), and foods (milk, banana).This chapter attempts to generalize the analysis reported in Tardif et al. (1999), asking more broadly whether words tend to be learned in the same order across languages.

The trouble is that the precise words that children learn in different languages are (of course) language-specific. Thus, we really want to know whether the concepts that are being talked about are the same — or at least similar. As detailed in the Chapter 2, the items on each language's form are adaptations and not translations: They are intended to capture the spirit of the items on the English form rather to replicate them exactly. So conceptual mappings are approximate, rather than exact. Nonetheless, when these approximate translation equivalents appear on multiple forms we can look at the variability in how quickly they are acquired across languages. Thus, we assume for simplicity that dog, chien, and perro name (roughly) the same concept.

To estimate the similarity of each item's trajectory, we use a single measure of its difficulty: age of acquisition (AoA) — the age at which 50% of children in each language are estimated to have acquired it (Appendix D). We analyzed consistency in both comprehension and production, with production ages of acquisition estimated by stitching across both forms. Consequently, we analyzed only the 29 languages for which data for both forms was available.

In total, we estimated ages of acquisition for 945 total words spread across the 29 languages. Unfortunately, not every word appeared on all forms. Figure 8.1 shows the cumulative proportion of forms on which every word appears. For our consistency analysis, we considered only the 335 words that appeared in at least 8 of the 15 languages.

8.1 The first 10 words

Following Tardif et al. (2008), we begin by examining the first 10 words acquired by children across the 15 languages we measured (Tables 8.1 and 8.1). Similar words appeared in the top 10 across languages, especially in children's first productions. These words consist primarily of important

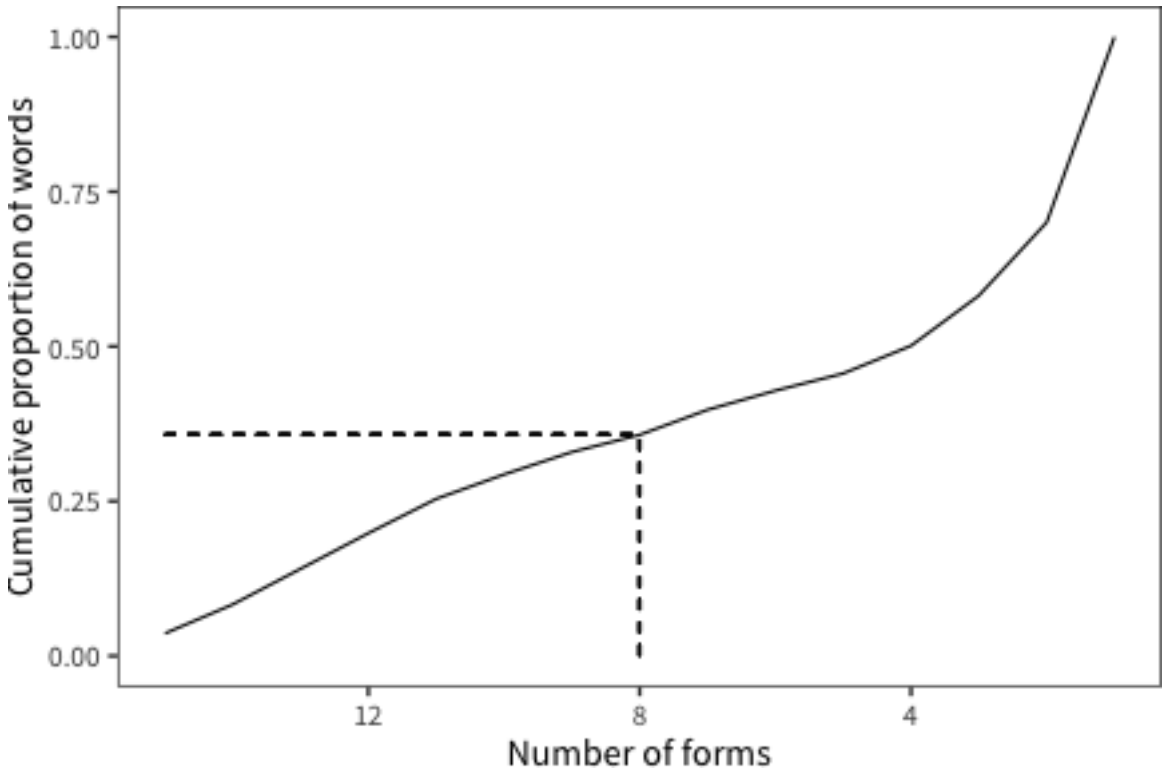


Figure 8.1: The proportion of words found on a given number of language's CDI forms. The dashed line shows the cutoff value we chose.

family members (mommy, daddy, grandma), social routines (hi, bye, peekaboo), and sounds (yum yum, vroom, woof woof).

The 10 earliest words that children produce in each language.

Croatian	Danish	English (American)	French (French)	French (Quebecois)	Hebrew	Italian	Kiswahili	Korean	Norwegian
mommy	hi	mommy	daddy	mommy	mommy	mommy	mommy	mommy	vroom
daddy	woof woof	daddy	mommy	daddy	yum yum	daddy	daddy	daddy	mommy
grandma	thank you	ball	baby	no	grandma	woof woof	car	peekaboo	yum yum
bye	mommy	bye	bye	bye	vroom	grandma	cat	woof woof	hi
woof woof	no	hi	thank you	baby	grandpa	water (beverage)	meow	cracker	daddy
baby	bye	no	bread	ball	daddy	hi	motorcycle	water (beverage)	bye
no	daddy	dog	peekaboo	vroom	banana	grandpa	baby	baby	thank you
yes	vroom	baby	ball	sock	this	meow	bug	yes	woof woof
grandpa	yes	woof woof	sock	peekaboo	bye	no	banana	ball	yes
aunt	food	banana	shoe	moo	car	shoe	baa baa	no	peekaboo

The 10 earliest words that children comprehend in each language.

Croatian	Danish	English (American)	French (French)	French (Quebecois)	Hebrew	Italian	Kiswahili	Korean	Norwegian
grandma	daddy	bottle	no	milk	yum yum	mommy	baa baa	food	child's own name
mommy	child's own name	daddy	mommy	mommy	cat	daddy	meow	daddy	mommy
bye	mommy	mommy	daddy	daddy	doll	peekaboo	car	mommy	daddy
daddy	peekaboo	child's own name	peekaboo	child's own name	balloon	child's own name	bug	peekaboo	bye
peekaboo	yum yum	bye	bye	bye	ball	hi	cat	no	hi
vroom	no	no	bath	peekaboo	head	woof woof	doll	dirty	peekaboo
grandpa	bye	peekaboo	hi	no	light (object)	dog	ball	bath	no
cat	hi	hi	good night	bathtub	daddy	water (beverage)	milk	ball	yum yum
child's own name	woof woof	dog	yes	bath	mommy	bottle	medicine	cracker	good night
woof woof	ball	ball	meow	ball	grandpa	grandma	spoon	water (beverage)	grandma

Unfortunately, we cannot determine if the greater consistency found in early production is a real regularity about children's lexical development, or is instead a measurement artifact arising from the greater difficulty of reporting on a child's comprehension (see Chapter 4).

8.2 Global cross-linguistic similarity

Despite these differences between comprehension and production, words reported to be acquired early in one are also generally reported to be acquired early in the other. The Figure below shows the relationship between the mean age of acquisition in production and the mean age of acquisition of the each of these 335 words across the 15 languages (Figure 8.2). The correlation between the two measures was quite high: $r = 0.8$ ($p < 0.001$). Taken together, these analyses suggest that the rate at which the words on the CDI, and by inference the processes that underpin them, are highly similar across languages.

The source of this similarity is hard to pin down. One possibility is that the difficulty of learning a word is determined predominantly by the complexity of the concept denoted by that word, and thus that variability in linguistic and cultural features play a relatively small role in determining the difficulty of learning a word (Gentner and Boroditsky, 2001). Alternatively, the primary driver of difficulty could be linguistic, but the dimensions of linguistic variability could be orthogonal to the difficulty of learning. For instance, verbs may be more difficult that nouns because they are relational, and thus that learning nouns makes learning verbs relatively easier than learning verbs makes learning nouns (Gleitman, 1990). In this case, the linguistically relevant dimensions would be relatively invariant across languages (Snedeker et al., 2007). (Finally, because the words on the CDI are not a random sample of words in each language, it could be that generalization from these analyses overestimates the degree of cross-linguistic similarity.)

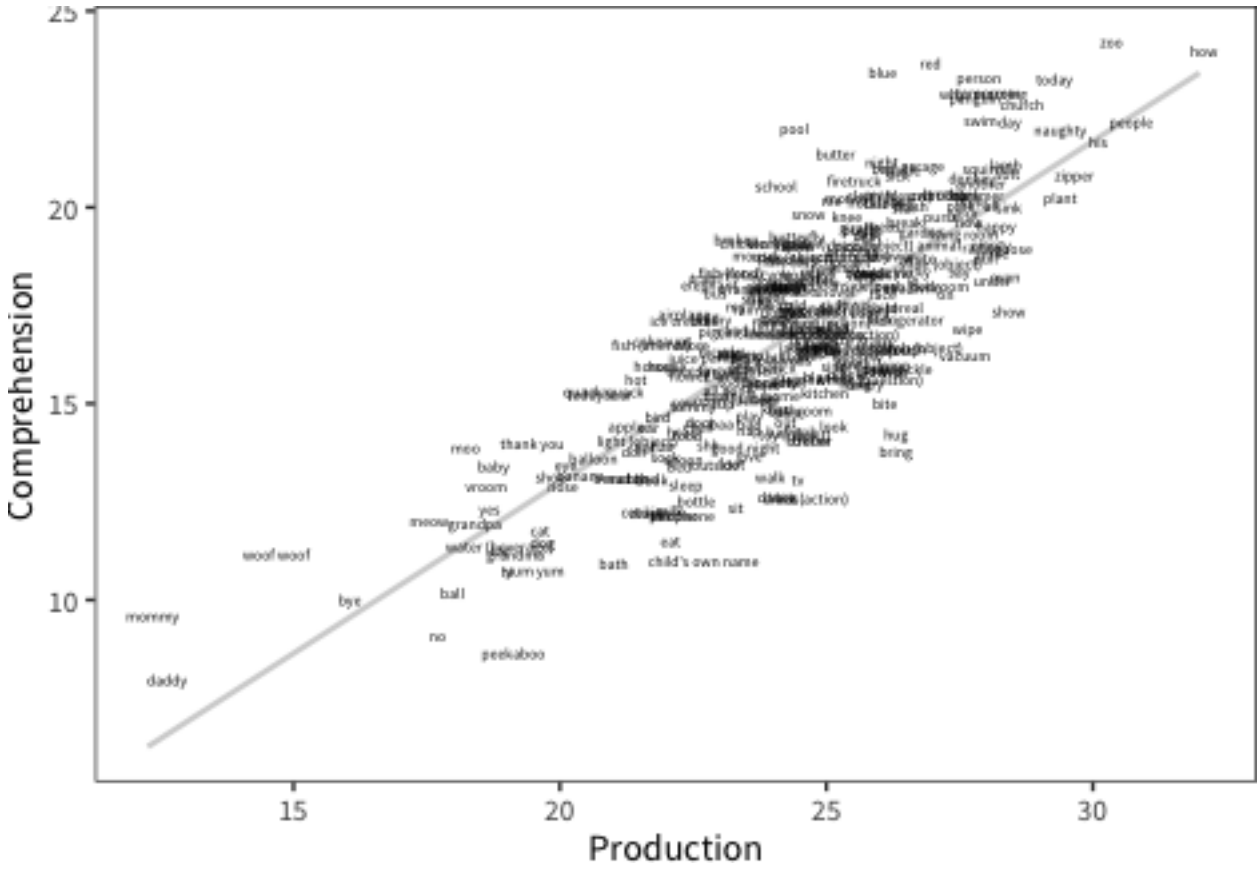


Figure 8.2: Correlation between the average age of acquisition in comprehhension and production for each measured word.

In Chapter 10 we begin to take up these questions using predictive models. Prior to taking this step, however we consider cross-linguistic ordering more holistically. In the remainder of the chapter, we address this problem from two directions: (1) Does similarity in order of acquisition vary with language-to-language similarity, and (2) Does similarity in order of acquisition change over development?

8.3 Acquisition similarity and linguistic similarity

Unfortunately, the 15 languages in our analyses are both a small and non-representative sample of the world's languages, and thus do not have sufficient power to detect typological features of language that might be responsible for differences in the similarity of acquisition across languages. Nonetheless, the languages do come from different languages families, and do vary in their phylogenetic distance. We leverage this variability to ask whether the similarity between two languages is related to similarity in how quickly words for the same concepts are learned in those two languages.

Instead of correlating the average similarity of age of acquisition across all languages, we consider the pairwise similarities in the age of acquisition of each of the 335 words in each language. Figure 8.3 shows these pairwise correlations for both comprehension and production as matrices in which each cell shows a single pairwise correlation. These correlation matrices appear to contain a significant amount of structure, with languages that are from the same language family (e.g. Norwegian and Danish) showing higher correlations between the ages of acquisition for the same concepts. Perhaps unsurprisingly from the high average correlation between production and comprehension, pairwise correlations were nearly identical for production and comprehension ($r = 0.98$, $p < 0.001$). Figures 8.4 and 8.5 respectively show dendrograms produced by hierarchically clustering these pairwise correlations.

These dendrograms show high similarity within the North Germanic, Slavic, and Romance language families. Some relationships resist straightforward linguistic explanations (e.g. the relationship of Quebec French to other languages). These may non-uniform sparsity of data across these languages, or may instead reflect interesting cultural or other source of variability. Despite these cases, the structure of ages of acquisition appear to a high degree to reflect the structure of the languages that children learning these words speak. To confirm this observation quantitatively, we borrowed an established measure for measuring linguistic similarity: the lexical similarity of words for the same meaning (Wichmann et al., 2010). Using a set of 40 words for meanings common to all of the words languages, Holman et al. (2008) were able to use a string-edit distance measure recover linguistic similarity measures that correlated highly with geographic distance and also several typological systems. This method is appealing for our purposes as it is relatively agnostic as to the processes of language contact and change that have produced modern-day languages and instead tracks the similarity of wordforms themselves. The language distance measures produced by this method were highly correlated with pairwise correlations in acquisition trajectories for both production ($r = -0.44$, $p < 0.001$) and comprehension ($r = -0.41$, $p < 0.001$).

We also applied this same analysis to the words on the words on the CDI themselves. For each language, we compute the average normalized Levenshtein (1966) distance between the word for each of the 335 common words in our analyses. Levenshtein distance is a measure of the minimum number of insertions, deletions, or substitutions required to transform one string into another. For instance, the distance between the Italian and Norwegian words for dog (cane and hund) is 3. We computed this measure pairwise for all words, and then divided it by the number of characters in the longest

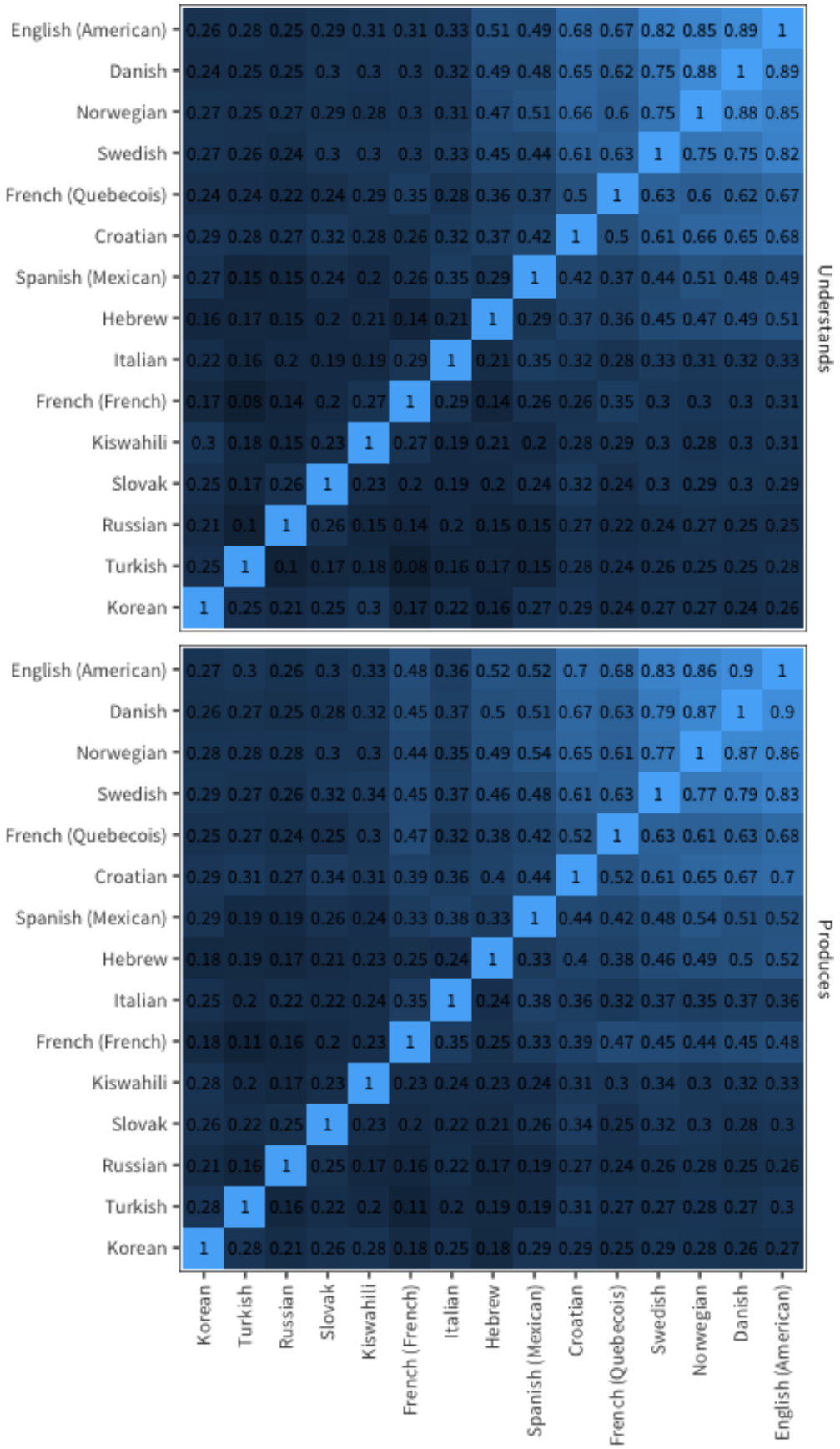


Figure 8.3: Correlation matrics showing pairwise correlations in words' age of acquisition. Languages that are more similar have more similar acquisition orders.

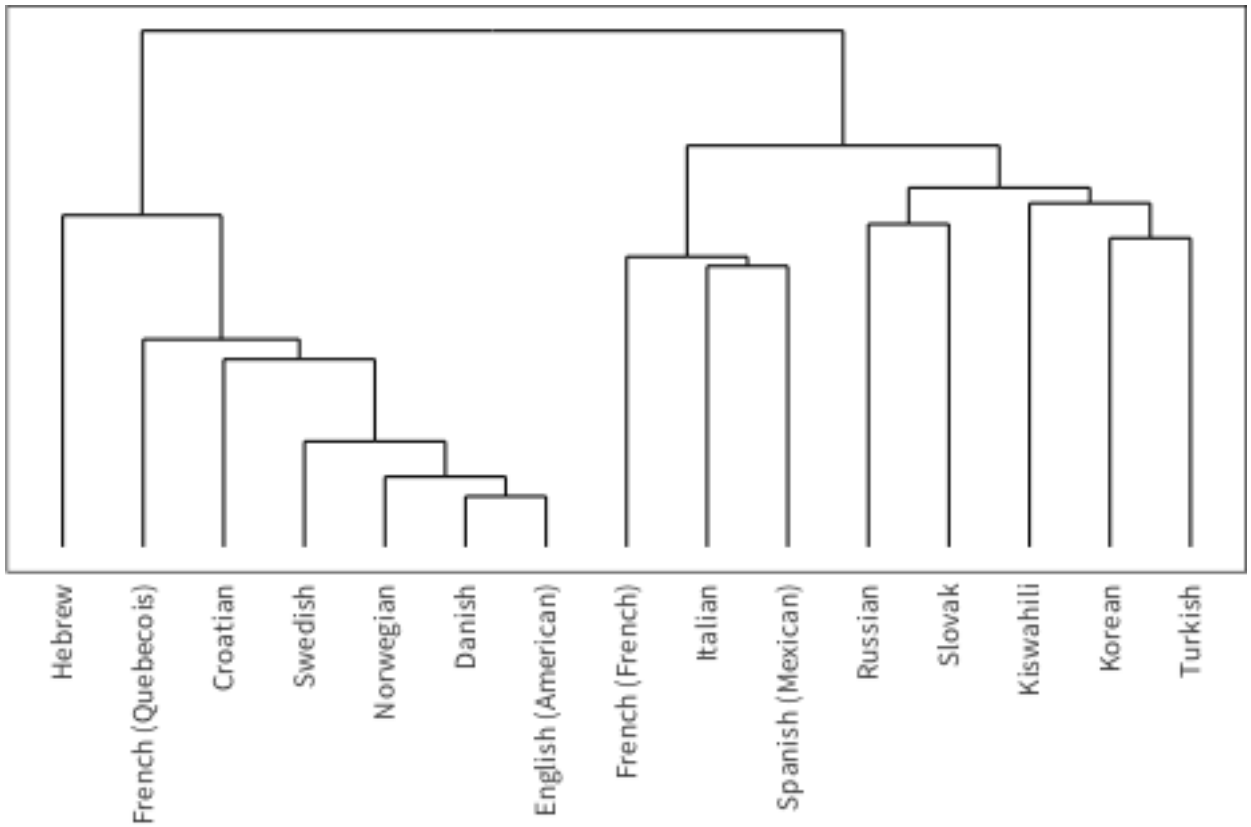


Figure 8.4: A hierarchical clustering of the similarity in the ages of words' first production cross-linguistically.

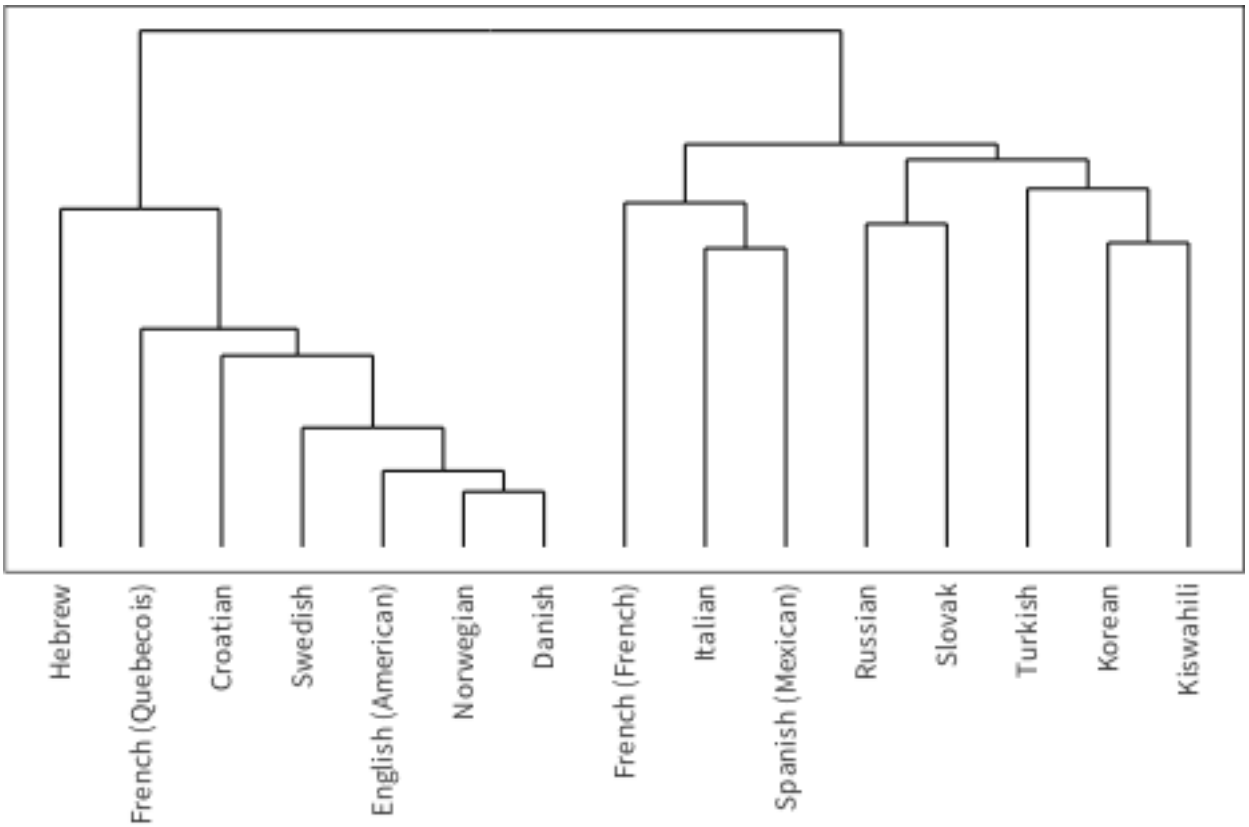


Figure 8.5: A hierarchical clustering of the similarity in the ages of words' first comprehension cross-linguistically.

word in order to get the edit distance per character (0.75 for cane and hund). This measure was even more correlated with pairwise acquisition trajectories as similarity computed using the 40 words identified by Holman et al. (2008) for both production (r = -0.57, $p < 0.001$) and comprehension (r = -0.52, $p < 0.001$).

Because this analysis likely overestimates the dissimilarity of langues written in different scripts–as every word receives a normalized Levenshtein distance of 1 — we replicated this analysis at the phonemic level. We used eSpeak to compute phonetic transcripts of each word and repeated the same analysis on distance between words' phonetic units in the International Phonetic Alphabet (IPA; Decker et al., 1999). These correlations between IPA distance and pairwise age of acquisition trajectories were again reliable although slightly attenuated for both production (r = -0.29, p = 0.009) and comprehension (r = -0.26, p = 0.020). The robustness of these correlations across a variety of methods suggests that in addition to the high degree of general cross-linguistic similarities in the order of acquisition of words, the dissimilarities between them likely likely reflect reflect differences in the target languages being learned.

In our final analysis, we ask whether similarities in ages of acquisition are constant over the course of acquisition, or whether the similarity across languages changes over development. If variability in acquisition trajectories across languages reflects variability in those languages, we might expect that children's trajectories diverge over the course of language acquisition as the structure of their target language or their cultural milieu play a stronger role in guiding which words are easy or important to learn. Our analyses of the first 10 words above shows striking similarity in the earliest words. Does this similarity decrease for the next 300 words?

8.4 Consistency across development

In order to measure change in cross-linguistic consistency over development, we extend the aoa-correlation approach we have used throughout this chapter. For each concept that appeared in at least 8 languages, we computed its average age of acquisition across all languages in whose CDIs it appeared in both comprehension and production. We then ordered these words from the earliest learned word on average (mommy to the latest learned word how. For each measure, we then computed the average cross-linguistic in ages of acquisition for each increasingly large set of words starting with 5 words to 335 words. If the correlation increases over acquisition, we can infer that acquisition trajectories become more similar as more words are learned, and thus that the hardest to learn words are learned in more similarly. In contrast, if the correlation decreases, we can infer that there is children start out learning similar concepts regardless of their native language, but that linguistic and cultural variability plays a greater role in the learning of more complex words.

Figure 8.6 shows these correlations for both comprehension and production over the course of acquisition. In addition, the gray shaded region shows a 95% confidence interval for a random baseline in which the concepts were ordered randomly rather than in average acqusition order. This baseline is important to control for changes in measurement error that arise from changing numbers of concepts in the correlation. For both comprehension and production, the trajectories are reliably above the shuffled baseline, and decrease over the course of acquisition. These results suggest that, indeed, there is significantly more similarity in the earliest learned words than in later learned words cross-linguistically, especially in production. This is exactly the pattern of results we would predict if language and culture produce more variability in the forms, frequencies, and contexts of use of later learned words.

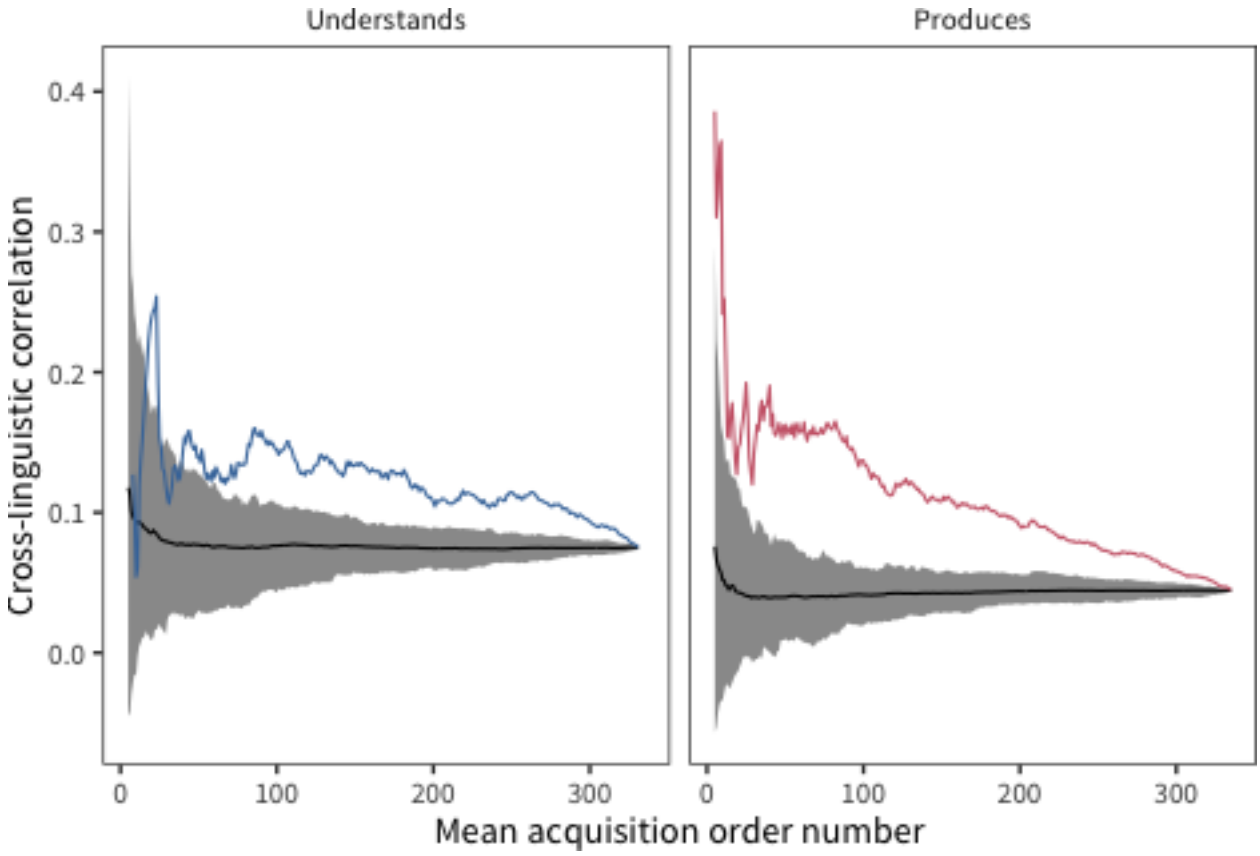


Figure 8.6: Cross-linguistic correlation ages of words' acquisition over the course of language development. Colored lines show empirical correlations, the gray area shows a 95 percent confidence interval for a randomly shuffled baseline. Especially in production, cross-linguistic similarity declines over the course of language development.

8.5 Conclusions

Children in all languages and culture learn language, but the languages they learn vary, and the cultures into which they are born may have quite different cultural practices around both language and cognitive development. Nonetheless, the order in which children learn the word for different concepts shows a great degree of cross-linguistic similarity, and dissimilarities are well explained by measurable linguistic dissimilarity.

In addition, we found that the degree of cross-linguistic similarity decreases over the course of acquisition. While the first ten words acquired in each language were highly consistent, the last ten words were quite different. We take these results to indicate a shared core of concepts–e.g., social routines, important people, and some early foods and household animals–that are perhaps especially communicatively important independent of their linguistic realization. As acquisitio unfolds, however, the features that make languages (and cultures) different from each other play an ever increasing role in driving acquisition. In Chapter 9, we explore demographic differences in acquisition that help to explain why two children learning the same language may acquire different words at different rates.

Chapter 9

Demographic Variation in Individual Words[1]

In Chapter 6, we documented demographic differences in total vocabulary. But where do these differences reside? Concretely, if girls say more words than boys, which words do they say? Is it the case that they simply produce each word more with some probability, or are there individual words that are more likely to be produced? Or are both true? In this chapter, we consider the possibility that individual words carry this demographic signal. We assess which words are learned differentially earlier or later by girls vs. boys, by first-born vs. later-born children, and by children with different levels of maternal education.

9.1 Methods

9.1.1 Data

Various subsets of the datasets in Wordbank are coded for one or more demographic variables. Here we examine the child's birth order, level of maternal education, and assigned sex at birth. For these analyses we extract all of the instruments with demographically coded data and combine them into two datasets: comprehension from WG forms, and production from both WG and WS forms. (We use the "by item stitching" approach described in Appendix C).

This approach creates six different analyses, one for each combination of demographic variable and measure. We exclude a language from a given analysis if it has fewer than 50 children for that demographic variable and measure. The demographic variables are coded into the values First / Second / Third+ for birth order, Below Secondary / Secondary / College and Above for maternal education, and Female / Male for sex.

Each dataset yields a trajectory for each word, created by smoothing the number of children that are are reported to understand or produce the word over age. These trajectories can be computed separately for each value of the demographic variable. For example, in Figure 9.1, these are the trajectories for some sample items in English for production data split by birth order. Note that the word brother is spoken much later by first-born children than by later-born children, whereas green

[1] An earlier version of the gender analyses below was presented to the Boston University Conference on Language Development in 2016

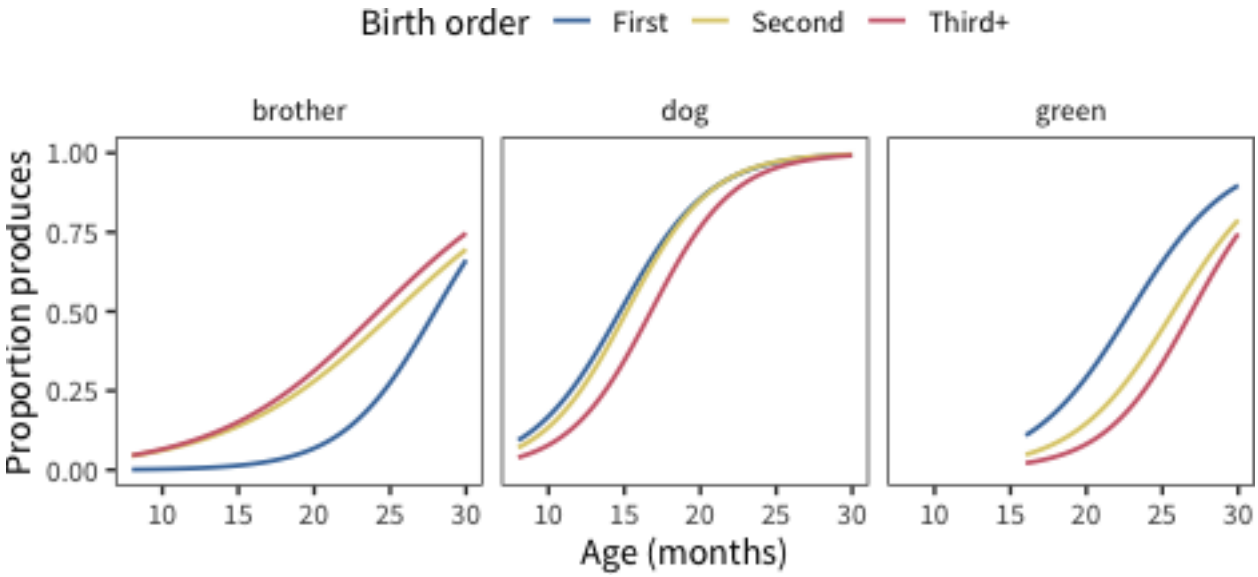


Figure 9.1: Developmental trajectories for three example American English words by birth order.

is spoken much later by later-born children. Averaging all of these trajectories together reproduces the general demographic achievement curves reported in Chapter 6.

The goal of the analyses is to quantify the overall effect of each demographic variable, i.e. the differences among the above curves, and the individual contribution of each item to that effect.

9.1.2 Models

There are a number of complementary methods to estimate individual item effects. In Chapter 6, we explored a simple, non-parametric approach to estimating demographic effects across groups. Here we are interested in estimating these effects for individual items, and thus data are sparser for each individual item. Thus, it is more effective to use a multi-level, model-based analysis in which demographic effects are estimated both at the level of all items and specifically for individual items.

In particular, we use a mixed-effects logistic regression to predict how many children produce/understand items from their age and their level for a given demographic variable, with a random effect for item. A model of this type is fit separately for the data for each language and measure. For example, the model for birth order would be specified as:

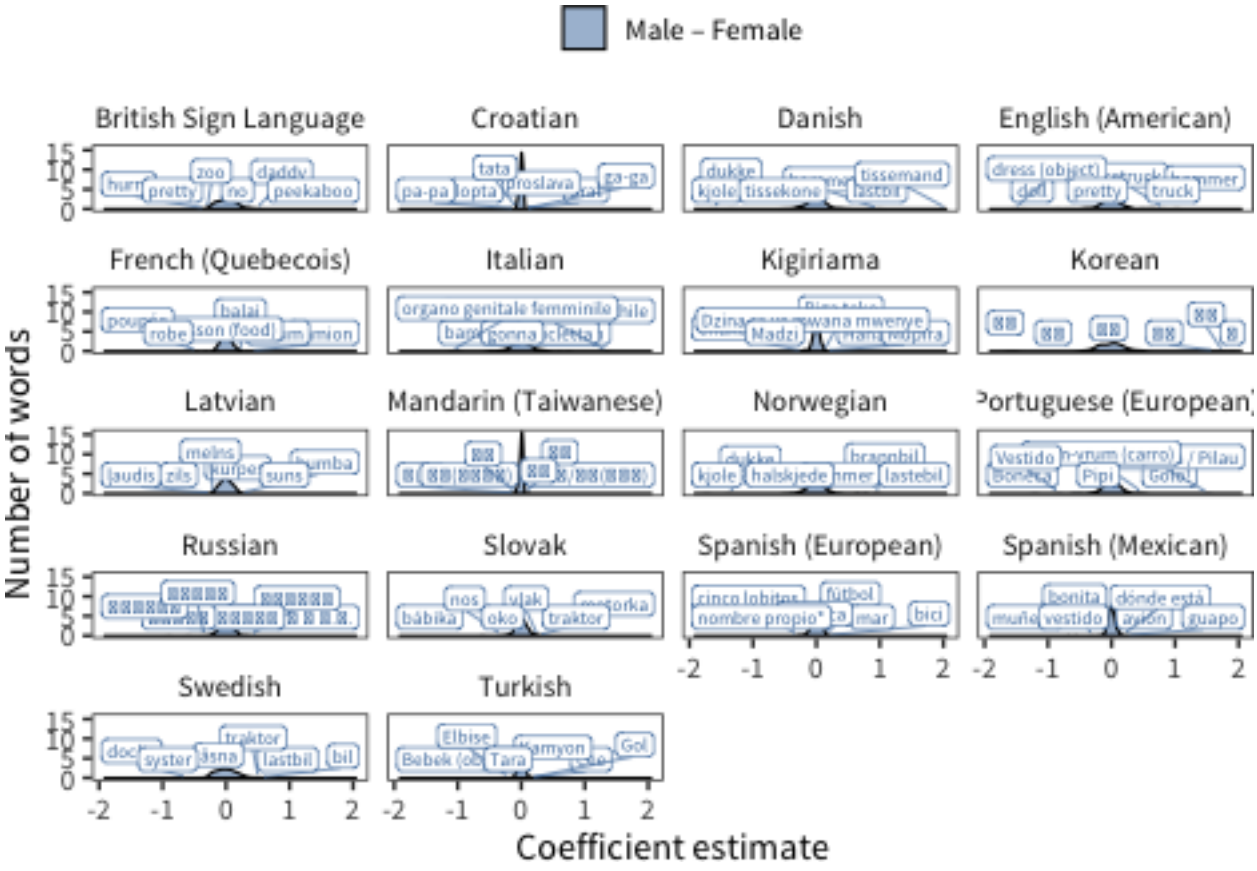
cbind(num_true, num_false) ~ (age + birth_order | item) + age + birth_order

For each demographic variable, we specify the contrasts such that their coefficient compares each level of the variable to the previous level. For example, the coefficients for birth order reflect the overall difference between second-born children as compared to first-born and the overall difference between third- (and later-) born as compared to second-born. The items' random slopes for each demographic indicate for each individual item, the contribution to those same differences over the main effect.

9.2 Results

The primary target of our analysis are the item random effects for each demographic variable, indicating our best estimate of the specific effect of a particular demographic on a particular item. These item random effects factor out the fixed, main effect of the demographic (the effects we reported in Chapter 6), thus they are centered at zero. But their magnitude and direction can be interpreted for individual effects.

Each of the plots below show the distribution of item random effects for each demographic variables and measure, with the top and bottom 3 items labelled. As well as the general qualitative shape of the distributions, it is these extreme items that we are most interested in.



One interesting question is the extent to which extreme items differ. Figure 9.2 shows the distribution of demographic random effects across all languages (selecting only sex effects for production), using a quantile-quantile (QQ) plot. In QQ plots, points on a diagonal line indicate conformity to the standard normal distribution, while deviations suggest differences in distributional form. Looking at the resulting plot yields a broad, low-slope diagonal (a normal distribution) with skewed tails. Further, the majority of coefficients are within a very tight range: only 1.6% of coefficients are outside of .5 logistic units in magnitude. Thus, as hypothesized, all of the action is in the tails of the distribution: a few words vary substantially in how often they are produced according to some demographic feature.

In the following subsections we examine the coefficients and their distributions for individual words/languages.

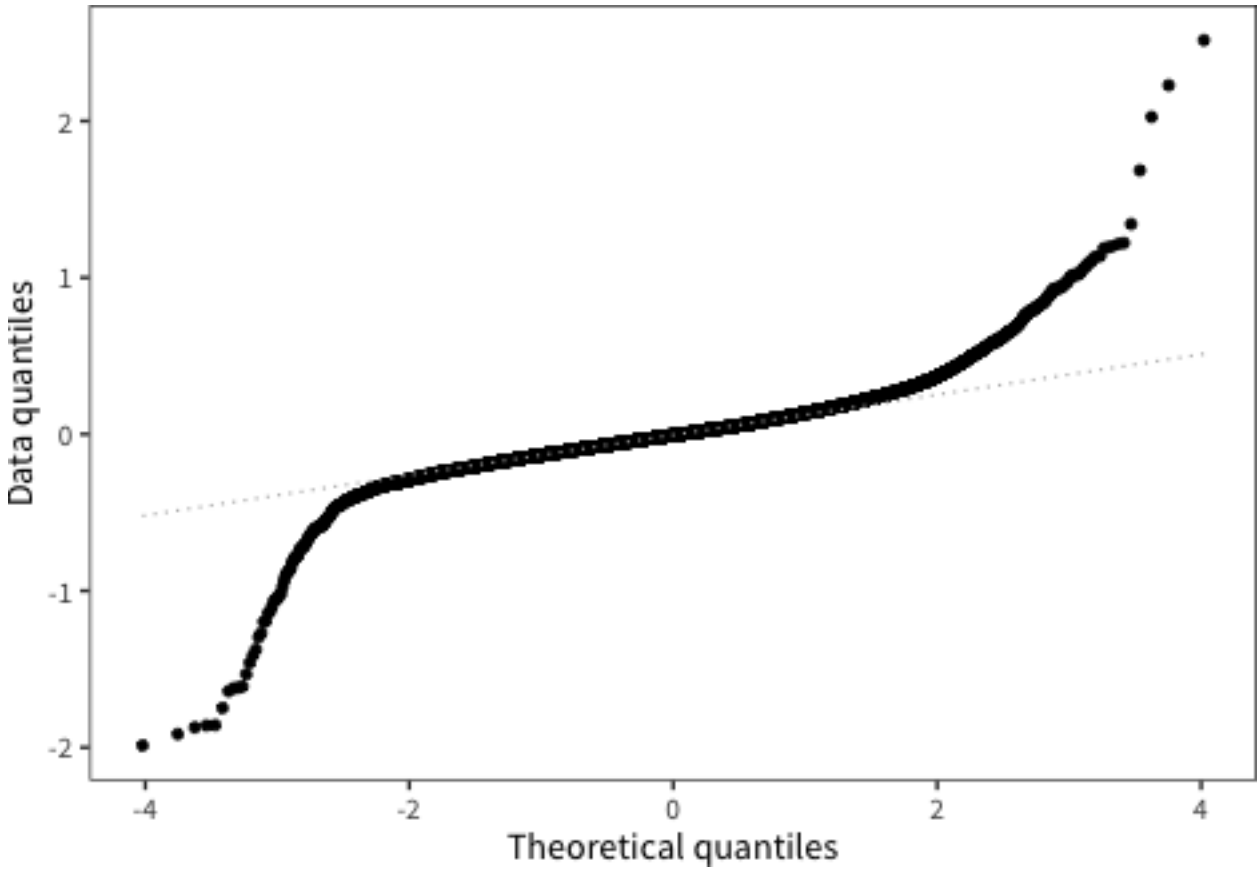


Figure 9.2: Quantiles of sex-based item random effects for production data compared to theoretical quantiles of a normal distribution.

9.2.1 Sex

As shown in Chapter 6, there is a highly consistent advantage for females in language production. This advantage is slightly less pronounced for comprehension but still present. However, independently of this advantage, we also see specific items emerge as understood differentially for males or females.

Sex item random effects with magnitude at least 0.5 for comprehension data in each language.

Language	Item	Male advantage
British Sign Language	peekaboo	0.54
Danish	tissemand	2
Danish	lastbil	1
Danish	hammer	0.96
Danish	brandbil	0.82
Danish	årnnn (bil-lyd)	0.75
Danish	motorcykel	0.64
Danish	kost	0.52
Danish	tog	0.51
Danish	halskæde	-0.53

Showing 1 to 10 of 69 entries Previous 1 2 3 4 5 6 7 Next

Figure 9.3 gives the full distribution for comprehension, and Table 9.2.1 gives the items outside of the .5 logistic units threshold, across all languages. These are almost exclusively traditionally gendered items — for English, for example, the words with a substantial male advantage are vehicle-related and hammer, while the female advantage words are purse and necklace. Thus, our first impression is that these tend to be specific content items associated with gendered play.

Sex-based item random effects with magnitude at least 0.5 for production data in each language.

Language	Item	Male advantage
Cantonese	Bi Boo (救傷車聲)	0.5
Cantonese	裙	-0.65
Croatian	brr-brr	0.73
Croatian	grr	0.63
Croatian	Ga-ga	0.56
Croatian	haljina	-0.6
Czech	vrn vrn (auto)	0.8
Czech	pindík (i jiné označení)	0.72
Czech	šššš hů	0.56
Czech	bagr	0.5

Showing 1 to 10 of 284 entries Previous 1 2 3 4 5 … 29 Next

Figure 9.4 and Table 9.2.1 give the same measures for production. There are considerably more words per language with substantial gender biases for production (11.36) than for comprehension (3.83). But the content of these is extremely similar. For English, we see a male bias for vehicles and objects associated with traditionally male activities (e.g., sports), and a female bias for genital

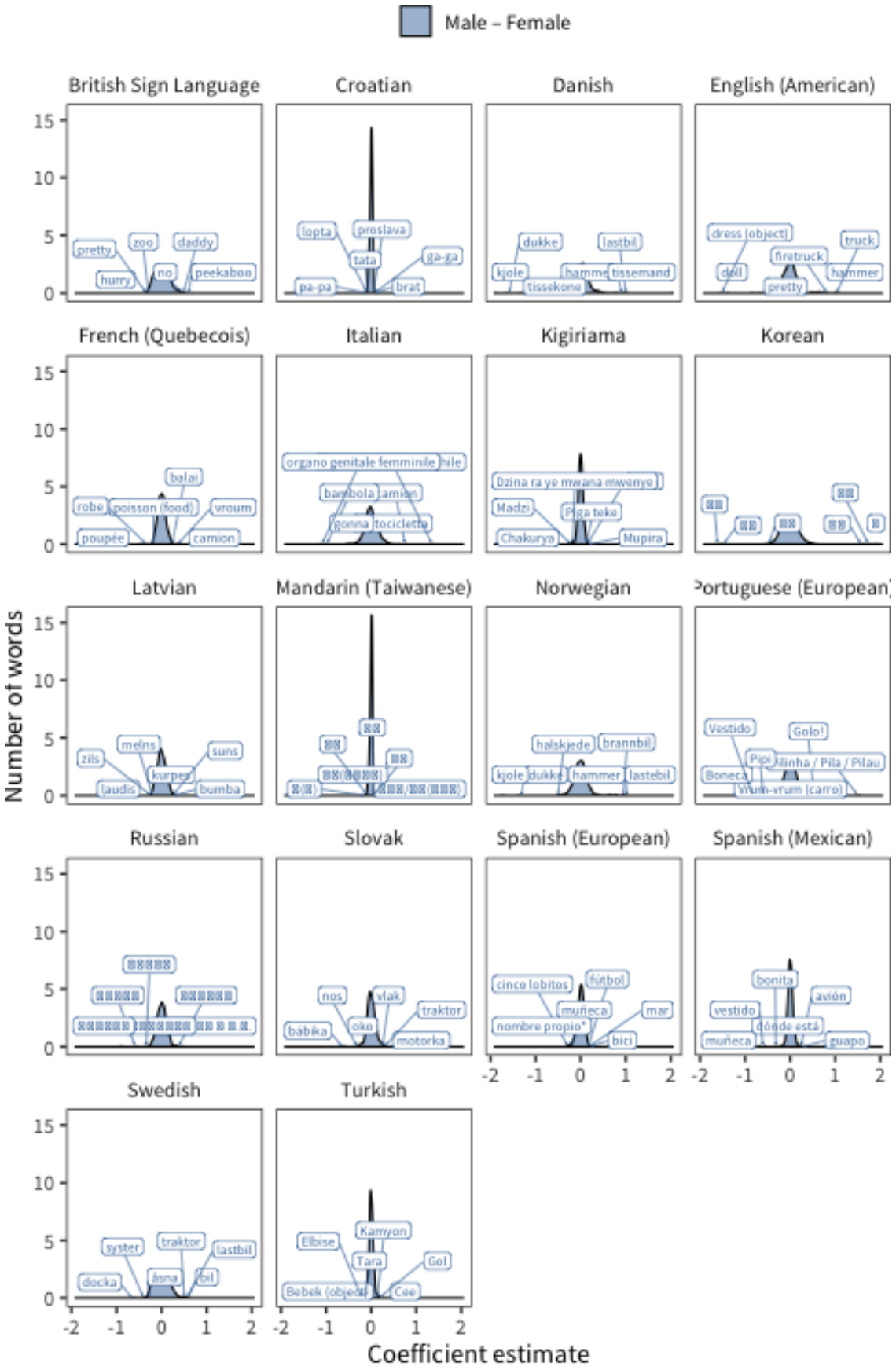


Figure 9.3: Distribution of sex item random effects for comprehension data in each language (most extreme 3 items are labelled).

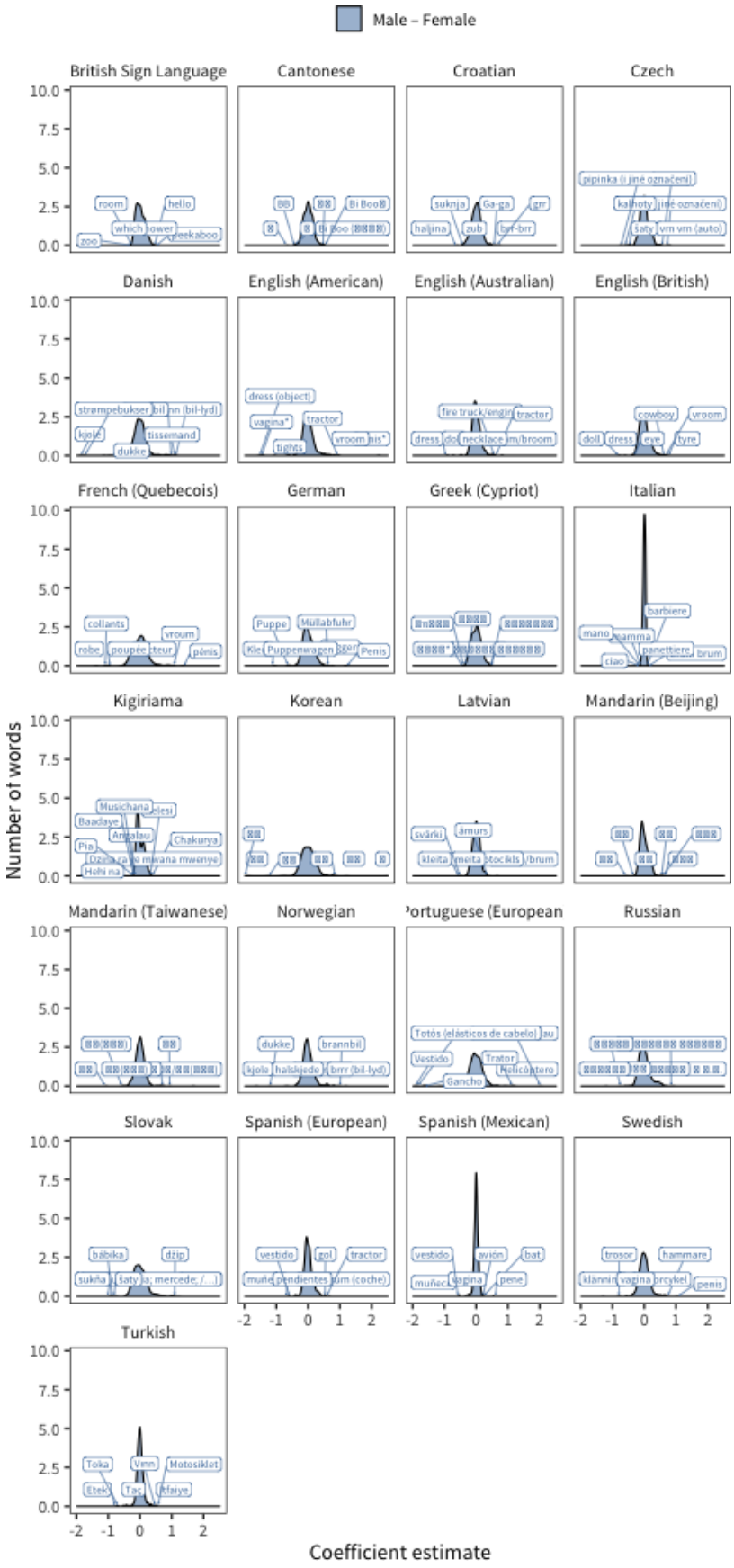


Figure 9.4: Distribution of sex item random effects for production data in each language (most extreme 3 items are labelled).

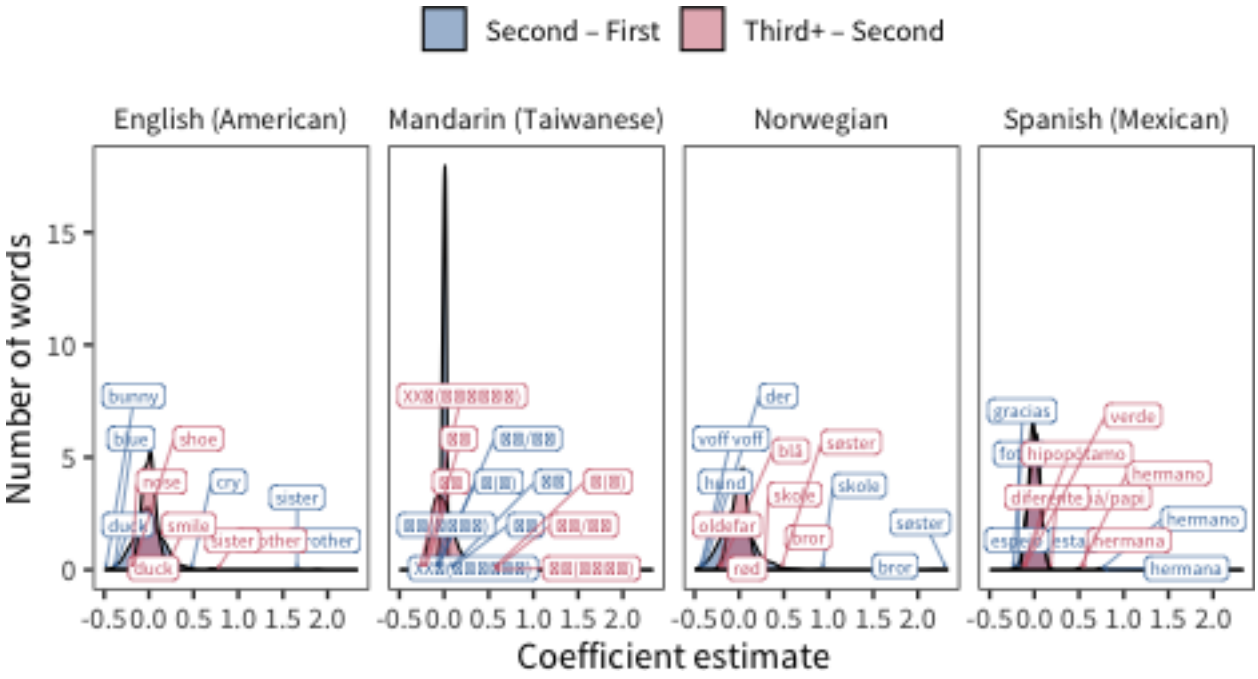


Figure 9.5: Distribution of birth order item random effects for comprehension data in each language (most extreme 3 items are labelled).

words and clothing. This pattern is replicated quite robustly across languages, although with varying magnitudes.

In sum, there appear to be two different processes at work in the gender effects we observe. The first is a general shift in the the probability that any word will be produced or understood such that females are slightly more likely to produce it. The average magnitude of this fixed effect is -0.42. In other words, if a female had a 50% chance of saying a word, a male would on average have a 40% chance of saying it. However, beyond this fixed effect, there are also variable effects for individual words. Most of these effects are small, but a few of them are quite large. For example, if an English-speaking female child had a 50% chance of saying the word dress, a male child would have an 12% chance of saying it.

9.2.2 Birth order

Again following Chapter 6, we consider individual items that are more or less likely in the vocabularies of first-born vs. later-born children. Here we consider both the contrast between second- and first-born children as well as between third- or later-born and second-born children. The number of languages for which we have birth order data is dramatically smaller, however, so conclusions are necessarily more tentative.

Birth order item random effects with magnitude at least 0.5 for comprehension data in each language.

Language	Item	Secondborn advantage	Laterborn advantage
English (American)	brother	1.9	0.73
English (American)	sister	1.7	0.69
Mandarin (Taiwanese)	抱(抱)	-0.096	0.53
Mandarin (Taiwanese)	媽媽/媽咪	-0.099	0.54
Mandarin (Taiwanese)	牛奶(ㄋㄟㄋㄟ)	-0.099	0.54
Norwegian	søster	2.3	0.45
Norwegian	bror	2.2	0.53
Norwegian	skole	0.92	0.27
Norwegian	sukkertøy	0.66	0.11
Norwegian	corn flakes	0.61	-0.027

Showing 1 to 10 of 14 entries Previous 1 2 Next

Figure 9.5 and Table 9.2.2 again represent random effects coefficients for particular items in comprehension. In general there are few surprises here: the words for brother and sister are much more likely for second-born children to understand, and even more likely for later-born children. The Norwegian data additionally show a few other words that second- and later-born children might be more likely to be exposed to via their siblings, including skole (school) and sukkertøy (sweets, hard candy).

Birth order item random effects with magnitude at least 0.5 for production data in each language.

Language	Item	Secondborn advantage	Laterborn advantage
Danish	skole	0.79	0.2
Danish	bror	0.7	0.19
Danish	søster	0.51	0.14
English (American)	brother	1.1	0.5
English (American)	sister	1.1	0.46
English (American)	gum	0.9	0.39
English (American)	hate	0.61	0.32
English (American)	popsicle	0.53	0.18
English (American)	donut	0.52	0.22
English (American)	candy	0.51	0.25

Showing 1 to 10 of 22 entries Previous 1 2 3 Next

The same general patterns are present in the production data (Figure 9.6 and Table 9.2.2), with further evidence that having elder siblings appears to be related exposure to sweets, at least in some cultures: popsicle, donut, and candy all appear now in the English data, and tyggegummi (gum) and several soda- and candy-related words appear in the Norwegian data. (Hate also appears in the English data, suggesting some emotional expressions due to having a sibling). We interpret this pattern with caution, however, as birth order is likely partially confounded with socio-economic status, and so later-born children might also be from low-SES families who have more environmental exposure to "junk foods" like soda and candy.

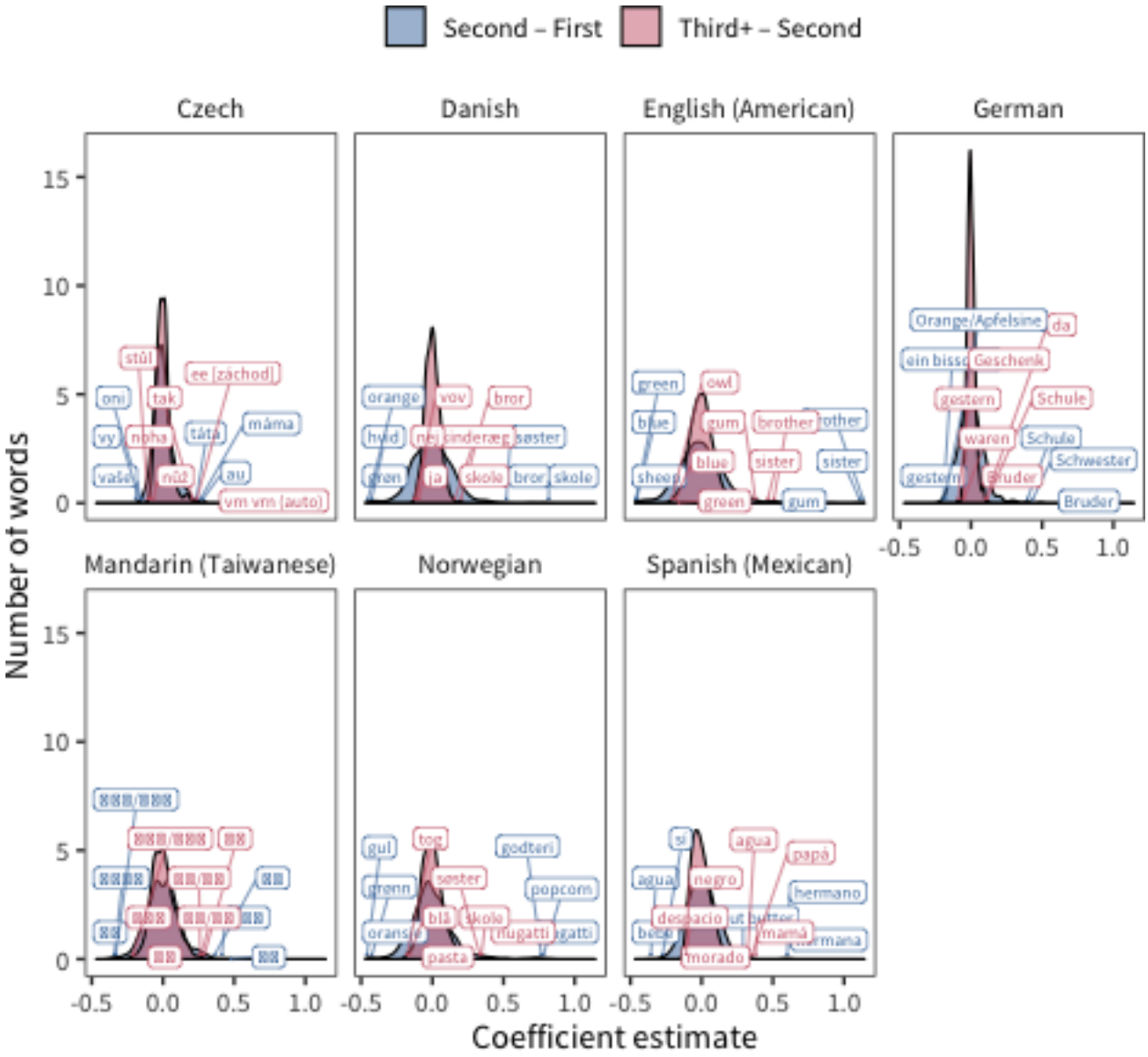


Figure 9.6: Distribution of birth order item random effects for production data in each language (most extreme 3 items are labelled).

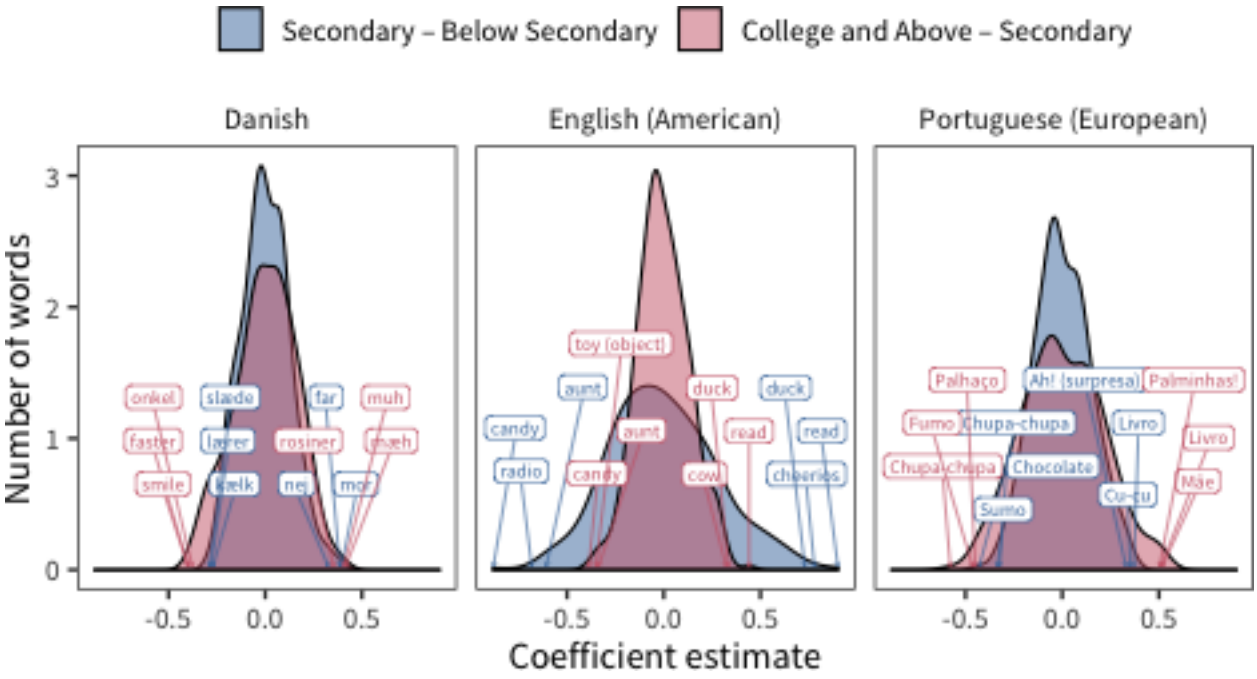


Figure 9.7: Distribution of maternal education item random effects for comprehension data in each language (most extreme 3 items are labelled).

9.2.3 Maternal education

Our final set of analyses examine vocabulary items that are differentially present in the vocabulary of children with lower maternal education. As noted in Chapter 6, there are substantial cross-linguistic differences in how large the overall socioeconomic stratification is. For example, we observe large differences in children's vocabulary size in the English (American) data, with children of less educated mothers reporting substantially lower production vocabulary.

Maternal education item random effects with magnitude at least 0.5 for comprehension data in each language.

Language	Item	Secondary advantage	College advantage
English (American)	read	0.9	0.44
English (American)	cheerios	0.78	0.3
English (American)	duck	0.73	0.34
English (American)	bird	0.69	0.29
English (American)	daddy*	0.67	0.13
English (American)	bunny	0.66	0.28
English (American)	cow	0.65	0.33
English (American)	quack quack	0.63	0.27
English (American)	gentle	0.61	0.28
English (American)	where	0.61	0.26

Showing 1 to 10 of 37 entries Previous 1 2 3 4 Next

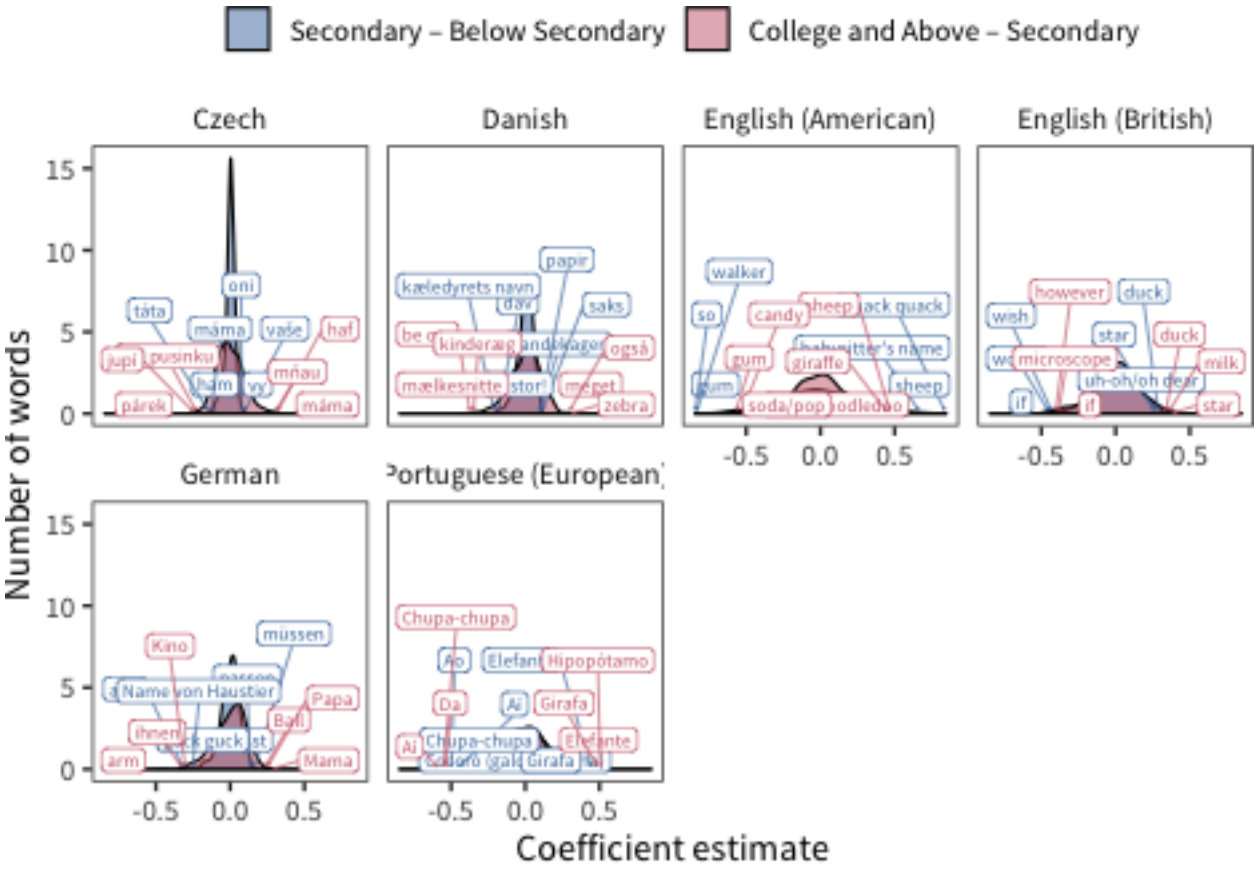


Figure 9.8: Distribution of maternal education item random effects for production data in each language (most extreme 3 items are labelled).

Figure 9.7 and Table 9.2.3 show comprehension results. The majority of words that exceed our (somewhat arbitrary) .5 bound come from the English (American) data. This finding is consistent with the idea that there may be more substantial maternal education effects in this dataset more generally. The words that are more likely to be understood by children of college- and secondary-educated mothers are often animal-related and may speculatively be related to reading books about animals (since these farm animals might not be prominent in all children's experience). Read is also on this list, perhaps related to reading practices (or the perception of the importance of these practices). Negatively linked words include cake (supporting the speculation above) and a number of other items that are perhaps harder to interpret as being SES-linked.

Maternal education item random effects with magnitude at least 0.5 for production data in each language.

Language	Item	Secondary advantage	College advantage
English (American)	quack quack	0.85	0.45
English (American)	sheep	0.71	0.48
English (American)	babysitter's name	0.68	0.42
English (American)	cockadoodledoo	0.67	0.45
English (American)	duck	0.67	0.3
English (American)	giraffe	0.65	0.47
English (American)	moo	0.64	0.29
English (American)	owl	0.62	0.41
English (American)	yogurt	0.58	0.35
English (American)	zebra	0.57	0.39

Showing 1 to 10 of 54 entries Previous 1 2 3 4 5 6 Next

Production data show a similar but more extreme picture (Figure 9.8 and Table 9.2.3), with a larger number of words linked to maternal education. Examination of the English data suggests that animal vocabulary is again more prevalent for the children of more educated parents (as are babysitters). Again supporting the birth-order/maternal education link, brother is less common for the children of more highly educated moms, as are candy, gum, and soda. Again, the most extreme linkage to maternal education was found in the English (American) sample.

9.3 Conclusions

Demographic factors like sex, birth order, and maternal education are related to children's vocabulary size. But in addition to these more global associations, they appear to be specifically associated with particular vocabulary items. Many of these are straightforwardly explicable in terms of differences in the environmental frequency (and importance) of particular lexical items for children in different circumstances. For example, there are many reasons why second-born children should say brother or sister more frequently than first-born children!

More generally, item level variation relates to two issues of interest within the context of our project. The first is the validity of CDI-based measurement. From a psychometric perspective, the sort of variation reported here is known as "differential item function" (Hambleton et al., 1991) and is a negative characteristic of tests that impairs their validity. Thus, from a test-design perspective, items like babysitter (or even brother) should probably not be included. (See Chapter 4 for more details on this issue).

The second broader issue is the question of mechanisms responsible for the demographic associations documented in Chapter 6. Sex differences in vocabulary appear quite consistent across languages. Why is this? We gain one small piece of leverage on the issue by noticing that there appear to be two qualitatively different processes involved in the demographic effects we observed: first, girls have a small bump in their probability of producing almost every word, and second, there are a small number of particular words for which their production probability is substantially different. To the extent these are separable, we might look for causal mechanisms that would provide a broader boost to language (rather than trying to explain the small number of specifically gender-linked items identified above). Such hypotheses might appeal to dyadic factors like differences in amount of

language input directed to girls, or learner-internal factors like stronger social cognition.

Chapter 10

Predictive Models of Individual Words[1]

As discussed in Chapter 1, one classic approach to word learning focuses on the specific mechanisms that children bring to bear on the learning problem. For example, across many laboratory experiments, a variety of mechanisms have been identified as plausible drivers of early word learning, including co-occurrence based and cross-situational word learning (Schwartz and Terrell, 1983; Yu and Ballard, 2007); social cue use (Baldwin, 1993); and syntactic bootstrapping (Gleitman, 1990; Mintz, 2003). The ability to identify which of these mechanisms is most explanatory has been challenging.

Indeed, many theories of early word learning take multiplicity of cue types and mechanisms as a central feature (e.g., Hollich et al., 2000; Bloom, 2000). As important as this work is, though, these studies typically are aimed at understanding how one or a small handful of words are learned in the laboratory under precisely-defined learning conditions. They do not directly address questions regarding the developmental composition and ordering of growth in the lexicon across many different children in their natural environments nor whether these patterns are consistent across different languages.

Why are some words learned so early and some much later? This question about the order of the acquisition of first words can provide a different window into the nature of children's language learning. Posed as a statistical problem, the challenge is to find what set of variables best predicts the age at which different words are acquired. Previous work using this approach has revealed that, in English, within a lexical category (e.g., nouns, verbs), words that are more frequent in speech to children are likely to be learned earlier (Goodman et al., 2008). Further studies have found evidence that a variety of other semantic and linguistic factors are related to word acquisition, such as salience and iconicity (Hills et al., 2009; Stokes, 2010; Perry et al., 2015; Roy et al., 2015; Swingley and Humphrey, 2017).

These exciting findings are limited in their generality because each study used a different dataset and focused on different predictors. In addition, nearly all studies to date have exclusively analyzed data from English-learning children, providing no opportunity for cross-linguistic comparison of the relative importance of the many relevant factors under consideration. Such cross-linguistic comparisons are critical. Identifying commonalities (and differences) across languages is our best strategy for uncovering the universal mechanisms that are in play for all children and differentiating them from patterns of acquisition that emerge due to the particulars of a given language or culture (Slobin, 1985; Bates and MacWhinney, 1987). In this chapter, we use the Wordbank data to extend

[1]The contents of this chapter are lightly adapted from Braginsky et al. (sion).

these classic approaches and assess the degree to which the predictors of word learning are consistent across different languages and cultures, as well as whether there are similar patterns across different word types (e.g., nouns vs. verbs).

We conduct cross-linguistic comparisons of the age of acquisition of particular words. We integrate estimates of words' acquisition trajectories from the Wordbank data with independently-derived characterizations of the word learning environment from other datasets. The use of secondary datasets for these analyses is warranted because no currently available resource provides data on both children's language environments and their learning outcomes for more than a small handful of children. In particular, we derive our estimates of the language environment from transcripts of speech to children in the CHILDES database (MacWhinney, 2000). This data-integration methodology was originated by Goodman et al. (2008); it relies on large samples to average out the (substantial) differences between children and care environments. While introducing additional sources of variability, it also allows for analyses that cannot be performed on smaller datasets or datasets that measure only child or environment but not both.

As our particular measures of environmental input, we estimated each word's (a) frequency in parent speech to children, (b) mean length of the parent utterances containing that word (MLU), (c) frequency as a sole utterance constituent, and (d) frequency in utterance-final position. While these measures are crude, they are both easy to compute and relatively comparable across the languages in our sample. To derive proxies for the meaning-based properties of each word, we accessed available psycholinguistic norms using adult ratings of each word's (a) concreteness, (b) valence, (c) arousal, and (d) association with babies. Integrating these two groups measures, which are based respectively on estimates of children's linguistic environment and words' meaning, we predict each words' acquisition trajectories. We assess the relative contributions of each predictor, as well as how those predictors change over development and interact with the lexical category of the word being predicted. Since vocabulary composition differs in comprehension and production (e.g., Benedict, 1979), we conduct our analyses on measures of each.

These analyses address two questions. First, we ask about the degree of consistency across languages in the relative importance of each predictor. Consistency in the patterning of predictors would suggest that similar information sources are important for learners, regardless of language. Such evidence would suggest that superficial linguistic dissimilarities (e.g., greater morphological complexity in Russian and Turkish, greater phonological complexity in Danish) do not dramatically alter the course of acquisition. Conversely, variability would show the degree to which learners face different challenges in learning different languages, posing a challenge for more universalist accounts. Further, systematicity in the variability between languages would reveal which languages are more similar then others in the structure of these different challenges.

Second, we ask which lexical categories are most influenced by linguistic environment factors, like frequency and utterance length, compared with meaning-based factors like concreteness and valence. Division of dominance theory suggests that nouns might be more sensitive to meaning factors, while predicates and closed-class words might be more sensitive to linguistic environment factors (Gentner and Boroditsky, 2001). And on syntactic bootstrapping theories (Gleitman, 1990), nouns are argued to be learned via frequent co-occurrence (operationalized by frequency) while verbs might be more sensitive to syntactic factors (operationalized here by utterance length) (Snedeker et al., 2007). Thus, examining the relative contribution of different predictors across lexical categories can help test the predictions of influential theories of acquisition.

Table 10.1: Statistics for data from Wordbank and CHILDES. N indicates number of children.

		Production		Comprehension		CHILDES	
Language	CDI items	N	Ages	N	Ages	Types	Tokens
Croatian	388	627	8-30	250	8-16	12,064	218,775
Danish	381	6,112	8-36	2,398	8-20	4,956	195,658
English	393	7,312	8-30	1,792	8-18	45,597	7,679,042
French	396	1,364	8-30	537	8-16	28,819	2,551,113
Italian	392	1,400	7-36	648	7-24	7,544	188,879
Norwegian	380	7,466	8-36	2,374	8-20	10,670	231,763
Russian	410	1,805	8-36	768	8-18	5,191	32,398
Spanish	399	1,891	8-30	788	8-18	33,529	1,609,614
Swedish	371	1,367	8-28	467	8-16	8,815	359,155
Turkish	395	3,537	8-36	1,115	8-16	6,503	44,347

10.1 Methods

10.1.1 Acquisition trajectories

Since analyses in this chapter rely on unilemma mappings (see Chapter 9.1), the set of languages represented is smaller than in other chapters.

We use data from the items on WG forms for our comprehension measure, and data from the items in common between WG and WS forms for our production measure. Table 10.1 gives an overview of our acquisition data. Each of the datasets were conducted in contexts in which the particular language was the language of the community, e.g., the Mexican Spanish CDI data were collected in several areas of Mexico; longitudinal administrations were excluded.

See Figure 10.1 for example item curves of the type being predicted in our subsequent analyses.

10.1.2 Word properties

For each word that appears on the forms in each of our 10 languages, we used corpora of child-directed speech in that language from CHILDES to obtain an estimate of its frequency, the mean length of utterances in which it appears, its frequency as the sole constituent of utterance, and its frequency in utterance final position (with frequency residualized out of solo and final frequencies). Additionally, we computed each word's length in phonemes.

To capture meaning-based factors in acquisition, we included ratings of each word's concreteness, valence, arousal, and relatedness to babies. All of these ratings were compiled based on previous studies using adult raters. In addition, since existing datasets for all of these ratings are primarily available for English, we mapped all the words in our datasets onto translation equivalents across CDI forms, verified by native speaker judgement, allowing us to use the ratings for English words across languages. Of the resulting translation equivalent meanings, 35% occur only in one language, 51% occur in more than one but not all languages, and 14% occur in all languages. While necessarily imperfect, this method allows us to examine languages for which limited resources exist. Example words for these predictors in English are shown in Table 10.2.

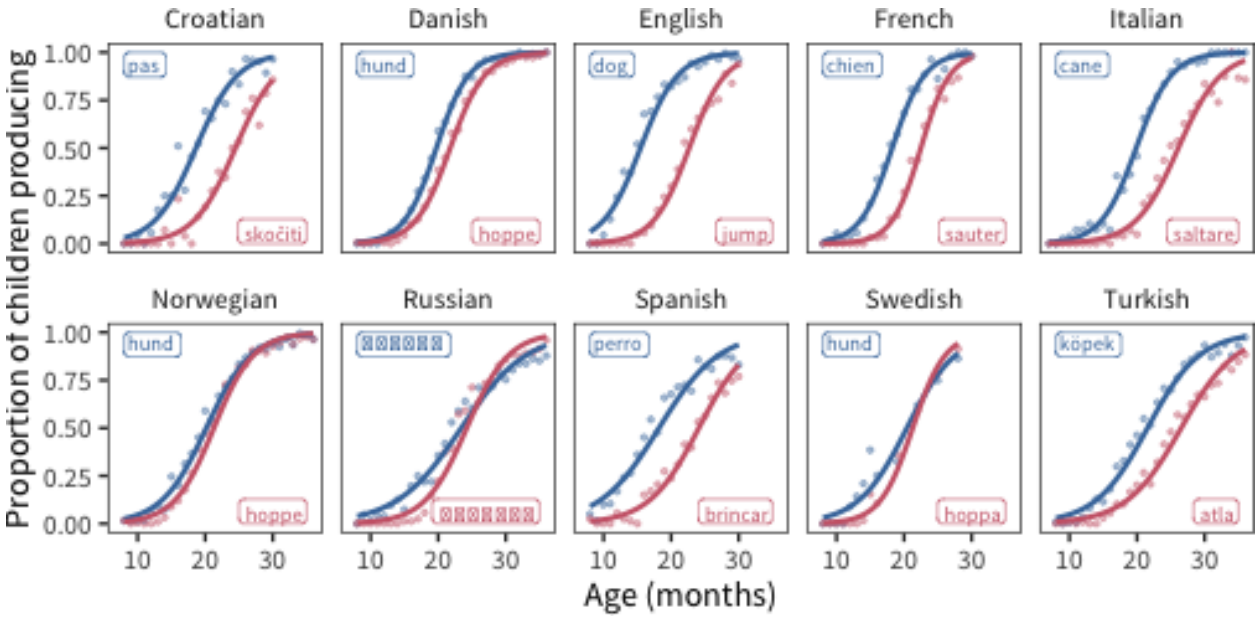


Figure 10.1: Example production trajectories for the words "dog" and "jump" across languages. Points show the proportion of children producing each word for each one-month age group. Lines show the best-fitting logistic curve. Labels show the forms of the words in each language.

Previous studies have shown robust consistency in the types of words that children learn very early (Tardif et al., 2008). These words seem to describe concepts that are important or exciting in the lives of infants in a way that standard psycholinguistic features like concreteness do not. Capturing this intuition quantitatively is difficult, but Perry et al. (2015) provides a proxy measure as a first step. This measure is simply the degree to which a particular word was "associated with babies." Intuitively, we expect this measure to capture the degree to which words like ball or bottle feature heavily in the environment (and presumably, mental life) of many babies.

Each numeric predictor was centered and scaled so that all predictors would have comparable units. For each predictor, missing values (CDI items that were not in the relevant corpus or norms) were imputed from the mean for their respective language and measure. Placeholder items, such as child's own name, were excluded.

Frequency. For each language, we estimated word frequency from unigram counts based on all corpora in CHILDES for that language. Frequencies varied widely both within and across lexical categories. Each word's count includes the counts of words that share the same stem (so that dogs counts as dog) or are synonymous (so that father counts as daddy). For polysemous word pairs (e.g., orange as in color or fruit), occurrences of the word in the corpus were split uniformly between the senses on the CDI (there were only between 1 and 10 such word pairs in the various languages; in the absence of cross-linguistic corpus resources for polysemy sense disambiguation, this is a necessary simplification). Counts were normalized to the length of each corpus, Laplace smoothed (i.e., count of 0 were replaced with counts of 1), and then log transformed.

Solo and Final Frequencies. Using the same dataset as for frequency, we estimated the frequency with which each of word occurs as the sole word in an utterance, and the frequency with which it appears as the final word of an utterance (not counting single-word utterances). As with frequency, solo and final counts were normalized to the length of each corpus, Laplace smoothed, and log

Table 10.2: Items with the highest and lowest values for each predictor in English.

Predictor	Highest	Lowest
Arousal	naughty, money, scared	blanket, asleep, shh
Babiness	baby, bib, bottle	jeans, penny, donkey
Concreteness	apple, baby, ball	that, now, how
Final frequency	book, it, there	put, when, give
Frequency	you, it, that	babysitter, rocking chair, grrr
MLU	when, day, store	ouch, thank you, peekaboo
Number phonemes	refrigerator, cockadoodledoo, babysitter	i, eye, ear
Solo frequency	no, yes, thank you	feed, bathroom, tooth
Valence	happy, hug, love	ouch, hurt, sick

transformed. Since both of these estimates are by necessity highly correlated with frequency, we then residualized unigram frequency out of both of them, so that values reflect an estimate of the effects of solo frequency and final frequency over and above frequency itself.

MLU. MLU is only a rough proxy for syntactic complexity, but is relatively straightforward to compute across languages (in contrast to other metrics). For each language, we estimated each word's MLU by calculating the mean length in words of the utterances in which that word appeared, for all corpora in CHILDES for that language. For words that occurred fewer than 10 times, MLU estimates were treated as missing.

Number of phonemes. In the absence of consistent resources for cross-linguistic pronunciation, we computed the number of phonemes in each word in each language based on phonemic transcriptions of each word obtained using the eSpeak tool (Duddington, 2012). We then spot-checked these transcriptions for accuracy.

Concreteness. We used previously collected norms for concreteness (Brysbaert et al., 2014), which were gathered by asking adult participants to rate how concrete the meaning of each word is on a 5-point scale from abstract to concrete.

Valence and Arousal. We also used previously collected norms for valence and arousal (Warriner et al., 2013), for which adult participants were asked to rate words on a 1-9 happy-unhappy scale (valence) and 1-9 excited-calm scale (arousal).

Babiness. Lastly, we used previously collected norms of "babiness", a measure of association with infancy (Perry et al., 2015) for which adult participants were asked to judge a word's association with babies on a 1-10 scale.

Lexical category. Category was determined on the basis of the conceptual categories presented on the CDI form (e.g., "Animals", "Action Words"), such that the Nouns category contains common nouns, Predicates contains verbs and adjectives, and Function Words contains closed-class words (following Bates et al., 1994).

Collinearity. A potential concern for comparing coefficient estimates is predictor collinearity. Fortunately, in every language, the only relatively correlations were between MLU and solo frequency (mean over languages r = -0.54), as expected given the similarity of these factors, along with modest correlations between frequency and concreteness (mean over languages r = -0.39) and between frequency and number of phonemes (mean over languages r = -0.35), a reflection of Zipf's Law (Zipf, 1935). More importantly, the variance inflation factor for each of the predictors in each language is

no greater than 2.4550255, indicating that multicollinearity among the predictors is low.

10.1.3 Analysis

We used mixed-effects logistic regression models (fit with the MixedModels package in Julia; Bates et al., 2018) to predict whether each child understands/produces each word from the child's age, properties of the word, and interactions between age and each property of the word. Each model was fit to all data from a particular language and included a random intercept for each word and a random slope of age for each word. We also fit such models separately to the words in each lexical category. The magnitude of the standardized coefficient on each feature gives an estimate of its effect on whether words are learned earlier or later. Interactions between features and age give estimates of how this effect is modulated for earlier and later-learned words. For example, a positive effect of association with babies ("babiness") means that words associated with babies are learned earlier; a negative interaction with age means that high babiness primarily leads to higher rates of production and comprehension for younger children.

10.2 Results

Predictor effects.

Figure 10.2 shows the coefficient estimates for English comprehension data, while Figure 10.3 shows the coefficient estimate for each predictor in each language. We find that frequency (mean over languages and measures $\bar{\beta} = 0.29$), babiness ($\bar{\beta} = 0.27$), concreteness ($\bar{\beta} = 0.23$), and solo frequency ($\bar{\beta} = 0.21$) are relatively stronger predictors of acquisition across languages (as well as all having significant effects at $\alpha = 0.05$ in at least 15 of the 20 languages and measure). These effects, along with final frequency and valence, are positive in all or almost languages (so words with higher babiness tend to be known by more children); while the effects of number of phonemes and MLU are negative in all or almost all languages (so longer words tend to be known by fewer children).

Given the emphasis on frequency effects in the language acquisition literature (Ambridge et al., 2015), one might have expected frequency to dominate, but several other predictors are just as strong in this analysis. In addition, some factors previously argued to be important for word learning, namely valence and arousal (Moors et al., 2013), appear to have limited relevance when compared to other factors (both have $\bar{\beta} < 0.06$ and are only significant in 4 languages and measures). These results provide a strong argument for our approach of including multiple predictors and languages in our analysis.

Consistency. Overall, there is considerable consistency in the magnitudes of predictors across languages. In almost all, babiness and frequency were highest, while valence and arousal were smaller. A priori it could have been the case that different languages have wildly different effects of various factors (e.g., due to linguistic or cultural differences in acquisition process), but this pattern is not what we observe. Instead, Figure 10.4 shows the mean pairwise correlation of predictor coefficients across languages (i.e., the correlation of coefficients for English with coefficients for Russian, for Spanish, and so on). These means are far outside of bootstrapped estimates for the average pairwise correlation in a randomized baseline created by shuffling predictor coefficients within language, meaning that coefficient estimates are far more consistent across languages than would be expected by chance.

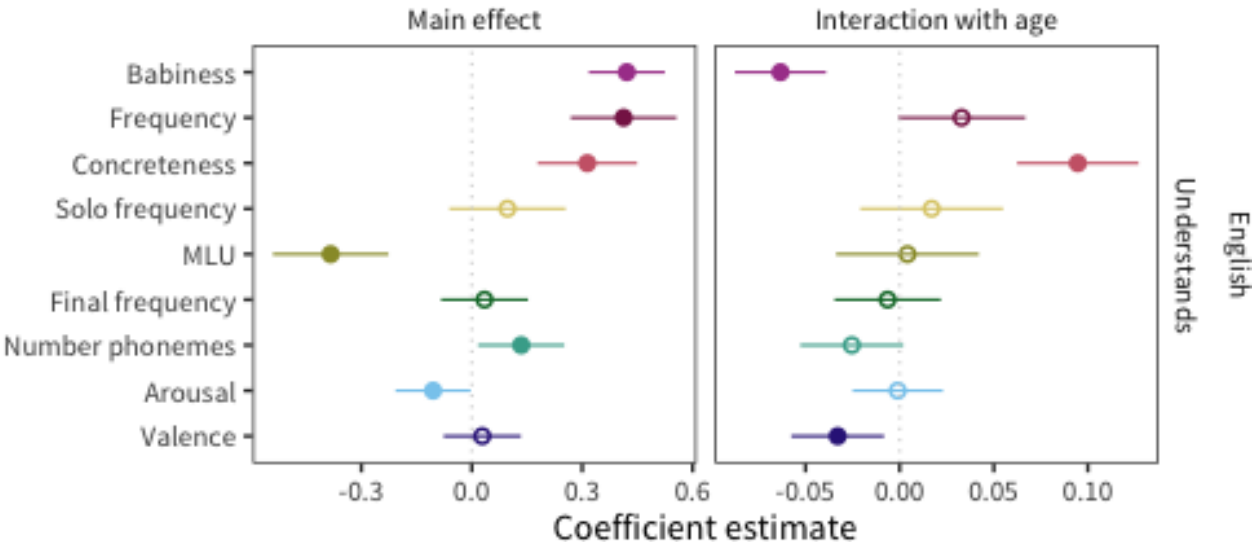


Figure 10.2: Estimates of coefficients in predicting words' developmental trajectories for English comprehension data. Larger coefficient values indicate a greater effect of the predictor on acquisition: positive main effects indicate that words with higher values of the predictor tend to be understood by more children, while negative main effects indicate that words with lower values of the predictor tend to be understood by more children; positive age interactions indicate that the predictor's effect increases with age, while negative age interactions indicate the predictor's effect decreases with age. Error bars indicates 95% confidence intervals; filled in points indicate coefficients with $p < 0.05$.

Variability. While some particular coefficients differ substantially from the trend across languages (e.g., the effect of frequency for Spanish is near 0), these individual datapoints are difficult to interpret. Many unmeasurable factors could potentially account for these differences. For example, Spanish frequency estimates could be less accurate due to corpus sparsity or idiosyncrasy, the samples of children in the Spanish CDI data and CHILDES data could differ more demographically, or Spanish-speaking children could in fact rely less on frequency in acquisition. Rather than attempting to interpret individual coefficients, we instead ask how the patterns of difference among languages reflect systematic substructure in the variability of the effects.

To examine the substructure of predictor variability, we used hierarchical clustering analysis to find the similarity structure among the pairwise correlations between languages' predictors. The resulting dendrograms are shown in Figure 10.5, which broadly reflect language typology (especially for production data). This result suggests that some language-to-language similarity data is captured by the profile of coefficient magnitudes our analysis returns.

Comprehension vs. production. Word length is the one predictor of acquisition that varied substantially between measures: it is far more predictive for production than comprehension. Thus as measured here, length seems to reflect effects of production constraints (i.e., how difficult a word is to say) rather than than comprehension constraints (i.e., how difficult it is to store or access). This result may explain why the hierarchical clustering analysis above appears more similar to linguistic typology in the production case than the comprehension case: the role of production difficulty may be more similar for more typologically related languages.

Developmental change. We also wanted to examine how the relative contributions of the predictors changes over development. For both comprehension and production, positive age interactions can be

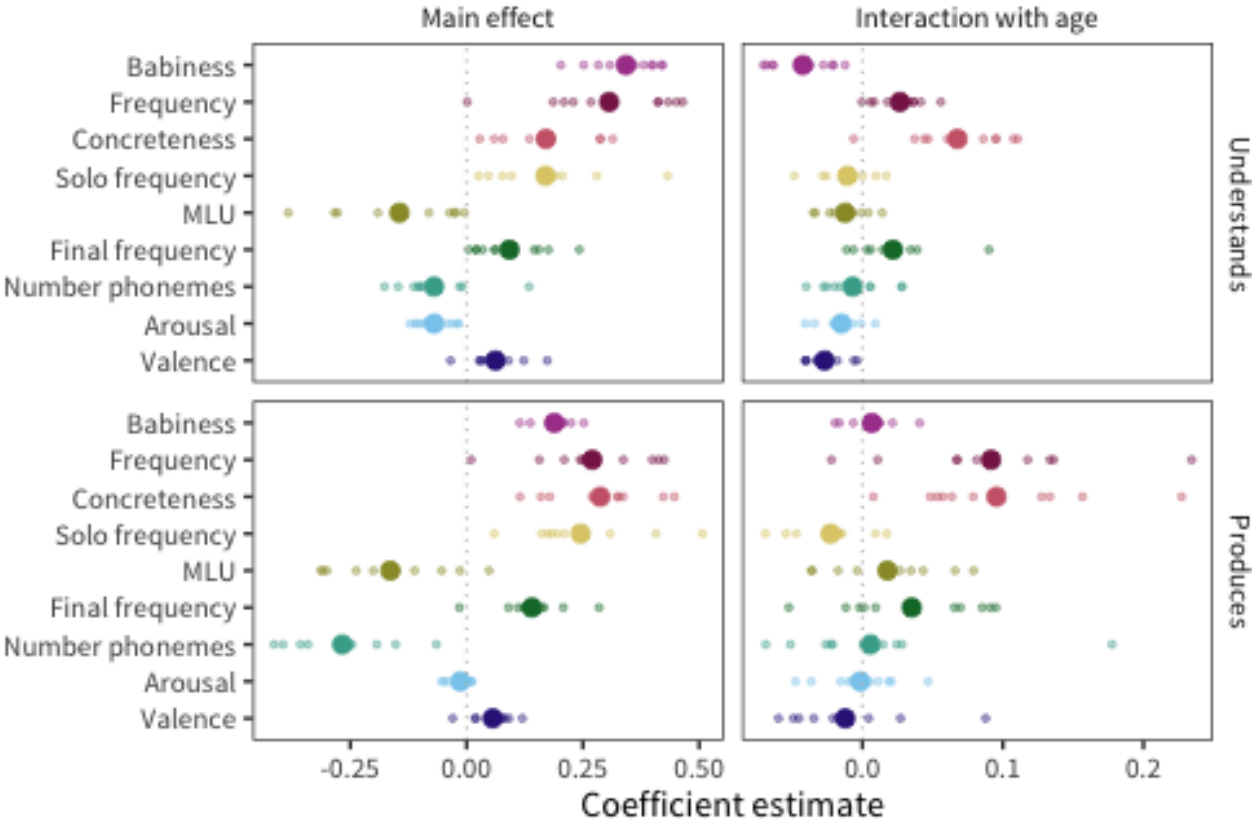


Figure 10.3: Estimates of coefficients in predicting words' developmental trajectories for all languages and measures. Each point represents a predictor's coefficient in one language, with the large point showing the mean across languages.

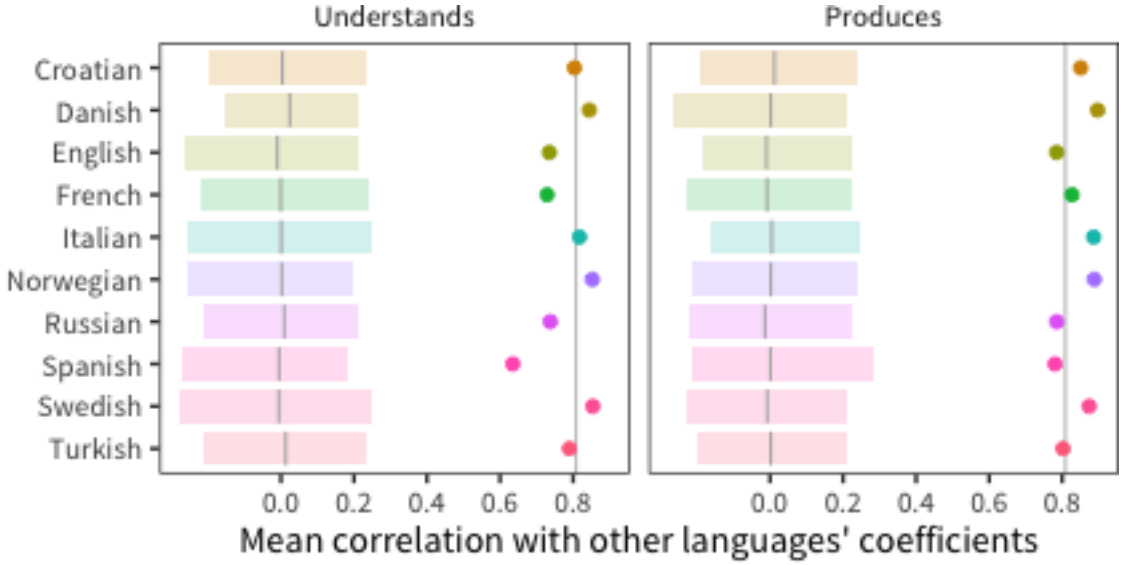


Figure 10.4: Correlations of coefficients estimates between languages. Each point represents the mean of one language's coefficients' correlation with each other language's coefficients, with the vertical line indicating the overall mean across languages. The shaded region and line show a bootstrapped 95% confidence interval of a randomized baseline where predictor coefficients are shuffled within language.

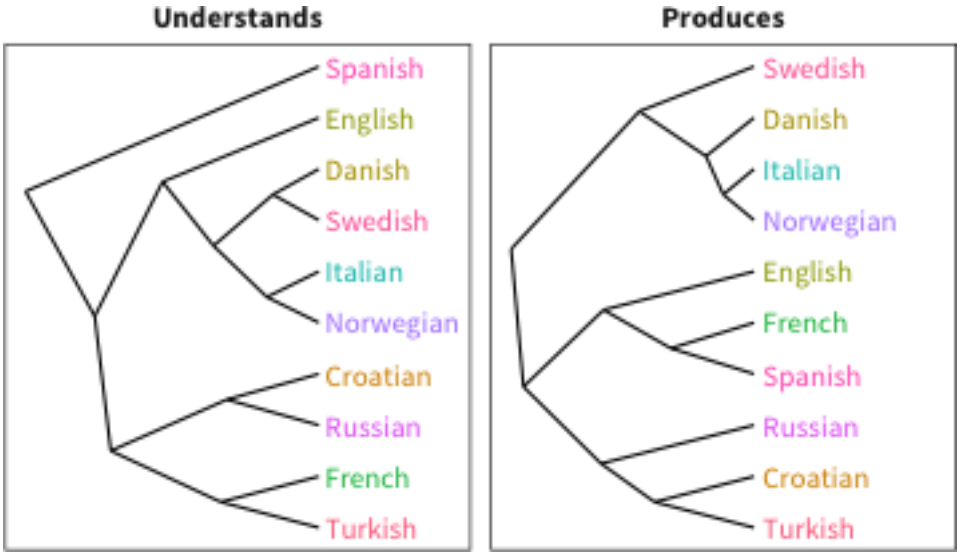


Figure 10.5: Dendrograms of the similarity structure among languages coefficients.

seen in at least 9 out of 10 languages for concreteness and frequency. Conversely, there are negative age interactions for babiness, valence, and arousal for comprehension in at least 9 out of 10 languages. This suggests that while concreteness and frequency facilitate learning, they tend to do so more later in the development; and while babiness, valence, and arousal facilitate learning as well, they then to do so more earlier in development. This result is consistent with the speculation above that the babiness predictor captures meanings that have special salience to very young infants.

Lexical categories. Previous work gives reason to believe that predictors' relationship with age of acquisition differs among various lexical categories (Goodman et al., 2008). To investigate these effects, we separated our data by lexical category and fit separate models for each category. Figure 10.6 shows the resulting coefficient estimates. Across languages, frequency had the highest magnitude for nouns and a lower magnitude for function words. In contrast, MLU was almost irrelevant for both nouns and predicates, but highly predictive for function words. These patterns are supportive of the hypothesis that different word classes are learned in different ways, or at least that the bottleneck on learning tends to be different, leading to different information sources being more or less important across categories.

Additionally, the mean pairwise correlation of coefficients between languages is much larger for nouns (0.7) and predicates (0.57) than for function words (0.3). The higher between-language variability for function words suggests the learning processes differ substantially more across languages for function words than they do for content words.

10.3 Discussion

What makes words easier or harder for young children to learn? Previous experimental work has largely addressed this question using small-scale lab studies. While such experiments can identify sources of variation, they typically do not allow for different sources to be compared directly. In contrast, observational studies allow the effects of individual factors to be measured across ages and lexical categories (e.g., Goodman et al., 2008; Hills et al., 2009; Swingley and Humphrey, 2017).

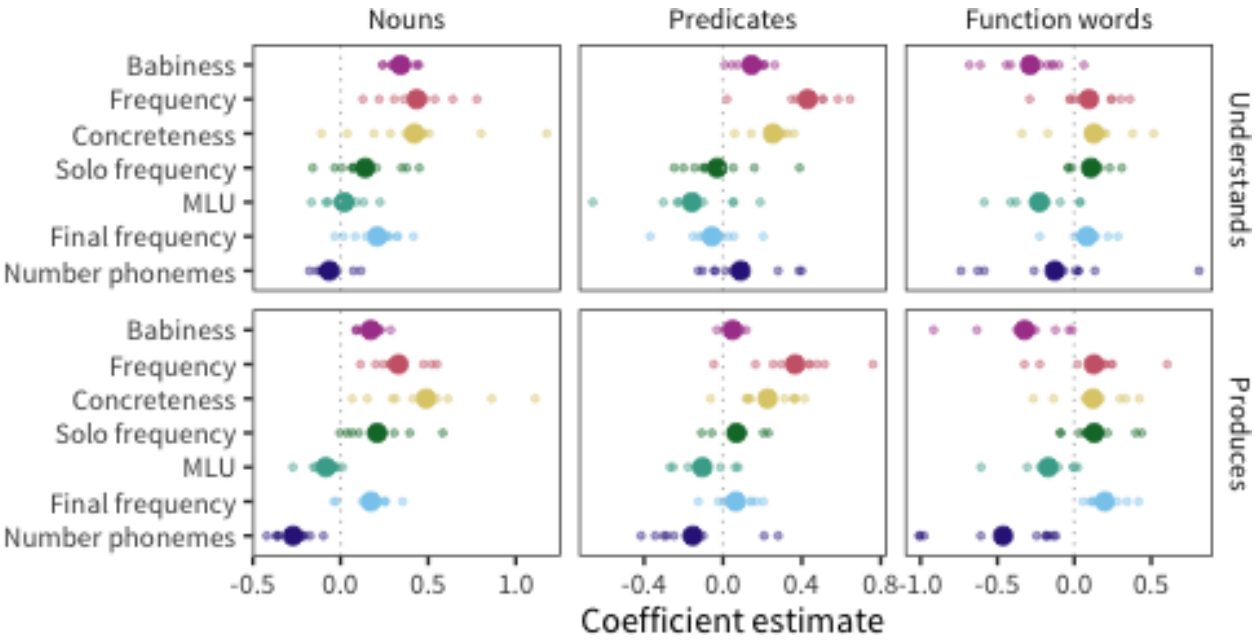


Figure 10.6: Estimates of coefficients in predicting words' developmental trajectories for each language, measure, and lexical category.

Such work has identified a number of candidate predictors of word learning. Our work expands the scope of these studies dramatically, leading to several new findings.

First, we found consistency in the patterning of predictors across languages at a level substantially greater than the predictions of a chance model. This consistency supports the idea that differences in culture or language structure do not lead to fundamentally different acquisition strategies, at least at the level of detail we were able to examine. Instead, they are likely produced by processes that are similar across populations and languages. Such processes could include learning mechanisms or biases internal to children, or interactional dynamics between children or caregivers. We believe these consistencies should be an important topic for future investigation.

Second, predictors varied substantially in their weights across lexical categories. Frequent, concrete nouns were learned earlier, consistent with theories that emphasize the importance of early referential speech (e.g., Baldwin, 1995). But for predicates, concreteness was somewhat less important. And for function words, MLU was more predictive, perhaps because it is easiest to decode the meanings of function words that are used in short sentences (or because such words have meanings that are easiest to decode). Overall, these findings are consistent with some predictions of both division of dominance theory, which highlights the role of conceptual structure in noun acquisition (Gentner and Boroditsky, 2001), and syntactic bootstrapping theory, which emphasizes linguistic structure over conceptual complexity in the acquisition of lexical categories other than nouns (Snedeker et al., 2007). More generally, our methods here provide a way forward for testing the predictions of these theories across languages and at the level of the entire lexicon rather than individual words.

In addition to these new insights, several findings emerged that confirm and expand previous reports. Environmental frequency was an important predictor of learning, with more frequently-heard words learned earlier (Goodman et al., 2008; Swingley and Humphrey, 2017). Predictors also changed in relative importance across development. For example, certain words whose meanings were more

strongly associated with babies appeared to be learned early for children across the languages in our sample (as in Tardif et al., 2008). Finally, word length showed a dissociation between comprehension and production, suggesting that challenges in production do not carry over to comprehension (at least in parent-report data).

Chapter 11

Vocabulary Composition: Syntactic[1]

This chapter focuses on splitting vocabulary data into syntactic categories and analyzing consistency and variability across languages in the acquisition of these. We quantify the "noun bias" across languages. In addition, we report the degree of bias for or against verbs and closed-class words. This chapter deals primarily with the aggregate trends across the population, but in Chapter 14, we consider variation of this sort within individuals.

11.1 Introduction

As we reviewed in Chapter 8, the first words children utter are strikingly consistent and primarily composed of names for people and things and words related to social routines (see also Tardif et al., 2008; Schneider et al., 2015). Soon after, however, they begin to add verbs (go) and adjectives (pretty) in greater proportions than earlier in development and may even begin to use closed-class forms, such as determiners (the). These patterns seem to suggest a developmental course that follows distinct "waves" of learning for words from different classes. That is, along with early social routines, nouns tend to predominate early vocabularies, while other types of words, such as predicates and closed class forms, are learned later. This pattern may be further qualified by differences in the types of words learned in comprehension vs. production (Benedict, 1979).

The composition of early vocabulary is complicated by the fact that we categorize words by their adult syntactic category. We do so in the discussion below without presupposing that children themselves do this categorization, however (Tomasello, 2000). Children may be sensitive to these categories very early in development (Valian, 1986; Yang, 2013) or they may discover them either gradually (Pine and Lieven, 1997) or more quickly (Meylan et al., 2017). Importantly, though, we treat adult syntactic categories as an analytic convenience that describes certain regularities in how groups of words are distributed in language samples and how they function in different contexts, rather than as an ontological fact about children's knowledge. Chapter 10 breaks down these categories further, asking what sorts of information predicts the order of acquisition for individual words, both within and across categories.

Bates et al. (1994) characterized these patterns of vocabulary composition in the following way.

[1] An earlier version of this work was reported to the Boston University Conference on Language Development in 2015 by Braginsky, Marchman, Yurovsky, & Frank.

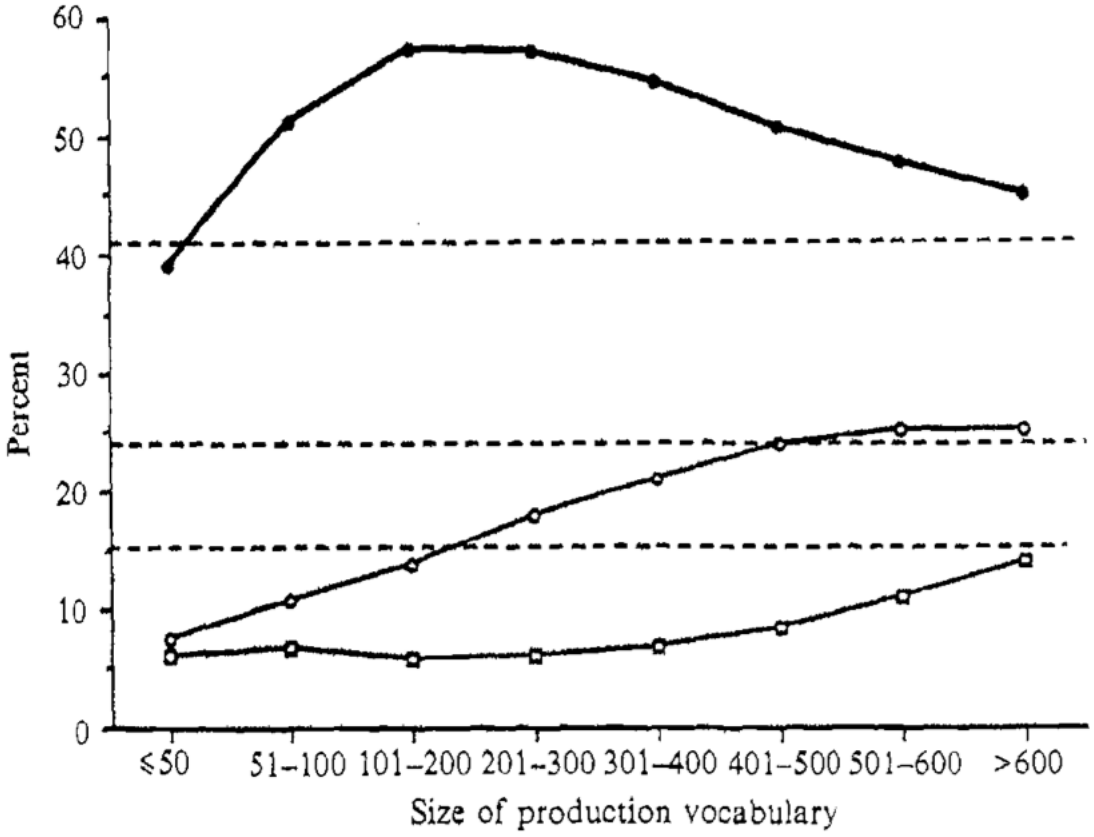


Figure 11.1: Figure from Bates et al. (1994), showing developmental trends in the categorical composition of early vocabulary.

Figure 11.1 (reprinted from that paper) shows average vocabulary composition of nominals, predicates and closed class forms as a function of children's vocabulary size for English-speaking children from the original norming study of CDI: Words & Sentences form (Fenson et al., 1994). Note that when children only know a few words (e.g., fewer than 50 words), the nominals comprise the greatest proportion of the words that children are reported to produce, with very few predicates or closed class forms ($< 10\%$). As the children learn the next hundred words or so, the proportion of nominals increases even more dramatically with a gradual increase in the proportion of children's vocabularies that are predicates. Closed class forms remain a much smaller proportion over the period. Yet after about 300 words or so, children tend not add nouns to their vocabularies at the same pace that they did earlier in development, reflected in the proportion of nominals tending to decrease.[2] It is during this developmental period that proportion of predicates tend to increase, followed by a growing proportion of closed class forms.

Why do children learn nouns before verbs and other types of words? This question has received a great deal of attention in the literature, and we can briefly summarize some of the major issues here. One reason for this "noun bias" could be that nouns are simply more frequent in the talk to young children. It is well-established that children learn the words that they hear more often (e.g., Hart and Risley, 1995). Many observational studies of English-speaking caregivers have demonstrated that caregivers use more nouns than verbs (types or tokens) with their children (e.g., Fernald and Morikawa, 1993; Goldfield, 1993; Gopnik et al., 1996; Kim et al., 2000; Poulin-Dubois et al., 1995; Tardif et al., 1997).

Other researchers have framed the "noun bias" in terms of universals about what and how different words "partition" things in the world. For example, Gentner (1978) has argued that children learn nouns before verbs because the meanings of nouns are easier to encode since they identify things that can be differentiated in the world (e.g., common everyday objects). Verbs and other predicates, in contrast, express relations among things in the world. Hence, the meanings of verbs are less accessible to children through common, everyday experiences and hence, are more difficult to map onto word forms without additional linguistic or social support.

Other reasons that nouns might be easier than verbs for young children is that nouns tend to be less morphologically complex than verbs (e.g., Tardif et al., 1997). For example, in many languages, nouns are typically marked only for number, whereas, verbs carry both person and tense information. In English, at least, verbs might also be harder to learn because they tend to occur in sentence-medial position (rather than sentence final), which make verbs less salient in the input that children hear (Slobin, 1985; Caselli et al., 1995).

Finally, differences in children's preferences for nouns vs. verbs might result from differences in what contexts children hear nouns vs. verbs in the speech from caregivers (e.g., Choi and Gopnik, 1995; Tardif et al., 1999). Several researchers have examined what caregivers talk about using naturalistic data of caregiver-child interactions. For example, caregivers in some cultures tend to emphasize the names for objects, spending a great deal of time labeling objects for their children. In other cultures, caregivers do so much less frequently, instead focusing on the actions in which those objects engage (e.g., Fernald and Morikawa, 1993; Gopnik et al., 1996). These differences in input to children can influence the words that are salient for children, and hence, the words that they are most likely to learn.

What is the evidence that a noun bias is a universal feature of children's vocabularies? Documenting

[2]This effect may also reflect aspects of CDI form design, e.g. "running out of nouns" to learn: children may increasingly learning nouns that are not on the forms.

the extent to which the noun bias is universal is relevant to understanding mechanisms of language learning, in particular, the presence of conceptual biases in early acquisition and the role of cross-cultural variability in the input that children receive from caregivers. The evidence varies across languages, as well as across methodologies (for example, naturalistic observation vs. parent report). Some studies find consistent evidence for a noun bias in English, as well as in Korean and Italian (Bates et al., 1994; Au et al., 1994; Caselli et al., 1995; Kim et al., 2000). Other studies do not find evidence of a noun bias in languages as varied as French, German, Chinese, Estonian, and Korean (Bassano, 2000; Bloom et al., 1993; Choi and Gopnik, 1995; Kauschke and Hofmeister, 2002; Tardif, 1996; Tardif et al., 1999; Schults and Tulviste, 2016).

In sum, identifying the extent of cross-linguistic variation vs. universals has been difficult since variation across studies may be due to the different methodologies that are used. For example, even within a single language, for example, Korean, parent reports of children's first words find a noun bias (e.g., Au et al., 1994), whereas, studies using direct observational methods find less evidence for this pattern (e.g., Gopnik et al., 1996). Further, few studies have had the scope to directly compare the extent of the noun bias across multiple languages using a common methodology.

One notable exception in a literature where samples have been small — in terms of both languages and children — is Bornstein et al. (2004a), in which the researchers compared vocabulary composition in seven different languages. In this chapter, we follow this comparative approach (see also Tardif et al., 2008). Since we have access to many more observations, our approach offers a more comprehensive approach than these earlier studies. Moreover, we attempt to quantify the estimates of the extent to which languages show a noun bias: we develop a statistical method for quantifying the extent of the noun bias across the entire developmental range in which a particular form is used.

11.2 Methods and data

Each CDI form contains a mixture of words in different classes. We adopt the categorization of Bates et al. (1994), categorizing words into nouns, predicates (both verbs and adjectives), function words (also referred to as "closed class" words), and other words. For each child's vocabulary, we compute the proportion of the total words in each of these categories that they are reported to produce. Following the approach developed by Bates et al. (1994), for each of the languages in our sample, we plot these proportions against total vocabulary. As shown in the Figure 11.2, if every time a child learns a word, that word is sampled randomly from the different words available on the form, then the proportion of nouns in vocabulary should track perfectly with the proportion of total vocabulary (the diagonal). In contrast, the extent to which each child's vocabulary has more words in a category than expected, the child's datapoint would be plotted above the diagonal; to the extent that a child's vocabulary contains words that are below the diagonal, they are reported to produce fewer words in that category than expected.

We limit our analysis to traditional WS and WG forms (along with variants in these classes) because short forms like the British English TEDS don't typically include category information. The sample sizes included in this analysis are given in Table 11.2.

Number of CDI administrations in every instrument included in these analyses.

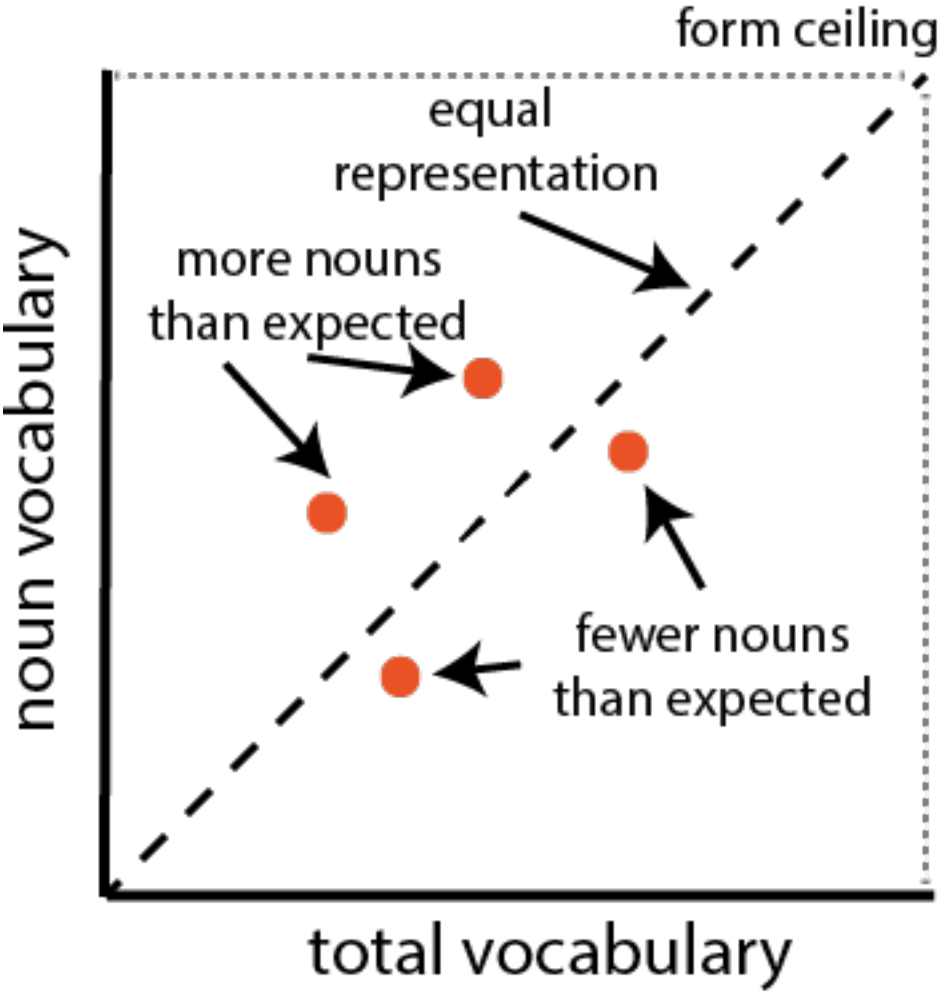


Figure 11.2: Schematic of our vocabulary composition analysis.

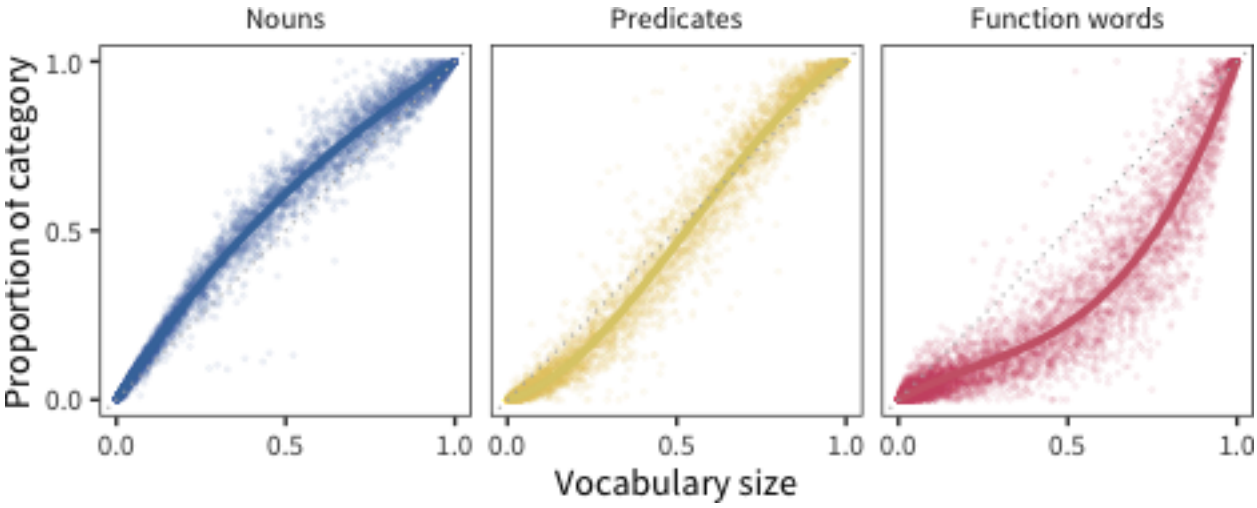


Figure 11.3: For American English WS data, proportion of each lexical category produced by each child as a function of the proportion of all vocabulary items produced by that child. Lines show model fits.

Language	Form	N
British Sign Language	WG	161
Cantonese	WS	987
Croatian	WG	250
Croatian	WS	377
Czech	WS	493
Danish	WG	2398
Danish	WS	3714
English (American)	WG	2454
English (American)	WS	5846
English (Australian)	WS	1520

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Figure 11.3 shows this analysis, carried out with English (American) WS data. Each point shows an individual child's vocabulary, and each panel shows a different lexical class (thus each child is represented once in each panel). The curves show the relationship between a class and the whole vocabulary. We capture the overall trend in this plot by estimating a linear model over the data, predicting category proportion as a function of total production. This model is fit with third-order polynomials (so as to allow both concave/convex functions and also changes in convexity). We fit these models with the constraint that they must predict the point [1,1] so that they are guaranteed to arrive at the diagonal point in the special case that all words on a form are checked. These model fits are shown by the lines.

The final step in our method is to capture the overall bias in a particular sample by estimating the difference in area between the curve and the diagonal. If the curve is substantially above the diagonal, this difference will be positive (indicating e.g., a positive noun bias). In contrast, if the curve is below the diagonal, the difference will be negative. To capture uncertainty in this area

estimate, we conduct a resampling analysis where we randomly resample the population 1000 times with replacement, then recompute the area measurement. Confidence intervals below are based on this resampling procedure.

Critically, this analysis controls for a number of confounds in previous analyses. First, because our interest is in the shape of the overall curve, under-representation of children in some age-band should add uncertainty but not bias. Of course, if data are too sparse, estimates will be unconstrained, but particulars of age sampling should not bias our estimates. Second, in principle, the analysis should not be biased by the number of items in a particular category, as the analysis is relative to the numerical representation of a particular class on the form. Thus, in principle we should be able to compare across forms with larger or smaller numbers of items in particular sections.

11.3 Results

We present results of this analysis across languages, beginning with comprehension on WG-type forms and moving to production in WS-type forms. We do not analyze WG production data here. For the most part, production estimates on WG forms are quite low, and hence curves are relatively unconstrained (or determined by a small number of children who are reported to have very large early vocabulary sizes).

11.3.1 Comprehension (WG)

Comprehension results are shown in Figure 11.4 Overall, the largest trend that is visible in these plots is the under-representation of function words, with nouns and predicates appearing quite close to one another. For further comparison, we show summaries of curve areas for each language in Figure 11.5.

Nouns are over-represented in many — but not all — languages. Portuguese, Turkish, Korean, and Slovak, a set of typologically- and culturally-distinct languages, show slight under representation, with BSL and British English showing the largest over-representation of nouns. Predicates are under-represented in some languages and over-represented in others.

As seen in Figure 11.6, there is a strong anti-correlation between noun and predicate bias measures ($r(21) = -0.83$). (This correlation should be interpreted with caution as nouns + predicates + function words are constrained to sum to 1, so some degree of correlation is built into the measure).

Function words are substantially under-represented across nearly every language in our sample (except Slovak). Note the scale difference — the function-word values are far more extreme than the noun and predicate values. These results likely reflect some combination of true under-representation of function-words as well as the difficulty of reporting on function-word comprehension in very early language (see Chapter 4 for more details on this issue). Such issues may also vary across cultures, languages, and administration methods. Function-word representation might plausibly differ due to linguistic factors such as morphological complexity, pronoun dropping, agreement, etc. However, it is also notable that the two lowest function-word scores come from Kiswahili and Kigiriama (Alcock et al., 2015), a study in which the parents may have had substantially less meta-linguistic awareness as a result of many being illiterate.

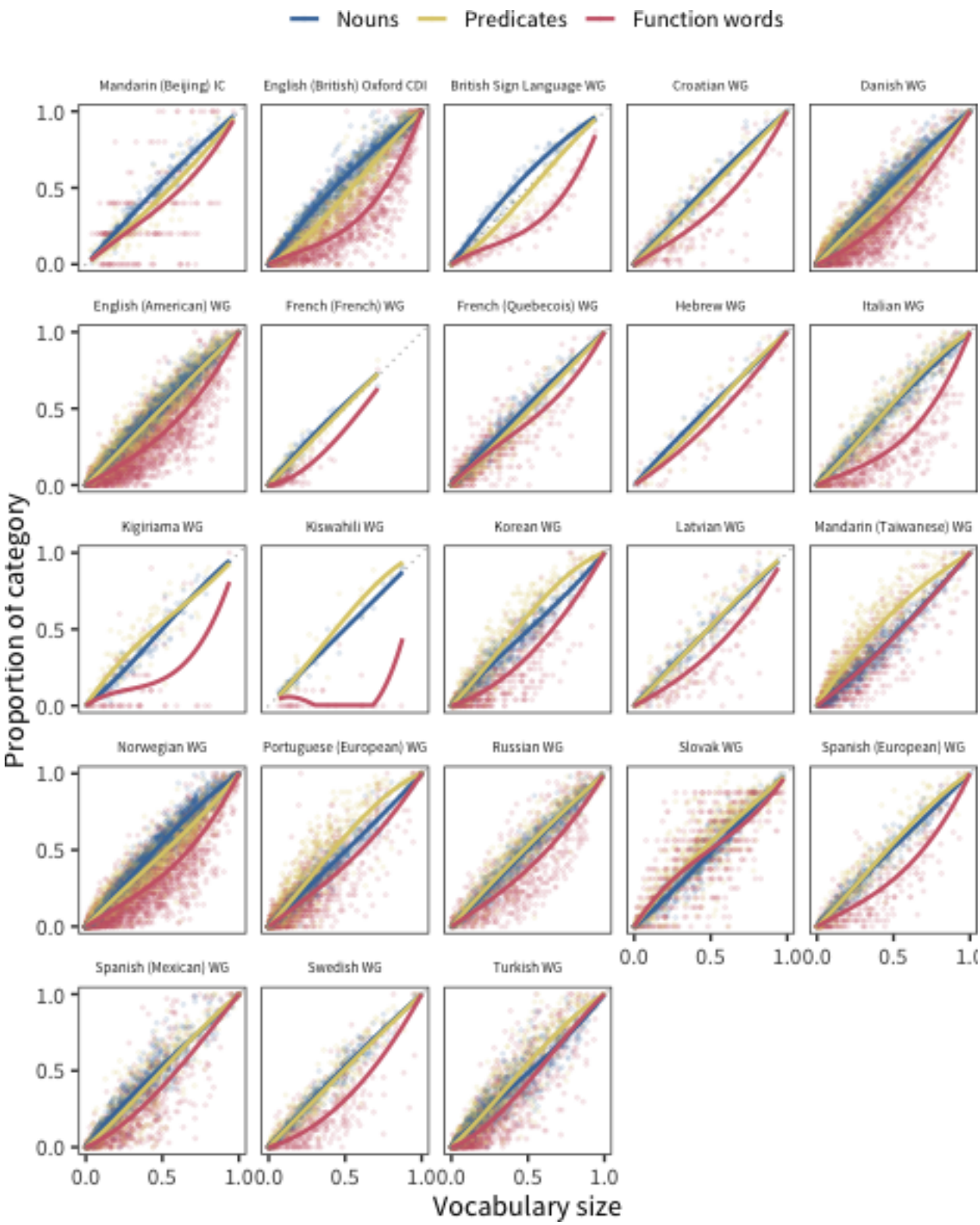


Figure 11.4: For each language's comprehension data, proportion of each lexical category produced by each child as a function of the proportion of all vocabulary items produced by that child. Lines show model fits.

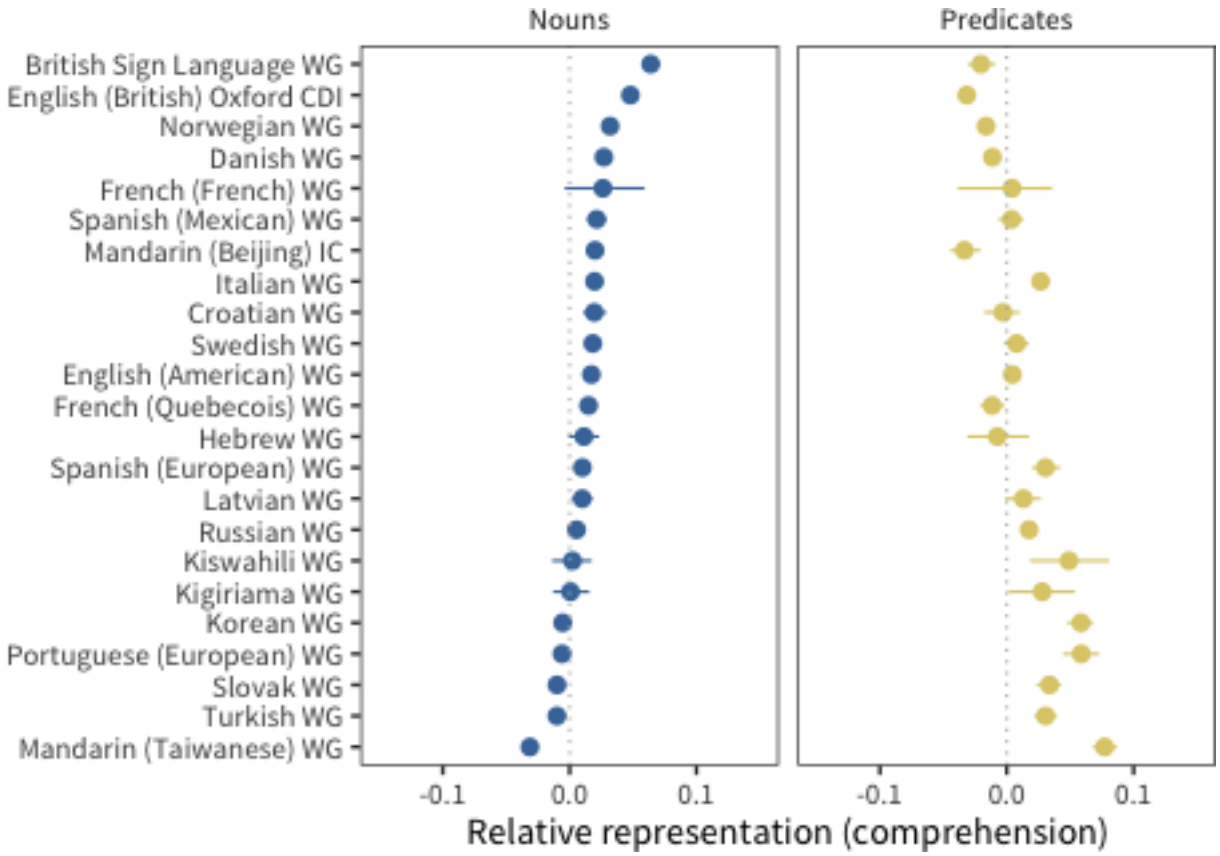


Figure 11.5: Relative representation in vocabulary compared to chance for nouns and predicates for comprehension data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

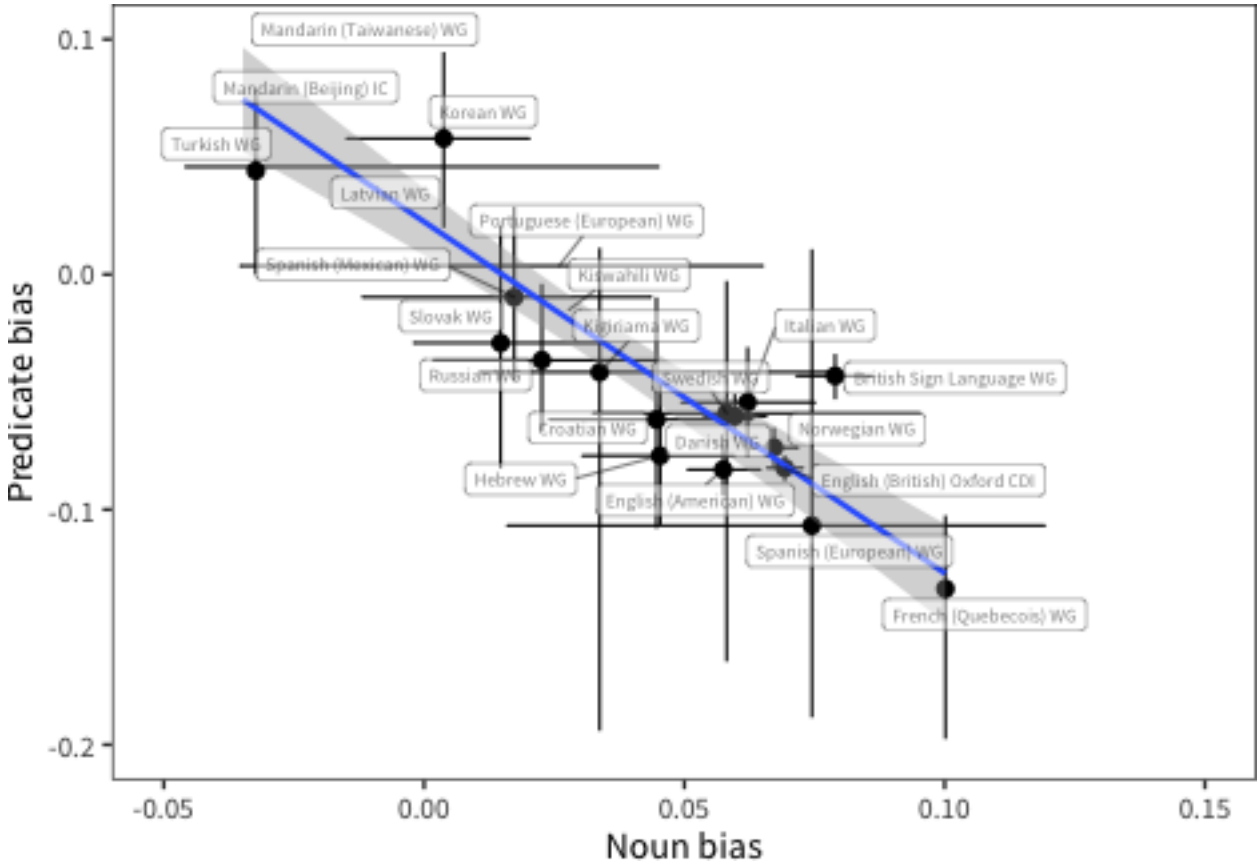


Figure 11.6: Relative representation in vocabulary of predicates compared with nouns for comprehension data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

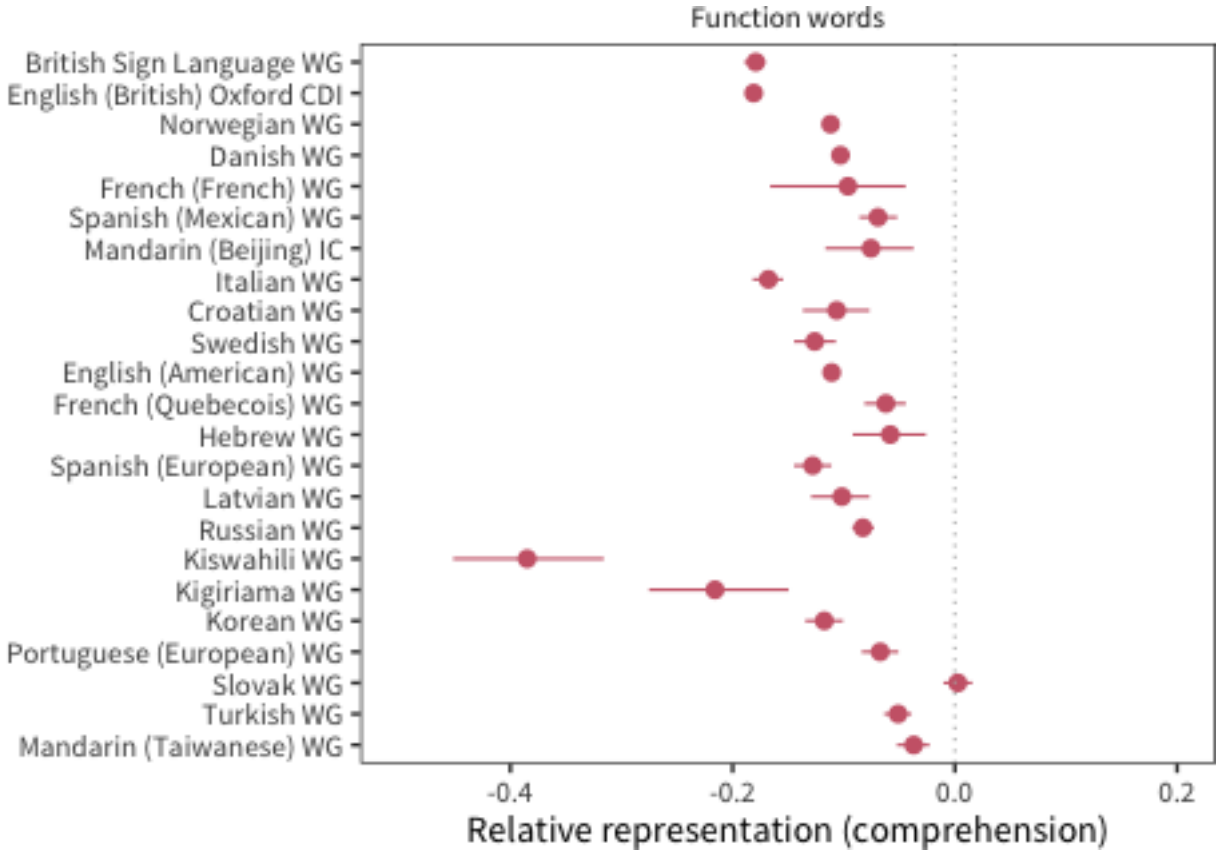


Figure 11.7: Relative representation in vocabulary compared to chance for function words for comprehension data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

11.3.2 Production (WS)

We next turn to production data from WS-type forms (Figure 11.8). We can immediately see the same function-word trend as was visible in comprehension. In addition, in many but not all languages, a noun bias is evident.

Turning to the language summaries (Figure 11.9), we see a larger pattern of variation in nouns and predicate representation. Every language has a relative over-representation of nouns, though the degree of this over-representation varies, with German and Korean especially high, and Mandarin and Cantonese especially low (we return to this trend below). Overall this trend is more extreme and more consistent in production data than comprehension. Predicate representation is both more variable and more negative than we observed for comprehension. Here we see Mandarin and Cantonese are the only two languages with substantial over-representation for predicates.

Noun and predicate representation is again anti-correlated (Figure 11.10), at r(25) = -0.65, though this correlation is somewhat weaker than for comprehension.

The negative representation of function words (Figure 11.11) is relatively consistent in overall magnitude with that seen in comprehension. Across all languages, children are reported to produce fewer function words than would be expected by chance sampling.

11.3.3 Reliability of bias estimates

One natural question is how consistent these estimates of bias are across comprehension and production. Figure 11.12 shows — for the sample of languages in which we have data from matched WG- and WS-type instruments — the relative bias we recovered in the analysis above. Somewhat surprisingly, correlations between these different instruments are quite low. Function word bias is negatively correlated between production and comprehension (r(19) = -0.39, p = 0.078). This result is likely due to Kiswahili and Kigiriama, discussed above, the lowest points for function word comprehension. But predicate bias estimates are close to uncorrelated with one another (r(19) = 0.031, p = 0.89), and the correlation between noun bias estimates is modest though positive (r(19) = 0.46, p = 0.034).

This analysis is conducted with only 19 languages and hence is relatively low power (despite the many thousands of children necessary to carry it out). Despite this, it raises some important questions. A number of explanations of the data are consistent:

1. Bias differs between comprehension and production, such that differences are related to type of response.
2. Estimates of bias are influenced by the composition of specific forms, so much so that WS- and WG-type forms yield radically different estimates of bias.
3. Bias differs developmentally. Perhaps different biases are evident earlier vs. later in acquisition.

We assess each of these explanations in turn.

In order to assess comprehension/production differences as a source of bias, we examine the Oxford CDI (shown in Figure 11.13), which is relatively unique in that it includes comprehension questions even later in development. Data on production from standard WG forms is simply too sparse to perform our bias assessment method; since most children do not produce half of the words on the

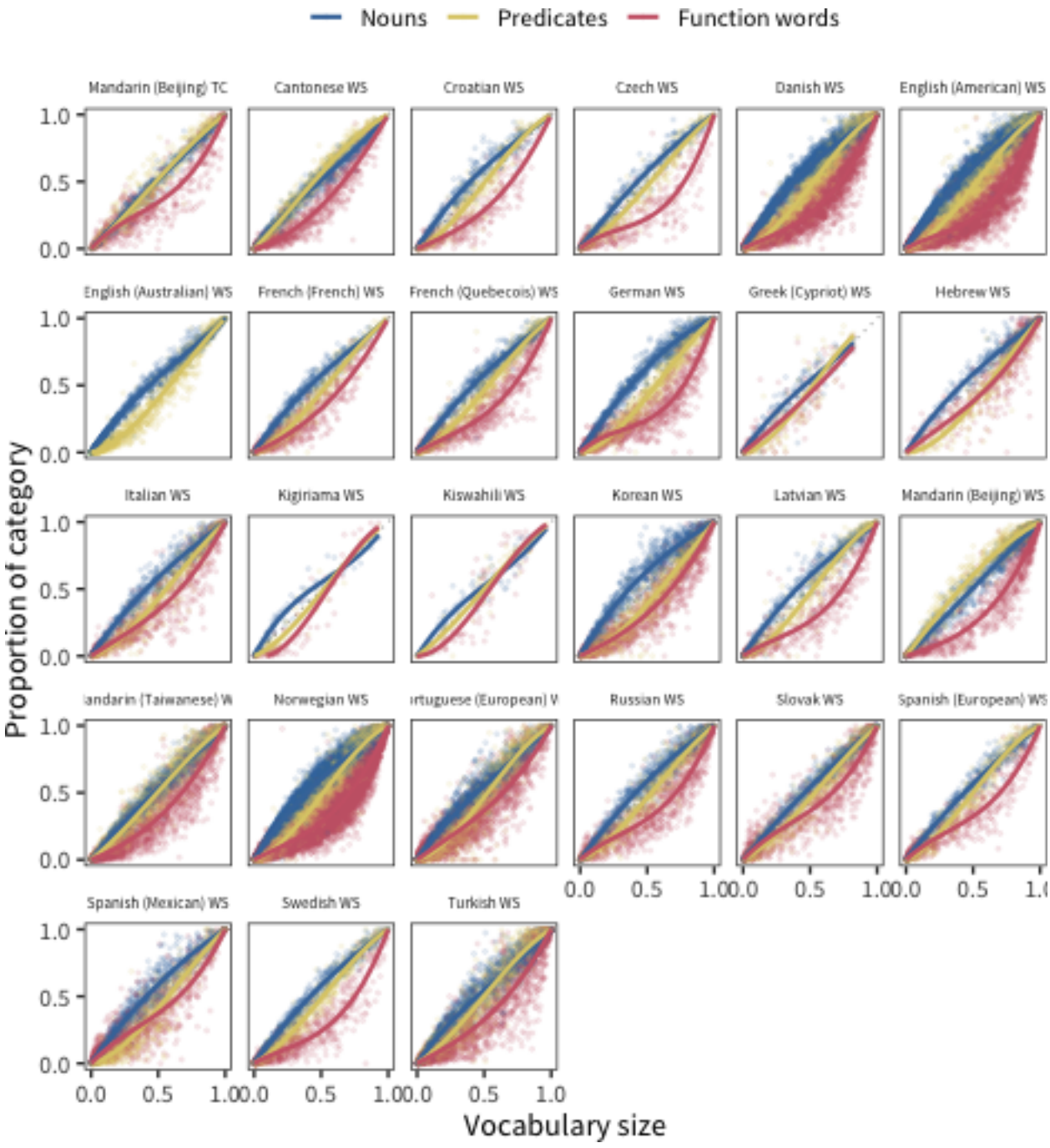


Figure 11.8: For each language's production data, proportion of each lexical category produced by each child as a function of the proportion of all vocabulary items produced by that child. Lines show model fits.

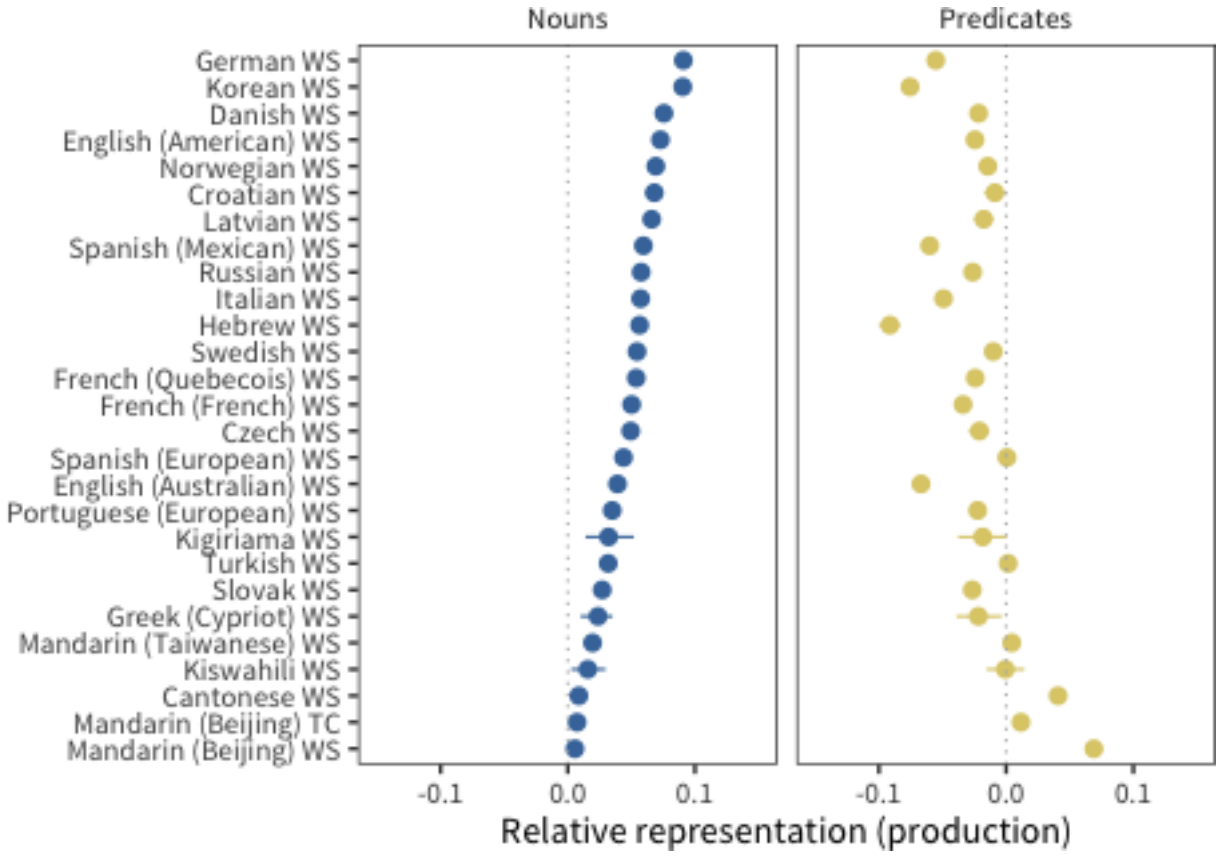


Figure 11.9: Relative representation in vocabulary compared to chance for nouns and predicates for production data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

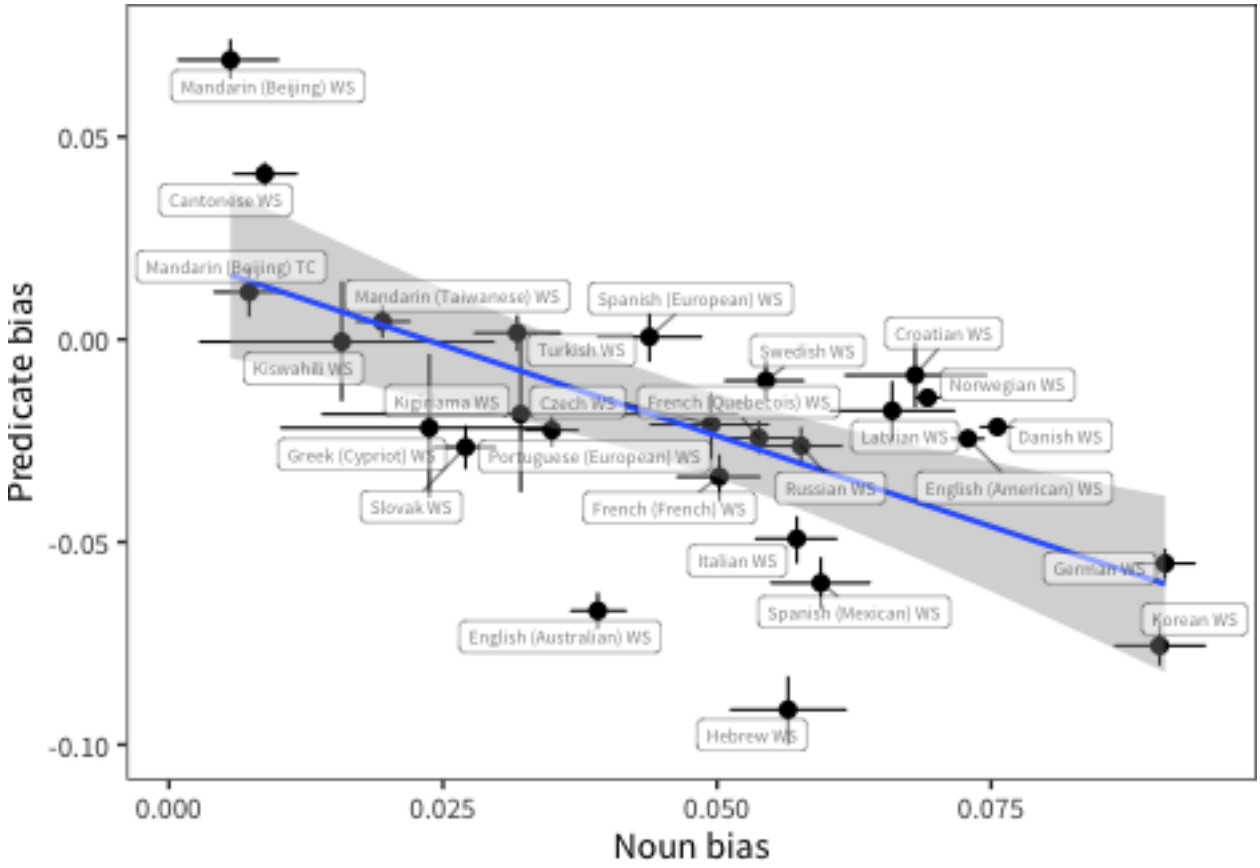


Figure 11.10: Relative representation in vocabulary of predicates compared with nouns for production data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

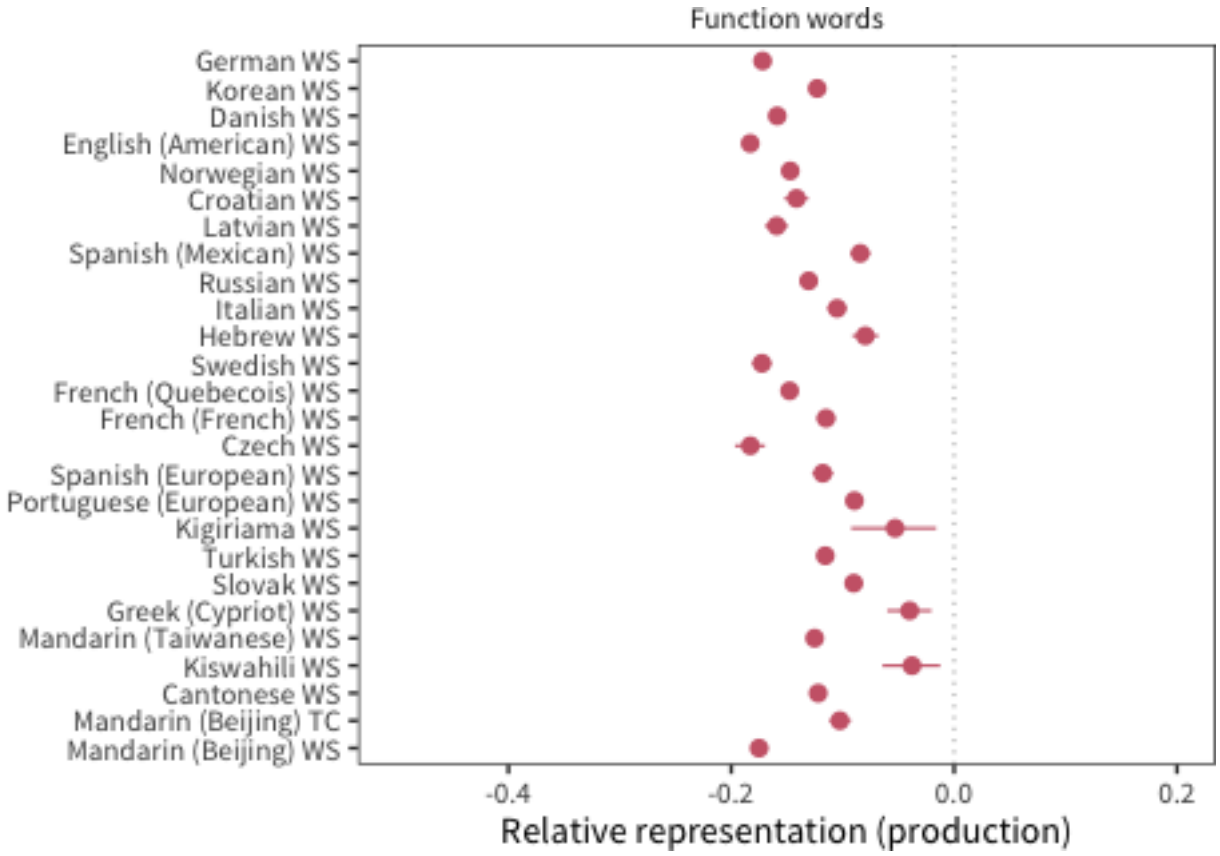


Figure 11.11: Relative representation in vocabulary compared to chance for function words for production data in each language (line ranges indicate bootstrapped 95 percent confidence intervals).

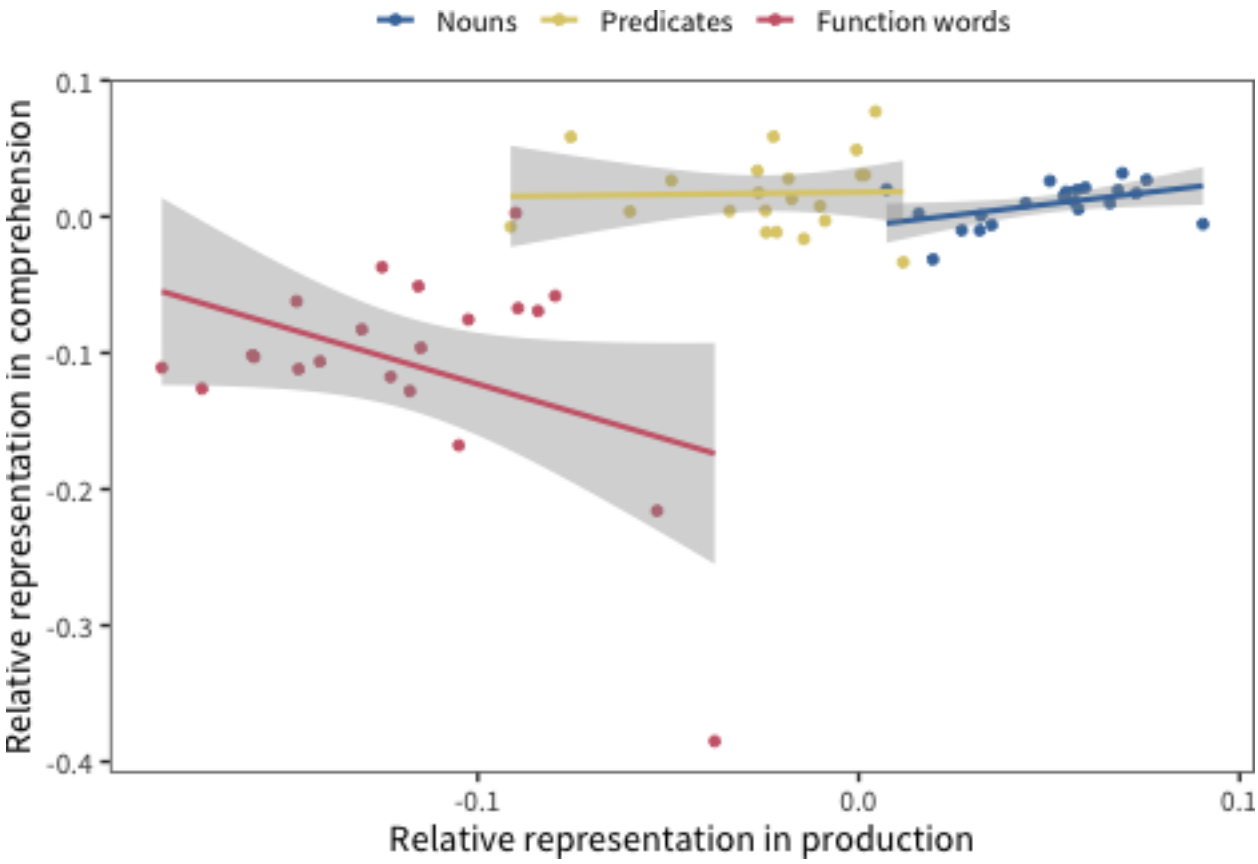


Figure 11.12: Relative representation in vocabulary for each lexical category for comprehension data compared to prooduction data in each language (lines indicates linear regression fits).

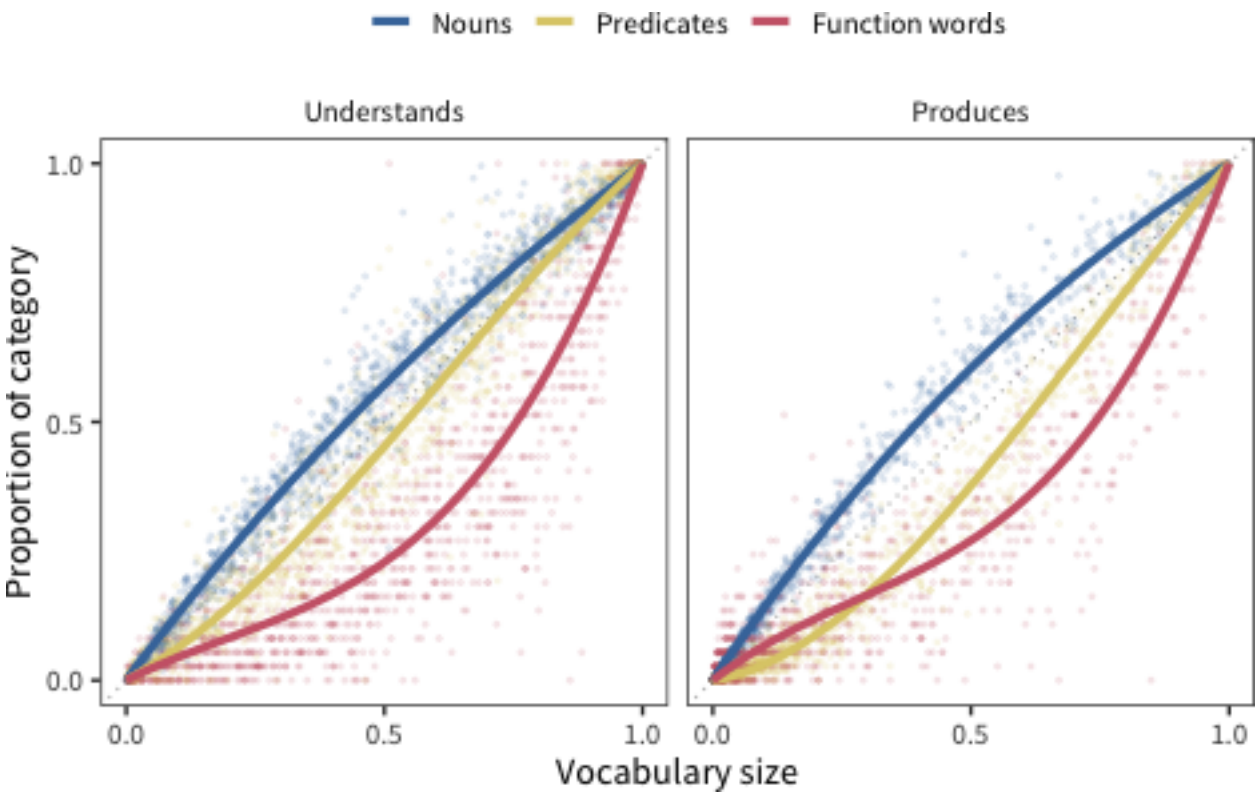


Figure 11.13: For Oxford CDI data, proportion of each lexical category produced by each child as a function of the proportion of all vocabulary items produced by that child. Lines show model fits.

form, the shape of the bias curves is driven primarily by older children. In contrast, for the Oxford CDI, it is possible to compare directly across younger and older children. The measured noun bias for comprehension is 0.048; for production it is 0.069. These values for predicates are -0.032 and -0.082, respectively. These values are somewhat similar to one another, but they do vary beyond the average confidence interval on each (+/- 0.0054). Thus, there is some evidence for production/comprehension asymmetries.

What might be the mechanism for these asymmetries? For languages that do not allow argument dropping, producing a predicate typically requires producing other words as well — English-speaking children do not often produce bare predicates, for example. Thus, predicate production may be limited by other constraints on production in such languages; in contrast, predicate comprehension has no such limits. Further, felicitous predicate production requires some ability to combine words syntactically; in contrast, comprehension of predicates (especially verbs) can often be accomplished by guessing based on known arguments (e.g., Gillette et al., 1999). For these reasons, there may be a greater bias against predicates in production compared with comprehension. This explanation is consistent with our data, in which the average production predicate bias is -0.021, while the average for comprehension is 0.013. Thus, production comprehension asymmetries likely explain some part of the differences we observed above.

Another potential explanation is that form composition relates to bias estimates. For example, a form with more predicates might actually show a lower degree of predicate bias — more predicates on the form would imply that some of those predicates are relatively more difficult (simply because the form designer had "run out" of easy predicates) and hence would not be checked as frequently

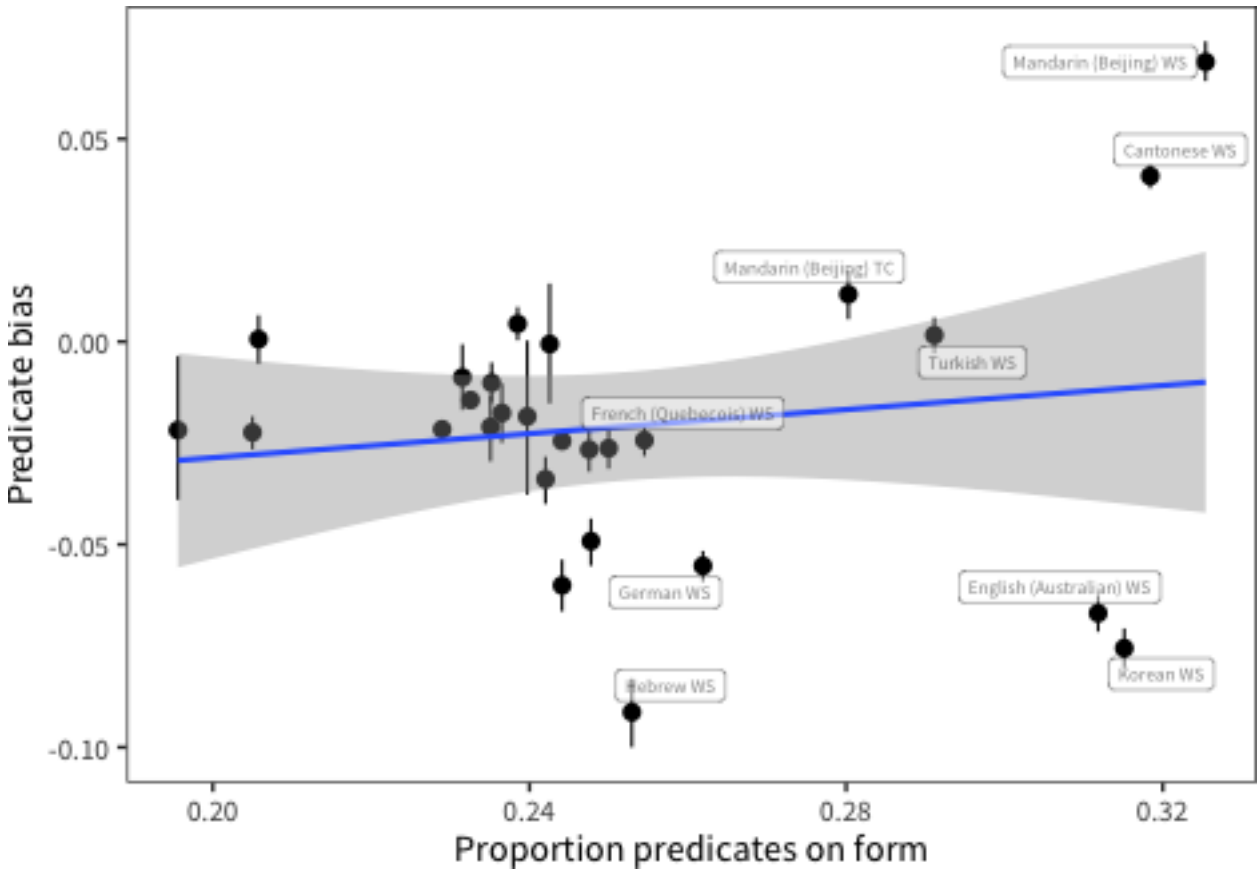


Figure 11.14: Relative representation in vocabulary for predicates as a function of proportion of predicates on form for comprehension data in each language, with languages labelled whose proportion of predicates is greater than 0.25 (line ranges indicate bootstrapped 95 percent confidence intervals).

by parents. We assess this hypothesis in two ways. First, we examine the relationship between predicate bias and predicate representation (making use of predicates because they are a minority on the form). Second, we consider the case of Mandarin where we have data from two forms with different compositions.

As shown in Figure 11.14, there is no reliable relation between the proportion of predicates on a form and the predicate bias that is demonstrated ($r(25) = 0.15$, $p = 0.45$). Thus, a simple relation between form composition and bias is not supported. On the other hand, it does appear that there is greater variance in this area for those languages with larger numbers of predicates on the form, and those languages with the highest predicate representation do have the highest number of predicates on the form as well. Perhaps the causality is reversed: Greater numbers of predicates have been included in forms for languages like Cantonese, Mandarin, and Korean where the predicate bias is an open theoretical question (or where the acquisition of predicates is of special interest).

The existence of two different forms for Mandarin opens the possibility of a further, more direct test of this issue. The Mandarin WS form (Tardif et al., 2009) and the Mandarin TC (Toddler Checklist; Hao et al., 2008) are completely independent forms but represent large datasets collected on Beijing Mandarin specifically. The Mandarin WS form is 33% predicates and shows overall a 0.07 predicate preference, while the TC form is 28% predicates and shows overall a 0.01 predicate

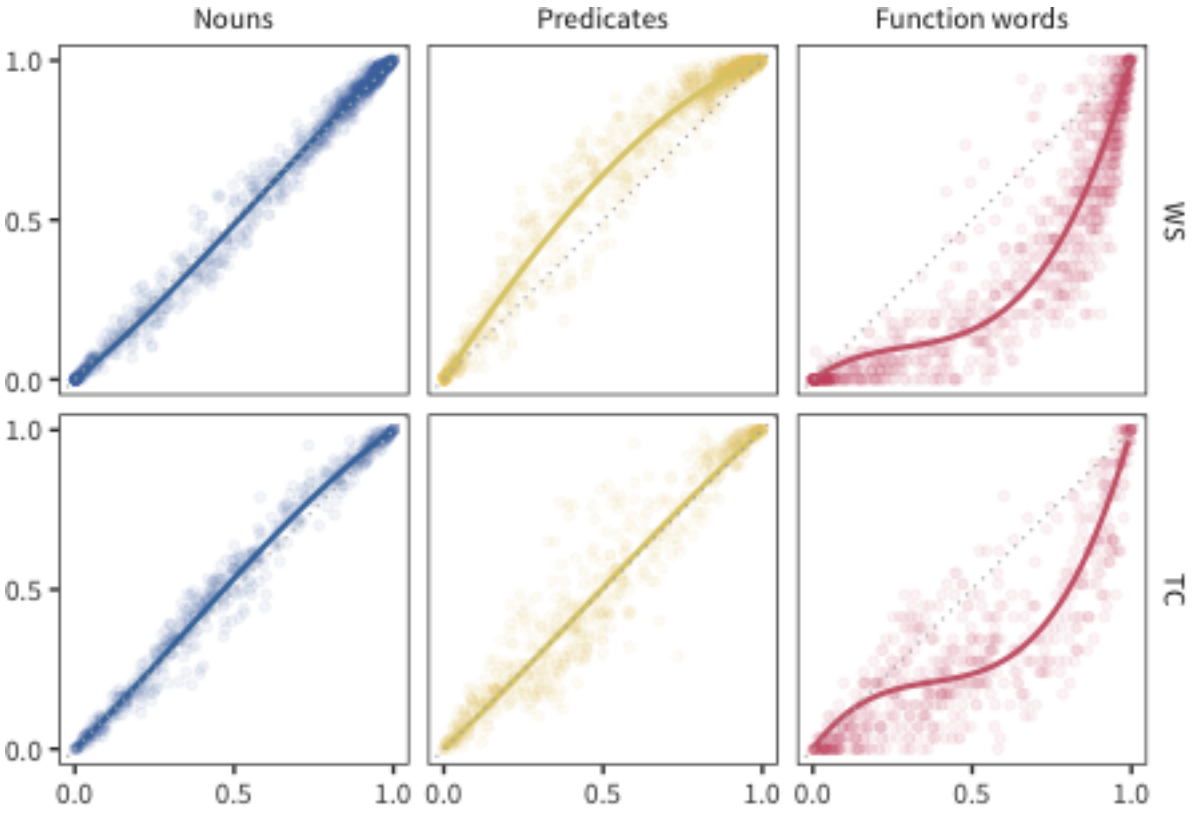


Figure 11.15: For Mandarin WS and Mandarin TC data, proportion of each lexical category produced by each child as a function of the proportion of all vocabulary items produced by that child. Lines show model fits.

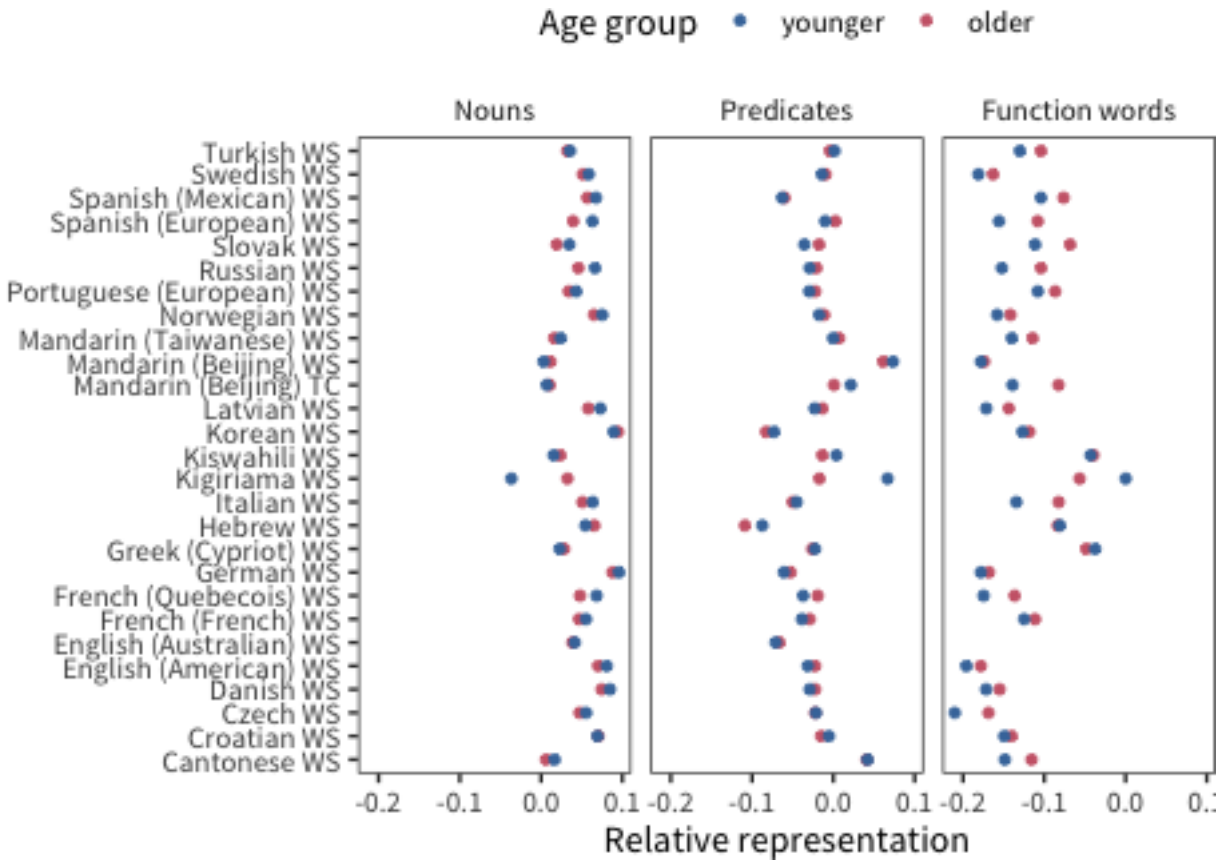


Figure 11.16: Relative representation in vocabulary for each lexical category per age group for production data in each language.

preference. The intersection of these forms yields 330 items, with 30% predicates. Interestingly, the predicate representation for the TC and WS samples, analyzing only shared predicates still differs, if anything more substantially: for WS it is 0.095 and for TC, 0.0036.

This result is worrisome — these are samples from the same city and using the same items. They even include very similar age ranges: 16–30 month-olds and 17–30 month olds respectively, with approximately uniform sampling. The suggestion is then that differences in predicate bias can be substantial based on relatively minor details, such as specifics of administration or form context or specifics of sample composition. Despite the apparent stability of these estimates under resampling (confidence intervals for bias estimates are around +/- .005 in the analysis above), we should be cautious in over-estimating our degree of certainty in particular bias estimates. Further, these data provide more data against the notion that form composition (or at least the specific sample of predicates being assessed) is the primary determinant of bias.

The final hypothesis that we examine is that there are developmental differences in bias. Such differences would help to explain the observed differences between bias in early comprehension and later production. To address this question, we split data from each language and form into older and younger groups at the median of the data for that sample. We then recomputed our bias estimates, shown in Figure 11.16. There was a developmental difference such that older children showed less of a negative function word bias, but differences in noun and predicate bias were very slight for most languages. Thus, overall we do not see evidence that bias estimates for nouns and verbs are globally

different for older vs. younger children.

In sum, we did not find strong support for effects of form composition or age on bias estimates. Each of these factors, of course, could contribute in part to the mismatch between WG comprehension and WS production estimates, but neither one was particularly strong. On the other hand, data from the Oxford CDI suggested some differences in bias estimates between comprehension and production on the same form, with the bias against predicates and for nouns being substantially more pronounced for production than for comprehension. Further, data from two different Beijing Mandarin datasets suggest possible factors relating to population and administration.

11.4 Discussion

This chapter presented a comprehensive examination of the issue of biases for and against particular syntactic categories in acquisition. Building on earlier work by Bates et al. (1994), we created a quantitative measure of noun, predicate, and function word bias and examined variability in these measures across languages. Overall, a number of generalizations emerged.

- Nearly every language showed a positive bias for nouns, though the degree of this bias varied.
- Every language showed a substantial bias against function words, supporting the generalization that these are acquired much later than content words, despite their typically higher frequency (see Chapter 10). This bias was larger on average than biases in type of content words.
- In comprehension there was variability in the degree of predicate representation; in production, as has previously been reported, languages were mostly biased against predicates. There were a few notable exceptions among the East Asian languages.
- Measures of bias in production and comprehension were not highly correlated with one another, especially for predicates and function words. There are likely many causes of these, but greater predicate comprehension compared with production appears to be one likely culprit.

Chapter 12

Vocabulary Composition: Semantic

Following the approach in the previous chapter, we investigate the consistency of semantic content categories across languages. By analogy with the "noun bias," are some languages "vehicle focused" or "animal focused"? These analyses are expected to reveal cultural and linguistic differences in the specific words learned by children (perhaps due to differences in the content of their environment).

12.1 Introduction and methods

In contrast to the "noun bias" literature, where a wide variety of hypotheses have been articulated over the preceding decades, differences in content have been less explored and so these analyses are far more exploratory. Analyses of cognitive biases in the early language of international adoptees by Snedeker et al. (2012) are a notable exception, but these analyses are limited to English data due to the complexity of gathering such data. To limit the scope of our current exploration, we focus on WS-type forms and production measures, which we have reason to believe will be most reliable.

In these analyses, we take advantage of the fact that CDI forms are typically structured into semantic categories (e.g., Animals or Body Parts). As Figure 12.1 shows, while some semantic categories are shared across many instruments, there are others that are quite rare (many corresponding to specific categories that are of interest in particular languages). We focus on those semantic categories with greater representation in the data. Further, to avoid duplicating our analysis in Chapter 11, we focus on those semantic categories that fall into "nouns" and "other" lexical classes. (In general, Action Words and Descriptive Words tend to be broad predicate classes without as much clear semantic differentiation). This filtering step leaves 14 categories: Animals, Body Parts, Clothing, Food & Drink, Furniture & Rooms, Games & Routines, Household, Outside, People, Places, Sounds, Time Words, Toys, Vehicles. Samples included in this analysis are shown in Table 12.1.

Number of CDI administrations in every instrument included in these analyses.

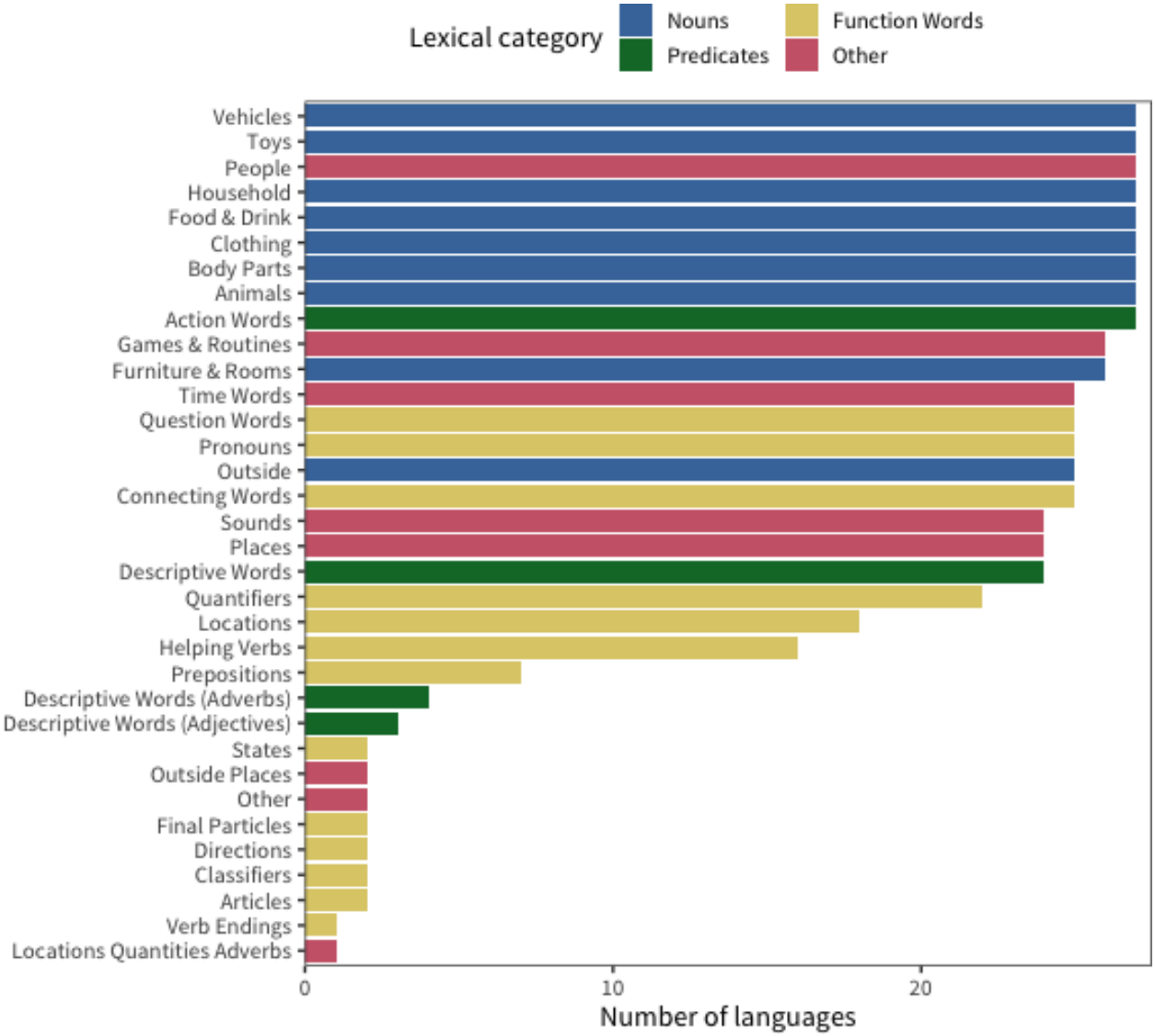


Figure 12.1: Number of languages whose forms contain each semantic category.

Language	Form	N
Cantonese	WS	987
Croatian	WS	377
Czech	WS	493
Danish	WS	3714
English (American)	WS	5846
English (Australian)	WS	1520
French (French)	WS	665
French (Quebecois)	WS	827
German	WS	1181
Greek (Cypriot)	WS	176

Showing 1 to 10 of 27 entries Previous 1 2 3 Next

We first illustrate this approach using data from the English WS form alone. Analogous to the plots in Chapter 11, Figure 12.2 shows areas where the data deviate from the pattern of category acquisition predicted by random item sampling. The size of the shaded region above vs. below the diagonal gives evidence of over- vs. under-sampling for a particular semantic category.

Many of the results of this analysis for English are expected. Sounds items are heavily over-represented, as are Body Parts, Games & Routines, and to a slightly lesser extent, Toys, Animals, and Vehicles. These particular biases are likely related to particular parenting practices, cultural emphases (for example, on animal names), and young children's' idiosyncratic interests. For a more in-depth examination of the consistencies in very early vocabulary, see Chapter 8; for more detail on what makes particular words easier or harder to learn, see Chapter 10.

The largest under-representation across categories is Time Words. This pattern is consistent with a body of work on children's acquisition of the semantics of time words that suggests that children struggle with understanding these complex terms through age five (Tillman and Barner, 2015; Tillman et al., 2017).

We next turn to how this pattern varies across languages.

12.2 Global results

Because there are so many different languages represented in this analysis, the simplest analysis examines the spread of languages across categories (Figure 12.3). Somewhat surprisingly, the ordering of categories looks quite similar to what was observed in English. Sounds, Games & Routines, and Body parts are all over-represented. Vehicles, Food & Drink, Animals, and Clothing all are variable across cultures, as is People. Small Household Items, Outside Things, and Furniture & Rooms show variability but overall less bias. Finally, Places To Go and Time Words are both under-represented systematically across all languages.

We can zoom in on the most highly over-represented categories (Figure 12.4). The highest mean comes from Body Parts, which are over-represented in just about every language. Interestingly, the three datasets with the lowest proportion of Body Parts are the two Mandarin datasets (WS and TC) and the Cantonese WS data. Games & Routines are generally over-represented but somewhat

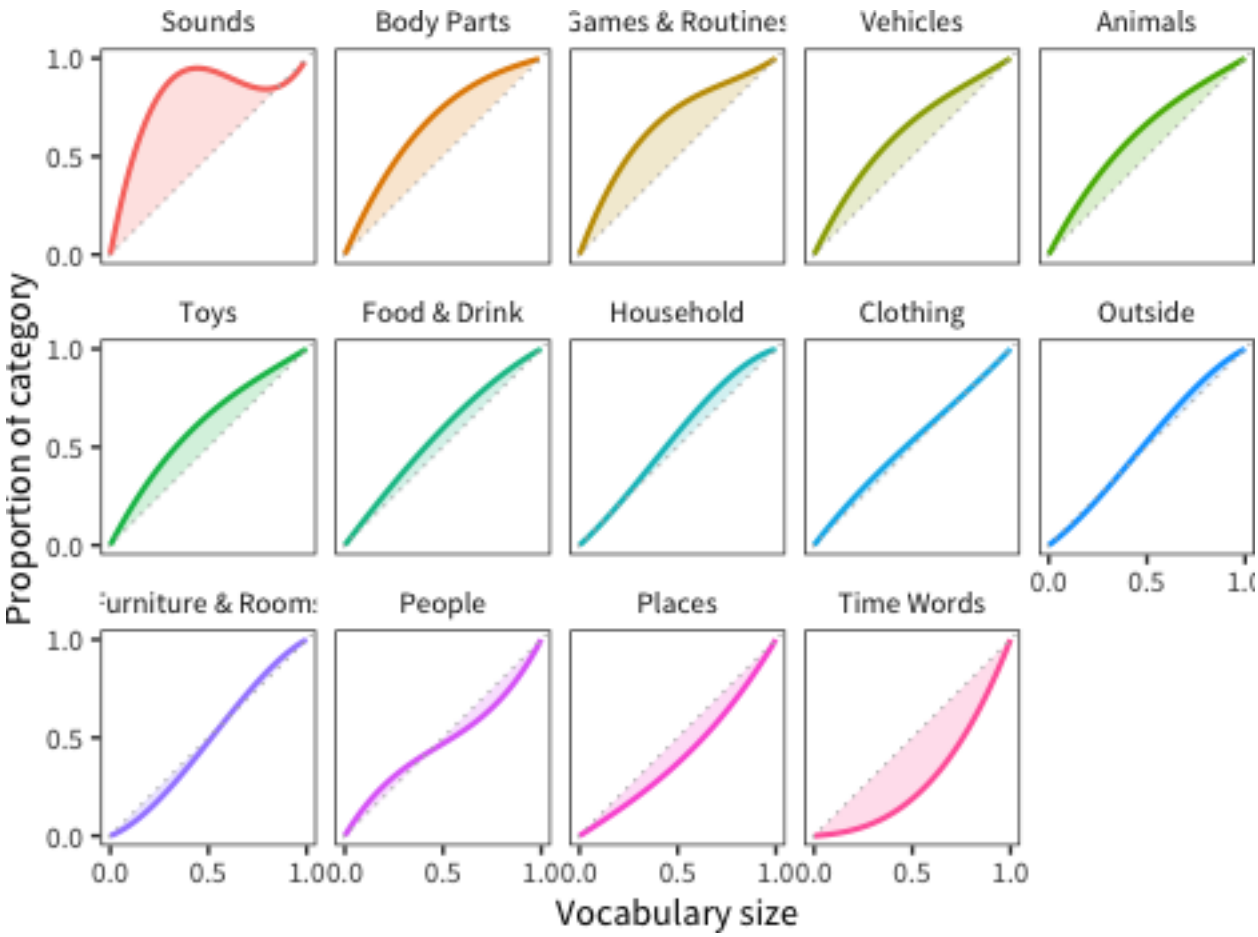


Figure 12.2: For American English WS data, model fit curves for proportion of each semantic category produced by each child as a function of the proportion of all vocabulary items produced by that child.

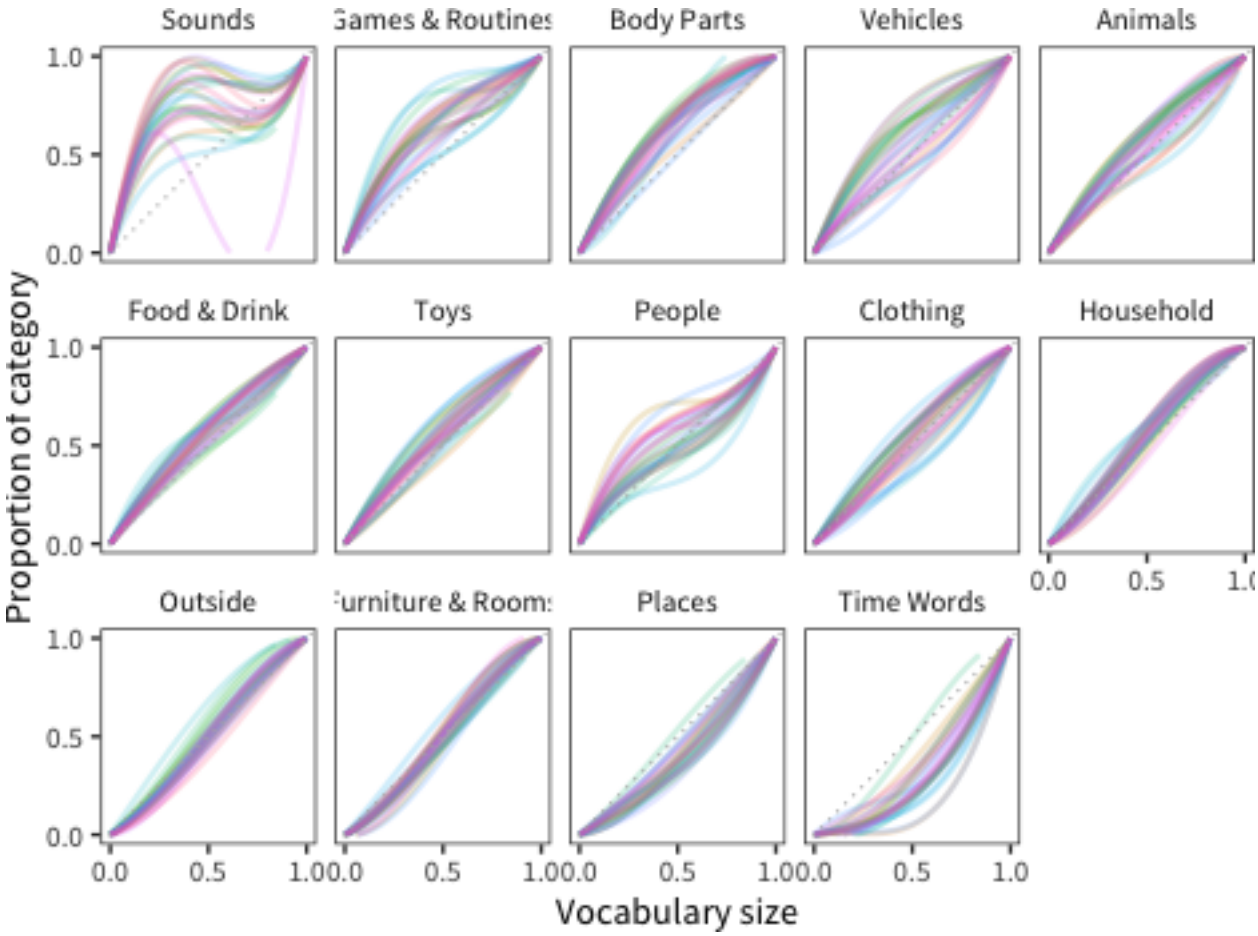


Figure 12.3: Model fit curves for each semantic category as a function of vocabulary size for each language.

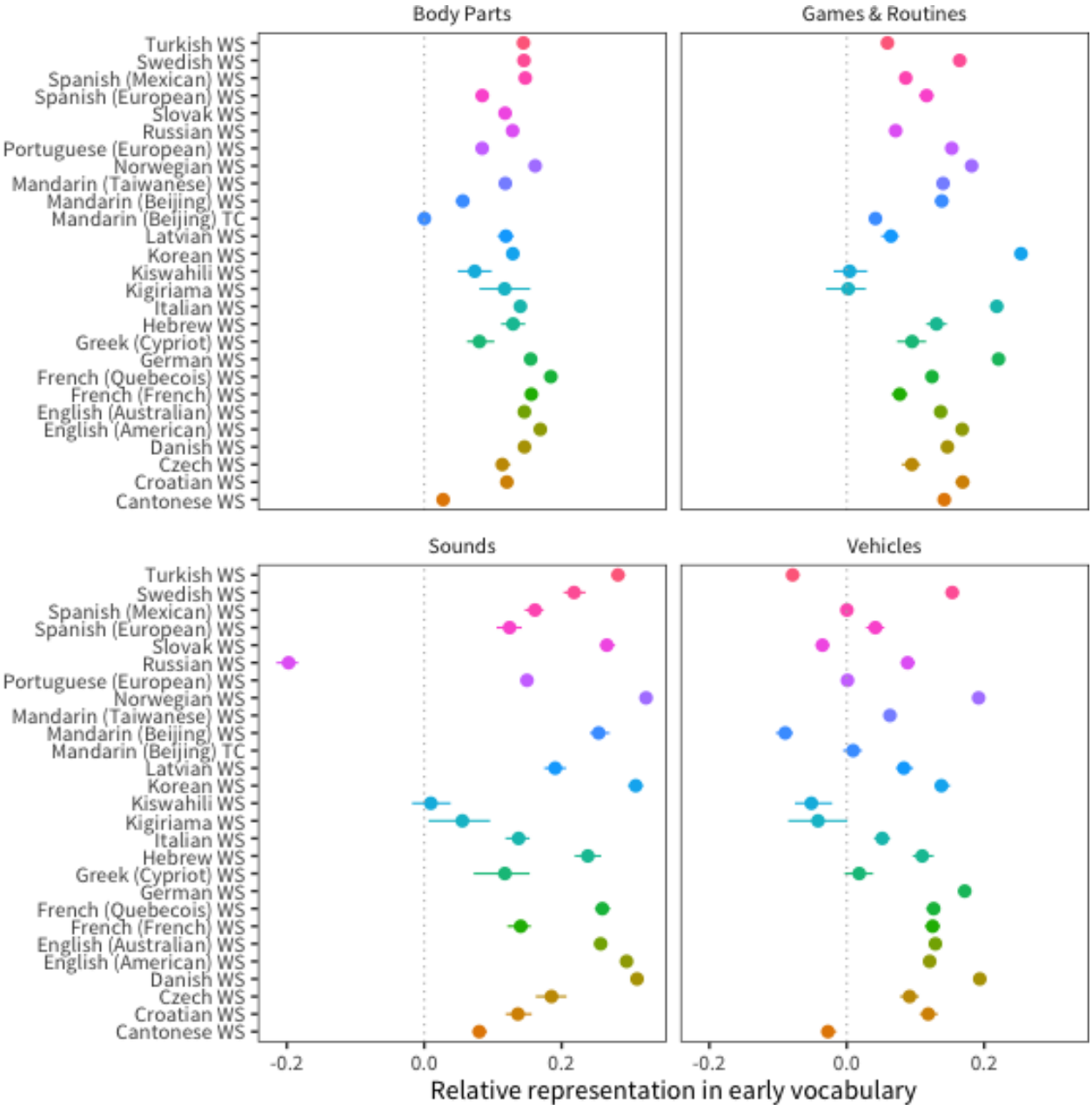


Figure 12.4: Relative representation in vocabulary compared to chance for categories that tend to be over-represented across languages (line ranges indicate bootstrapped 95 percent confidence intervals).

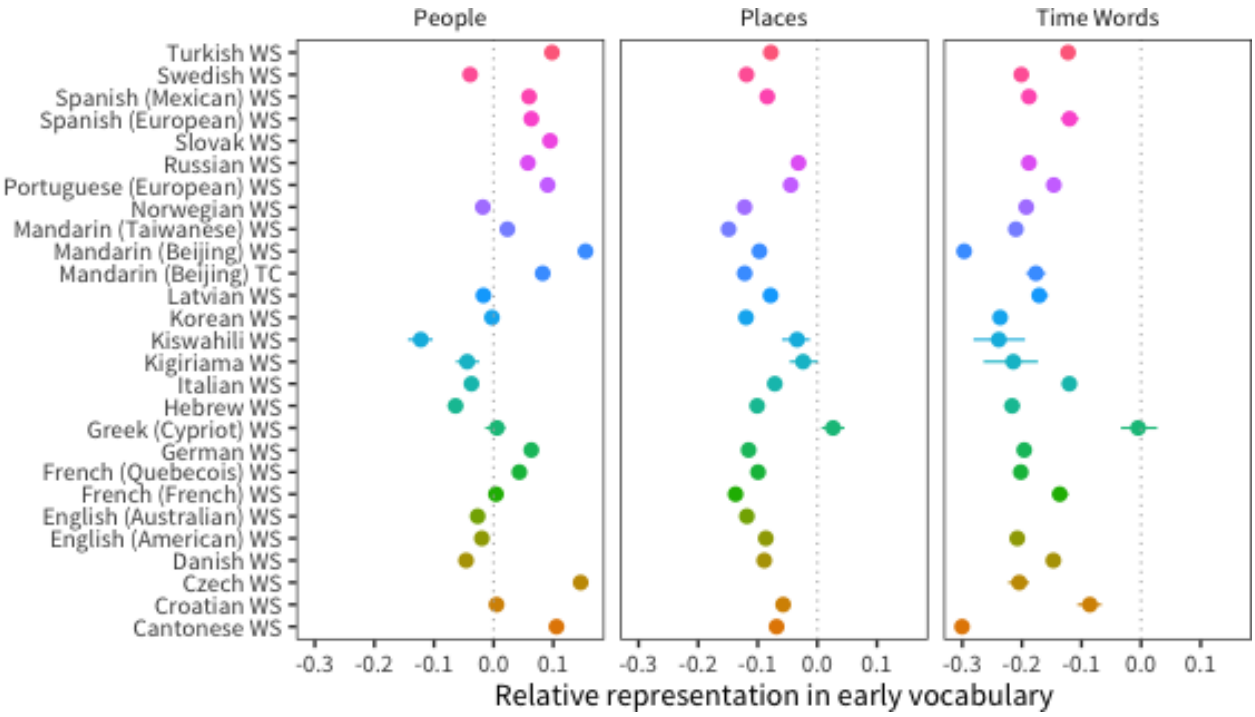


Figure 12.5: Relative representation in vocabulary compared to chance for categories that tend to be under-represented or highly variable across languages (line ranges indicate bootstrapped 95 percent confidence intervals).

more variable, with Kiswahili, Kigiriama, and Mandarin TC data lowest. Sounds are quite highly variable but almost all positive, with Russian being the outlier. Inspection of these items shows negative developmental trajectories for a number of words in the Sounds category. We believe these data are likely an artifact of parents feeling that they should "trade off" with noun labels in the Animals category, and hence items in Sounds category should be discounted. Finally, words in the Vehicles category appear more variable with positive preferences across language families.

We end by considering People, Places To Go, and Time Words (Figure 12.5). People is a highly-variable category, with some languages under-representing and others over-representing. Tardif et al. (2008) speculated that names for people were a substantial part of children's earliest words, but that may reflect that study's use of Mandarin and Cantonese data where people terms are very over-represented due to cultural emphasis on family connections. Surprisingly, despite the relatively multi-generational and family-centric nature of children's experience in Kenya (Alcock and Alibhai, 2013), people words were relatively under-represented in Kiswahili and Kigiriama.

In contrast to the heterogeneity in people words, words in Places To Go and, especially, Time Words were almost uniformly under-represented in children's vocabulary. As noted above, time is known to be conceptually difficult for children. Interestingly, though, less has been written about children's understanding of geographical vocabulary. Time words offer a number of conceptual challenges in terms of mapping an ordered set of durations (second $<$ minute $<$ hour $<$ day, etc.) to a set of concepts that do not map cleanly onto perceptual experience. In some sense, many of the same conceptual difficulties hold true for larger locational/geographical hierarchies (neighborhood $<$ city $<$ state $<$ country). Or alternatively, the under-representation of places in children's early vocabulary may simply reflect the relative lack of diversity of their experiences with some of the items that

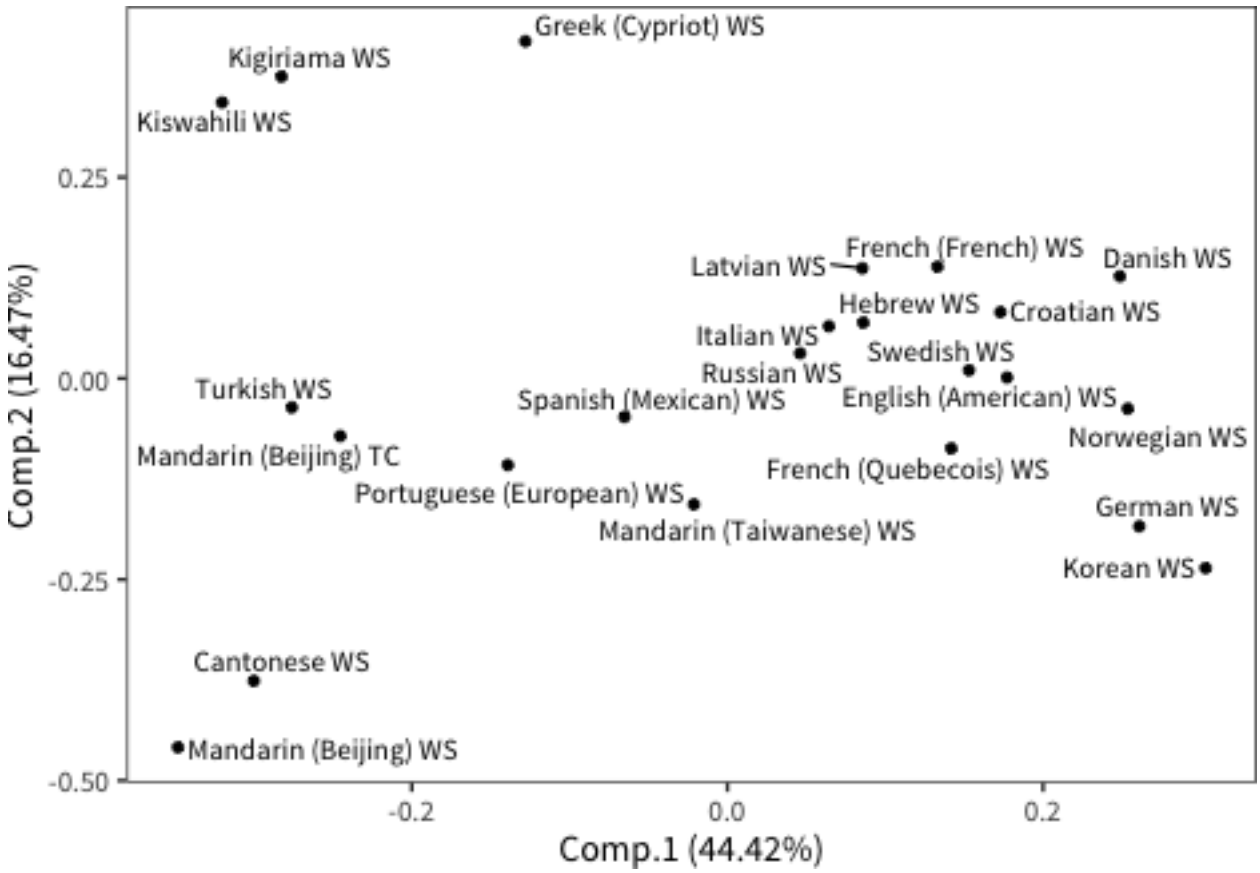


Figure 12.6: First two principal components for each language.

traditionally populate this section (e.g., beach, camping, church, circus to name the first four). See Chapter 4 for some evidence that camping especially may be variable in children's experiences.

12.2.1 Dimensionality reduction

Our next analysis of these data takes an exploratory dimensionality-reduction approach. Rather than examining each semantic category individually, we consider the space defined by variation in semantic preferences by running principal components analysis (PCA) on these data. PCA is a dimensionality reduction technique that projects high-dimensional data (e.g., bias by semantic category for each language) into a set of orthogonal dimensions where lower dimensions capture as much of the variance as possible.

Standard PCA requires no missing data, thus we removed languages with missing categories. This analysis thus includes 23 language/form combinations and 13 categories. (We exclude words from Sounds because of the issue with Russian in this category and other missing data).

Figure 12.6 and Figure 12.7 show the data projected into the space of the first two principal components and the loadings of semantic categories on these two components, respectively.

Several observations emerge: Mandarin and Cantonese WS data are very far towards the direction of people (indicating that these datasets are unusual in this respect). Second, Kiswahili and Kigiriama

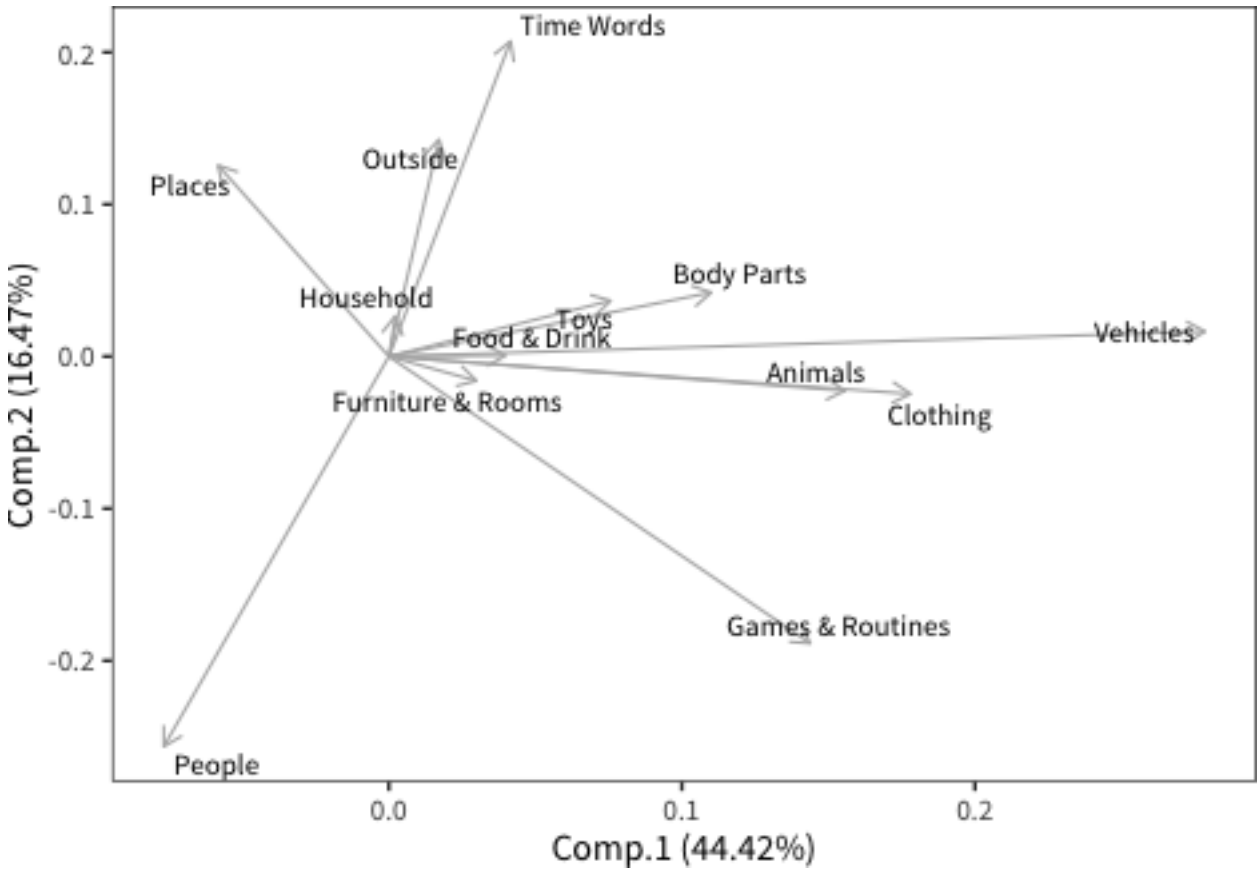


Figure 12.7: Loadings of each semantic category.

are especially far in the direction of Outside and Places To Go words, perhaps consistent with the datasets being collected in rural and semi-rural areas. Many Northern European datasets (as well as Korean) are clustered at the far left, with high scores on Vehicles, Clothing, and Animals. Overall, this analysis reveals some interesting structure, but care should be taken not to over-interpret. In particular, within culture differences (e.g., Mandarin TC vs. Mandarin WS) are as large in size as between-culture differences.

12.3 Individual conceptual items

In this section, we isolate individual items from specific domains of interest. Our approach is to use the "universal lemma" mappings (see Chapter 9.1) to find matching lexical items across languages. The specific domains we consider are time, color, body parts, and logical words. We also investigated spatial prepositions and number words, but do not include them here. Spatial prepositions present a wide variety of mapping issues since lexical items "cut up" space differently across languages (see e.g., Bowerman, 1996). Number words are not found on enough CDI forms to have sufficient data for inclusion.

12.3.1 Time

As discussed above, the semantics of time words are very challenging for children through middle childhood (Tillman and Barner, 2015; Tillman et al., 2017). Despite this, parents report that children do produce them by age 2.5. The set of words with sufficient translation equivalents for inclusion was after, day, later, morning, night, now, today, tomorrow, yesterday.

Figure 12.8 shows trajectories for these lexical items across languages, sorted by difficulty. Because night is typically signaled by darkness, it is perceptually very concrete and likely easier than other time words. Similarly, now seems relatively more straightforward given that it has a common imperative meaning in sentences like "give me that right now." In contrast, the latest-acquired is yesterday, which is highly abstract and requires a sort of "mental time travel" in thinking retrospectively beyond the "here and now" (Busby and Suddendorf, 2005). While tomorrow shares those same features, it does not appear to pose similar challenges for children in these data.

12.3.2 Color

Color word acquisition has been a focus of interest at least since early work by Carey (1978)'s influential study of "fast mapping." Although early work suggested that color words were learned almost simultaneously (Bartlett, 1977), more recent studies have described a more protracted trajectory of partial knowledge. Many children learn some color words and overextend these to cover the rest of color space (Wagner et al., 2013). Adding to the complexity of this issue is substantial cohort changes in the age at which colors are learned: while school-aged children struggled with their colors 50-100 years ago, more recently children learn colors in the age range spanned by the CDI forms (Bornstein, 1985).

There is tremendous cross-linguistic variation in color vocabulary (Kay et al., 2009). We take advantage of the fact that most of the languages in our dataset have relatively larger color vocabularies, which

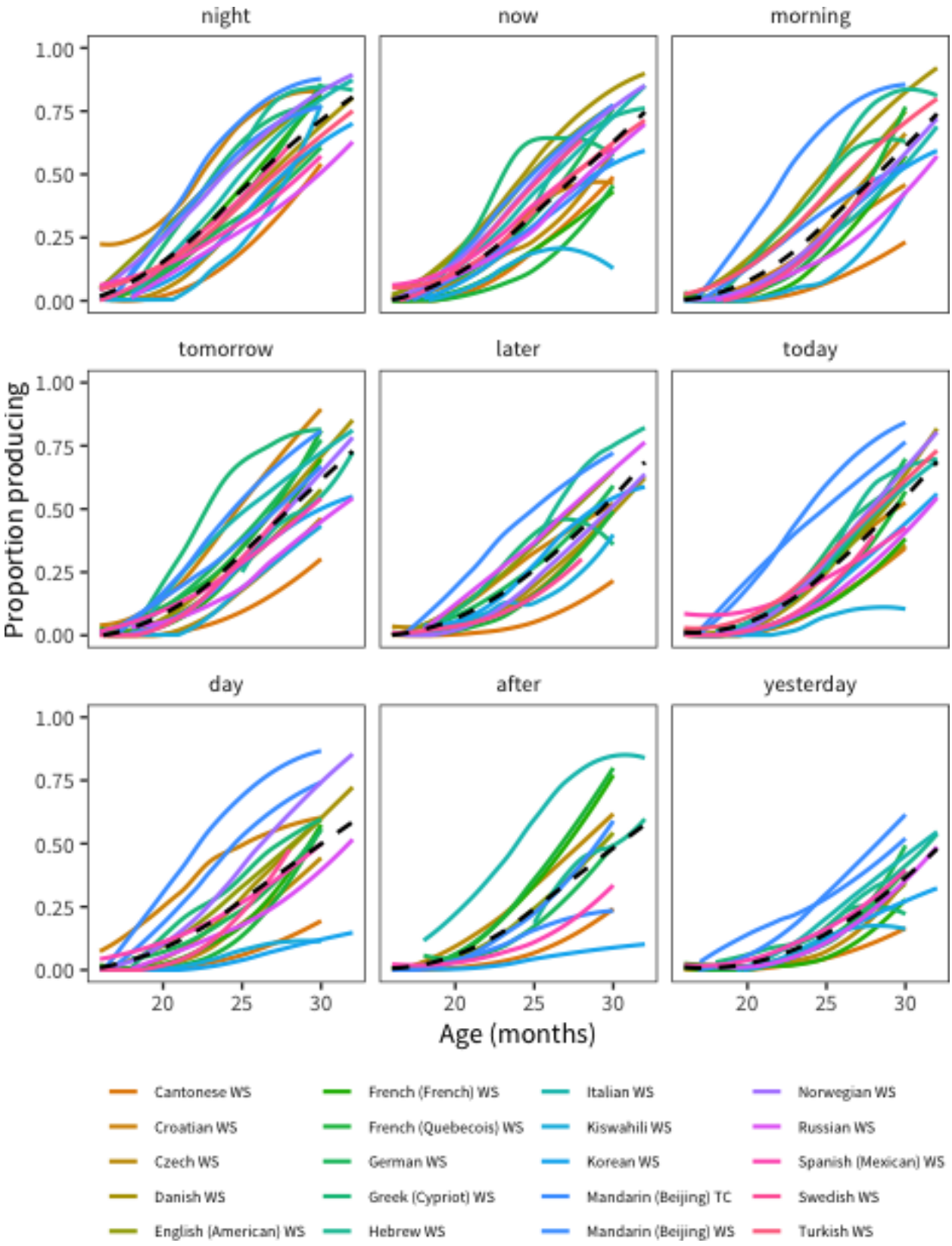


Figure 12.8: Developmental trajectory of each time word in each language.

we can assume means that individual colors probably have relatively similar extensions.[1] Despite this, most CDI forms do not include all the basic level color words. The set of color words with sufficient translation equivalents for inclusion was black, blue, green, red, white, yellow.

In this set of words (Figure 12.9), we see that red is typically the first learned, although there is substantial variability in when it is learned. It is followed by yellow, blue, and green, with black and white following behind, consistent with reports by Wagner et al. (2013).

We additionally see an ordering across languages which have higher rates of color word production reported (Figure 12.10). As in other analyses (see Chapter 5), Mandarin WS has the highest level of production. American and Australian English also tend to have high levels of color production. Interestingly, Kiswahili has by far the lowest level of color production, perhaps related to the limited availability of manufactured toys of contrastive colors (Bornstein, 1985).

12.3.3 Body parts

Words for Body Parts (Figure 12.11) are produced very early by most children, and variance is quite low across languages (with the exception of a few terms in Cantonese and Cypriot Greek). One interesting pattern that is visible in these data is the ordering of hand and foot before leg and arm, perhaps due to the fact that these particular body parts are more central to the activities in which young children engage.

12.3.4 Logic

Finally, we examine words for logical operators (Figure 12.12). The only items that are available across significant samples of languages are all, and, because, none, some, no, not. The negative words are learned early, with an ordering consistent with Bellugi (1967) and Pea (1982). No is very early, and not later. Interestingly, the quantifiers are not ordered as shown by Katsos et al. (2016) in a massive cross-linguistic study. In that study — as well as in our own work in English (Horowitz et al., 2017) — all was found to be understood better than none. In contrast, here we tend to find none is learned earlier than all and definitely learned earlier than some. One possibility is that these uses are only found in a restricted set of cases. Another is that contextualized production of negation is simpler than de-contextualized comprehension, as we have found in some of our work on the comprehension of negation in context (Nordmeyer and Frank, 2014, 2018).

12.3.5 Category variability

Finally, we quantify the variability across languages for each of these restricted sets of lexical items. For 22–26 month-olds (chosen somewhat arbitrarily to be an age range of high coverage across forms that does not encompass too much developmental change), we compute the coefficient of variation for children at each age on each lexical item. (We first average across ages and then across lexical items; reported Ns are for the average number of contributing languages). We additionally add words from the Animals category for the sake of comparison. Table 12.1 gives the coefficient of variation for each category.

[1] Such an assumption would not be warranted if we were considering languages with just a handful of color terms, in which the extension of a term like red would be much larger than in English.

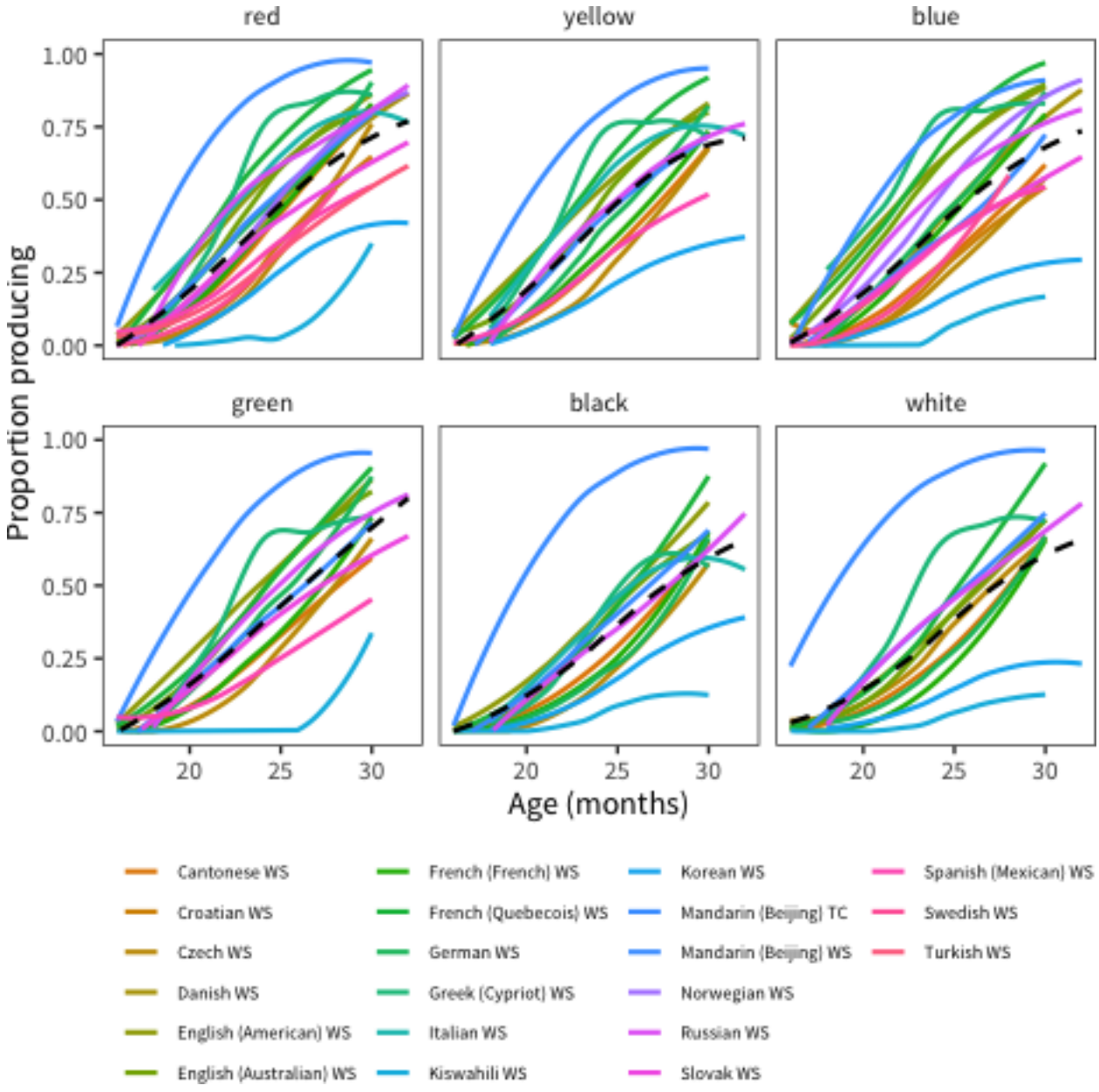


Figure 12.9: Developmental trajectory of each color word in each language.

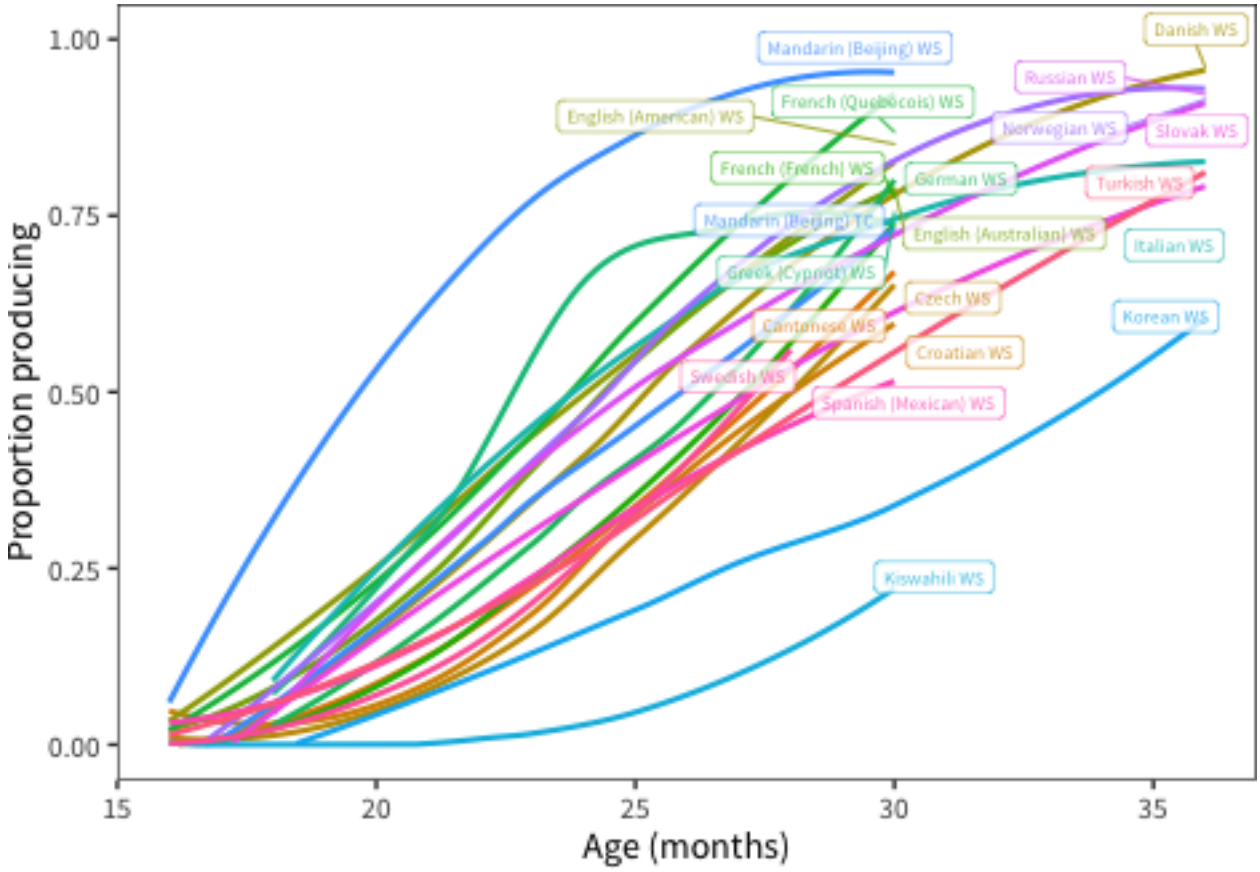


Figure 12.10: Mean developmental trajectory of color words in each language.

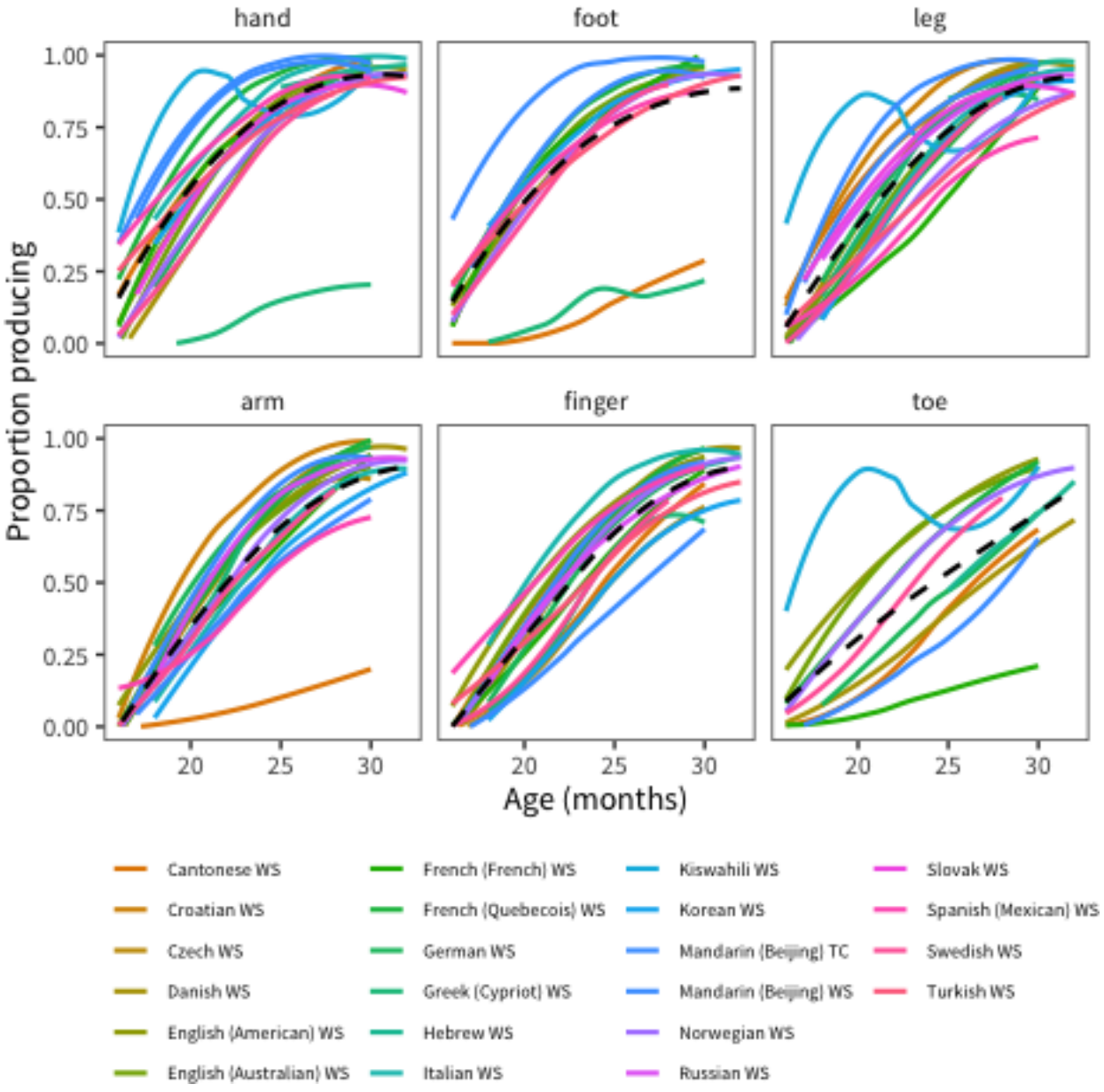


Figure 12.11: Developmental trajectory of each body part word in each language.

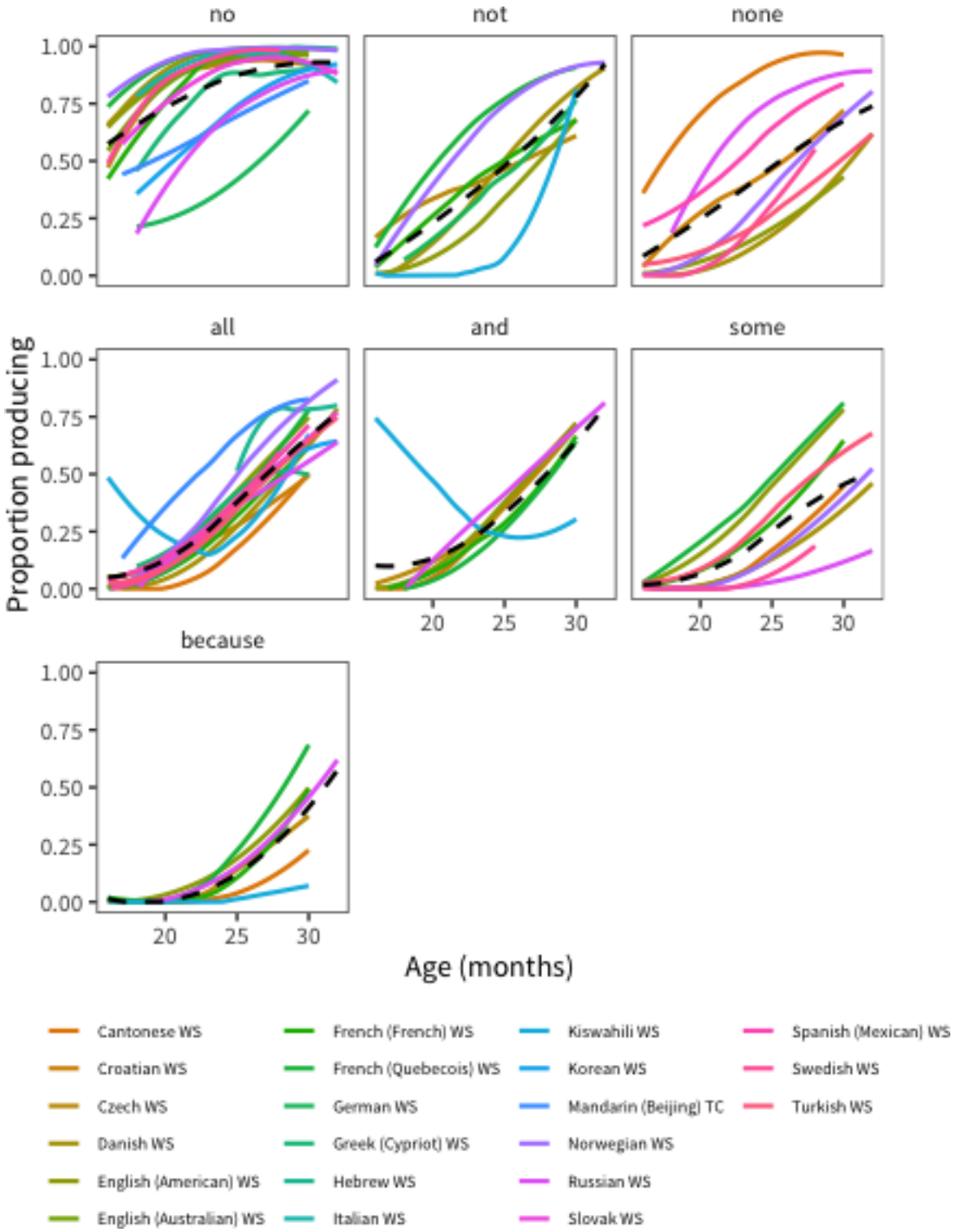


Figure 12.12: Developmental trajectory of each logic word in each language.

Table 12.1: Mean coefficient of variation for each semantic category.

Category	CV	SEM	N
Animals	0.32	0.06	18
Body	0.27	0.05	16
Color	0.47	0.09	15
Logic	0.47	0.11	11
Time	0.53	0.09	16

The acquisition of words from the Body Parts, as well as Animals, categories are highly consistent across languages. In contrast, color words, logic words, and time words are far less consistent cross-linguistically. These effects are likely somewhat affected by floor and ceiling effects, but inspection of individual items confirms the robustness of the general conclusion.

12.4 Discussion

In these exploratory analyses, we considered representation of different semantic categories across the different languages in our dataset. We found some surprising consistencies. Words from the Places to Go and Words About Time categories were under-represented, while words from the Sounds, Games & Routines, and Body Parts categories were over-represented. These consistencies were also contrasted with some areas of greater variability: for example, the preference for words from the Vehicles, Clothing, and Animals categories appeared to be a somewhat coherent dimension in our data, with many (northern) European languages higher on this dimension than non-European languages. Still, substantial caution is necessary in interpreting these results as the sample of non-European languages is small. Finally, we found that acquisition of complex conceptual words reflecting colors, time, and logical constructs was highly variable across languages.

Chapter 13

Morphology, Grammar, and the Lexicon[1]

How does abstract structure emerge during language learning? On some accounts, children's early syntax emerges from direct generalizations from particular lexical items, while on others, syntactic structure is acquired independently and follows its own timetable. CDI data can help us decide between these two views. In this chapter, we summarize the state of grammatical development across languages (noting the challenges posed by radically different representations of grammar across CDI forms). We also replicate and generalize analyses linking grammatical generalization to children's vocabulary size. We end by investigating the idea that that age modulates the relationship between grammar and the lexicon.

13.1 Introduction

For many children, their first words are spoken in isolation. While these single word utterances sometimes seem to be picking out objects in the world (e.g., ball!), others seem to convey more complex ideas or desires (e.g., up! for Mommy, pick me up!). But by two years of age, many children have acquired a large repertoire of words, and are beginning to use them in two- or three-word combinations (e.g., Mommy up! or kitty sleep here). These utterances will gradually increase in length and complexity in various ways, forming sentences that increasingly reflect the grammatical structure of their native language (e.g., Mommy, the kitty is sleeping here). Children also begin to add more verbs, adjectives and other predicates to their working vocabularies (see Chapter 11), and substantively increase their use of prepositions, articles and other closed class forms that do grammatical work, including the productive use of inflectional morphemes (e.g., English past tense -ed or -ing).

Understanding the origins of grammar is critical because children's ability to use morphosyntactically-rich language is thought to reflect the uniquely-human mental machinery that enables speakers to produce novel utterances that have never been heard in the input (Berko, 1958; Pinker, 1991). The questions surrounding the development of grammar are challenging. How do abstract morphosyntactic structures emerge during language learning? What mechanisms underlie the formation of generalizations that support such inferences and allow children to apply them during language production? Does an understanding of the abstract rule-structure of language emerge from the interactions of individual words, or is that structure independently acquired and represented separately?

[1]Material in this chapter first reported in Braginsky et al. (2015).

Broadly speaking, theoretical views on grammatical development generally take one of two forms. On nativist theories like Principles and Parameters (Chomsky, 1981; Baker, 2005), grammar emerges independently from lexical knowledge following its own, largely maturational, timetable. Moreover, grammatical regularities are mentally represented in a format that is distinct from that used by the lexical system. In contrast, according to lexicalist theories, mental representations of morphosyntactic structure generally emerge from graded generalizations on the basis of lexical items, and at least early in development, there may be little or no representation of morphosyntactic rules or regularities per se (Tomasello, 2003; Elman et al., 1996). Even when syntactic structures are eventually represented, these representations are directly related to more concrete lexical structures (Bannard et al., 2009).

Historically, the study of individual differences has been critical to this debate. While variation in word learning is generally uncontroversial, individual differences in grammatical development are less clearly predicted under a universalist, nativist perspective. In contrast, lexicalist theories predict that variation in grammatical development should be tightly yoked to variation in lexical development (Bates and Goodman, 1999). Research has shown that, as with lexical development, there is sizable variation in exactly when and how children move into using more grammatically complex utterances in their everyday speech. While some children use primarily multi-word phrases and many closed class forms by 24 months, other children are still primarily producing nouns in single word utterances at that same age (e.g., Bates et al., 1988; Bates and Goodman, 1999). Moreover, there is also variation in the kinds of multi-word utterances that children produce. For example, some children build up sentences from individual words (e.g., want dat!), whereas other children seem to produce utterances that reflect "unanalyzed" chunks of more complex speech (e.g., iwantdodat!).

Associations between individual differences in lexical and grammatical development have been robustly substantiated in the literature. In the original norming data from the English CDI: Words & Sentences, children with more sophisticated grammatical productions were also those children with the largest vocabularies (Bates et al., 1994). Using that same dataset, Marchman and Bates (1994) found that size of verb vocabulary was concurrently related to children's overregularization of past tense inflections (e.g., daddy goed), productions that are viewed as a major milestone in the development of grammatical rule-based knowledge. Links between lexical development and grammar have also been reported longitudinally (Bates et al., 1988; Bates and Goodman, 1997), in late talkers (e.g., Paul, 1996; Rescorla et al., 2000, 1997; Thal et al., 1997), early talkers (Thal et al., 1996, 1997), and children with neurodevelopmental disorders, such as Williams syndrome (e.g., Singer Harris et al., 1997). Similar relationships have also been demonstrated in many other languages, including Slovenian (Marjanovič-Umek et al., 2013), Hebrew (Maital et al., 2000), Icelandic (Thordardottir et al., 2002), Italian (Caselli et al., 1999; Devescovi et al., 2005), Bulgarian (Andonova, 2015), Finnish (Stolt et al., 2009), Spanish (Mariscal and Gallego, 2012; Thal et al., 2000), and German (Szagun et al., 2006).

Finally, and perhaps most intriguingly, in behavioral genetic studies of monozygotic and dizygotic twins, the relation between lexical and grammatical level has been found to be strongly heritable (Dale et al., 2000; Dionne et al., 2003). In other words, even though genetic factors contribute relatively weakly to each aspect of language as assessed individually, the genetic factors that influence lexical growth are the same as those that influence grammatical growth, perhaps operating in a bidirectional manner.

While these studies substantiate that vocabulary and grammar development are strongly associated developmentally, the interpretation of these relations is still under debate. Some researchers have interpreted these links to suggest that domain-general learning mechanisms guide the child's construction of a working linguistic system at many different levels, in this case, learning words and

learning grammatical rules (e.g., Elman et al., 1996). As Bates and MacWhinney (1987) proposed many years ago, "the native speaker learns to map phrasal configurations onto propositions, using the same learning principles and representational mechanisms needed to map single words onto their meanings" (p. 163). Other proposals suggest that the process of learning words involves learning both their lexical-semantic and their morphosyntactic properties (e.g., in what constructions they can legally appear and what inflectional morphemes are required), and that grammatical knowledge is generally built up on a case-by-case basis (Tomasello, 2003). Early word combinations are often highly routinized and situation specific, suggesting that learning grammar, like word learning, is guided by learning mechanisms that are item specific and frequency dependent. It is only later that grammatical structures become encoded in terms of their abstract syntactic form (e.g., Lieven et al., 1997; Tomasello, 2003). Yet other accounts view the relation as reflecting mechanisms that operate in the opposite direction. On these views, grammatical analysis is a driving force behind word learning, such that the process of analyzing sentences into their constituent grammatical parts facilitates the further acquisition of lexical-semantic knowledge (Anisfeld et al., 1998; Naigles, 1990). Finally, other studies have proposed that the lexical-grammar relations may not be as direct as previously proposed, actually being driven by common third-party influences such as the speech that children hear (Hoff et al., 2017).

In this chapter, we explore relations between estimates of children's vocabulary size based on the vocabulary checklist and responses on other sections of the Words & Sentences instruments. Many versions of the instruments provide indices of grammar learning by asking about children's use of inflected forms (e.g., walked), children's use of overgeneralizations (e.g., goed), and the complexity of their multi-word combinations (e.g., kitty sleeping / kitty is sleeping). While many studies have examined associations between lexical and grammatical development cross-linguistically, the scope and power of these early studies were limited, with few opportunities for direct comparisons of the nature or extent of these relations across multiple languages at the same time. In contrast, our data allow analyses of lexical-grammar relations with enhanced statistical power and broader cross-linguistic representation.

In addition, we explore a hypothesis that was not explicitly tested in these earlier studies: that there remains age-related variance in grammatical development unexplained by vocabulary development. While the overall relationship between grammar and the lexicon provides support for lexicalist theories, the identification of age-related variance would suggest the presence of developmental processes that regulate grammar learning, above and beyond those captured by measures of vocabulary size. Such age-related processes could be either maturational or experiential, and either domain-general (like working memory) or language-specific (like grammatical competency). Importantly, since both nativist and constructivist theories could in principle predict age-linked variance in grammatical development, our goal is not to differentiate these theories, but instead to test this novel prediction and explore its implications for future work on understanding the processes of grammatical development.

An additional contribution of work is that, due to the size of our dataset, we are able to make more fine-grained distinctions than the initial cut between grammar and the lexicon. In particular, we distinguish morphology from multi-word syntax, since morphological generalizations might be more specifically dependent on vocabulary size than those requiring more global, sentence-level syntactic regularities.

13.2 Methods

In all 12 languages included in these analyses, the CDI forms contain both vocabulary checklists and other questions relevant to the child's linguistic development. All of the data reported here come from Words & Sentences type forms, administered to children ages 12–36 months (most in the 16–30 month range). In addition to the vocabulary checklist items, these forms typically contain a single item asking whether the child is combining words yet at all; Word Form section, which asks whether the child produces each of around 30 morphologically inflected forms of nouns and verbs (e.g., feet, ran); and a Complexity section, which asks whether the child's speech is most similar to the syntactically simpler or more complex versions of around 40 sentences (e.g., two foot / two feet, there a kitty / there's a kitty).

Importantly, each instrument for languages other than English is not just a translation of the English form, but rather was constructed and normed to reflect the lexicon and grammar of that language. Thus, there are substantial differences in the content of these items and their coverage of different morphological and grammatical phenomena. The major commonality is that the form developers believed that they provided a good survey of important developmental phenomena in their language.

Word Form items are shown in Table 13.2.

All Word Form items included in these analyses.

Language	Form	Item
Danish	WS	børn
Danish	WS	heste
Danish	WS	mænd
Danish	WS	fødder
Danish	WS	hunde
Danish	WS	skibe
Danish	WS	(flere) får
Danish	WS	(flere) mus
Danish	WS	(flere) sko
Danish	WS	blev

Showing 1 to 10 of 189 entries Previous 1 2 3 4 5 … 19 Next

Complexity items are shown in Table 13.2.

All Complexity items included in these analyses.

Language	Form	Item
Cantonese	WS	Sentence Structure Markers-classifiers
Cantonese	WS	Sentence Structure Markers-verbs
Cantonese	WS	Sentence Structure Markers-possessions
Cantonese	WS	Sentence Structure Markers-verb + resultative verb complement
Cantonese	WS	Sentence Structure Markers-past events
Danish	WS	To bil / To biler
Danish	WS	Det min bil / Det er min bil
Danish	WS	Tænd lys / Tænd lyset så jeg kan se
Danish	WS	To fod / To fødder
Danish	WS	Du ordne det? / Kan du ordne det?

Showing 1 to 10 of 272 entries Previous 1 2 3 4 5 … 28 Next

To analyze lexical and grammatical development, we derive several measures. Each child's Vocabulary Size is computed as the proportion of words on the corresponding CDI form that the child is reported to produce. Similarly, each child's Word Form score is the proportion of word forms they are reported to produce, and their Complexity score the proportion of complexity items for which they are reported to use the more complex form. We compute all of these quantities as proportions to make the scales comparable across languages.

13.3 Results

We present four sets of results. First, we show analyses of the "combines" item, which is a binary item in which parents indicate whether their child is combining words. Second, we give analyses of the relationship between vocabulary size and Word Form items and Complexity items. Third, we follow up on a pattern found in the "combines" item, namely age-related modulation of the grammar-lexicon relationship. Finally, we investigate the degree to which the age-related pattern is found in individual items.

13.3.1 Combine

Figure 13.1 shows the probability of a parent checking that their child combines words, plotted by the child's chronological age (left) and raw productive vocabulary size (right). As can be seen, across 8 languages, there is some consistency in the chronological trajectories for this item. By 24 months, around 75% of children are reported to be combining words, though this estimate is substantially earlier in Quebec French. One possibility is that the phrasing of the "combines" item contributes — some forms (including Quebec French, but also Norwegian and Danish) give examples of simple combinations, which could encourage earlier reporting.

Vocabulary-related trajectories were more variable, however. In general, children who were marked as combining had vocabularies larger than around 100 words. However, as noted in Chapter 5, raw Mandarin vocabulary in the WS form is unusually high, but the "combines" item does not appear to be comparably accelerated. Thus, Mandarin children appear to be producing words only after

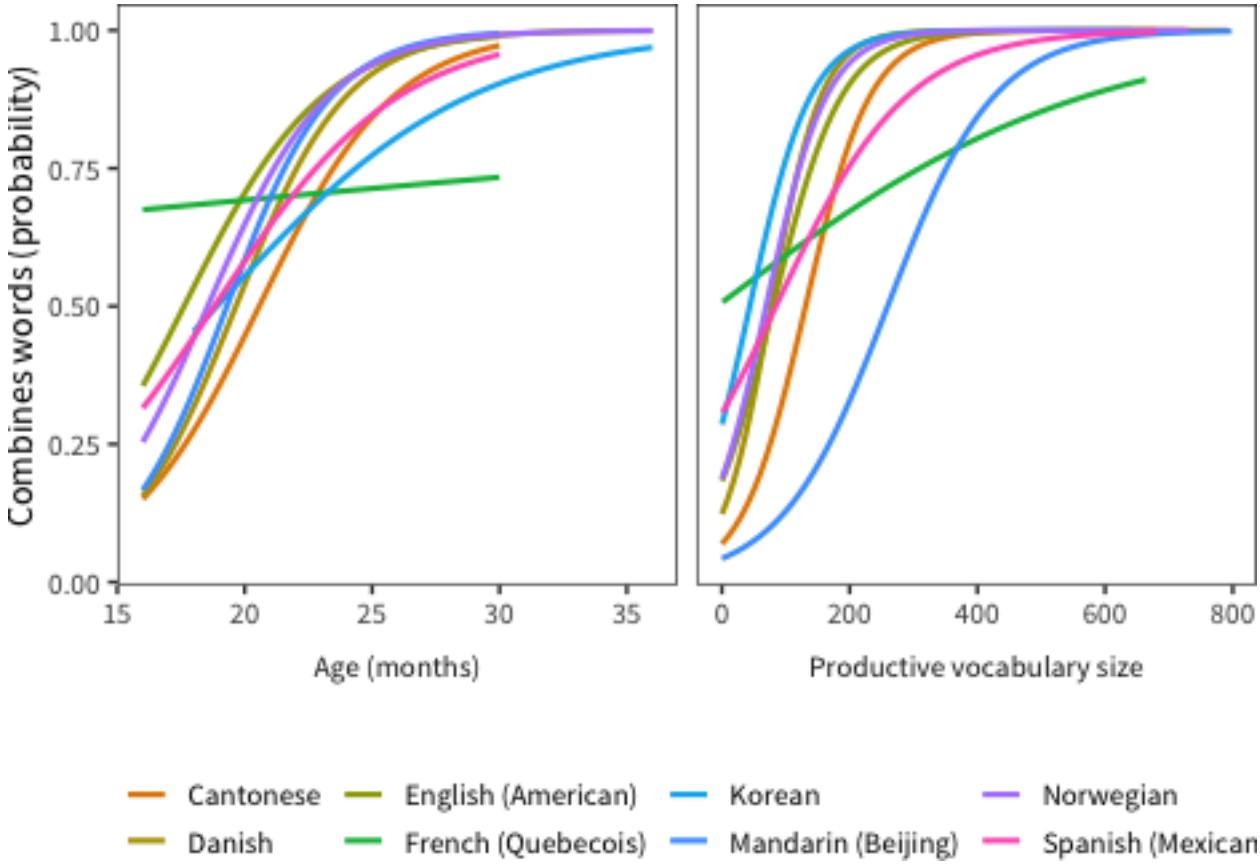


Figure 13.1: Trajectory of the Combines item in each language across age (left) and vocabulary size (right).

Table 13.1: Coefficient estimates from Combines model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.33	1.011	-4.3	0.000
production_prop	18.68	1.777	10.5	0.000
age	0.13	0.042	3.2	0.001
production_prop:age	-0.34	0.048	-7.0	0.000

producing substantially more vocabulary items. On the opposite side, children learning Quebec French and Korean were reported to be combining with quite small vocabularies.

To investigate the quantitative relationship between word combination (as measured with this item), age, and vocabulary, we fit a linear mixed effects model predicting combination as a function of vocabulary (proportion), age, and their interaction. We also included random intercepts by language and random slopes for both vocabulary and age. Coefficient estimates are shown in Table 13.1.

This model shows an extremely large effect of vocabulary, with a relatively smaller amount of variance due to age. In addition, there was a substantial negative interaction of vocabulary and age, reflecting that older children were more likely to be combining words, even with less vocabulary. This result parallels others reported below suggesting that there are age-related components in grammatical performance that are unaccounted for by vocabulary. All coefficients were highly significant due to the large amount of data.

Overall, although there was some cross-linguistic variation — perhaps due to true variation and perhaps due to idiosyncrasies of individual forms or datasets — word combination emerged around 24 months and 100 words for most children.

13.3.2 Grammar and lexicon relationship

We next examine the correlation between the proportion of Word Form and Complexity items completed and the proportion of vocabulary items completed. First reported by Bates et al. (1994), these correlations are extremely robust, and can be observed in essentially all of our datasets. Figure 13.2 shows this relation for Word Form items. We fit linear regressions predicting vocabulary as a function of linear, quadratic, and cubic predictors (subtracting the intercept to ensure that the function passed through the origin). The total r^2 for these relationships ranged from 0.82 to 0.93.

Complexity items show the same relationship (Figure 13.3), typically with equal or greater strength (depending on data density and number of items). r^2 values varied from 0.56 to 0.94.

Overall, these data provide strong cross-linguistic support to the contention of Bates et al. (1994) and others that the emergence of grammatical competence in production is related across individuals to the size of the productive vocabulary.

13.3.3 Age effects

In our next analysis, we follow up on the relationship between age and grammatical ability found in the "combines" analysis above. In that analysis, we noted that less vocabulary was needed for older children to be marked as combining words. We investigate this pattern in the full Word Form and

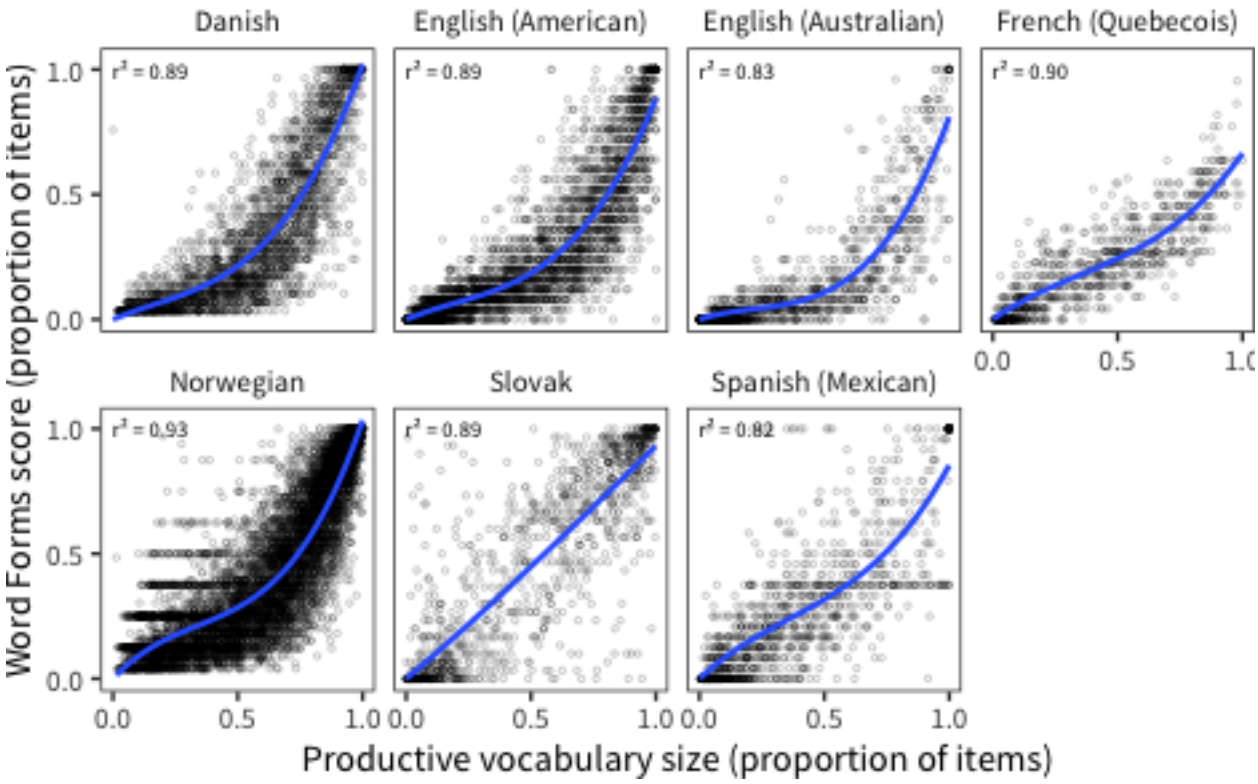


Figure 13.2: Each child's Word Forms score as a function of their vocabulary size in each language (curves show model fits).

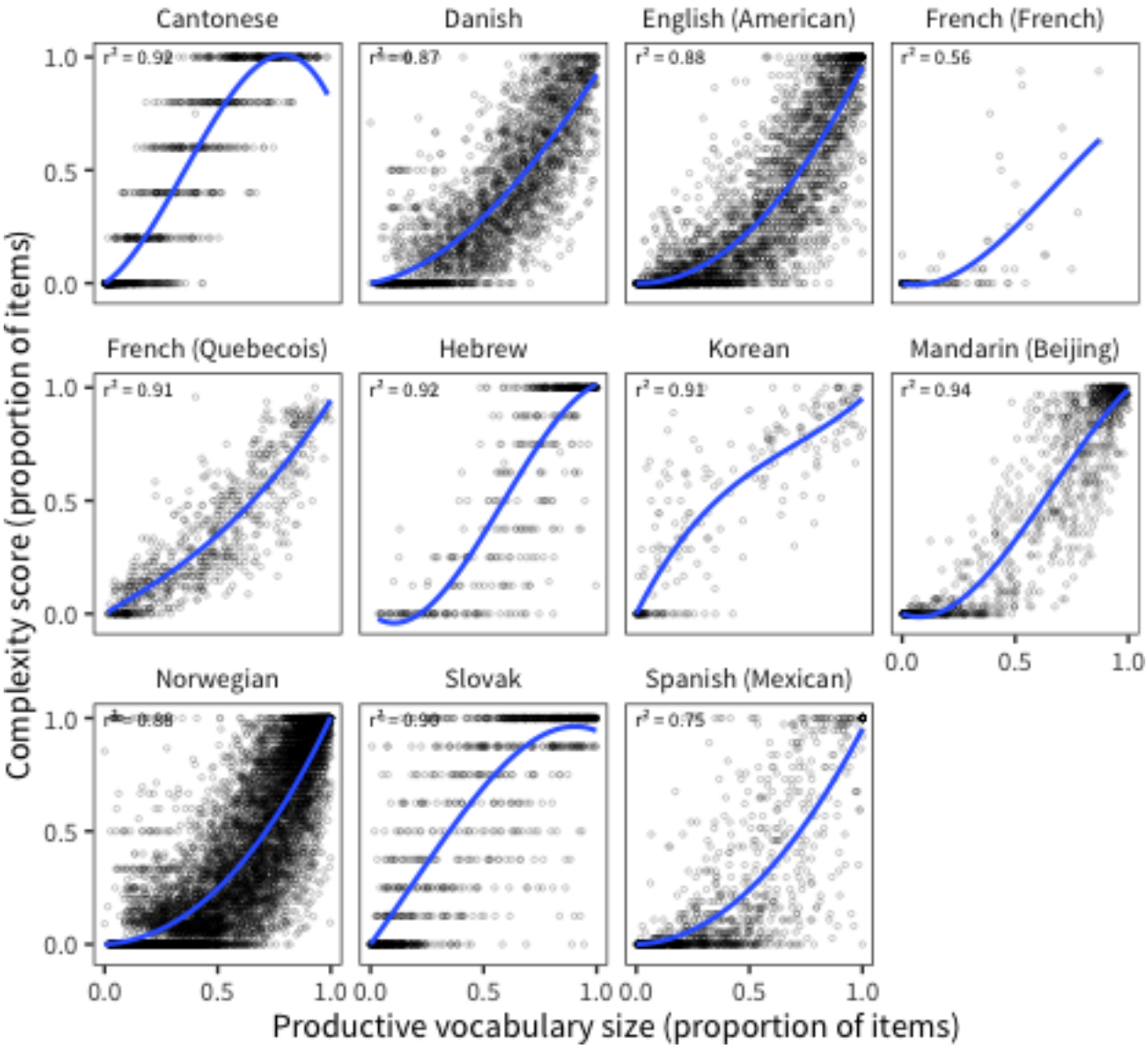


Figure 13.3: Each child's Complexity score as a function of their vocabulary size in each language (curves show model fits).

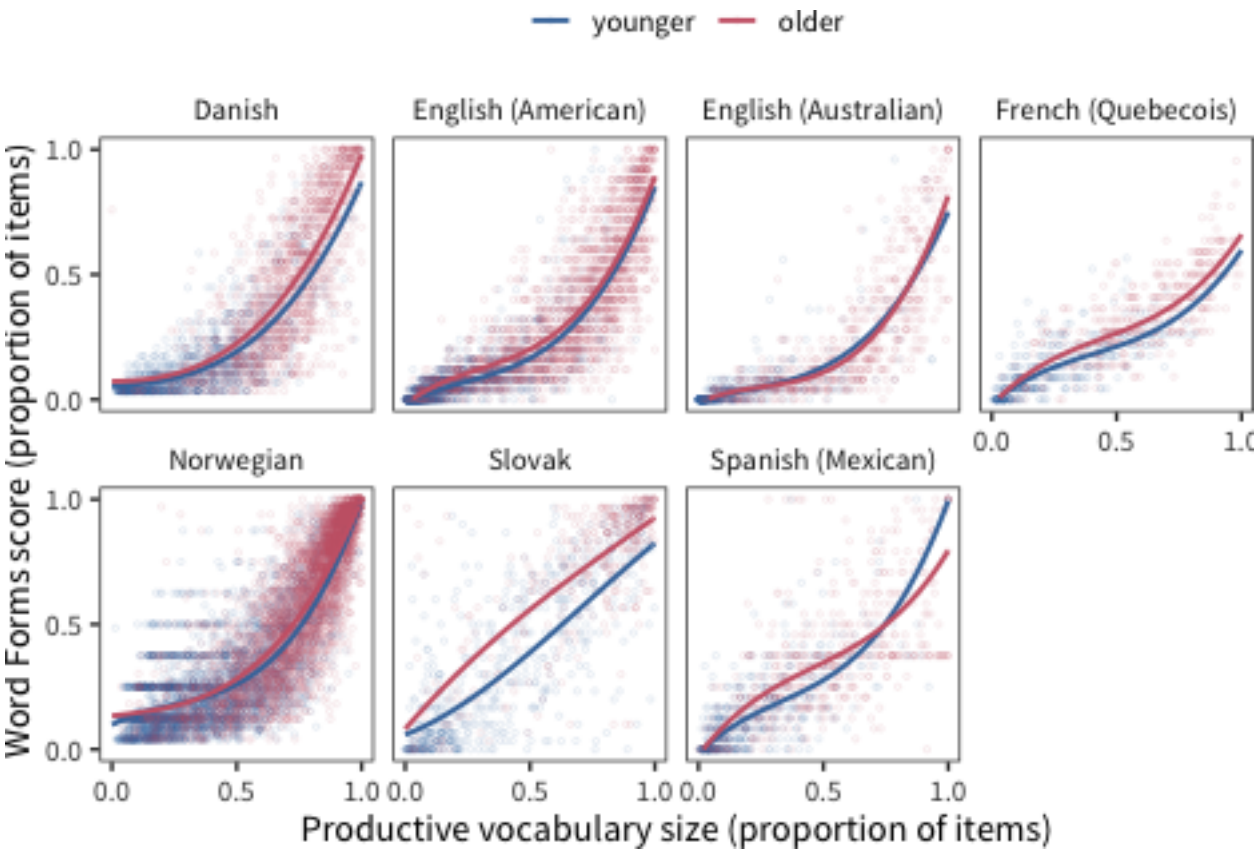


Figure 13.4: Each child's Word Forms score as a function of their vocabulary size in each language for younger and older children (curves show model fits).

Complexity item set by splitting data from each language by age. We plot the same curves as above, but separately for children older and younger than the median.

In essentially every language, for both Word Form items (Figure 13.4) and Complexity items (Figure 13.5), we see a higher curve for older children than younger. This finding is consistent with the idea that older children have less vocabulary per unit grammar (mirroring the negative interaction shown for the "combines" item).

This pattern is further summarized in Figure 13.6, where we show the area under the grammar/lexicon curve for younger and older children. The upward slope of nearly every line demonstrates the consistency of the age effect, which we discuss further below. In addition, there is a trend for age effects to be larger in Complexity rather than Word Forms, suggesting a more syntactic locus for the effect.

13.3.4 Individual items

In our final analysis, we examine the individual items on the Word Form and Complexity sections. Given the heterogeneous nature of the CDI instruments, particularly in the Complexity sections, we attempted a more fine-grained item-analysis by classifying items as capturing either more morphological or more syntactic phenomena. Items for which the difference between the simple and

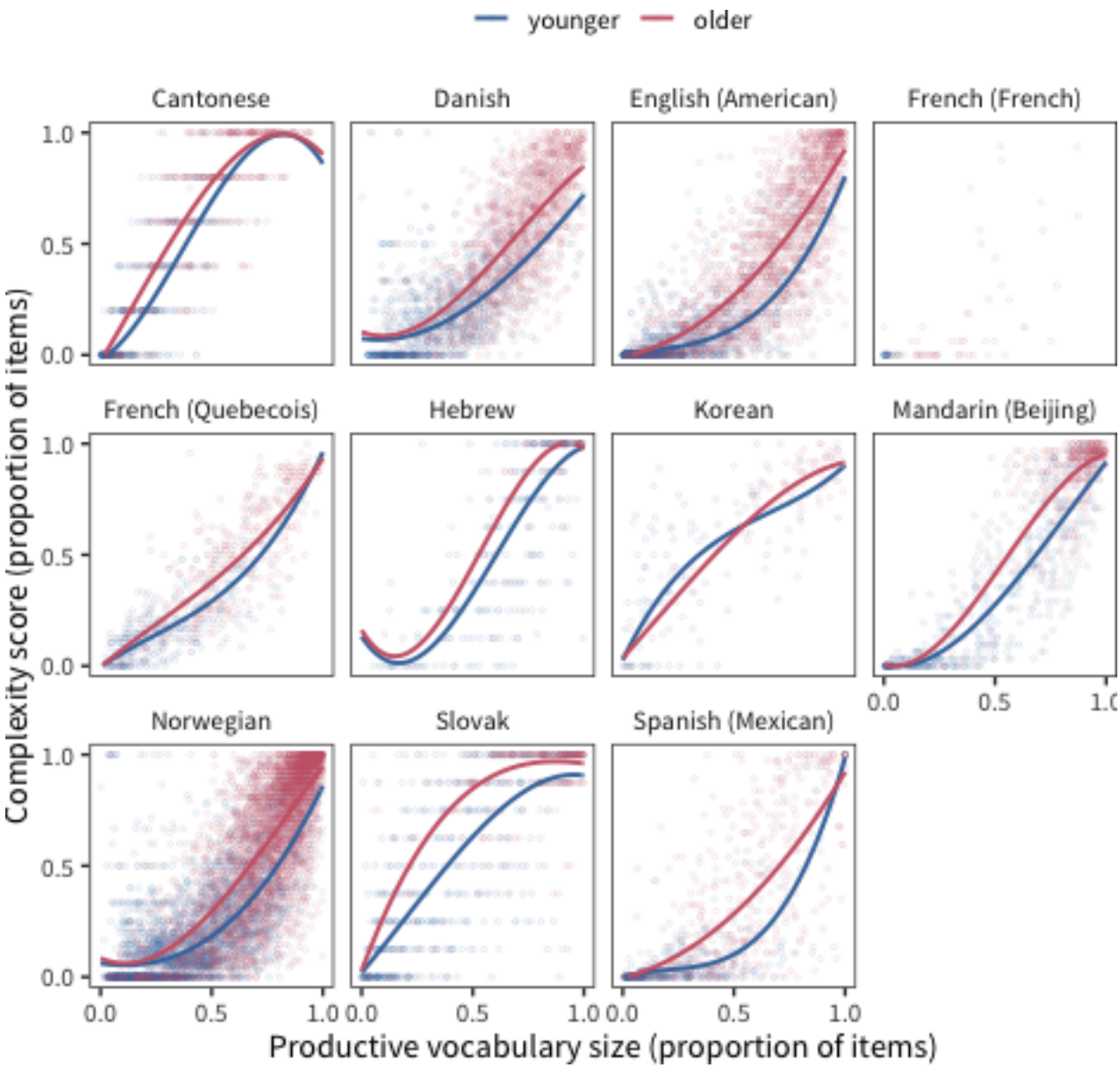


Figure 13.5: Each child's Complexity score as a function of their vocabulary size in each language for younger and older children (curves show model fits).

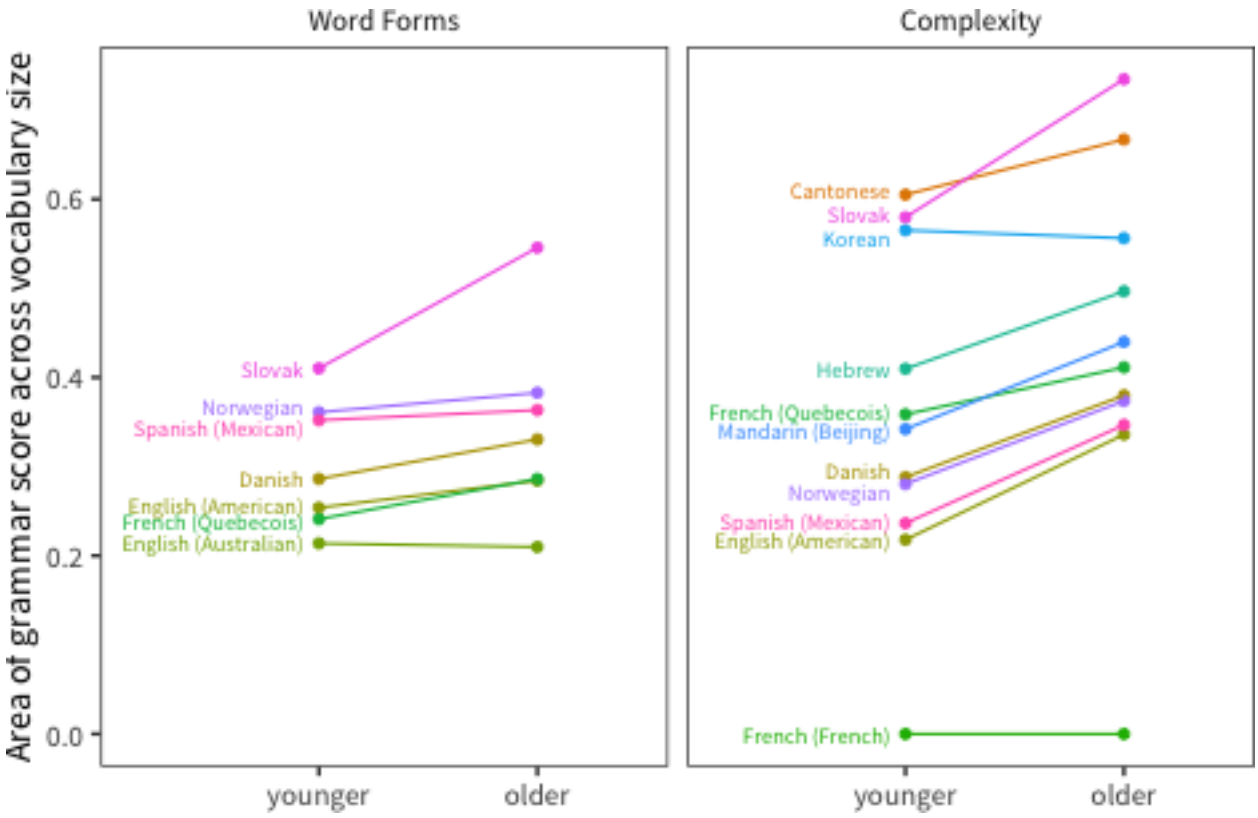


Figure 13.6: Area under model fit curve for Word Forms score and Complexity score as a function of vocabulary size in each language for younger and older children.

complex sentences is in the inflection of a noun or verb (such as doggie kiss me / doggie kissed me) were coded as Morphological. The remainder of the items were coded as Syntactic, since they involved the use of some sentence-level syntactic construction (such as doggie table / doggie on table).

We then fit logistic regression models with linear, quadratic, and cubic predictors (as above) separately for every item. Figure ?? shows the age effect coefficient of each item. In general, age effect were smaller for Word Form items, then Morphological Complexity items, and largest for Syntactic Complexity items, suggesting that more syntactic phenomena likely have greater age contributions.

13.4 Discussion

We revisited classic findings on the relationship between grammar and the lexicon, further exploring novel questions regarding the role of age in this relation. Our results provide general support for a lexicalist view, in that, in 12 languages, variance in vocabulary production strongly aligned with variance in grammar. However, we also estimated additional age-related contributions, specifically contrasting the links to morphological forms vs. syntactic constructions, and for different lexical categories. In general, we find that measures of grammar that are more closely aligned with syntax are modulated by age to a greater extent than those reflecting inflectional morphology.

As with the correlations reported in Chapter 7, parent bias is a potential confound for these correlational analyses. If some parents tend to over-report on their child's language than others — whether for reasons of sensitivity, optimism, greater time spent with the child — then this over-reporting would likely extend across linguistic domains. Thus, in principle an observed correlation between two sections of a single parent report form could be driven by parent bias acting independently on each section without any connection.

Two studies provide evidence against this deflationary hypothesis. First, Moyle et al. (2007) used a variety of instruments to provide evidence for the same relation in both typically-developing and late-talking children. Second, Brinchmann et al. (2018) give a similar analysis using cross-lagged structural equation models. Critically their work relied only on direct testing of the child (not parent report) using standardized instruments. In their model, they found a strong correlation between the time-invariant (trait-like) components of vocabulary and grammar ($r = 0.72$). While this correlation is smaller than the correlations we report, it is still quite large — and it appears in a model that also controls for a number of other relationships. Intriguingly, however, this study suggests that as grammar-lexicon correlations are cross-lagged, grammar to vocabulary links are stronger than vocabulary to grammar links, and neither are that strong This finding suggest that syntactic bootstrapping effects — as well as correlations driven by shared input may be responsible for the relationship between the two abilities. Regardless, these two studies suggest that reporting bias is very likely not the sole cause of the correlations we observed.

Our analyses go beyond earlier work by also investigating the relationship of age to vocabulary and grammar. One possibility is that age-related developments are dependent on maturational factors that operate on grammatical development in a domain-specific way, independent of lexical-semantic processes. Another possibility is that age-related effects represent more domain-general learning mechanisms, such as attention or working memory, that provide differential support for sentence-level processes than word-internal ones (Gathercole et al., 2013). Future studies should also explore the extent to which lexical and age-related processes are shaped, either independently or in tandem, by

features of the learning environments that children experience (e.g., Weisleder and Fernald, 2013; Hoff et al., 2017; Brinchmann et al., 2018).

Questions about the nature of morphosyntactic representations in early language have often seemed deadlocked. But by mapping out developmental change across large samples and multiple languages, our findings here challenge theories across the full range of perspectives to more fully describe the mechanistic factors underlying the interaction of vocabulary, grammar, and development.

Chapter 14

Individual Variation in Vocabulary

> Of all the individual differences described to date in the literature on early child language, variations in rate present the least interesting challenge to traditional 'universalist' models of development, If it can be shown that all children go through the same basic sequence, activating a common set of structures and processes, then small variations in the onset time for specific language milestones might represent little more than a minor perturbation to a maturational theory (like variations in the onset of puberty). Putative variations in style of development are more problematic, because they raise questions about the order in which structures are acquired, and the mechanisms used to acquire those structures. (Bates et al., 1994)

Preceding chapters have dealt with the degree of variability between individuals in Chapter 5 and the stability of individuals' learning in Chapter 4. Then Chapters 6 and 9 used demographic factors to explain variability. In general, our treatment of variability has focused on issues of learning rate in the sense discussed by Bates et al. in the quote above (or, to be more precise, vocabulary size). With the exception of Chapter 9, which examined the rate of learning individual words, we have not dealt yet with issues of style.

What does it mean for two children to show differences in language learning style? Intuitively, children who vary in style of learning should differ in some aspect(s) of the learning process. Unfortunately, from the data that we use here, we cannot observe process — we can only observe the outcome of learning: knowledge. Thus, children must exhibit some differences in the way their vocabulary grows. This difference could be distributional and inferred indirectly from cross-sectional data or it could be shown directly through longitudinal data (though this move limits the number of datasets we can use from Wordbank).

One prominent candidate for a stylistic difference in language acquisition is the distinction between "referential" and "expressive" children. In an early report on individual differences in vocabulary acquisition, Nelson (1973) noticed that there was substantial variation in how many nouns children had in their vocabularies. Children who had more than half of their vocabulary devoted to nouns were named "referential" children. They tended to have speech that was less syntactically complex than other children and showed faster vocabulary growth. In contrast, "expressive" children had speech that was more syntactically complex and had fewer nouns. Dore (1974) proposed a related version of this distinction, focusing on speech acts from two children in the middle of the second year and labeling them as "code oriented" (focused on labeling, similar to "referential") vs. "message

oriented" (instrumental, social requests, similar to "expressive").

Since this seminal work, the referential/expressive distinction has become enshrined in the literature as a canonical aspect of variation in children's style. Yet further debate about the consequences of this stylistic distinction continued in the literature. For example, Bates et al. (1988) observed a correlation between the proportion of nouns in the children's vocabulary and their vocabulary size and suggested the possibility that a referential style might be more effective for learning. Reacting to this claim, Pine and Lieven (1990) argued that the direction of causality might be reversed, however: perhaps having a bigger vocabulary (at least at a certain point) would lead to a greater representation for nouns.[1]

While exploring the referential/expressive phenomenon more rigorously, Bates et al. (1994) presented analyses largely supporting that view: much systematic variation in the "referentiality" (proportion nouns) in children's vocabularies was due to the size of their vocabulary. As children's vocabularies grow, they tend to increase in their over-representation of nouns (as shown also in Chapter 11). Thus, two children of the same age who have different proportions of nouns may also have different overall vocabulary sizes. After controlling for this factor, Bates et al. (1994) found only more limited evidence for stylistic variation. In the first subsection of this chapter, we pick up these analyses and apply them broadly to our dataset.

A second area of stylistic variation that has been much discussed is whether some children go through a vocabulary "spurt," defined as a change in the rate of vocabulary growth. The idea of a "spurt" or "explosion" occurs in a wide variety of discussions of early vocabulary (e.g., Nelson, 1973; Kamhi, 1986; Bates et al., 1988; see e.g., Dapretto and Bjork, 2000, for review). Clearly, from the average growth rates observed in Chapter 5, children's vocabulary growth accelerates dramatically during the period following their first birthday and the emergence of first word production. Investigations of this acceleration have focused on two distinct features: its explanation and its variation across children.

Although there has been a tremendous amount of discussion of the explanations for accelerations in vocabulary growth, we will not investigate this topic further here. Our data do not allow us to investigate issues of mechanistic process directly or of what cognitive or social changes are co-contemporaneous with the vocabulary spurt. In addition, a short but convincing analysis by McMurray (2007) suggests that such acceleration is over-determined in the sense that it will likely emerge from almost any plausible acquisition mechanism(s). Assuming that words vary in difficulty as a normal distribution, vocabulary growth will naturally accelerate with increases in a child's ability. Thus, making reverse inferences from the presence of acceleration to a particular mechanism is unwarranted.

The second issue — variation in acceleration across children — is related to our aim here, however. Does every child's vocabulary acceleration follow the same general pattern, or is there substantial variation in the type of growth that is followed, for example, in the point at which acceleration begins, or the specific shape of the growth curve? Ganger and Brent (2004) report a systematic study of a sub-part of this issue: whether there is a discontinuity in growth rate for individual children.

[1] Caught up in this discussion is the question of whether there is a route into language via the memorization of "frozen phrases." This is an independent theoretical question that is difficult to address with CDI data as it is a question about the repetitive use of chunks of language in production. One observation is that, due to evidence of early verb generalization (e.g., Gertner et al., 2006), the original discussion about limited generalization in children's early syntax has been somewhat subsumed into a discussion about differences between general comprehension and conservative production.

Defining a spurt specifically as a change in the rate of acceleration, they conduct a quantitative analysis of whether such a change occurs. Our second set of analyses follows up on this general issue.

The final area of stylistic variation that we address in our third subsection is the dissociation between comprehension and production. It has been long acknowledged that children tend to understand more than they can say (e.g., Clark and Hecht, 1983). But, to what extent does this generalization vary across children? That is, are there some children who are more fluent producers vs. others who have a large vocabulary in comprehension but more limited production? At the clinical extreme, significant variation in this respect must be present, because there are some children with various apraxias of speech that have strong comprehension but serious production difficulties. At the other extreme, delays in early comprehension have clinical relevance, and are often thought to be a more reliable indicator of language delays compared to production alone (e.g., Rescorla, 2009; Thal et al., 1991). While both of these extremes are interesting and important, the question for us is whether, in a typically-developing population, we can detect systematic variation on this dimension over the full range of language abilities.

14.1 Variation in vocabulary composition

In this subsection we examine the question of variation across children in referential vs. expressive vocabulary. Following Bates et al. (1994), we operationalize the notion of a "referential" vocabulary as one that has a greater proportion of nouns relative to other classes (the definition of "noun" is the same here as in Chapter 11. While there might be other more nuanced measures that we could construct, this one has the advantage of being directly related to the framework in Chapter 11; thus, we use that same framework to investigate vocabulary composition in individuals here.

The proportion of children's vocabulary that is made up of nouns is shown in Figure 14.1. There is a general trend for an over-representation of nouns, as shown by the blue line (representing the smoothed mean proportion nouns) being above the red dashed line (total proportion nouns on the form). The size of this over-representation is the topic of Chapter 11. Here we examine its variability across children.

Is a referential style associated with having a larger vocabulary? The proportion of nouns in a child's vocabulary should then be a predictor of vocabulary size, over and above age.

A simple model of this hypothesis is a GLM predicting the number of CDI words a child produces as a function of age and proportion nouns.[2] The coefficients of such a model, fit to the data from each language, are shown in Figure 14.2. Age coefficients are positive, indicating more words with age. Proportion of nouns is also negative, indicating that having more nouns is related to a smaller vocabulary, controlling for age. (Confidence intervals are plotted, but are typically tiny and hence invisible.) This result appears to provide support for the opposite to the claimed relationship between the referential/expressive distinction and vocabulary size. Those children with more referential vocabularies have smaller vocabularies, controlling for age.

The trouble is that these variables — age, noun bias, and total vocabulary size — have a complex relationship to one another. For a young child, having a bigger noun bias is correlated with having a bigger vocabulary (because they are on the early part of the "noun over-representation" curve shown

[2]We omit interactions from most of the models below for interpretability; including interactions leads to unstable coefficient estimates.

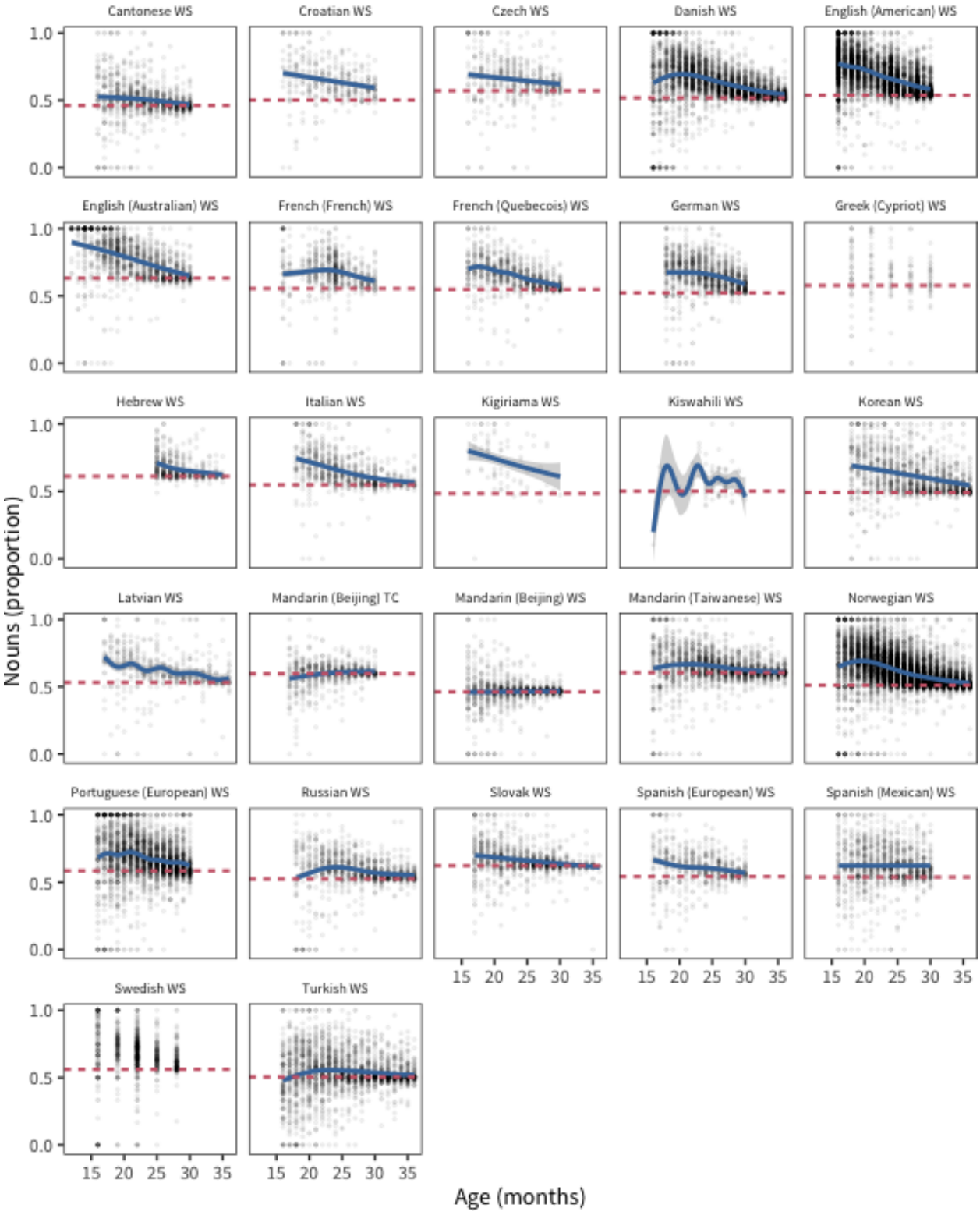


Figure 14.1: Proportion of nouns that each child produces as a function of age, with the blue curve showing a smoother fit and the red line indicating the proportion of nouns on the form.

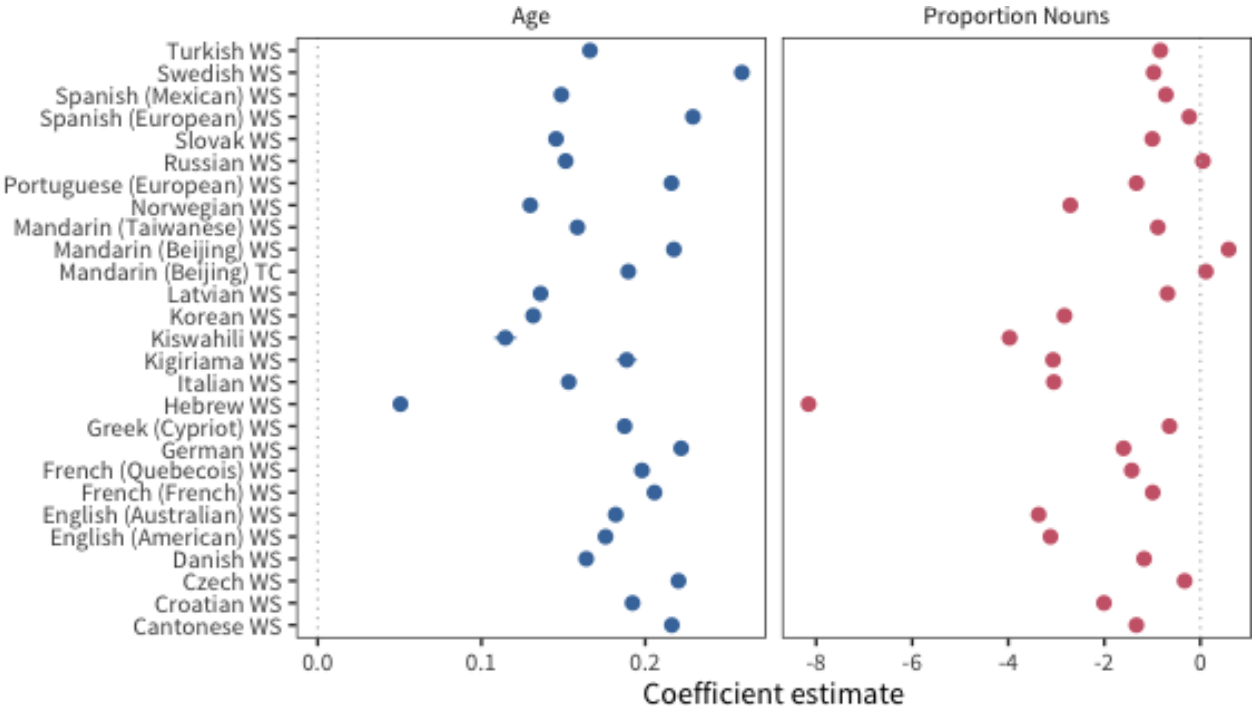


Figure 14.2: Effects of age (left) and proportion of nouns produced (right) on vocabulary size in each language.

in Chapter 11). In contrast, for an older child, having a bigger noun bias is correlated with having a smaller vocabulary because they are on the later part of the curve. Thus, the directionality of the relationship that you discover is largely determined by what part of your sample is densest.

Put another way, proportion nouns could be predictive of vocabulary size not because children with a particular style have bigger vocabularies, but because having more nouns in your vocabulary tends to indicate that you are further along a standard progression. Put another way, perhaps all children follow the same trajectory through the noun bias. Even in this scenario, knowing the size of a child's noun bias will tell you something about vocabulary size, without that implying that the child is following a different trajectory. This point is made in different ways by both Bates et al. (1994) and Lieven et al. (1992).

One way to circumvent this critique statistically is to measure whether a particular child has a greater-than-average, vocabulary-adjusted noun-bias. In other words, if we remove the average correlation with noun bias and vocabulary, can we still find a relation with individuals' degree of noun bias and vocabulary size?

Figure 14.3 shows both the English noun-proportion data (plotted now by vocabulary size) and the residuals of that distribution when fit via a cubic model. The next question we can ask is how this "residualized style" relates to other variables like age, grammatical ability, and (in longitudinal data) further vocabulary growth. Note that we cannot compare this variable to other aspects of concurrent vocabulary size because features like, for example, number of closed class items in the vocabulary are non-independent (since the more nouns you have, by definition the fewer closed class items).

Our first analysis looks at the correlation between vocabulary-residualized noun bias and age. Figure 14.4 shows coefficient estimates on this analysis. Now most age coefficients are reliably negative,

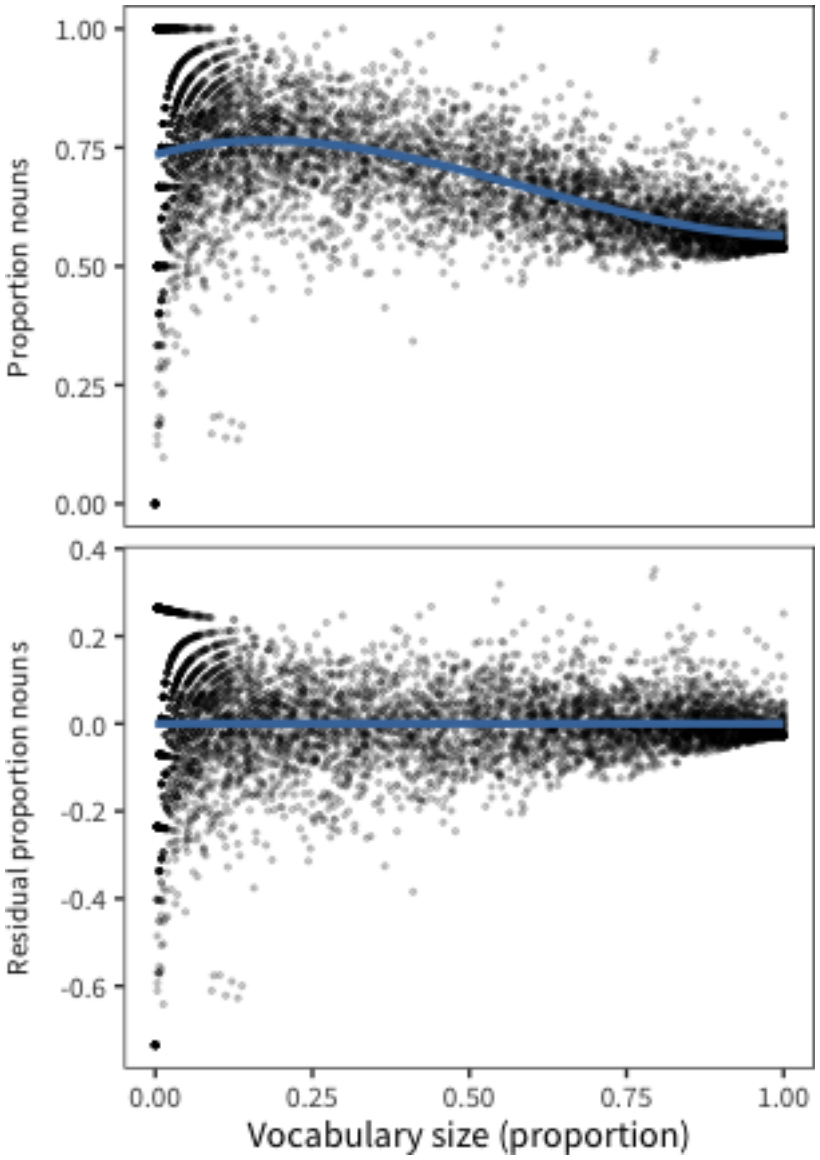


Figure 14.3: For American English data, proportion of nouns produced by each child as a function of their vocabulary size (top) and the residual proportion of nouns regressing out vocabulary size (bottom).

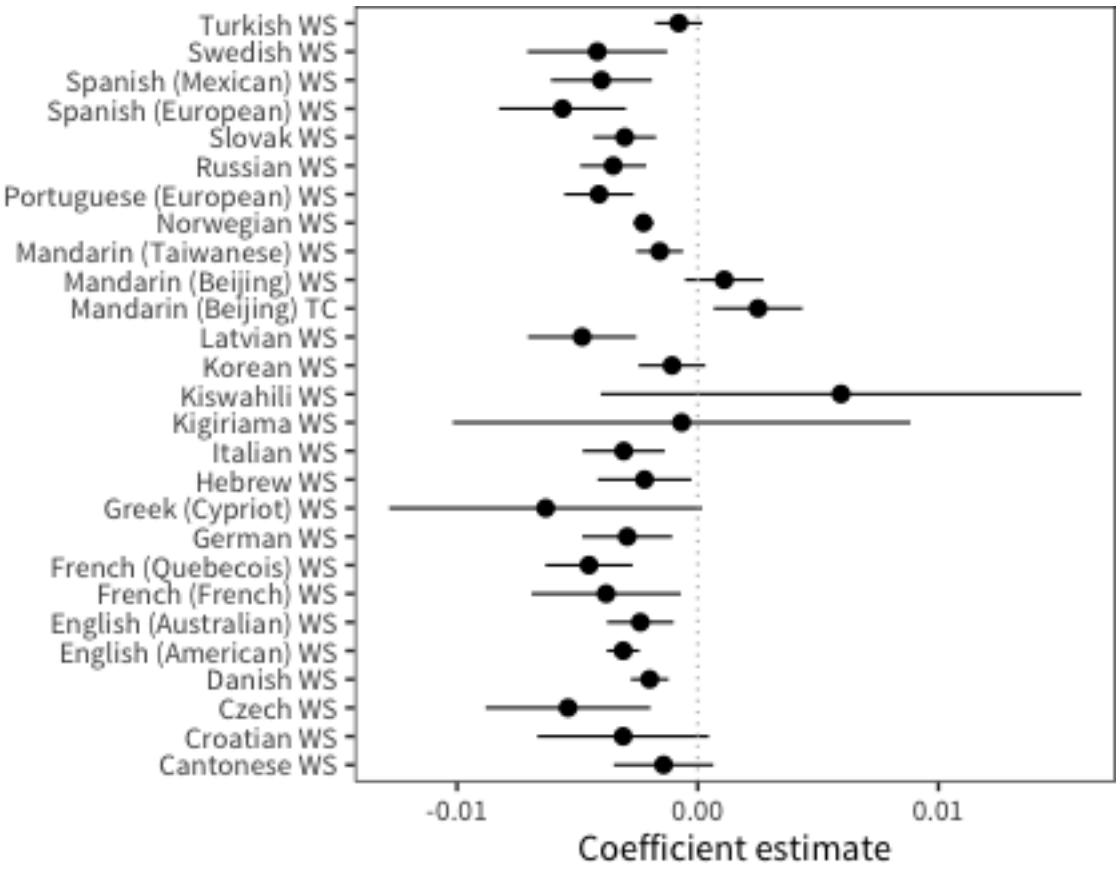


Figure 14.4: Effect of age on vocabulary-residualized noun bias in each language.

suggesting that a greater residual noun bias is associated with being younger. Two of the main outliers here are from Mandarin, which, as discussed in Chapter 11, has the smallest noun bias of any of the languages in our dataset.

Next we examine correlations with grammatical ability. Lieven et al. (1992) was interested in the possibility that an alternative route into language from a referential style would be the use of construction-based generalizations.

For grammatical complexity scores (Figure 14.5), we see that vocabulary-residualized noun bias is reliably related to grammatical complexity. Because of the residualization procedure, this relationship is over and above the correlations between grammar and lexicon (as reported in Chapter 13). Thus, those children with more nouns than expected for their vocabulary size produce less complex language. (Again, Mandarin is an exception). As seen above, they are also younger than expected.

We next repeat the analysis of complexity while controlling for age and total production. The coefficients are shown in Figure 14.6. Production of course has a positive relationship to grammatical complexity, as does age (see Chapter 13). Even controlling for these two factors, however, we still observe a consistent negative relationship between residual noun bias and complexity.

Summarizing, we were interested in this subsection in whether we found cross-linguistic evidence for different styles of language learning, in particular, individual differences that mapped onto the referential vs. expressive distinction. Operationalizing this distinction, we asked whether children

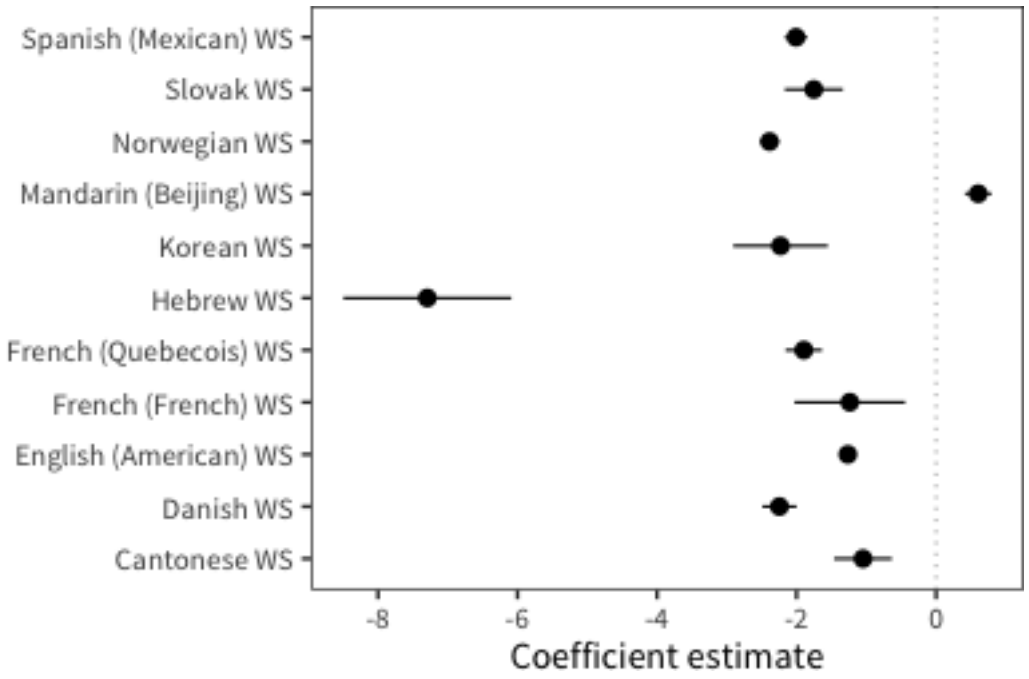


Figure 14.5: Effect of vocabulary-residualized noun bias on complexity score in each language.

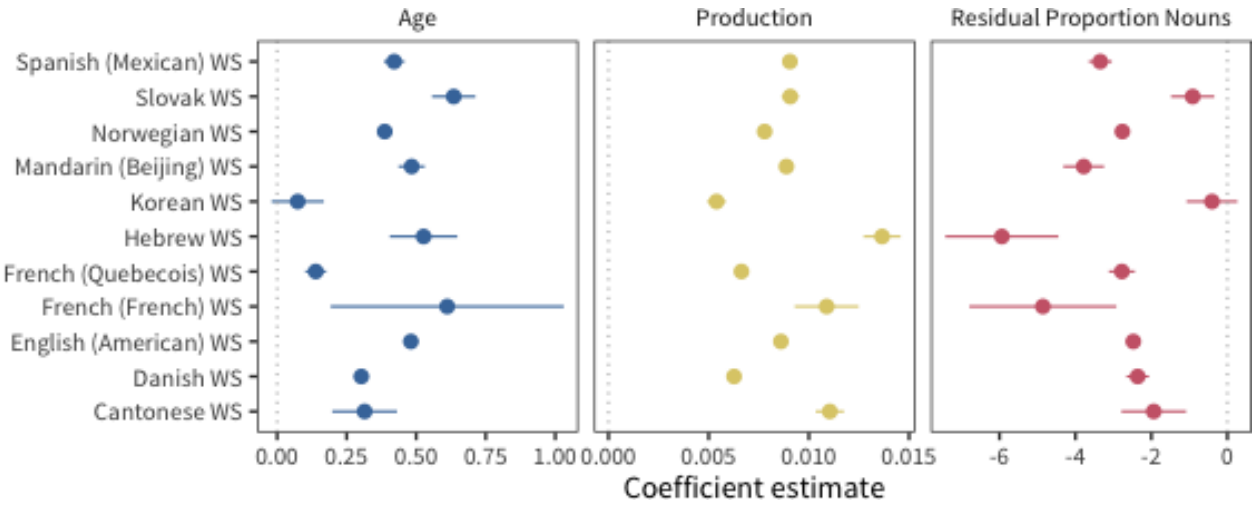


Figure 14.6: Effects of age (left), vocabulary size (middle), and vocabulary-residualized noun bias (right) on complexity score in each language.

with a larger noun bias show differences in their language learning trajectory than children with a smaller noun bias. The relations between overall noun bias and vocabulary are complex to interpret due to the non-linear relationships between these variables and age. To circumvent this, we examined a residualized noun bias measure that controls for total vocabulary. This measure was related to age and to grammatical complexity: children with relatively more nouns in their vocabulary tended to be younger and to be producing less complex speech, across languages. Taken together, these data provide some cross-linguistic support for the idea that children show variability beyond differences in rate of vocabulary acquisition. One dimension of variability is that, for a particular level of vocabulary size, there appear to be children who are younger, know more nouns, and combine words less (perhaps the "referential" children referred to in previous literature); other children will tend to be older, have a more diverse vocabulary (including more predicates), and will tend to use these to combine words more.

14.2 "Spurts" in vocabulary

Perhaps the most obvious aspect of early vocabulary is that it picks up speed in growth over the second year after birth. First words are hard-won but soon children's language appears to "explode." This acceleration is easily visible in Figure 14.7, which shows the median productive vocabulary from 8–18 months (for American English Words & Gestures data). (We could measure the rate of learning by taking the derivative of this curve; but this rate calculation is slightly misleading for cross-sectional data and so we postpone this analysis until below). This increase in the rate of vocabulary learning has been much remarked on in the literature, as discussed above.

Our starting point here is a study by Ganger and Brent (2004), who provide a detailed, curve-fitting analysis of the question of vocabulary "spurts." They evaluate whether individual children's longitudinal growth patterns are better fit by a model with constant acceleration in growth rate, or whether some children have a discontinuous "step" in terms of their growth rate.

In our view, this is a productive approach, but presupposes some descriptive facts about children's growth rate generally. For example, using longitudinal data, we can simply examine features of rate and acceleration and how they change with time. For these analyses we focus on longitudinal production data from Norwegian and English WS and WG forms. Because we are interested in computing (potentially non-linear) changes in rate, we need four datapoints from each child as a minimum, and because we are interested in early changes, we set the restriction that the first time-point reported should have fewer than 50 words reported (50 words is often used as a semi-arbitrary cutoff for the vocabulary spurt; Dapretto and Bjork, 2000; Ganger and Brent, 2004).

The decision to exclude children with larger vocabularies at their first recorded measurement means that we have some bias present to include children with slower vocabulary growth. In particular, from the WS data, we exclude a substantial proportion of children even from the youngest groups (e.g., 22% of Norwegian 16-month-olds). So as not to bias the analysis further by including a large proportion of older, slower learners, we only include children younger than 21 months in this sample.

We now have a population of children for whom we can evaluate the rate of vocabulary growth and how it varies as a function of age. This winnowing leaves us with data from 290 children, whose growth curves are shown in Figure 14.8.

We exclude datapoints associated with large decreases in vocabulary (negative rates). Although some small negatives would be expected based on measurement error or forgetting, large negative spikes

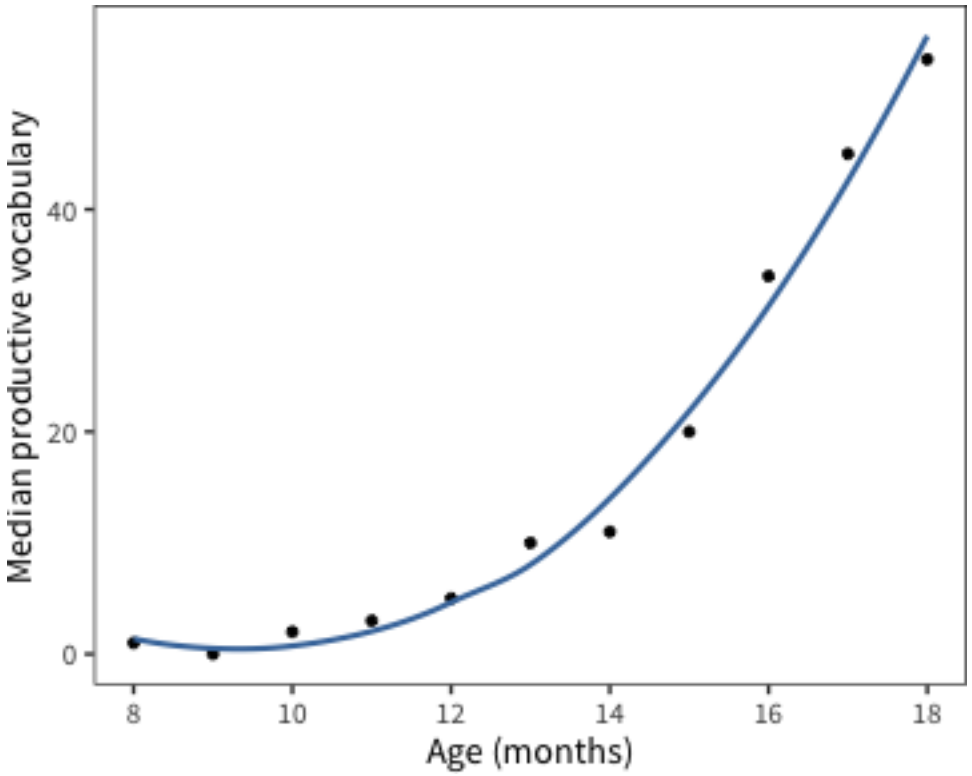


Figure 14.7: For American English data, median productive vocabulary as a function of age (curve shows smoothed fit).

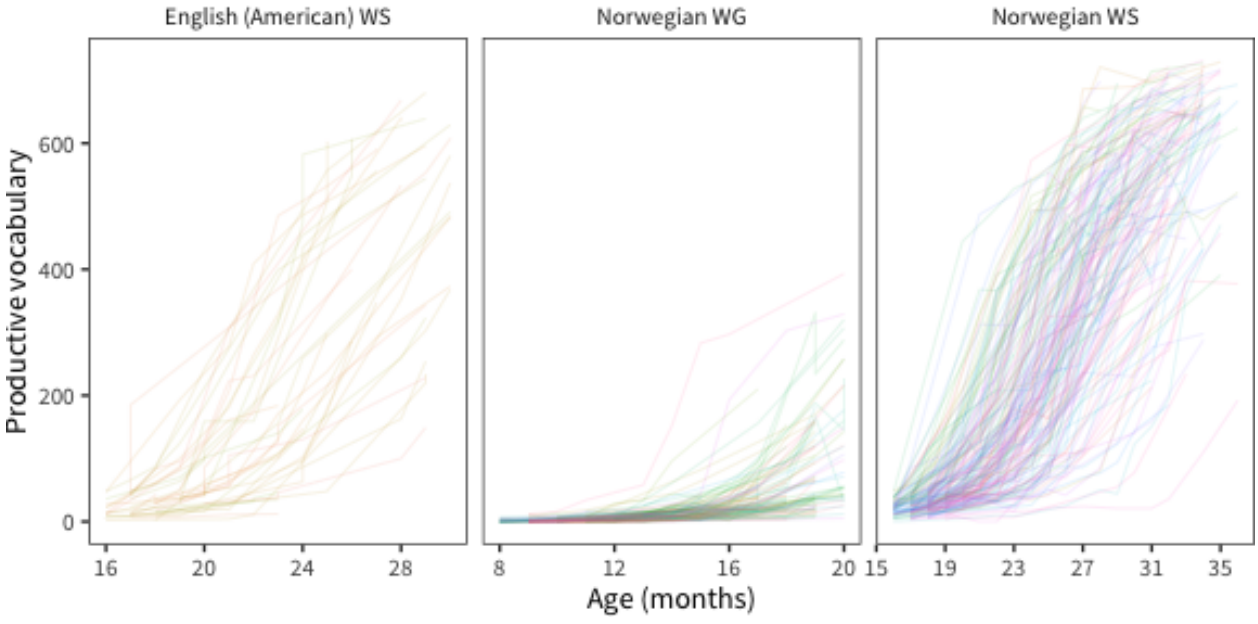


Figure 14.8: Vocabulary size as a function of age for included longitudinal administrations.

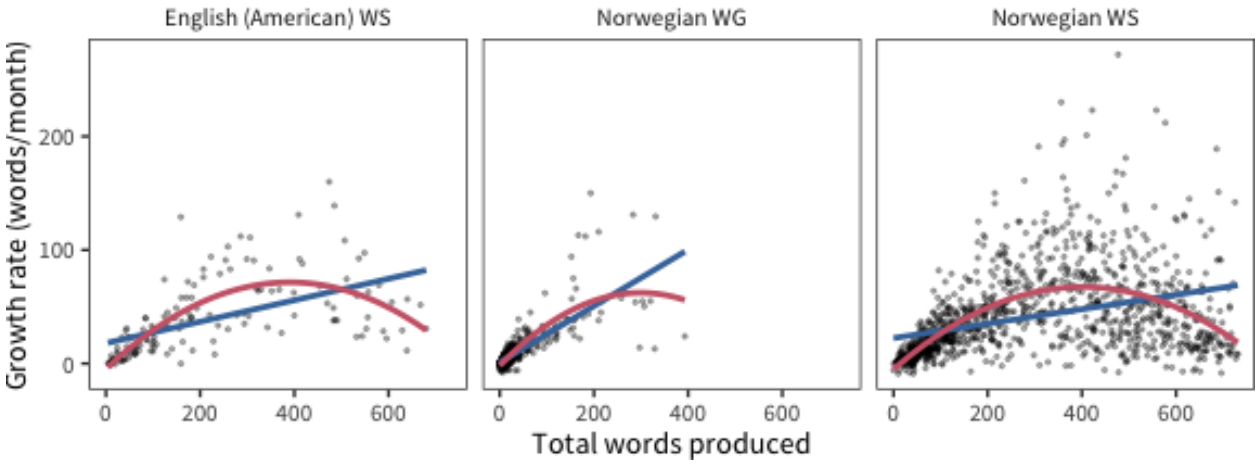


Figure 14.9: Vocabulary growth rate as a function of vocabulary size for all included longitudinal administrations (red line shows quadratic fit, blue line shows linear fit).

are rare and likely due to errors in the data (e.g., a partially filled form). We exclude rates of -10 words/month and below (1.3% of data).

Figure 14.9 shows the child's estimated growth rate in the measurement period leading up to that month (number of words learned since the last longitudinal measurement divided by number of months since that measurement) plotted by total words produced in that month. There is a clear quadratic shape to this pattern, almost certainly caused by ceiling effects for children who are "running out of words" on the form.

The question of vocabulary spurts, as posed by Ganger and Brent (2004), is the way that vocabulary growth rate increase for individual children. Clearly it increases (because vocabulary growth picks up speed generally) — but does it grow smoothly, indicating constant acceleration? Or does it move from one equilibrium to another (indicating an initial "spurt")? Ganger and Brent (2004) propose analyzing children's growth rate as a function not of age, but of total vocabulary. To examine this question, we need to focus in on the initial 250 words when the average rate appears to be increasing linearly (before ceiling effects are found; see red curve above), and identify children with more than 4 CDIs before this time (to ensure sufficient density). Further, we need to examine the rate trajectories of individual children. Figure 14.10 shows this analysis for a randomly sampled subset of children in our available datasets.

Ganger and Brent (2004) analyzed the question of developmental spurts by fitting different curve types to the rate function in their data. They compared the likelihood ratio of quadratic and logistic rate curves for each child's data. The quadratic curve represented the hypothesis of smooth growth in rate (smooth acceleration). In contrast, the logistic curve was of the form

$$R \sim \frac{\alpha}{1 - e^{-\beta(W-\gamma)}}$$

where R is the rate of acceleration, and it is assumed to be distributed as a function of α, the asymptotic rate, β, the slope of the change between the initial and final rates, W (the total vocabulary), and γ, the point at which the spurt occurs. This curve captures a discrete "spurt" — a movement from one equilibrium to another.

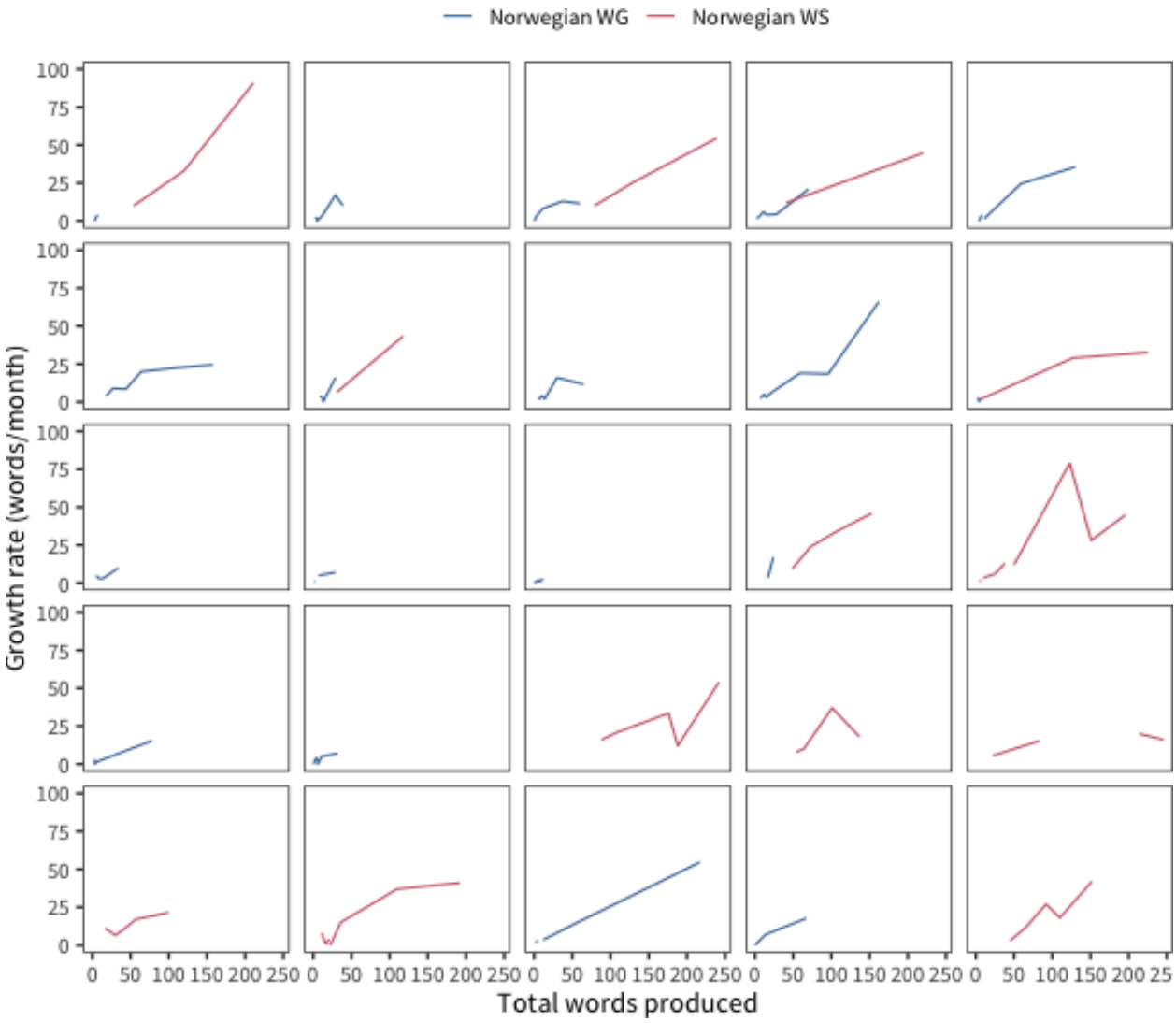


Figure 14.10: Vocabulary growth rate as a function of vocabulary size for 25 randomly sampled children.

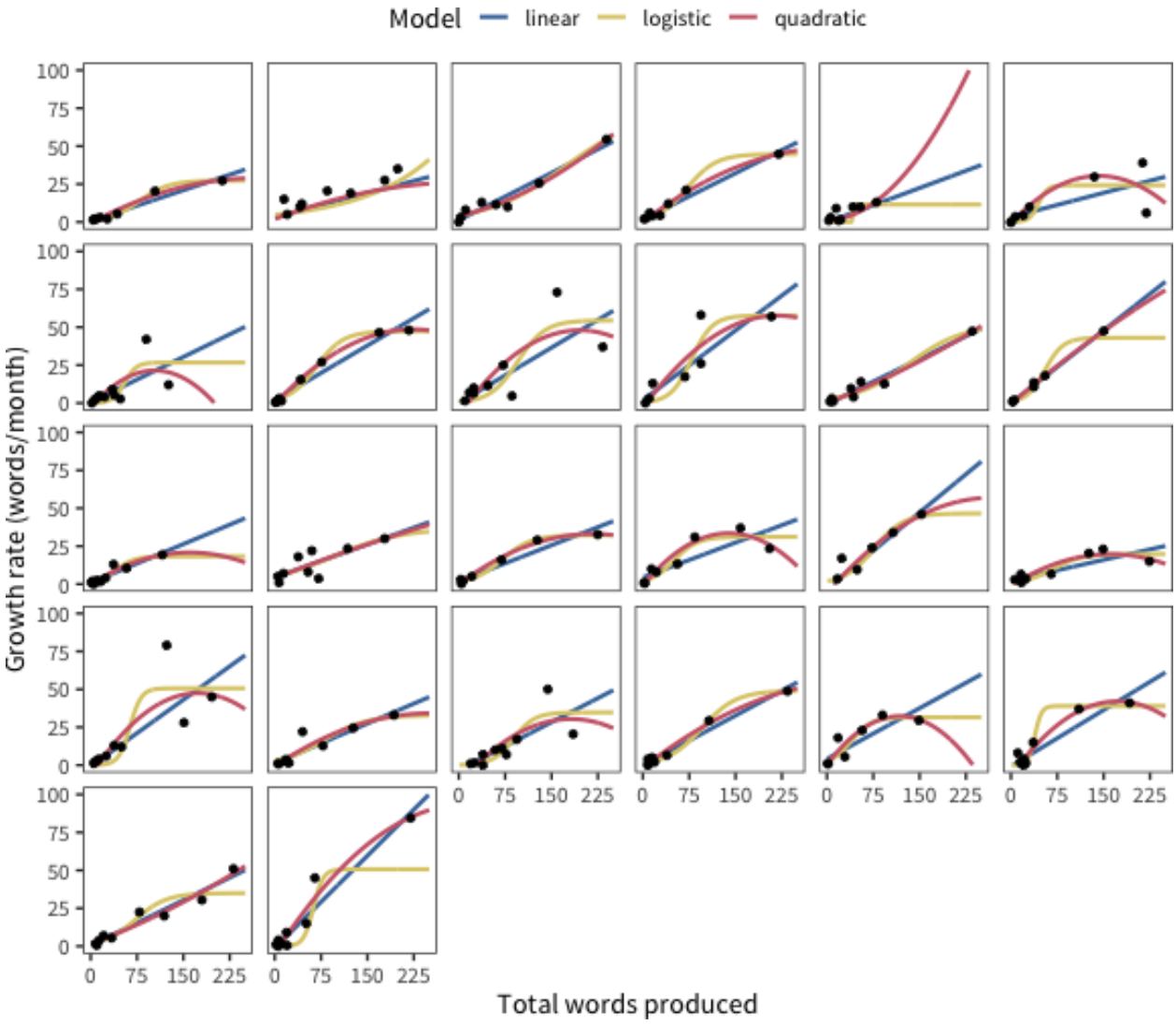


Figure 14.11: Vocabulary growth rate as a function of vocabulary size for children with more than 7 datapoints, with curves showing various model fits.

We fit these functions (as well as a simple linear function) to our data. Fitted curves for children with more than 7 datapoints are shown in Figure 14.11. The basic visual impression from these data is that, even with the longitudinal depth we have for individual children, there is substantial uncertainty in the best-fitting curve. However, it does not appear that there are many children who show something that looks clearly like a spurt. A few children rise quickly and then show one datapoint that levels off. But there is a puzzle. We can only confirm that the rate has leveled off if we have data at higher levels of production. But for vocabulary size over 300 words, every child will level off in their rate because they have "run out of words" on the form. This example illustrates some of the difficulties in making strong inferences from data of this type.

We next conducted the full model comparison analysis that Ganger and Brent (2004) conducted, over the 101 separate longitudinal records included.[3] In Ganger and Brent's analysis, they used only

[3]This number consolidates data between WG and WS forms when a child has both, to maximize our use of the

Table 14.1: Number of children for whom the best fitting model falls into each model type.

Instrument	Linear (no intercept)	Linear	Quadratic	Logistic
English (American) WS	2	1	0	0
Norwegian WG	35	2	17	20
Norwegian WS	11	3	4	6

two models (linear and quadratic), which had the same number of parameters. Accordingly, they compared only the likelihoods of the data under these models. In contrast, we compared a linear function with intercept at 0 (1 parameter), standard linear (2 parameters), quadratic (3 parameter), and logistic (3 parameter) curves. To make up for the difference in parameters, we computed Akaike's Information Criterion (a measure of model goodness of fit, where smaller is better) for each model for each participant.

Table 14.1 shows the proportion of children in each dataset for which different model types fit best. Overall, children were split between models, with some children best fit by the logistic. The linear functions, which were simpler, however, fit more participants better, with the 1-parameter linear model with no intercept best fitting more children than any other.

The 21 children (out of 101 total) with a best-fitting quadratic model are shown in Figure 14.12. Some of these children do appear to have data that are well-fit by the quadratic model. But for many, this fit appears to be the product of a single datapoint; assuming some error, a more parsimonious model (e.g., simple linear) might do just as well. Thus, with more children but less density, our conclusion tends to be similar to Ganger and Brent (2004): there is limited evidence for a vocabulary spurt in most children.

To further examine this issue in a denser dataset, we used data from Roy et al. (2015)'s in-depth study of a single child. This is an ultra-dense dataset with millions of words of transcribed speech and hand-checked age-of-acquisition data for over 600 words up to the child's second birthday. The comparable curves for this dataset are shown above. Using the same AIC method, the quadratic model fits best, but the linear model is clearly close as well.

Stepping back, in this subsection we examined the growth rate of children's vocabulary. To a first, group-wise approximation, children's vocabulary growth accelerates linearly with vocabulary size during the initial period (up to around 250 words). After this point, we run into substantial measurement issues because the CDI does not contain enough words to be certain of the pattern of growth. Further, when we examined individuals' growth, it also often appeared to be linear or quadratic; only in a minority of individuals was there any evidence for a "spurt" (a discrete change). This conclusion was tempered, however, by the difficulty of drawing conclusions without even denser longitudinal data concerning the very beginnings of language.

14.3 Variation in production vs. comprehension

Our next investigation concerns the question of how tightly comprehension and production are yoked within CDI data. Our assumption is that there is variability between children on this dimension — while some children can say a large portion of the words that they understand, others appear

available longitudinal data.

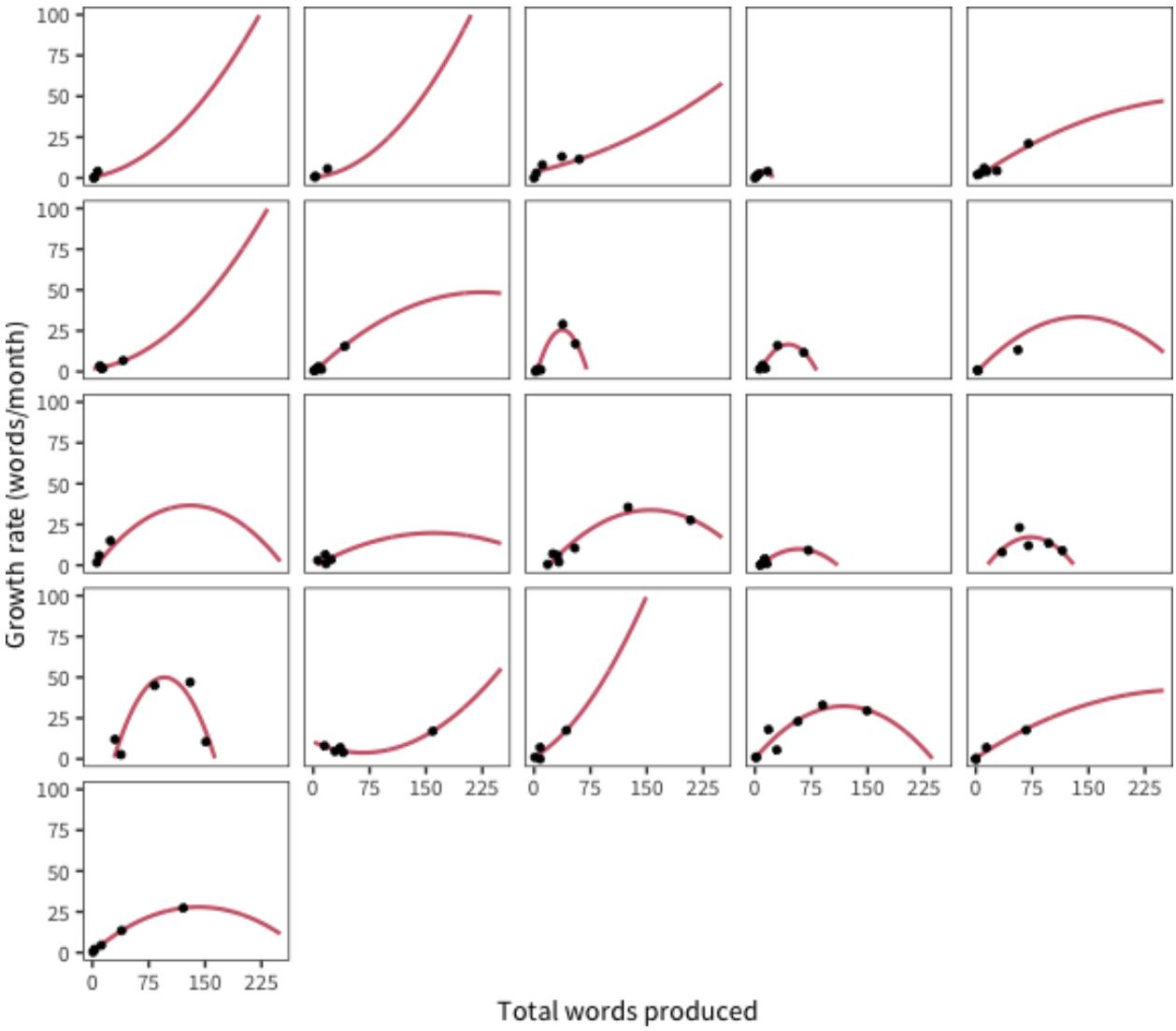


Figure 14.12: Vocabulary growth rate as a function of vocabulary size for children for whom the best fitting model is quadratic, with curves showing quadratic model fits.

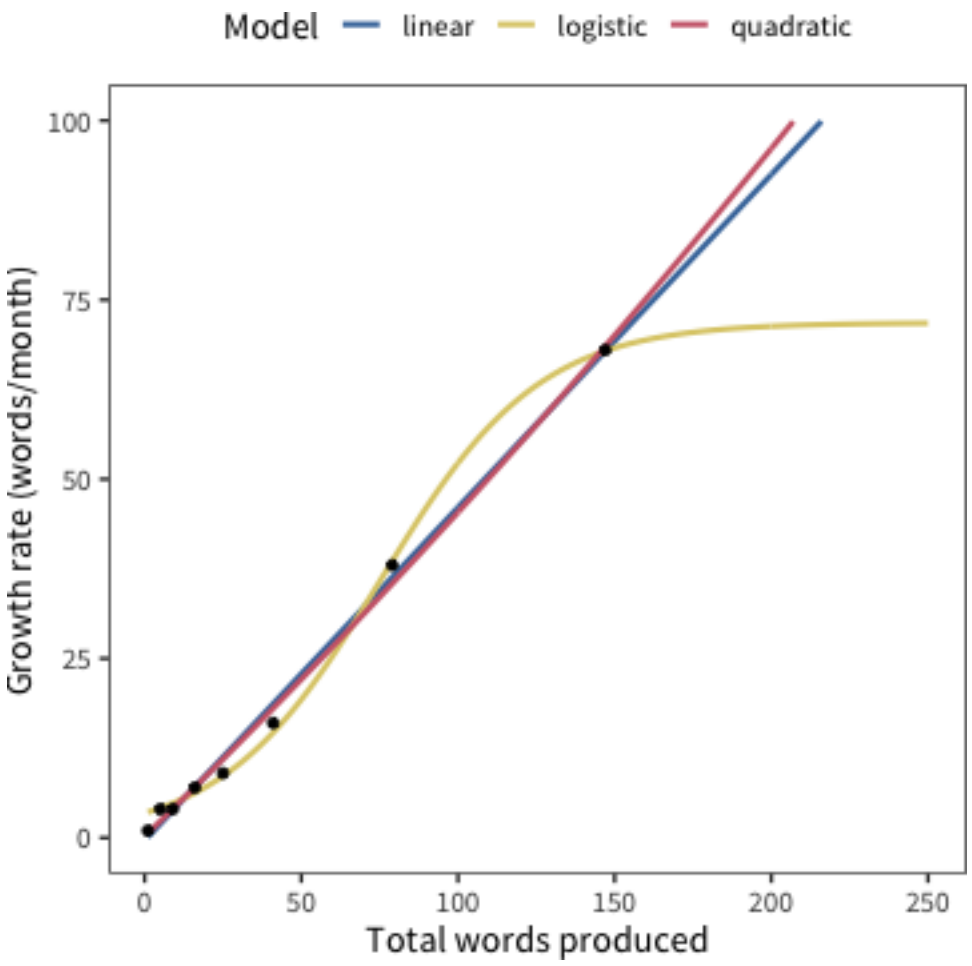


Figure 14.13: Vocabulary growth rate as a function of vocabulary size for the @roy2015 data.

to produce less but still understand substantial amounts. How does the ratio of production to comprehension vary across ages, and across languages?

It is important to be clear that some of the pattern in this variable could be due to variation between parents in under- or over-reporting comprehension (or for that matter, production, but we assume — and Chapter 4 confirms — that production reports likely carries more signal). For example, we might be detecting variation in the threshold at which parents assume that a response indicates comprehension. Some parents might be very liberal and recall a generally-understood story that included a particular word, while others might be searching for a specific anecdote that clearly illustrates comprehension of that word. Compared to the more observable facts of word production, the variation might also reflect that comprehension is a more difficult concept for parents to grasp, so there will be differences in across parents in what "comprehends" means. Nevertheless, there is evidence that children's level of early comprehension is a useful metric for identifying possible language delays beyond production data alone. That is, children with only a few words who also have low comprehension scores are at increased risk for language delays compared to children with a few words who nonetheless appear to be understanding language well (e.g., Rescorla, 2009).

Of course, this type of analysis can only be conducted on WG-type forms, because of the presence of comprehension information. We begin by investigating the American English WG data as an example.

Figure 14.14 shows individual children's comprehension and production plotted against one another. The diagonal indicates a child who comprehends and produces exactly the same number of words. In practice, this measure is always below the diagonal because, by the design of the form, a child cannot "say but not understand" a particular word, they can only "understand" or "understand and say."

We can convert these data into a productivity ratio:

$$\text{productivity} = \frac{\#\text{produced}}{\#\text{understood}}$$

Figure 14.15 shows this ratio for all children.

The resulting scatterplot is quite interpretable. It contains a few outliers at the very top of the range for very young children (whose parents report them producing and comprehending the same number of words). But for most others, the ratio is low, increasing from about 10% to 30% by the top of the form.

Figure 14.16 plots these productivity ratios by language for an age-restricted subset between 8 and 18 months. Plots are sorted by the mean productivity ratio. While the majority of languages show the same pattern as English (an increase from around 10% to 30%) there are some outliers that show a flatter slope.

We can see this pattern even better by plotting the best-fit lines across languages (Figure 14.17 and Figure 14.18). Nearly all of these go up with age and have similar slopes.

However, in nearly every language, to one degree or another, we see some number ratios $> .95$, indicating that parents are essentially not using "understands" as a separate option. Table 14.3 shows the proportion of children showing more than 95% productivity. A number of samples have substantial proportions of parents reporting comprehension in this way. While it is possible that these numbers represent actual children whose production is synchronized with their comprehension, a more parsimonious explanation is that there are local variations in administration, leading to some

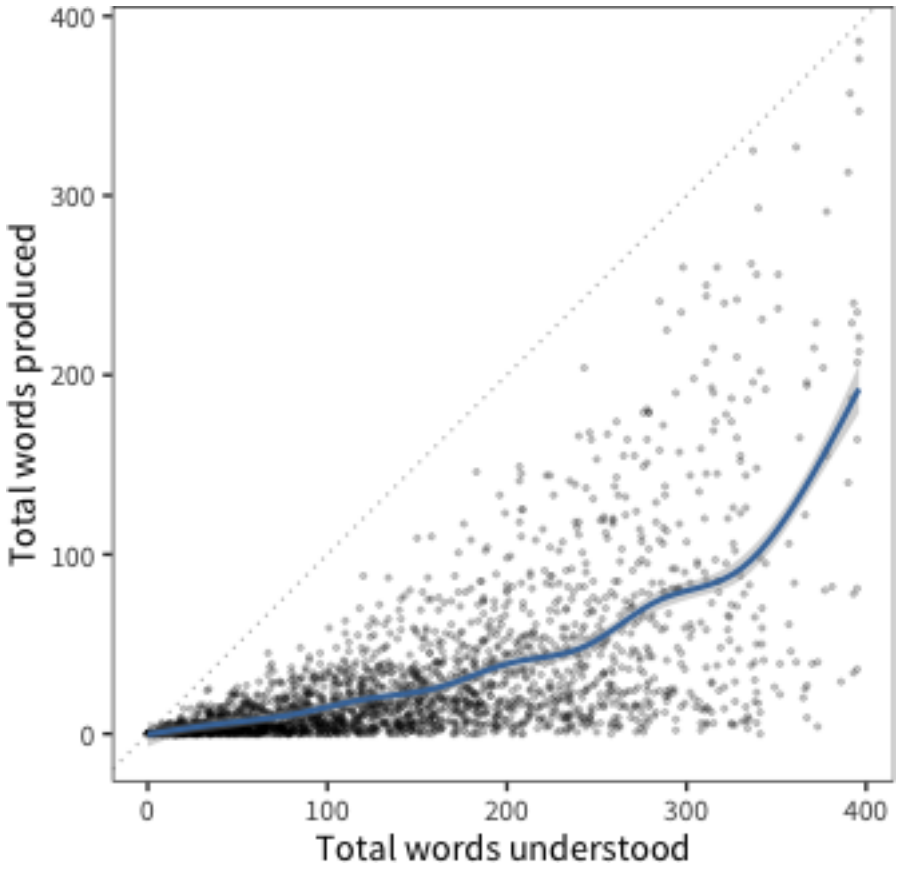


Figure 14.14: For American English data, each child's productive vs. receptive vocabulary size.

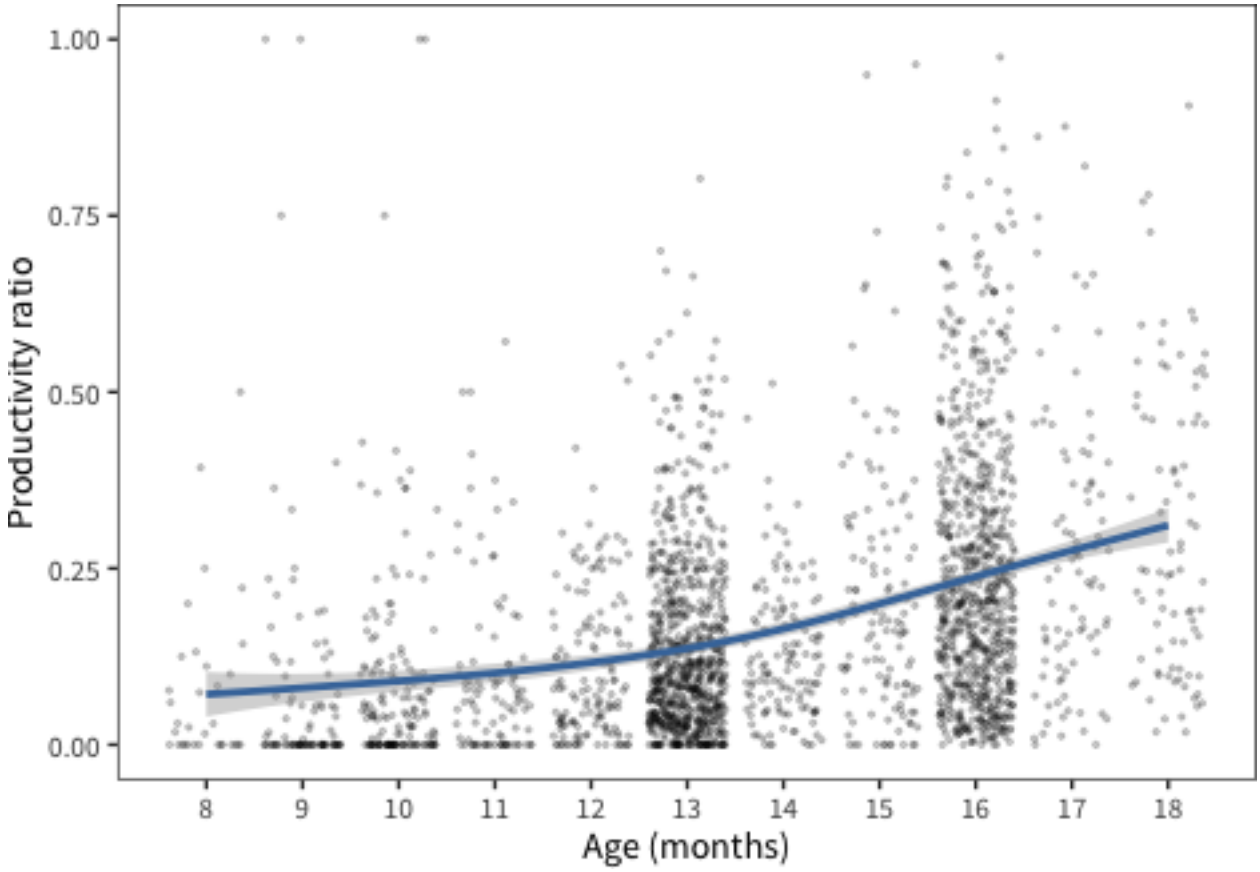


Figure 14.15: For American English data, each child's productivity ratio as a function of age.

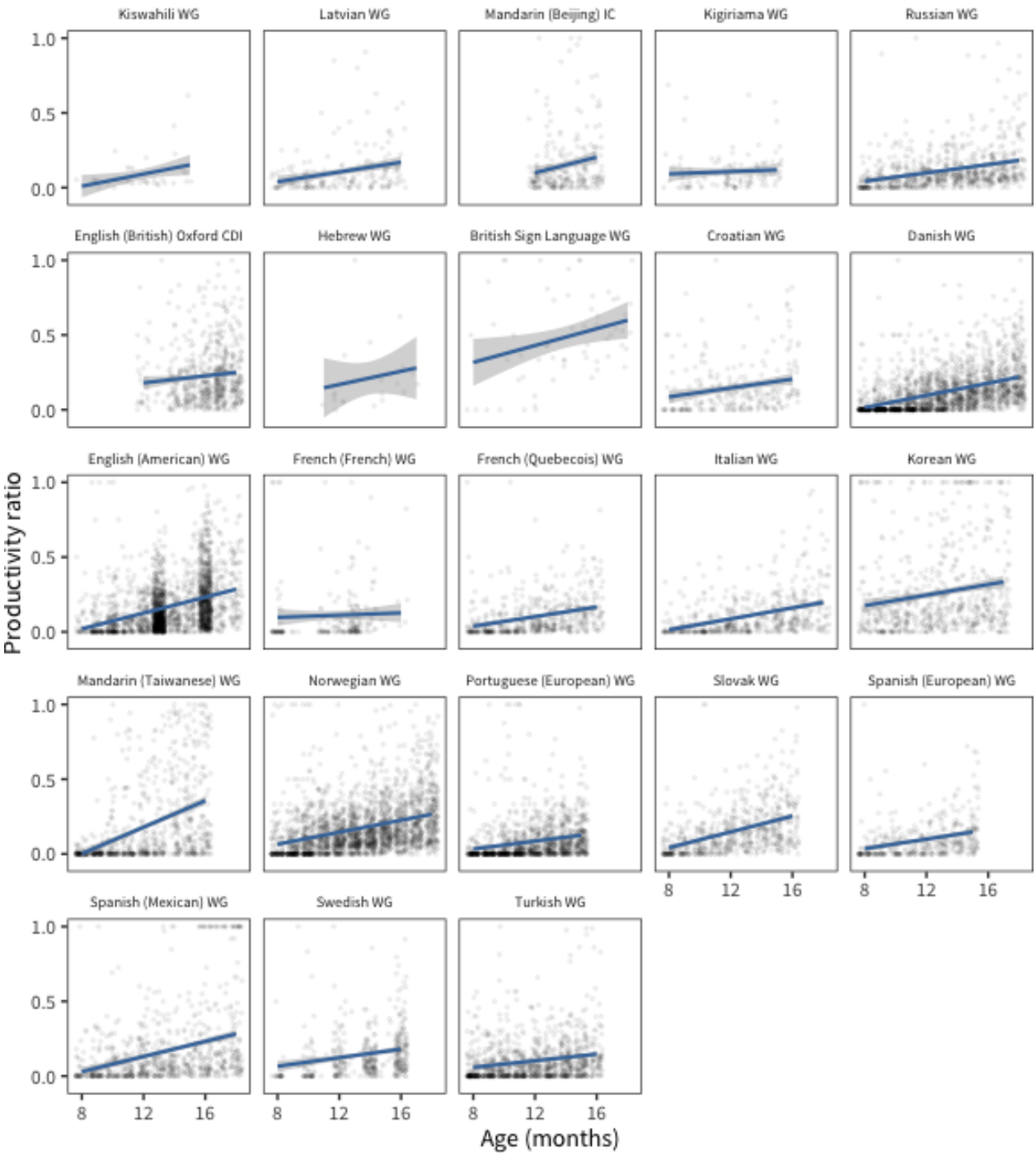


Figure 14.16: Productivity ratio as a function of age for each child in each language (lines show linear model fits).

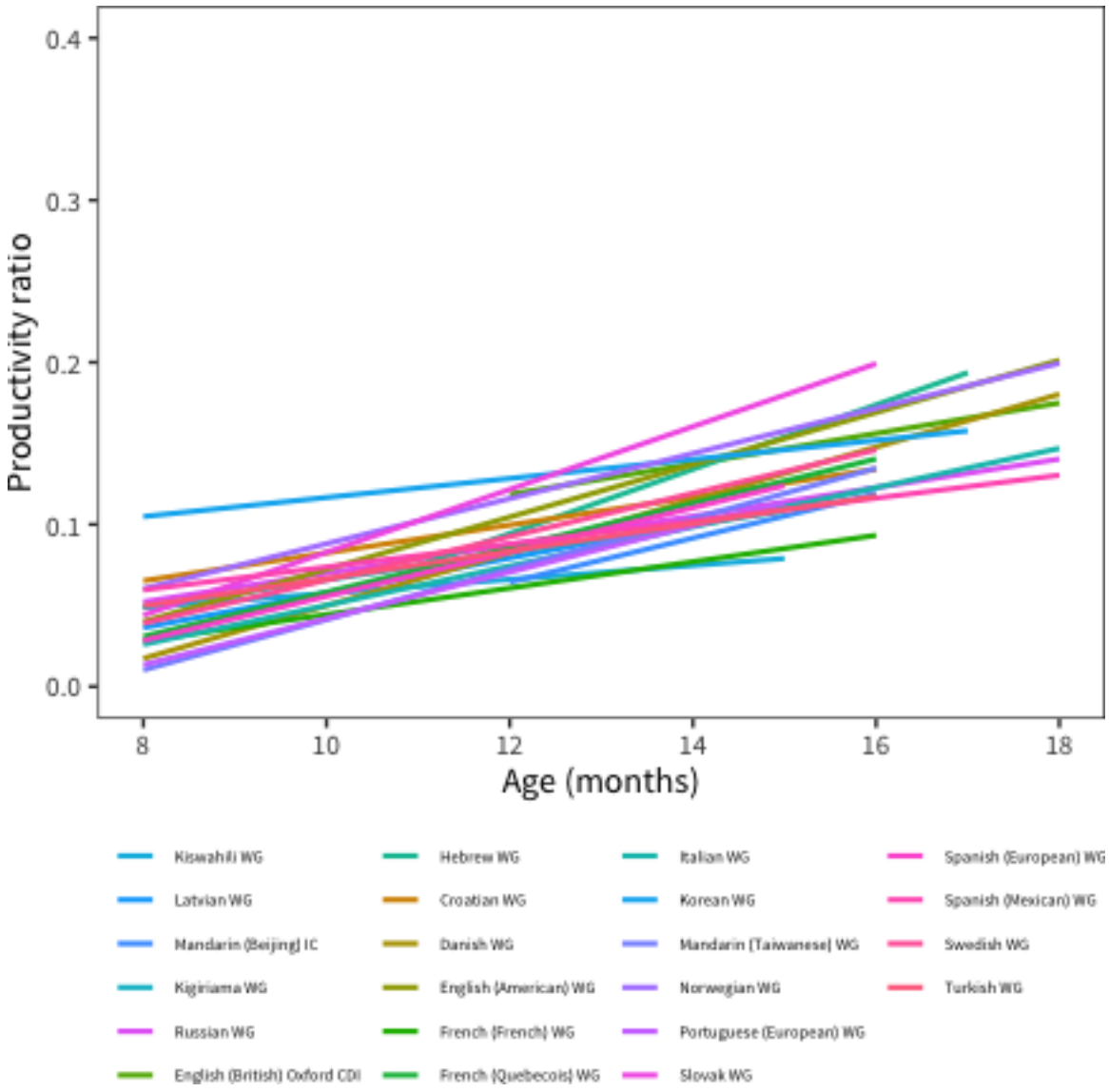


Figure 14.17: Linear model fits of productivity ratio as a function of age for each language.

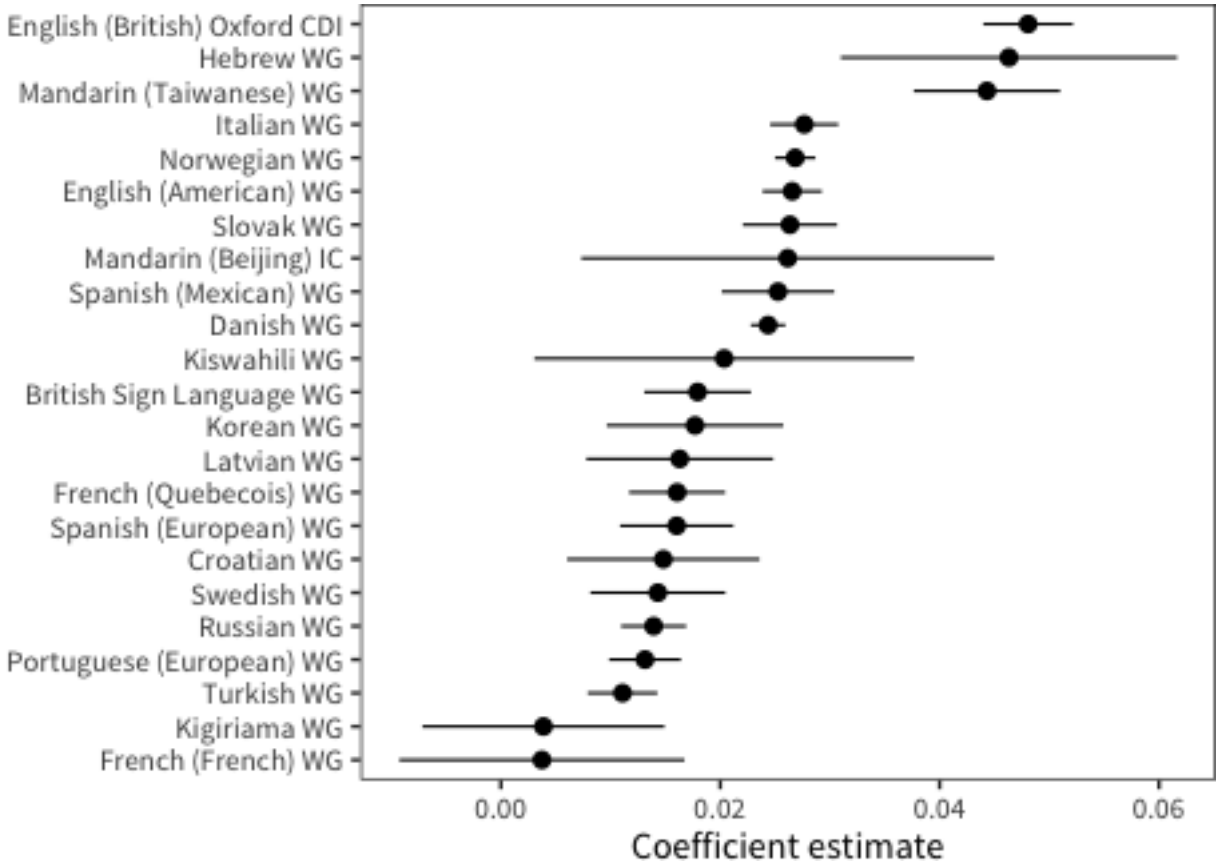


Figure 14.18: Effect of age on productivity ratio in each language (ranges indicates 95 percent confidence intervals).

fraction of parents who are not completing the form properly. In particular, it does not appear that these "no comprehension without production" children are the tail of a shifted distribution of productivity ratios; instead, they appear to be due to a separate small population. Yet despite that they appear to have an outsized effect on our estimates of the development of productivity ratios across languages.

Proportion of administrations in each language for which the productivity ratio is greater than 95 percent.

Instrument	Proportion
Hebrew WG	0.098
British Sign Language WG	0.089
Korean WG	0.069
Spanish (Mexican) WG	0.046
English (British) Oxford CDI	0.036
French (French) WG	0.029
Mandarin (Beijing) IC	0.017
Mandarin (Taiwanese) WG	0.015
Norwegian WG	0.01
Italian WG	0.0095

Showing 1 to 10 of 23 entries Previous 1 2 3 Next

In sum, although the relationship between production and comprehension is a fascinating locus for individual differences, we may not be able to measure this relationship effectively using cross-linguistic comprehension data. Further, these analyses underscore the importance of developing instructions for parents that more effectively convey the concept of what it means for a child to "understand" a word. Outside of these concerns, while there very well may be reliable variation in the production-comprehension ratio across children, we unfortunately do not have access to independent signals that could validate this quantity.

14.4 Discussion

In this chapter, we have explored three classic indices of individual variation in children's "style" of learning, that is, variation in how children approach the task of language learning. Compared to differences in rate/timing of learning, stylistic differences have been proposed to be a stronger test of the proposal that language acquisition is a constructed process. But, few studies to date have had the power and scope to explore the extent of these differences cross-linguistically.

We can summarize the conclusions from our various different sub-analyses. First, we were not able to substantiate differences between individuals in production/comprehension differences. It is tricky to use comprehension data to estimate variability between individuals in how much they produce vs. comprehend, due to likely cross-linguistic differences in the uptake of instructions regarding comprehension. In contrast, we did see support for the traditional referential vs. expressive distinction from cross-linguistic analysis. We found that, even at a given level of vocabulary size, some children tend to be younger, use more nouns, and combine less; others tend to be older, know more predicates, and combine words more. Finally, confirming previous work by Ganger and Brent (2004), we found

no evidence supporting a "vocabulary spurt." Instead, we found that, while vocabulary growth accelerates, that acceleration is approximately linear for the first 250 words, and hence a "vocabulary spurt" is not reliably observed across most children.

In sum, our analysis does confirm the existence of stable individual variation in language learning "style." Our work also highlights the methodological difficulty of this enterprise, however; despite the tremendous amount of data we have access to, our analyses still suffer from the limitations of the parent-report format of the CDI; it is often difficult to tell whether we are measuring children's style or parents'. Overcoming this limitation would require non-parent report measures with equivalent domain breadth, psychometric stability, and sample diversity to the CDI; at present to our knowledge, no publicly available datasets of this type exist.

Chapter 15

Variability and Consistency

In the preceding chapters, we have have presented evidence from a wide range of analyses of the Wordbank dataset. These analyses have revealed both striking variability across our units of sampling — children, primarily, but also words and even languages — but they have also revealed substantial consistency. In this short chapter we present a set of analyses of the consistency and variability of specific phenomena. The final chapter, Chapter 16, synthesizes these observations.

In each of the substantive chapters of this book (roughly speaking from Chapter 5 to 14), we have presented analyses of specific phenomena of theoretical interest. Wherever possible, we have generalized these analyses across languages so that their relative consistency can be examined and discussed. The goal of the current analysis is to bring together these analyses into a single meta-analysis (in the general sense, not the specific statistical sense).

This strategy executes the general idea discussed in Chapter 1: Our chapters have identified "signatures" of language development, measurements that we believe are theoretically interesting or central. We then quantify these signatures and their variation across languages. The resulting numbers represent an empirically-derived measure of which aspects of language development are more similar across different languages and contexts. We discuss the types of inferences licensed — and not licensed — by these analyses below.

15.1 Methods

We begin by identifying a small number of measures computed in each chapter to serve as the "signatures" to be promoted into this analysis. For each measure, we compute its cross-linguistic variability using a standardized measure of variance, the coefficient of variation (CV, the standard deviation divided by the mean). This measure can range from 0 (indicating a phenomenon that is completely invariant across languages) to infinity (with higher numbers indicating greater variation). These CV values provide a single common measure to allow comparability of otherwise very different quantities, allowing inferences across analyses and datasets — albeit with some cautions that we describe below.

Each measure for which we compute the CV will have both a different base unit and a different number of languages contributing. For example, when considering the correlation between grammar and the lexicon (Chapter 13), we will be looking at the CV of a set of correlations with one specific

set of languages contributing. In contrast, when we look at the size of the noun bias (Chapter 11), we will be looking at a group of bias estimates that have different units and a different set of languages. Thus, caution is warranted in interpreting these variability estimates, even though we believe that they indeed are informative. To assist in interpretation, we exclude measures that can be computed in fewer than 7 languages; provide the N contributing languages for all analyses; and compute an estimate of the standard error of the CV ($SEM \approx CV/\sqrt{2N}$).

The set of signatures we include in this analysis are necessarily a subjectively-determined subset of the possible measures we have examined in the book. And, of course, those in turn are a subset of the measures we could have computed. Wherever possible we have attempted to make reasonable decisions, but some of these are, by necessity, somewhat arbitrary. An example of such a decision comes from the summary of Chapter 5. In that chapter we noted that population variability appears quite consistent across languages. We summarized population variability in production via a statistic, MMAD — but what is the appropriate range of ages to include in a single estimate? In this chapter, we observed that there appears to be a ceiling effect in the later ages. Thus, we decided to include variability in production from 12 — 24 months. But this decision is data-dependent and so, of course, there is a risk of circularity. We point the issue out not to undermine this particular analysis; we believe the ceiling effect is quite clear and other aspects of the age choice do not lead to much change in the CV estimate. Rather we intend to highlight that the summary we give is not a theory-neutral estimate but rather a "best guess" — an attempt to navigate the myriad choices involved in our analysis in a reasonable way.

One example of such a choice is that we have made the decision to omit estimates of early production from WG-type forms. Our judgment was motivated by the fact that such estimates are routinely quite noisy and difficult to interpret, likely due to the sparsity of early production. In chapter after chapter, we found unreliable or uninterpretable results that are plausibly due to data sparsity; thus, we choose to omit these patterns.

15.2 Results and discussion

Figure 15.1 shows the coefficient of variation across languages for all measures in each of these categories. For the sake of our analysis, we have divided measures from the preceding chapters into four categories, corresponding to the panels of Figure 15.1. These are:

Measures of the composition of vocabulary, from Chapters 11 and 12. These measures describe the over- and under-representation of various word categories in vocabulary. The units over which CV is computed are bias scores; these are bounded from -.5 to .5 (deviation from unbiased acquisition of a particular category).

Predictors of word difficulty, from Chapter 10. The consistency of different regression predictors of age of acquisition are here represented by their cross-linguistic consistency. This analysis is distinct from the analysis presented in that chapter (which focused mostly on the magnitude rather than variability of the coefficients themselves). Despite that, we include it here for comparison with other signatures. The units over which CV is computed are standardized regression coefficients.

Relational measures, from Chapters 7 and 13. These measures are originally correlations between vocabulary size and other aspects of early language.

Vocabulary signatures, from Chapters 5 and 6. These measures document patterns in the overall

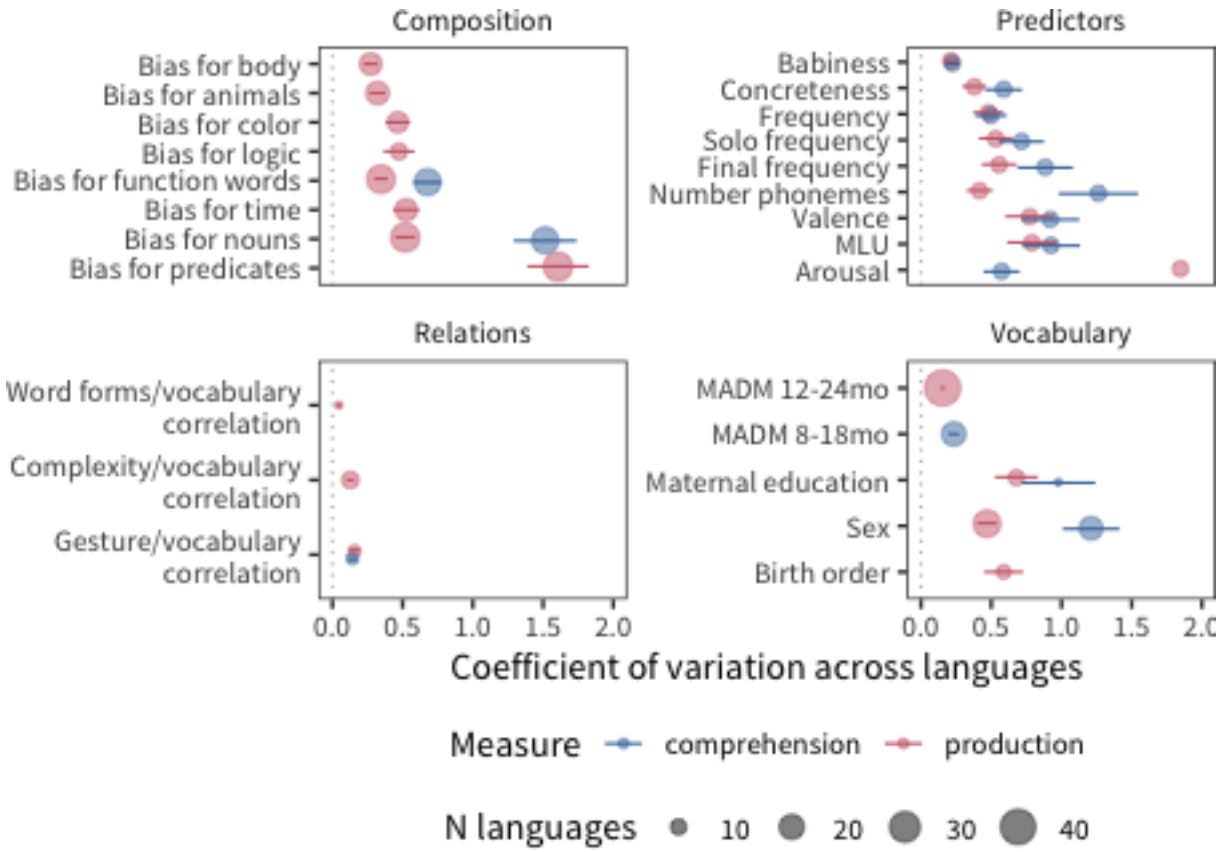


Figure 15.1: Coefficient of variation across languages for signatures of language development corresponding to four different categories (panels). Each point gives an estimate, with point size corresponding to number of languages. Color indicates whether comprehension or production is measured. Error bars give the standard error of the coefficient of variation.

size of vocabulary across individuals and demographic groups. The original units are themselves variability-based (MMAD scores).

A number of local patterns are immediately apparent; we discuss these briefly below before turning to larger-scale generalizations in the next section.

First, comprehension is almost always more variable than production, even when a comparable number of languages are included. The only obvious exception to this regularity is for coefficient weights on arousal in the Predictors category — and we can discount this example: arousal was on average one of the weakest predictors of acquisition order overall. Why would comprehension be more variable across languages than production, especially given evidence that comprehension vocabulary tends to be less idiosyncratic than production (Mayor and Plunkett, 2014)?

One strong possibility is the psychometric properties of comprehension vs. production reports. As described in Chapter 4, while comprehension scores are likely still a reliable and valid index of children's abilities in the aggregate, individual comprehension questions tend to carry less information. Thus, there may simply be more noise in these measurements, leading to less cross-linguistic stability. This regularity illustrates a point we have made earlier and will return to below: the inferences from consistency and variability are asymmetric. In the case of consistency, we can make relatively strong inferences about some kind of shared process or mechanism. In contrast, in the case of variability, there are many sources of variance (including measurement error) that can account for a specific pattern of performance.

Second, relational measures are highly invariant across languages. These relations include correlations between the size of children's lexicon (in production or comprehension) and their gesture, morphology, and grammatical complexity. These findings can be measured in a relatively small set of languages (due to limited data availability for the gesture and complexity items on the form). Nevertheless, the high level of consistency is striking.

Third, demographic predictors — birth order, maternal education, and sex — are somewhat variable, likely reflecting at least some cultural variation. The most variable of these is the relation of maternal education with vocabulary. Maternal education is plausibly a proxy for socioeconomic status (SES); in turn, the relation of SES to vocabulary is likely mediated by many local- and national-level policies including access to childcare, parent leave, pre- and post-partum education and services. Thus, we view variation on this dimension as highly plausible a priori. In contrast, as we asserted in Chapter 5, the consistency of children's variability is quite striking: the variability of children across instruments is almost completely constant, especially for production. Around the world, toddlers appear to be have similar levels of variability in their level of production.

Finally, while predictors of word difficulty are consistent relative to a chance baseline (see Chapter 10), they are also on the higher end of variability. Especially in comparison to some of the relational signatures (e.g., the correlation between grammatical complexity and vocabulary), their variability is high. Plausibly some of this variability is due to additional variation added to these estimates by the use of external resources (e.g., corpus counts). In particular, as we discussed in that earlier chapter, there is an unknown amount of noise added by estimating frequencies from smaller-scale cross-linguistic corpora.

Chapter 16

Language Development at Scale

This chapter synthesizes knowledge gained from the broader enterprise, in the sense of attempting to make generalizations about early language development that hold true across the different datasets in our sample. We begin by briefly reviewing the content of our findings across the preceding substantive chapters. Next, we discuss three synthetic generalizations from these findings. We then turn to the question of the particular process universals that might be compatible with this empirical picture.

16.1 Summary

We began with the question of what parent report can tell us about children's early language. We are now in a better position to answer this question, summarizing findings across the content chapters of the manuscript.

In Chapter 4, we reviewed evidence for the reliability and validity of parent report instruments for assessing children's early language, with generally positive conclusions. Further, we contributed two sets of novel analyses. First, examining large-scale longitudinal datasets, we estimated the fall-off in test-retest correlations over developmental time, which suggested a relatively high level of reliability overall. Second, we used item response theory (IRT) models to examine the measurement properties of individual words on the CDI. Overall most items had strong psychometric performance, and even comprehension reports appeared to have consistent informational value about a child's overall abilities (although individual items appeared less diagnostic on average than production reports). The broad picture from this chapter was positive, motivating further use of CDI data.

Chapter 5 then examined trends in the growth and distribution of individual vocabulary estimates. Across languages, and for production and comprehension, early vocabulary development starts slow and grows rapidly in the second year after birth. There are both absolute and relative differences between languages in the estimated rate of vocabulary growth, however. While some of these may reflect true differences (for example, Danish being famously difficult to acquire), others likely stem from differences in instrument, sample, and administration conditions. On the other hand, we observed a striking consistency across languages in the degree to which children varied. This finding suggests the operation of individual- and dyad-level processes that vary systematically in a way that is somewhat independent of the precise content being acquired and the cultural background of acquisition. On average, language emerges quickly — but despite the average pattern, toddlers

around the world are "all over the place" in their learning rate. Indeed, one of the most compelling universals in language development is the variation that is observed across children.

Chapter 6 considered demographic predictors of vocabulary size. We found a general female advantage in early vocabulary that was more pronounced for production than comprehension. This advantage did not appear to stem from reporting bias, and was relatively consistent in size across languages. We also observed a first-born advantage and an advantage for children with more-educated mothers (likely a proxy for general socio-economic status). Chapter 9 followed up these analyses by identifying whether specific words were the locus of some of these differences. The conclusion of these analyses were that there were both general demographic differences in vocabulary and word-specific differences. For example, girls, on average, both have a higher probability of knowing any word, and they also have a much higher probability of knowing words like "dress" and "doll."

Turning to gestural communication, Chapter 7 showed that children's uses of early gestures are both quite consistent with one another and quite consistent across languages. In other words, gestures appear to group together as part of a set of communicative abilities (rather than existing as isolated items). A small set of similar gestures are often the earliest emerging across languages, likely arising from relatively universal dyadic needs (e.g., a signal for "pick me up" or "give me"). Finally, variation in early gestures is quite well correlated with variation in early vocabulary within individuals.

Like early gestures, early words are often surprisingly similar across languages. In Chapter 8, we quantified this similarity and showed that it was related to typological relatedness. On the other hand, a number of words occur in children's earliest vocabulary, even across quite dissimilar languages. Like early gestures, these early vocabulary items likely reflect relatively universal communicative needs in infant-caregiver dyads. (We return to this theme below in Section @ref{generalizations}).

In Chapter 10, we extended this approach to the composition of vocabulary by creating predictive models of the ease or difficulty of individual words. These models, uniquely in our analyses, take into account the average environmental input for a learner of a particular language (estimated from corpus resources). Although a number of factors including input, conceptual factors, and phonological complexity interact to predict when words are first understood or produced, the profile of predictors across languages was quite similar. We interpret this finding to point to a relatively consistent set of processes that are operating in different linguistic contexts.

Following the thread of cross-linguistic differences, Chapter 11 examined the question of the syntactic composition of vocabulary across early development. Consistent with previous reports, we found that most languages exhibited a bias towards nouns, although the strength of that bias varied across languages. Further, essentially all languages showed a developmental bias against function words. On the other hand, the bias for predicates (verbs and adjectives) was more variable across languages, with most languages showing a negative bias, but a few (Mandarin and Cantonese, in particular) showing a neutral or positive bias. This pattern of variability points to some language-specific factors (perhaps syntactic structure, perhaps other linguistic or interactional factors) that influence learning within particular syntactic categories.

We next applied this same approach to semantic categories (Chapter 12). Computing the relative bias for or against particular semantic categories, we found some cross-linguistic consistency in early biases: biases for body parts, games, and onomatopoeia were quite consistent; biases against words for places and time were as well. These biases suggest that there are likely some attentional and conceptual biases that facilitate or inhibit word learning in similar ways across languages.

Turning next to syntactic development, Chapter 13 showed that children's morphological and

grammatical ability appears tightly coupled with their vocabulary size. This generalization holds very consistently across languages, validating Bates' speculation that there is a general law of language development that is shared across vocabulary and grammar learning. This chapter also extended earlier analyses of these relations by identifying a moderating effect of age on this relationship such that older children appear to gain more "units grammar" per "unit vocabulary" — that is, grammar emerges slightly faster for these older, compared to younger, children.

The last empirical chapter, Chapter 14, examined individual variability in language. This chapter presented evidence for the "referential" vs. "expressive" distinction — that, even controlling for overall developmental level, some children appear to have larger, "noun-ier" vocabularies while others have smaller vocabularies but are more likely to combine words. Similarly, some children certainly produce relatively more of their total vocabulary than others (though these measurements are difficult to validate independent of reporting variability). On the other hand, we did not observe strong evidence for the idea that some children's vocabulary growth is particularly "spurt-y" — rather, once statistical artifacts are accounted for, vocabulary growth appears to accelerate relatively smoothly in all children, varying quantitatively along a rate continuum.

Finally, in the previous chapter, Chapter 15, we attempted to consolidate the cross-linguistic variability in all of the findings above and place these on a single axis. This comparison revealed a number of patterns that we take up in our generalizations below.

16.2 Generalizations

What is the picture of language development that emerges from these individual findings? We now return to the theoretical project of Chapter 1, which was to use patterns of consistency and variability to provide constraints on theories of early language learning. We begin by considering three major generalizations from our data, and then turn back to the question of possible learning processes that could support these generalizations.

16.2.1 The language system is tightly woven

It could easily be the case that children are "pointy" with respect to language — that is, highly proficient in some sub-aspect of languages and poor at others. A naive examiner of a group of children in a toddler classroom might conclude that the variation that they observed was supportive of such a view. In such a classroom, some children will only be able to gesture and say a few words; while some others will appear to show a large noun bias; and still others will be fluently combining words. Furthermore, these children will be distributed across a range of ages, with this progress only somewhat predicted by relative age (especially in a tightly-grouped sample from say 18–30 months).

This viewpoint is largely incorrect, however. Instead, the language system is tightly woven together, such that, on the whole, children who are good at gestures have a larger vocabulary, and children who have a larger vocabulary also inflect and combine words more proficiently. These correlations within individuals are large and they are universal across those languages in which we can evaluate them. Further, while these correlations could, in principle, arise in CDI data from parent reporting bias across sections of a single form, they are also largely borne out in a variety of observational and behavioral datasets (e.g., Brinchmann et al., 2018; Rowe and Goldin-Meadow, 2009).

How can the consistence and coherence of these distinct aspects of early language be reconciled with the type of variation our preschool observer might see? Consider the following model of early language. The first posit of this model is that the trajectory of children's learning follows a single, coherent path through the stages of early language. Initially, gestures and early nouns and social routines are learned. Nouns follow, leading into verbs. Verbs in turn promote word combination and morphology. The second posit is that aspects of this trajectory may be non-linear: the noun bias increases and then decreases (Bates et al., 1994); the relation between lexicon and grammar may be flatter at one period than another (Dixon and Marchman, 2007). Finally, imagine a third posit: this non-linear trajectory is followed at different rates, with a child's current state partially a function of age and partially a function of their own idiosyncratic learning rate.

Under such a "unitary non-linear trajectory" model, we would still see "pointy" children in the preschool classroom. These children would be cross-sections cut from the same trajectory, but at different points along it. The younger child who knows a surprising number of names might be further along relative to age and hence firmly in the "noun bias" part of the general path. In contrast, the older child who primarily gestures might be a child with a slightly slower growth rate. Observed variation need not indicate that the subparts of language do not "hang together" (to return to the Batesian phrasing). We return below to the analyses of true variation in learning style that are reported in Chapter 14.

Alongside its support from the cross-linguistic consistency in cross-domain correlations, this unitary view of early language is supported by several other bodies of work. The first is a set of longitudinal analyses by Bornstein and colleagues. In a two-cohort longitudinal study, Bornstein and Haynes (1998) and Bornstein et al. (2004b) collected data from a sample of American English learning children at two and four years and measured language using a variety of instruments including parent report, transcripts, and behavioral tasks. In a re-analysis of these data, Bornstein and Putnick (2012) found that the core construct of early language that emerged from the sub-measures was both highly correlated with each of the sub measures and also highly stable over time (as we noted earlier in Chapter 4).

In addition, research on international adoption, a unique case study of changes in learning rate relative to input, provides support for this unitary trajectory view as well. International adoptees often have to let go of their native language completely, sometimes at a fairly old age, and begin learning a host language. A body of work by Snedeker, Geren, and Shafto (2007; 2012) suggests that these children go through the same general course of acquisition, including an early noun bias and early grammatical omissions, but they did so much more quickly on average then native learners. Although these children were more conceptually sophisticated and could learn faster, the informational challenges of breaking into the language were still similar on the whole.

In sum, the language system is much more tightly woven than casual observation might lead an observer to expect. Our work here confirms and extends previous investigations by Bates et al. (1994) and Bornstein and Putnick (2012). Although our data provided multiple opportunities for aspects of language development (communication, vocabulary, morphology, grammar) to pull apart across individuals into modular subsystems, they instead hung together very tightly. Further, the strength of these connections did not vary across every language examined: the nature of the trajectory was surprisingly consistent, even across languages with substantial differences in morphosyntactic structure.

16.2.2 Young children talk about similar things

A second broad generalization from our work here is the content similarities in children's early vocabulary across languages. Put simply, regardless of the language they are learning, children in the early stages of language learning appear to talk about similar things, especially people, social routines, small objects, and body parts. This work builds on analyses by Tardif et al. (2008), who compared the distribution of children's first ten words across three languages. Many of the content generalizations that held for those three languages in fact hold here as well in a much more broad array of languages.

As with the example above, it is useful to consider the counterfactual situations, which are very reasonable a priori. Without knowing anything about language acquisition in other languages, we might suppose that children in different cultures begin speaking about quite different things. We could easily imagine a culture in which children primarily began by describing the properties of objects: "soft!" "small!" or another in which children described places or another in which children talk about places they go or things around the house. We might think that American children are obsessed with animals, whereas, children in other cultures might abhor them. It turns out that all of these counterfactuals are largely not observed.

We observed three broad categories of findings that support this generalization. First, relating most directly to Tardif et al. (1999)'s work, in Chapter 8, we observed a strong correlation in the rank order of acquisition for the earliest words. Those words that were typically learned very early in one language tended to be very early learned in others. These correlations were attenuated for later-learned words. Second, we saw that some word categories — both syntactic and semantic categories — were typically over-represented across languages, while others were under-represented. Nouns generally were over-represented (with some cross-linguistic variability in the amount of over-representation). But beyond that, words for body parts, games and social routines, and sounds were all over-represented. In contrast, predicates and function words, as well as words for time and places were under-represented. And third, when we looked for systematic predictors of the ease of learning words across language, the predictor "babiness" — which captures association with infants — was among the strongest predictors, especially for early-learned words.

Where do these similarities come from? Before answering this question, we can say that they likely do not emerge from simple linguistic or environmental frequency. Not only is frequency statistically controlled in some of our analyses (e.g., those in Chapter 10), but also there are clear dissociations between linguistic frequency and children's comprehension and production. Although mothers and fathers are likely common in children's experience, "mom" and "dad" (or the local equivalent) are actually relatively infrequent in language directed to children — yet they are uttered very early, sometimes even first. Many function words are highly frequent in speech to children, yet are produced much later than expected. Finally, many omnipresent environmental stimuli from diapers to couches and carpets are referred to only much later than the people, animals, and small objects on and in them.

Instead, these similarities likely emerge from a combination of children's particular attention and interests; the communicative priorities of child-caregiver dyads; and the informational structure of language, which renders certain aspects of language easier to learn than others. We consider each of these in turn.

First, although we do not have an independent operationalization of children's specific interests, anyone who has spent time with toddlers can say that these interest are present and extremely

powerful. The red puppet Elmo, who has a squeaky (some might say annoying) voice, is almost designed to drive children's attention and interest, despite some parents' best efforts to minimize his frequency in the environment. More generally, toddlers tend to make a bee-line towards animals and toys that they perceive as interesting. These preferences may stem from general conceptual preferences for animates, or broader perceptual preferences, or likely some mixture of the two, but their impact on children's attention — and from there their vocabulary across languages — is undeniable.

Second, it is also true that children's own attention is far from the only driver of the language they hear. Children's language exposure takes place in an environment that is largely constructed by their caregivers, rather than them. Thus, much of the language they hear is "functional talk" that takes place in the context of routines like dressing, diapering, mealtime, and bathtime. These dyadic priorities mean that children's vocabulary often contains a wide variety of functional terms. In fact, Roy et al. (2015) even argued that the contextual distinctiveness of words that appear in these routines may lead to earlier acquisition. Similar dyadic priorities across cultures would thus lead to cross-linguistic consistency in early vocabulary.

Finally, some words may be learned earlier than others simply because of the informational structure of language. As argued by Snedeker et al. (2007), verbs may be later-learned simply because you need to know some nouns to figure them out. To the extent that these regularities extend across languages, they would impose a variety of ordering constraints on acquisition that could be reflected in our analyses.

Of course, it is always possible that we might observe more variation to the extent that relatively more variant languages exist. All of the typological caveats that we first stated in Chapter 1 hold still: most of our languages are WEIRD (Henrich et al., 2010), and the majority are Indo-European. Especially with respect to assessing semantic variation across languages, it will be critical for future work to incorporate a wider range of languages and language families.

16.2.3 Children take different routes into language

Despite these aggregate similarities across children and across cultures, children individually take different routes into language. The speed with which they acquire language is highly variable in the first years of life and this variability itself is a constant across cultures. Further, children vary stably in the nature of their acquisition path, some naming more objects and others combining words relatively more.

First, even controlling for the non-linear trajectory of acquisition mentioned above, we found stable differences in "referential style" among children. Thus, although the language system is tightly woven and moves through a relatively consistent learning trajectory across individuals, there is nevertheless an interesting, second-order component. Some children, especially the faster-learning ones, appear to learn more nouns. Others, often the slower-learning ones, tend to combine words more frequently. This pattern may relate to the age-moderation found in Chapter 13: it is precisely the older children who gain more "grammar per unit lexicon." Perhaps those children who are learning slightly more slowly then are able to bring more mature working memory to the task of grammatical induction. Or perhaps these children have heard more language and hence the natural statistics of language are easier to extract. Or perhaps these early word combinations are actually "unanalyzed wholes" that are syntactically less complex than they might appear. Regardless, these differences appear as a stable aspect of the relatively consistent developmental course we observed.

Beginning with the first monograph reporting the CDI norming study, Fenson et al. (1994) noted that variability is perhaps the primary and most striking fact about children's vocabulary learning. In practical terms, the huge variance across children accounts for the fact that while some typically-developing two year olds will talk your ear off, others will barely utter a handful of words (even if they understand more). From a biological perspective, this variability is quite unprecedented. As a comparison, variation in heights for toddlers is tiny compared with variation in vocabulary: the mean height for a 24-month-old is around 33 inches, with a standard deviation of a little more than an inch, leading to a coefficient of variation around .03. This measurement is almost two orders of magnitude smaller than the coefficient of variation on vocabulary.

Examining the variability in English-learning children's vocabulary documented by Fenson et al. (1994), it was easy to think that the spread of children's outcomes was due to the demographic and parenting variability found in the United States, which — even in the restricted set sampled in that initial study — was large. But a look at the variability estimates found in our broader sample dispels this hypothesis — the variability is surprisingly constant (and large) across every language we examined. If culture homogeneity was an important influence on the homogeneity of early language, we would expect a smaller variability measurement for the Norwegian, Korean, or Chinese populations, where cultural, educational, and socio-economic heterogeneity would be expected to be higher. Instead, these populations show a similar level of variation.

Where does this variability come from? We can only speculate. Some must come from variation in input across households, which has been amply documented to relate to children's early vocabulary (e.g., Hart and Risley, 1995; Hoff, 2003; Weisleder and Fernald, 2013). Further, some component of this input-correlated variation is likely to be genetic (e.g., Hayiou-Thomas et al., 2012), such that some children inherit a tendency towards slower vocabulary growth from parents who themselves talk relatively less and use relatively less diverse vocabulary.

Does this variation persist beyond the range of our study? One of the most intriguing aspects of our analysis is the suggestion that, accounting for ceiling effects, variation in ability does not compress in the range we measured. An important direction for future work would be to ask about the range of variation observed within and across cultures on other standardized measures of language in school-age children. One possibility is that variability compresses on most standard vocabulary and grammar measures simply because functional communication is possible for nearly all speakers, but variation transfers instead to more "leading edge" domains like literacy or discourse comprehension.

More generally, we are curious about how the range of variation we observed here relates to the variation in other non-trivial aspects of cognition. If we were to obtain stable measures of mathematical cognition, for example, would we observe the same degree of variation? This extension points to one of the under-appreciated virtues of scale of the type we have explored here: by virtue of the quantitative measurement that it affords, scale can lead to brand new questions.

16.3 Learning processes

In Chapter 1, we introduced the idea of "process universals." These cannot be universals of content as all of the content being reported by parents filling CDI forms is language-specific. Instead, the idea of process universals is that there are processes that operate in different language context to produce the observed pattern of phenomena. These processes could be — but need not be — learner-internal. They could also operate at the level of the child-caregiver dyad, for example. What more can we say

about them from the perspective of the discussion above?

As we have stressed throughout, the connections between the broad trends and tendencies that we observed and any specific learner- or dyad-level process are weak, at best. Many different processes can likely lead to the same outcome, thus the linkages we make here are abductive rather than deductive. Further, as with many abductive arguments, some amount of evidential weight is given to the prior probability of the theoretical proposal. Bearing all that in mind, we highlight three process-level connections that appear consistent with our data.

16.3.1 Language grows through interactional input

Without input, there can be no uptake. The importance of language input is a fundamental tenet of all models of word learning, from the simplest accumulator model (McMurray, 2007) through to more complex probabilistic and neural network models (e.g., McMurray et al., 2012; Frank et al., 2009; Fazly et al., 2010). In all of these models, what is learned by the model is a function of the frequency and statistical distinctiveness of the learner's input. This conclusion is amply supported by a body of correlational research linking observed speech from children's caregivers to their vocabulary size (e.g., Hart and Risley, 1995; Hoff, 2003; Huttenlocher et al., 1991; Hurtado et al., 2008).

But quantity is not enough. Beyond the first order correlation of input quantity to language outcomes, a body of research now provides nuanced qualification. In a number of studies across cultures, child-directed speech is a better predictor than overall speech, even in cultures where this kind of speech is relatively rare (Weisleder and Fernald, 2013; Shneidman and Goldin-Meadow, 2012). One hypothesis is that child-directed speech provides more grounded moments in which word meanings can be inferred from context (Cartmill et al., 2013). Indeed, in one study, a variety of measures of input quality — including joint engagement as well as the presence of rituals and routines — were better predictors of vocabulary size than pure quantity (Hirsh-Pasek et al., 2015). Of course, high-quality input must be developmentally appropriate, and for older children, language that is more syntactically complex supports complex syntax aqcuisition (Huttenlocher et al., 2002). Presumably the evidence for the importance of grounded, engaged communication is at least in part due to its specific importance for younger learners who are engaged in discovering word meanings; other aspects of language gain importance for other learners (Hoff, 2006; Hoff and Naigles, 2002; Rowe, 2012).

We see many of our conclusions as fitting well into this broader picture. While we do not have predictors of individual children's input, the research presented in Chapter 10 suggests that even aggregated, average measures of input are relatively powerful predictors of word-by-word uptake. Of course, word frequency is a useful predictor of age of acquisition, especially for nouns. Even if high-quality, grounded instances are the appropriate learning input, greater frequency overall will — all else being equal — lead to greater frequency in the appropriate learning context. We further observed effects of solo frequency (being used in a one-word sentence) and mean length of utterance. These both are consistent with the idea that it is easier to learn meanings for words in shorter sentences, which pose both fewer word segmentation challenges and fewer word-meaning mapping ambiguities. In addition, the relative cross-linguistic consistency in these predictors suggests that the input-uptake connections that have largely been documented for English learners are likely to be robust for learners of other languages as well.

The demographic differences in vocabulary we observed are also consistent with interactional-input theories of vocabulary development, in which the more higher-quality the input the child receives, the faster vocabulary grows. Under this hypothesis, children who are first-born and who have mothers

with more education are likely to receive more and more higher-quality input. First-born children receive more input through their greater allocation of parent attention (though as we note in Chapter 6, first-born children's parents may also be more aware of their vocabulary). Children with mothers who have more education receive more and more higher-quality talk through the availability of more parent time, different values around talk, greater awareness of the role of interaction for young children, and perhaps differing parental practices (Evans, 2004; Farkas and Beron, 2004; Rowe et al., 2012). Our data do not differentially support this interpretation over other possibilities, but prior work leads us to favor this interpretation. In contrast, it is an open question the extent to which the sex differences we observed emerge due to interactional/input differences vs. other learner-internal factors.

In addition, our analyses of demographic differences in the composition of children's early vocabulary are consistent with this view as well. Siblings have a high probability of knowing "brother" and "sister"; children from mothers with less education are more likely to know words for candy and sweets. Across cultures, the words that emerge from these analyses plausibly appear related to the ways that children's input differs across different learning contexts. In sum, across a number of our analyses, we see what are plausibly the residues of input through interaction, accreted across a range of different input conditions.

16.3.2 Individual word meanings must be inferred based on (cross-situational) evidence

Cross-situational learning is the proposal that children use the statistical properties of how words are used across contexts to help them infer meaning (Gleitman, 1990; Siskind, 1996). This specific proposal has been instantiated in a variety of associative learning experiments with both adults (Yu and Smith, 2007; Yurovsky and Frank, 2015) and children (Smith and Yu, 2008; Vlach and Johnson, 2013). The specific mechanisms underlying learners' performance in these tasks are still controversial (e.g., Medina et al., 2011; Trueswell et al., 2013; Yurovsky et al., 2013; Yurovsky and Frank, 2015). Despite this controversy, the general model of learning as proposed in these studies has rapidly become the default instantiation of how interactional input translates into learning (e.g., the models of Frank et al., 2009; Fazly et al., 2010; McMurray et al., 2012). The basic idea is that children bring multiple information sources — including statistical, social, and grammatical information — to bear on resolving the referent of each utterance. Then, this information is accumulated and brought to bear on resolving reference in subsequent utterances.

Several findings from our investigations are consistent with these general ideas. As noted above, the predictors of word difficulty we found in Chapter 10 are consistent with the general input-uptake viewpoint. In fact, it is very easy to see some of them playing out in the context of the general cross-situational learning proposal. Of course, frequency effects are ubiquitous in cross-situational learning models (though their specifics vary somewhat from model to model; e.g., Kachergis et al., 2012). Further, on the cross-situational perspective, shorter sentences and especially single-word utterances are less confounded in terms of the information they provide about the relationship between words and their meanings.

In Chapter 11, we observed a complicated cross-linguistic pattern of biases. Nouns were typically over-represented in early vocabulary to some exent, while predicate representation was quite variable. While this pattern likely emerges out of the contributions of a variety of different cognitive and cultural features, its broad outlines are in fact consistent with the cross-situational viewpoint. On this view, nouns are easy to learn because the more they are heard, the more opportunities children

get to build consistent mappings between the words and their contextual referents. This regularity should be especially true for concrete nouns that are more likely to be found in grounded contexts (Gentner and Boroditsky, 2001).

In contrast, verbs and other predicates cannot be acquired as easily from basic co-occurrence. Syntactically "light" verbs like "make" or "do" require some nominal information to constrain their meaning in context. And "heavier" verbs may still be systematically ambiguous without syntactic information (e.g., "chase"/"flee"; Gleitman, 1990). Thus, verb learning — and likely the learning of other predicates and many function words as well — relies on a base of nouns and a basic comprehension of the syntactic structures in which they appear in order to infer meaning in context. This informational sequencing viewpoint predicts that these categories should be acquired relatively later, with relatively more support from shorter, easier-to-parse utterances (MLU as a predictor for predicates, for example). This account would explain cross-linguistic exceptions like Mandarin as cases where many early-learned verbs are semantically-transparent enough to be learned cross-situationally without syntactic information (following Tardif, 1996).

16.3.3 Generalizations appear gradually

As they are instantiated in the CDI, syntactic and morphological structures are fundamentally "item-based" in the sense of Tomasello (2003). Since all syntactic information is instantiated in particular example items, every bit of knowledge that the CDI assesses is grounded in lexical items used in a specific construction. Although there may well be broader, more abstract generalizations that underly the growth of word order, we simply do not have the signal to address the presence or absence of such abstractions.

That said, at the level of the measurements we have, our evidence is broadly consistent with the view of language as growing through the gradual induction of syntactic regularities through a reciprocal interaction with the acquisition of individual content words. Versions of this viewpoints exist throughout the broad theoretical space of language acquisition (e.g., Yang, 2016; Tomasello, 2003; Meylan et al., 2017), but all proposals rely on a learning mechanism in which generalizations about structure are graded and rely on the amount of evidence available. All such mechanisms would presumably predict some relationship individuals' grammatical and lexical abilities; an important target for future theoretical work in this area is to explore how tight these correlations are predicted to beon different viewpoints.

Further complicating matters, the relationship between grammar and the lexicon is likely reciprocal. As content vocabulary grows, it provides both the groundwork for generalization of constructions (Tomasello, 2003; Goldberg, 2006) and also the specific content words necessary for the induction of predicate meanings (as described above in the case of verbs). In the other direction, as grammar grows, it also allows for better disambiguation of predicate meanings (Gleitman, 1990) and creates myriad other opportunities for learning. Both of these directions of longitudinal influence are present in at least one study (Brinchmann et al., 2018), although that study found that considerable shared variance between individuals was also due to the stability of both grammatical and lexical knowledge within individuals.

As we disussed in Chapter 1, we do not see our work here as resolving long-standing debates about the nature of abstract syntactic representations. Instead, we hope our contribution is somewhat different — we refocus the debate away from phenomena that make only occasional contact with the gross regularities of children's early language use in context. Instead, we focus on the pattern of

observable changes in complexity and diversity of early language. We hope that this focus leads to theorizing that takes these observables as its primary predictive target.

16.4 Methodological morals

As we noted in Chapter 1, psychology has recently been plagued by concerns about reproducibility (e.g., Hardwicke et al., 2018) and replicability (e.g., Open Science Collaboration, 2015). Our work here was inspired by considering these issues and their impact on the field of language development. The ultimate goal of research in this area is to create a quantitative theory that allows for precise predictions and principled explanations of developmental phenomena. Such a theory cannot be built on a series of non-reproducible findings and binary conclusions (Frank et al., 2017).

Wordbank is a reply to this situation: by compiling the extant CDI datasets into a single open database, researchers can reproduce previous and new research conclusions that use these data. The analyses we report here using the Wordbank data are computationally reproducible through the availability of the code necessary to build the book and all its figures and analyses. In addition, by seeking a level of scale beyond previous efforts, we have attempted to avoid the variability inherent in "small-N" studies.

Further, our work is built on the notion of replication. Nearly every one of the preceding substantive chapters is in some sense a "replication" of previous work — an analysis was taken from previous work with a particular CDI dataset and applied (sometimes with technical modifications) to other data. Yet the result is not a judgement on the original; we do not declare a binary success or failure of the replication attempt. Instead, we are interested in the degree to which a particular quantitative estimate varies across languages and cultures.

This sort of analysis is superficially similar to the idea of "hidden moderators" that has plagued the replication debate (Van Bavel et al., 2016). But, this effort has largely been an effort to contextualize failures to replicate particular experimental effects by invoking unknown sources of variability across contexts. In contrast, our efforts here allow us to quantify variation across "replications" of the same effect and use these estimates as the signal — rather than an un-measured source of noise.

One notable feature of our analytic strategy is that we try to rely very little on binary decision-theoretic inferences using null hypothesis significance testing. There are a handful p-values in occasional analyses, but few of these appear in any prominent inferential conclusion. Instead, our goal has been to measure quantities of interest with high precision, looking for statistical estimates that relate to particular theoretical goals. For example, the existence of a noun bias is a fascinating observation, but this observation gives limited leverage to differentiate theories. The magnitude of a noun bias provides more leverage for quantitative theorizing. And the distribution of magnitudes across many of the world's languages gives greater leverage still. Our hope is that, by generalizing and applying many influential analyses by many contributors, our work here can affect theory more broadly.

16.5 Conclusions

Developmental psychology often appears to be divided between two groups that do not communicate with one another. On the one hand, there are researchers interested in the ontogenetic and phylogenetic

origins of knowledge — the epistemological project of understanding how we reason about objects, communicate using language, or learn from other people (Carey, 2009; Spelke and Kinzler, 2007; Tomasello, 2010). On the other hand, there are researchers interested in growth, change, and variation across individuals in psychometric constructs like executive function, working memory, and personality (Diamond, 2013; Gathercole et al., 1994; Ainsworth et al., 2015). The first type of research has led to a productive union of philosophical and psychological ideas, addressing exciting questions in the history of philosophy using empirical methods (e.g., Gopnik et al., 2004; Carey, 2009). In contrast, the second type has had its impact in connections to clinical practice and educational policy (e.g., Nelson, 2007; Diamond and Lee, 2011).

"Origins" research and "variations" research have often appeared to be traditions at odds with one another. While there are, of course, exceptions to this split, in many cases researchers in these two traditions read different journals, record different measures, use different research designs and statistical models, and often appear to be pursuing different goals.

Yet, language learning is an area where these traditions come together. The knowledge being acquired — the stock of words in the child's lexicon — is both the epistemic construct whose origins we are studying and the psychometric construct whose reliability across individuals we wish to assess. By studying the variability and consistency in the early lexicon, we can both study the origins of knowledge and the processes and factors by which human beings differ from one another across different cultural contexts, family environments, and genetic endowments.

In our work here, we sought generalizations in the universals and variations that hold across cultures and languages — and took steps to tie these generalizations to processes of language learning. These generalizations are based in the "variations" approach via the Batesian project of using variability to constrain theory. Yet they are also about the "origins" of a very specific type of knowledge: knowledge of language. The ultimate aim of theorizing in this area must be broader than either "origins" or "variations" alone.

We hope that our work in compiling measurements of consistency and variability across languages is a beginning rather than an ending. At the end of this mansucript, we hope to have inspired potential for more new work than we did at the beginning. Although the number of children represented in our analyses is large, the number of languages, and their diversity across families, is quite small. And our models of acquisition — especially those linking input to uptake — are only a rough sketch of what is possible. In the end, however, our hope is that by showing what is possible in the quantitative study of language development, we illustrate a broader set of possibilities for a data-driven science of developmental change.

Appendix A

Individual Datasets

This appendix give the contributors and citations for each of the datasets used in this book.

English (American)

Form: WS Contributor: Larry Fenson, San Diego State University Citation: Fenson, L., Marchman, V. A., Thal, D., Dale, P., Reznick, J. S. & Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual. 2nd Edition. Baltimore, MD: Brookes Publishing Co.

Form: WS Contributor: Virginia Marchman, Stanford University

Form: WS Contributor: Virginia Marchman, Stanford University

Form: WS Contributor: Linda Smith, Indiana University

Form: WS Contributor: Linda Smith, Indiana University

Form: WS Contributor: Krista Byers-Heinlein, Concordia University

Form: WS Contributor: Donna Thal, San Diego State University Citation: Thal, D. J., Marchman, V. A. & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued delay. In L. Rescorla & P. Dale (Eds.). Late Talkers: Language Development, Interventions, and Outcomes. Baltimore, MD.: Brookes Publishing.

Form: WS Contributor: Donna Thal, San Diego State University Citation: Thal, D. J., Marchman, V. A. & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued delay. In L. Rescorla & P. Dale (Eds.). Late Talkers: Language Development, Interventions, and Outcomes. Baltimore, MD.: Brookes Publishing.

Form: WG Contributor: Larry Fenson, San Diego State University Citation: Fenson, L., Marchman, V. A., Thal, D., Dale, P., Reznick, J. S. & Bates, E. (2007). MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual. 2nd Edition. Baltimore, MD: Brookes Publishing Co.

Form: WG Contributor: Krista Byers-Heinlein, Concordia University

Form: WG Contributor: Donna Thal, San Diego State University Citation: Thal, D. J., Marchman, V. A. & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued

delay. In L. Rescorla & P. Dale (Eds.). Late Talkers: Language Development, Interventions, and Outcomes. Baltimore, MD.: Brookes Publishing.

Form: WG Contributor: Donna Thal, San Diego State University Citation: Thal, D. J., Marchman, V. A. & Tomblin, J. B. (2013). Late talking toddlers: Characterization and prediction of continued delay. In L. Rescorla & P. Dale (Eds.). Late Talkers: Language Development, Interventions, and Outcomes. Baltimore, MD.: Brookes Publishing.

Form: WG Contributor: Michael C. Frank, Stanford University

Form: WS Contributor: Anne Fernald, Stanford University Citation: Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. Developmental Science, 16, 234–248. http://doi.org/10.1111/desc.12019

Spanish (Mexican)

Form: WS Contributor: Donna Jackson-Maldonado, Universidad Autónoma de Querétaro Citation: Jackson-Maldonado, D., Thal, D., Marchman, V., Newton, T., Fenson, L, & Conboy, B. (2003). MacArthur Inventarios del Desarrollo de Habilidades Comunicativas. User´s Guide and Technical Manual. Brookes, Baltimore.

Form: WG Contributor: Donna Jackson-Maldonado, Universidad Autónoma de Querétaro Citation: Jackson-Maldonado, D., Thal, D., Marchman, V., Newton, T., Fenson, L, & Conboy, B. (2003). MacArthur Inventarios del Desarrollo de Habilidades Comunicativas. User´s Guide and Technical Manual. Brookes, Baltimore.

Form: WG Contributor: Anne Fernald, Stanford University Citation: Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. Psychological Science, 24, 2143–2152. http://doi.org/10.1177/0956797613488145

Form: WS Contributor: Anne Fernald, Stanford University Citation: Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. Psychological Science, 24, 2143–2152. http://doi.org/10.1177/0956797613488145

Danish

Form: WS Contributor: Dorthe Bleses, University of Southern Denmark Citation: Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008). The Danish Communicative Development Inventories: validity and main developmental trends. Journal of Child Language, 35, 619-650.

Form: WG Contributor: Dorthe Bleses, University of Southern Denmark Citation: Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. & Basbøll, H. (2008). The Danish Communicative Development Inventories: validity and main developmental trends. Journal of Child Language, 35, 619-650.

Norwegian

Form: WS Contributor: Hanne Simonsen and Kristian Kristoffersen, University of Oslo Citation: Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23. DOI: 10.1177/0142723713510997

Form: WS Contributor: Hanne Simonsen and Kristian Kristoffersen, University of Oslo Citation: Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The

Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23. DOI: 10.1177/0142723713510997

Form: WG Contributor: Hanne Simonsen and Kristian Kristoffersen, University of Oslo Citation: Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23. DOI: 10.1177/0142723713510997

Form: WG Contributor: Hanne Simonsen and Kristian Kristoffersen, University of Oslo Citation: Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23. DOI: 10.1177/0142723713510997

Croatian

Form: WG Contributor: Melita Kovacevic, University of Zagreb Citation: Kovacevic, M., Babic, Z., & Brozovic, B. (1996). A Croatian language parent report study: Lexical and grammatical development. Paper presented at the VIIth International Congress for the Study of Child Language, July 1996, Istanbul, Turkey.

Form: WS Contributor: Melita Kovacevic, University of Zagreb Citation: Kovacevic, M., Babic, Z., & Brozovic, B. (1996). A Croatian language parent report study: Lexical and grammatical development. Paper presented at the VIIth International Congress for the Study of Child Language, July 1996, Istanbul, Turkey.

German

Form: WS Contributor: Gisela Szagun, University College London Citation: Szagun, G., Stumper, B. & Schramm, A.S. (2009). Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform). Frankfurt: Pearson Assessment.

Italian

Form: WS Contributor: Christina Caselli, Institute of Cognitive Sciences and Technologies Citation: Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. Cognitive Development, 10(2), 159-199.

Form: WG Contributor: Christina Caselli, Institute of Cognitive Sciences and Technologies Citation: Caselli, M. C., Rinaldi, P., Stefanini, S., & Volterra, V. (2012). Early action and gesture 'vocabulary' and its relation with word comprehension and production. Child Development, 83(2), 526-542.

Russian

Form: WG Contributor: Stella Ceytlin, SPb Russian Pedagogical University Citation: Е.А.Вершинина, М.Б. Елисеева, Т.С. Лаврова, В.Л. Рыскина, С.Н. Цейтлин. Некоторые нормативы речевого развития детей от 8 до 18 месяцев// Специальное образование: традиции и инновации: Сборник научно-методических трудов с международным участием. — СПб.: Изд-во РГПУ им. А. И. Герцена, 2011.

Form: WS Contributor: Stella Ceytlin, SPb Russian Pedagogical University Citation: М.Б. Елисеева, Е.А. Вершинина. Некоторые нормативы речевого развития детей от 18 до 36 месяцев (по материалам МакАртуровского опросника) // Проблемы онтолингвистики – 2009. Материалы международной конференции 17-19 июбня 2009 г. Санкт-Петербург, С.72-78

Swedish

Form: WG Contributor: Mårten Eriksson, University of Gävle Citation: Eriksson, M., & Berglund, E. (2002). Instruments, scoring manual and percentile levels of the Swedish Early Communicative Development Inventory, SECDI. (FoU-Rapport 17). Gävle, Sweden: Institutionen för pedagogik, didaktik och psykologi.

Form: WS Contributor: Mårten Eriksson, University of Gävle Citation: Eriksson, M., & Berglund, E. (2002). Instruments, scoring manual and percentile levels of the Swedish Early Communicative Development Inventory, SECDI. (FoU-Rapport 17). Gävle, Sweden: Institutionen för pedagogik, didaktik och psykologi.

Turkish

Form: WG Contributor: Aylin Küntay, Koç University Citation: Acarlar, F., Aksu-Koç, A., Küntay, A.C., Maviş, İ., Sofu, H., Topbaş, S., Turan, F. (2009). Adapting MB-CDI to Turkish: The first phase. In S. Ay, Ö. Aydın., İ. Ergenç, S. Gökmen, S. İşsever, and D. Peçenel (Eds.) Essays on Turkish linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6-8, 2008. Harrassowitz Verlag: Wiesbaden, Germany.

Form: WS Contributor: Aylin Küntay, Koç University Citation: Acarlar, F., Aksu-Koç, A., Küntay, A.C., Maviş, İ., Sofu, H., Topbaş, S., Turan, F. (2009). Adapting MB-CDI to Turkish: The first phase. In S. Ay, Ö. Aydın., İ. Ergenç, S. Gökmen, S. İşsever, and D. Peçenel (Eds.) Essays on Turkish linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6-8, 2008. Harrassowitz Verlag: Wiesbaden, Germany.

Hebrew

Form: WG Contributor: Hila Gendler Shalev, Tel-Aviv University Citation: Gendler-Shalev, H. (2005). HCDI-WG .

Form: WS Contributor: Hila Gendler Shalev, Tel-Aviv University

Cantonese

Form: WS Contributor: Twila Tardif, University of Michigan Citation: Tardif, T., Fletcher, P., Liang, W., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. Journal of child language, 36(05), 1115-1144.

Mandarin (Beijing)

Form: WS Contributor: Twila Tardif, University of Michigan Citation: Tardif, T., Fletcher, P., Liang, W., & Kaciroti, N. (2009). Early vocabulary development in Mandarin (Putonghua) and Cantonese. Journal of child language, 36(05), 1115-1144.

Form: TC Contributor: Ping Li, Pennsylvania State University Citation: Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. Behavior Research Methods, 40, 728-733.

Form: IC Contributor: Ping Li, Pennsylvania State University Citation: Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. Behavior Research Methods, 40, 728-733.

British Sign Language

Form: WG Contributor: Bencie Woll, University College London Citation: Woolfe, T., Herman, R., Roy, P., & Woll, B. (2010). Early vocabulary development in deaf native signers: a British Sign

Language adaptation of the communicative development inventories. Journal of Child Psychology and Psychiatry, 51(3), 322-331.

French (Quebecois)

Form: WG Contributor: Natacha Trudeau, Université de Montréal Citation: Boudreault, M. C., Cabirol, E. A., Poulin-Dubois, D., Sutton, A., & Trudeau, N. (2007). MacArthur Communicative Development Inventories: Validity and preliminary normative data. La Revue d'orthophonie et d'audiologie, 31(1), 27-37.

Form: WS Contributor: Natacha Trudeau, Université de Montréal Citation: Trudeau, N., & Sutton, A. (2011). Expressive vocabulary and early grammar of 16-to 30-month-old children acquiring Quebec French. First Language, 0142723711410828.

Slovak

Form: WG Contributor: Svetlana Kapalková, Comenius University

Form: WS Contributor: Svetlana Kapalková, Comenius University

English (British)

Form: TEDS Twos Contributor: Philip Dale, University of New Mexico Citation: Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient difficulties at 3 and 4 years. Journal of Speech-Language-Hearing Research, 46, 544-560.

Form: TEDS Threes Contributor: Philip Dale, University of New Mexico Citation: Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of early language delay: I. Predicting persistent and transient difficulties at 3 and 4 years. Journal of Speech-Language-Hearing Research, 46, 544-560.

Form: Oxford CDI Contributor: Caroline Floccia, Plymouth University Citation: Floccia, C. (2017). Data collected with the Oxford CDI over a course of 5 years in Plymouth Babylab, UK. With the permission of Plunkett, K. and the Oxford CDI from Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. Journal of Child Language, 27, 689-705.

Form: Oxford CDI Contributor: Caroline Floccia, Plymouth University Citation: Floccia, C. (2017). Data collected with the Oxford CDI over a course of 5 years in Plymouth Babylab, UK. With the permission of Plunkett, K. and the Oxford CDI from Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. Journal of Child Language, 27, 689-705.

American Sign Language

Form: FormA Contributor: Diane Anderson, University of California, Berkeley Citation: Anderson, D., & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. Journal of Deaf Studies and Deaf Education, 7(2), 83–106.

Form: FormBOne Contributor: Diane Anderson, University of California, Berkeley Citation: Anderson, D., & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. Journal of Deaf Studies and Deaf Education, 7(2), 83–106.

Form: FormBTwo Contributor: Diane Anderson, University of California, Berkeley Citation: Anderson, D., & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. Journal of Deaf Studies and Deaf Education, 7(2), 83–106.

Form: FormC Contributor: Diane Anderson, University of California, Berkeley Citation: Anderson, D., & Reilly, J. (2002). The MacArthur Communicative Development Inventory: Normative data for American Sign Language. Journal of Deaf Studies and Deaf Education, 7(2), 83–106.

Greek (Cypriot)

Form: WS Contributor: Kleanthes K. Grohmann, University of Cyprus Citation: Taxitari, Loukia, Maria Kambanaros & Kleanthes K. Grohmann. 2015. 'A Cypriot Greek Adaptation of the CDI: Early Production of Translation Equivalents in a Bi(dia)lectal Context'. Journal of Greek Linguistics 15, 1–24.

Kigiriama

Form: WG Contributor: Katie Alcock, Lancaster University Citation: Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., Newton, C.R.J.C. (2015) Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. Journal of Child Language, 42, 763-785.

Form: WS Contributor: Katie Alcock, Lancaster University Citation: Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., Newton, C.R.J.C. (2015) Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. Journal of Child Language, 42, 763-785.

Kiswahili

Form: WG Contributor: Katie Alcock, Lancaster University Citation: Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., Newton, C.R.J.C. (2015) Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. Journal of Child Language, 42, 763-785.

Form: WS Contributor: Katie Alcock, Lancaster University Citation: Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., Newton, C.R.J.C. (2015) Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. Journal of Child Language, 42, 763-785.

Form: WS Contributor: Katie Alcock, Lancaster University Citation: Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., Newton, C.R.J.C. (2015) Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. Journal of Child Language, 42, 763-785.

Czech

Form: WS Contributor: Filip Smolik, Academy of Sciences of the Czech Republic Citation: Markova, G., Smolík, F. (2014). What Do You Think? The Relationship between Person Reference and Communication About the Mind in Toddlers. Social Development, 23, 61-79. DOI: 10.1111/sode.12044

English (Australian)

Form: WS Contributor: Marina Kalashnikova, MARCS Institute for Brain, Behaviour and Development Citation: Kalashnikova, M., Schwarz, I.-C., & Burnham, D. (2016). OZI: Australian English Communicative Development. First Language, 36, 407-427.

Latvian

Form: WG Contributor: Olga Urek, The Arctic University of Norway Citation: Urek, Olga, Anna Vulāne, Roberts Dargis, Agrita Tauriņa, Tija Zīriņa, Hanne Gram Simonsen (to appear) Latvian CDI: methodology, developmental trends and cross-linguistic comparison.

Form: WS Contributor: Olga Urek, The Arctic University of Norway Citation: Urek, Olga, Anna Vulāne, Roberts Dargis, Agrita Tauriņa, Tija Zīriņa, Hanne Gram Simonsen (to appear) Latvian CDI: methodology, developmental trends and cross-linguistic comparison.

Korean

Form: WG Contributor: Dongsun Yim, Ewha Womans University

Form: WS Contributor: Dongsun Yim, Ewha Womans University

Form: WS Contributor: Soyeong Pae, Hallym University Citation: Pae, S., & Kwak, K. (2011). Korean MacArthur-Bates Communicative Development Inventories (K M-B CDI). Seoul: Mindpress.

Form: WG Contributor: Soyeong Pae, Hallym University Citation: Pae, S., & Kwak, K. (2011). Korean MacArthur-Bates Communicative Development Inventories (K M-B CDI). Seoul: Mindpress.

French (French)

Form: WG Contributor: Christina Bergmann (Max Planck Institute for Psycholinguistics),Anne-Caroline Fievet (Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University)

Form: WG Contributor: Katie Von Holzen, University of Maryland Citation: Von Holzen, K., Nishibayashi, L.-L., & Nazzi, T. (2018). Consonant and vowel processing in word form segmentation: An infant ERP study. Brain Sciences, 8(24), 1–15. DOI: 10.3390/brainsci8020024.

Form: WS Contributor: Katie Von Holzen, University of Maryland Citation: Von Holzen, K., Nishibayashi, L.-L., & Nazzi, T. (2018). Consonant and vowel processing in word form segmentation: An infant ERP study. Brain Sciences, 8(24), 1–15. DOI: 10.3390/brainsci8020024.

Form: WS Contributor: Sophie Kern, Centre national de la recherche scientifique (CNRS)

Portuguese (European)

Form: WG Contributor: Irene Cadime, University of Minho

Form: WS Contributor: Irene Cadime, University of Minho

Spanish (European)

Form: WG Contributor: Alexandra Karousou, Democritus University of Thrace Citation: López Ornat, S., Gallego, C., Gallo, P., Karousou, A., Mariscal, S., & Martínez, M. (2005). MacArthur: Inventario de desarrollo comunicativo. Manual y Cuadernillos. Madrid, TEA Ediciones. ISBN: 84-7174- 820-7

Form: WS Contributor: Alexandra Karousou, Democritus University of Thrace Citation: López Ornat, S., Gallego, C., Gallo, P., Karousou, A., Mariscal, S., & Martínez, M. (2005). MacArthur: Inventario de desarrollo comunicativo. Manual y Cuadernillos. Madrid, TEA Ediciones. ISBN: 84-7174- 820-7

Mandarin (Taiwanese)

Form: WG Contributor: Huei-Mei Liu, National Taiwan Normal University Citation: Liu, H. M., & Tsao, F. M. (2010). The standardization and application of Mandarin-Chinese communicative developmental inventory for infants and toddlers. Formosa Journal of Mental Health, (4)23, 503-534. DOI: 10.30074/CJMH.201012.0001 Liu, H. M., & Chen, Y. (2015). Developmental changes in the content and composition of early expressive vocabulary in Mandarin-speaking infants and toddlers. Bulletin of Educational Psychology, 24(7), 217-242. DOI: 10.6251/BEP.20150205

Form: WS Contributor: Huei-Mei Liu, National Taiwan Normal University Citation: Liu, H. M., & Tsao, F. M. (2010). The standardization and application of Mandarin-Chinese communicative developmental inventory for infants and toddlers. Formosa Journal of Mental Health, (4)23, 503-534. DOI: 10.30074/CJMH.201012.0001 Liu, H. M., & Chen, Y. (2015). Developmental changes in the content and composition of early expressive vocabulary in Mandarin-speaking infants and toddlers. Bulletin of Educational Psychology, 24(7), 217-242. DOI: 10.6251/BEP.20150205

Appendix B

Measures of Variability

In Chapter 5, we make use of non-parametric measures of variability, especially MADM (mean absolute deviation from the median) rather than the more standard coefficient of variation. In this brief Appendix, we show that these are very similar in the limit with a large amount of data, although they can produce quite different answers for individual data points, especially those that are at the floor or ceiling of the particular form.

Figure B.1 shows coefficient of variation (CV) vs. MADM, with each point representing a single age group for a particular combination of form and language. The slope of the relationship between the two measures is 1, despite some considerable variation. Overall, it appears that for the majority of the data, CV is slightly lower than MADM, but that it goes dramatically higher for some individual datasets. We speculate that this is due to floor/ceiling effects and small sample effects. This analysis suggests that MADM, the non-parametric estimate we use, is less subject to extreme fluctuations than CV.

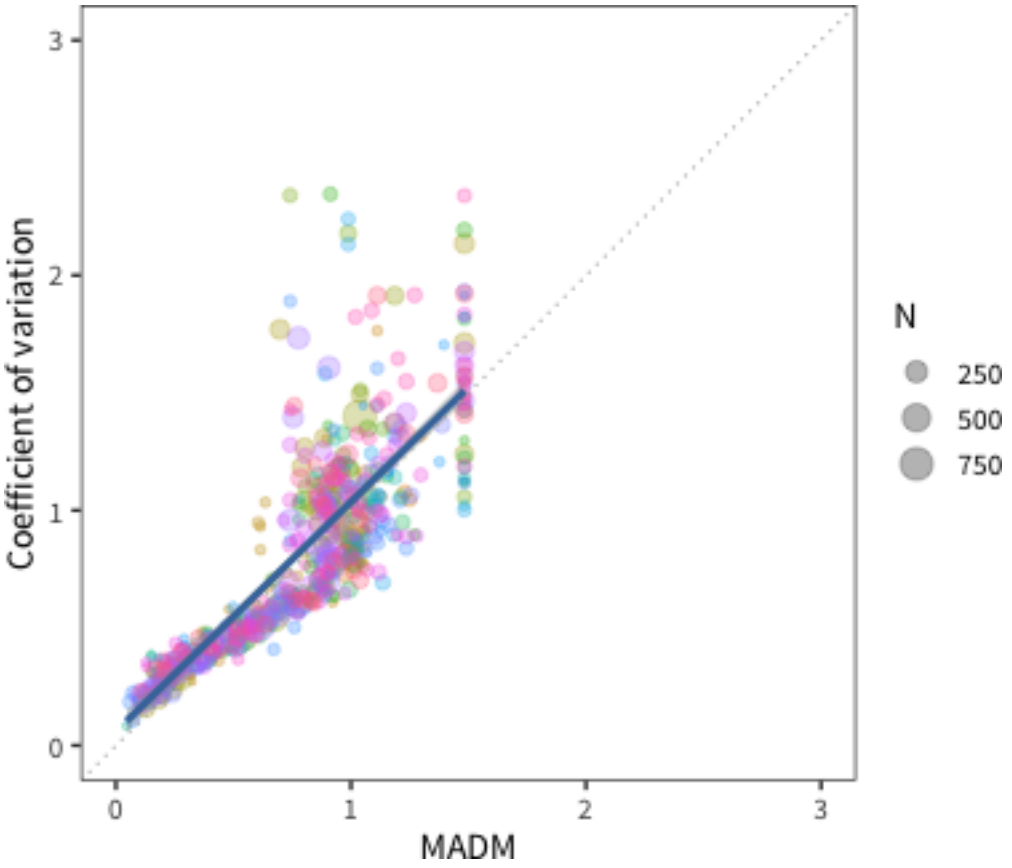


Figure B.1: Coefficient of variation and MADM, with each point showing a particular combination of language, form, and age and the line indicating a linear model fit.

Appendix C

Stitching Across Forms

Because we use different forms for different ages, there are sometimes good reasons to combine data across forms to get a broader range of ages in a particular analysis. We call this combination "stitching." This appendix provides some motivation for the practice. Figure C.1 shows 25 randomly-sampled items from the American English data. To a first approximation, production trajectories line up quite nicely with little or no visible gap between the two instruments.

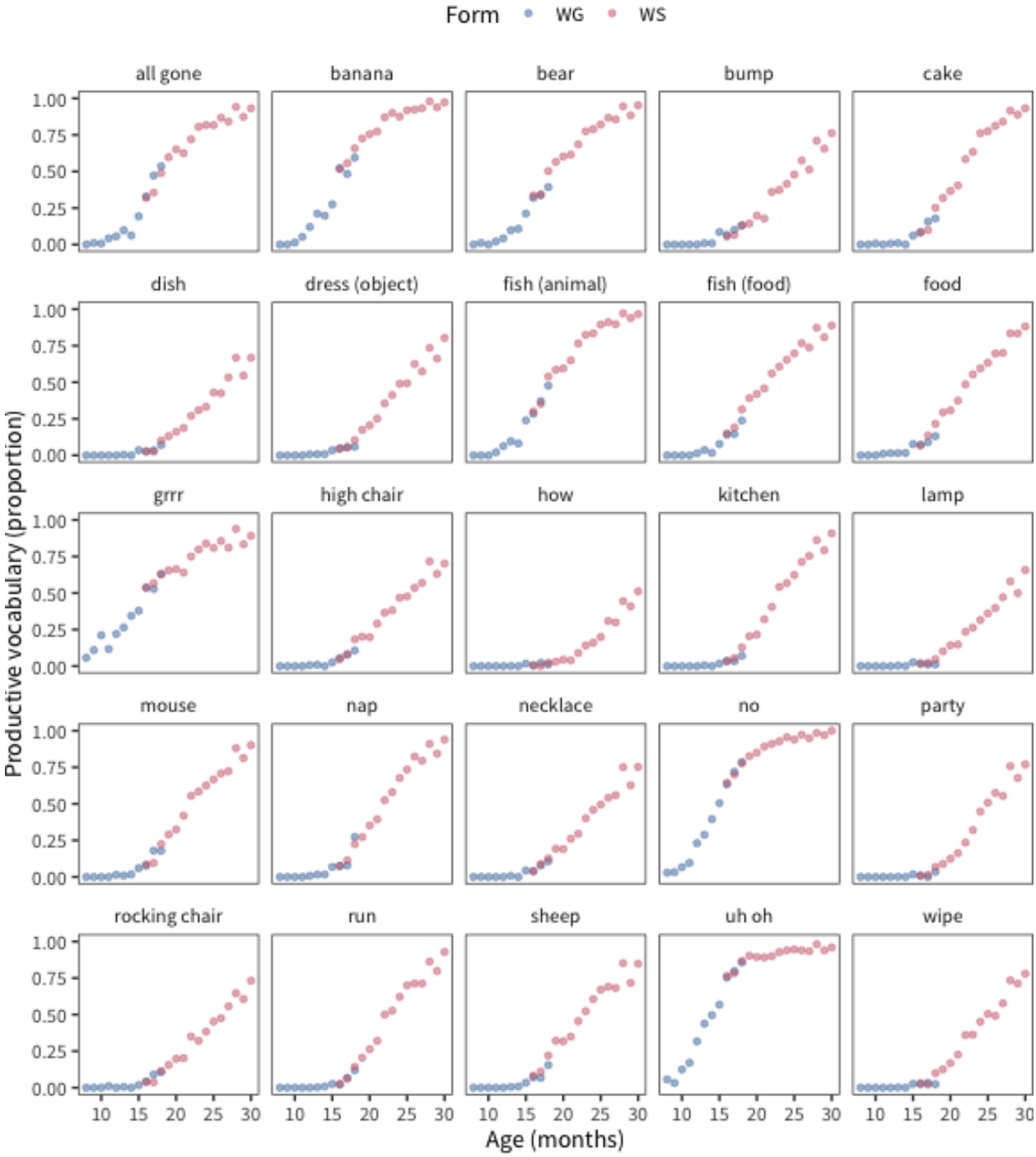


Figure C.1: WS and WG proportion production scores for a set of 25 randomly-sampled examples.

Appendix D

Estimating Age of Acquisition

It is frequently useful to have an estimate of the age at which children produce a particular word with a probability greater than some threshold; these are commonly referred to as the word's age of acquisition (AoA; Goodman et al., 2008). In this Appendix, we compare methods for estimating age of acquisition, using the English Words & Sentences data as a case study.

The simplest and most obvious measure is to use the empirically-determined first month at which the proportion producing a word exceeds the threshold (we will use 50% of children producing as our threshold in all subsequent discussion, following previous literature (Goodman et al., 2008). This approach is simple, but results in the exclusion of a large number of words. Of the total, 11% do not reach this rate by the ceiling of the form and must be discarded. Further, this method is very sensitive to sparse data. A dataset with highly clustered data will show clustered AoAs. For example, in the Swedish data, there are 307 22-month-olds and 0 21-month-olds. Thus, there will be no 21-month AoAs in this dataset. For these reasons, a model-based approach is likely to be more robust.

We initially investigated three model-based methods. Each of these models was fit to data from each word (proportion of children producing at each age) individually, resulting in a continuous curve that can be used to predict AoA more precisely. The three models we examined were:

1. Standard generalized linear model, logistic link (GLM)
2. Robust GLM
3. Bayesian GLM with hand-tuned prior parameters (Gelman et al., 2008)

For the Bayesian GLM, we experimented substantially with prior setting in order to incorporate information about the slopes and intercepts we expected for words. We used default (Cauchy) priors over coefficients. We used the empirical distribution of GLM slopes to set a strict prior over the age slopes we expected, while setting a much weaker prior over intercepts. In practice, in larger datasets, only the most extreme words (e.g. mommy, for which there is a ceiling effect for nearly every age) are affected by this choice.

In addition to these models, we investigated a hierarchical Bayesian model with shared distributional components across words. This model appeared to perform similarly to the simple word-based Bayesian model (at least in the presence of sufficient data) and was relatively expensive to fit in terms of computation, so we do not discuss it here.

Figure D.1 shows the results of these simulations, in the form of histograms of recovered 50% AoA

values. The empirical AoAs are clearly clumpy in precisely the way we describe above, even with a substantial amount of data in the analysis (N = 5846). In contrast, all three models smooth the distribution substantially, which is likely beneficial to downstream analyses. Although there are some subtle differences in the shape of the main distribution between models, the main action is found on the tails. The different models treat floor and ceiling items differently. Both the GLM and robust GLM recover two AoAs below zero (mommy and daddy); the priors of the Bayesian GLM regularize these AoAs to be 8 and 9 months respectively. Further, the Bayesian GLM estimates 10% of AoAs above 30 months (the max value in the data), while the other two methods estimate slightly fewer: 8% and 7% respectively. These Bayesian GLM results strike us as a priori more reasonable than those returned by the other methods (although they only affect a small minority of words).

To further test the Bayesian GLM approach, we tested the accuracy of the method in recovering AoAs for much smaller datasets. We did this by taking a subsample of only 100 children from the full English (American) WS dataset. We then fit standard and Bayesian GLMs to this sparse subsample. The resulting AoA estimates are plotted in Figure D.2. The Bayesian GLM shows the same minor bias to lower AoAs for hard words that the regular GLM does (a slightly below-diagonal slope), but the GLM shows much noisier estimates for the top and bottom words, suggesting that the regularization from the prior values in the Bayesian model is allowing it to deal with sparse data more effectively.

In sum, our analyses suggest that some sort of Bayesian approach is useful for estimating AoA values.

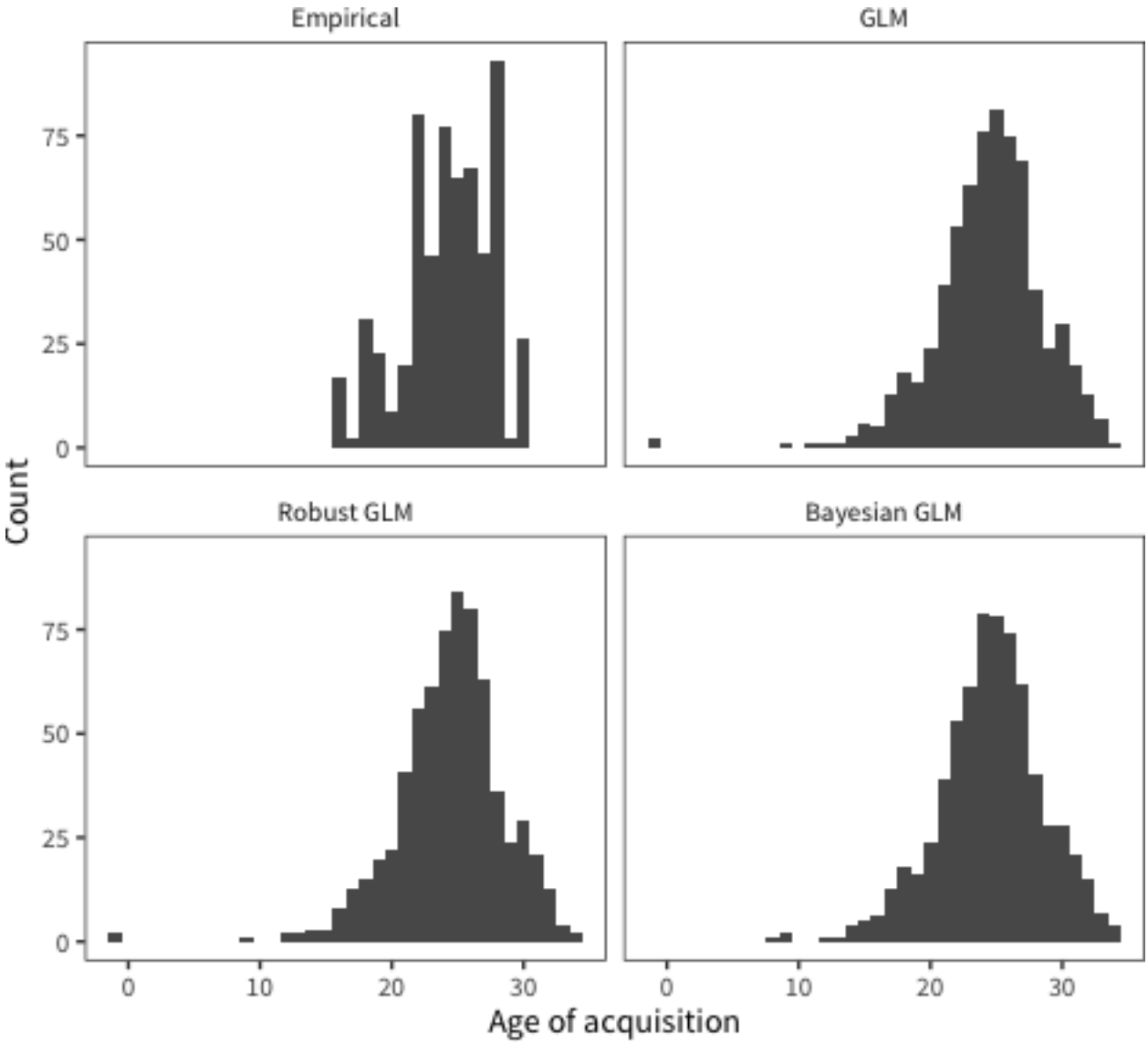


Figure D.1: Histogram of English (American) age of acquisition values as estimated via a variety of statistical methods (panels).

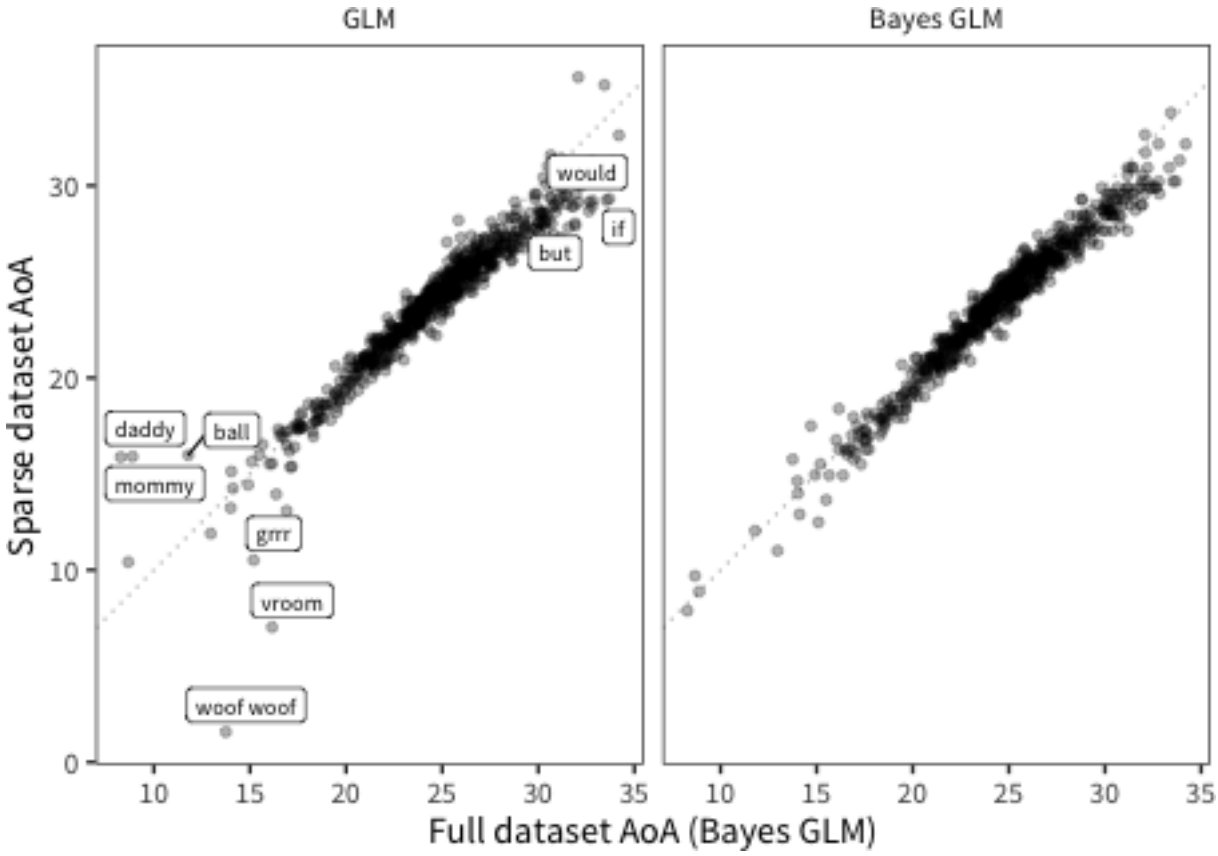


Figure D.2: Recovered AoAs from a sparse subsample (100 children), plotted by the Bayesian GLM AoAs from the full dataset. Left panel shows standard GLM, right panel shows Bayesian GLM. Differences of AoA $>$ 4 months between methods are labeled.

Bibliography

Acredolo, L. P. and Goodwyn, S. W. (1985). Symbolic gesturing in language development. Human development, 28(1):40–49.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. N. (2015). Patterns of attachment: A psychological study of the strange situation. Psychology Press.

Akhtar, N. (2005). The robustness of learning through overhearing. Developmental Science, 8(2):199–209.

Akhtar, N., Callanan, M., Pullum, G. K., and Scholz, B. C. (2004). Learning antecedents for anaphoric one. Cognition, 93(2):141–145.

Akhtar, N., Jipson, J., and Callanan, M. A. (2001). Learning words through overhearing. Child development, 72(2):416–430.

Alcock, K. and Alibhai, N. (2013). Language development in sub-saharan africa. In Neuropsychology of Children in Africa, pages 155–180. Springer.

Alcock, K., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., and Newton, C. (2015). Developmental inventories using illiterate parents as informants: Communicative development inventory (cdi) adaptation for two kenyan languages. Journal of child language, 42(4):763–785.

Alishahi, A. and Stevenson, S. (2008). A computational model of early argument structure acquisition. Cognitive Science: A Multidisciplinary Journal, 32(5):789–834.

Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisuisition. J Child Lang, 42(02):239–273.

Andonova, E. (2015). Parental report evidence for toddlers' grammar and vocabulary in bulgarian. First Language, 35(2):126–136.

Anisfeld, M., Rosenberg, E. S., Hoberman, M. J., and Gasparini, D. (1998). Lexical acceleration coincides with the onset of combinatorial speech. First Language, 18(53):165–184.

Arriaga, R. I., Fenson, L., Cronan, T., and Pethick, S. J. (1998). Scores on the macarthur communicative development inventory of children from lowand middle-income families. Applied Psycholinguistics, 19(2):209–223.

Au, T. K.-f., Dapretto, M., and Song, Y.-K. (1994). Input vs constraints: Early word acquisition in korean and english. Journal of Memory and Language, 33(5):567–582.

Baker, F. B. (2001). The basics of item response theory. ERIC.

Baker, M. C. (2005). Mapping the terrain of language learning. Language Learning and Development, 1(1):93–129.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. Developmental Psychology, 29(5):832–843.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. Joint Attention, pages 131–158.

Bannard, C., Lieven, E., and Tomasello, M. (2009). Modeling children's early grammatical knowledge. Proceedings of the National Academy of Sciences, 106(41):17284.

Bartlett, E. J. (1977). Semantic organization and reference: Acquisition of two aspects of the meaning of color terms.

Bassano, D. (2000). Early development of nouns and verbs in french: Exploring the interface between lexicon and grammar. Journal of child language, 27(3):521–559.

Bates, D., Kelman, T., Kleinschmidt, D., SimonAB, Mogensen, P. K., Bouchet-Valat, M., Hatherly, M., Saba, E., Baldassari, A., and Noack, A. (2018). dmbates/MixedModels.jl: Add adaptive Gauss-Hermite quadrature.

Bates, E. (1976). Language and context: The acquisition of pragmatics, volume 13. Academic Press, New York, NY.

Bates, E., Bretherton, I., and Snyder, L. (1988). From first words to grammar.

Bates, E., Bretherton, I., and Snyder, L. (1991). From first words to grammar: Individual differences and dissociable mechanisms, volume 20. Cambridge University Press.

Bates, E., Bretherton, I., Snyder, L., Shore, C., and Volterra, V. (1980). Vocal and gestural symbols at 13 months. Merrill-Palmer Quarterly of Behavior and Development, 26(4):407–423.

Bates, E., Camaioni, L., and Volterra, V. (1975). The acquisition of performatives prior to speech. Merrill-Palmer quarterly of behavior and development, 21(3):205–226.

Bates, E. and Goodman, J. (1999). On the emergence of grammar from the lexicon. In Macwhinney, B., editor, The Emergence of Language. Lawrence Erlbaum Associates, Mahwah, NJ.

Bates, E. and Goodman, J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. Language and cognitive Processes, 12(5-6):507–584.

Bates, E. and MacWhinney, B. (1987). Competition, variation, and language learning. Mech of Lang Acquis, pages 157–193.

Bates, E., MacWhinney, B., et al. (1989). Functionalism and the competition model. The crosslinguistic study of sentence processing, 3:73–112.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., and Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. J Child Lang, 21(01):85–123.

Bauer, D. J., Goldfield, B. A., and Reznick, J. S. (2002). Alternative approaches to analyzing individual differences in the rate of early vocabulary development. Applied Psycholinguistics, 23(3):313–335.

Bellugi, U. (1967). The acquisition of negation. Unpublished doctoral dissertation, Harvard University.

Benedict, H. (1979). Early lexical development: Comprehension and production. Journal of child language, 6(2):183–200.

Bergelson, E. and Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. Proceedings of the National Academy of Sciences, 114(49):12916–12921.

Bergelson, E. and Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. Proceedings of the National Academy of Sciences, 109(9):3253–3258.

Bergelson, E. and Swingley, D. (2013). The acquisition of abstract words by young infants. Cognition, 127(3):391–397.

Bergelson, E. and Swingley, D. (2015). Early word comprehension in infants: Replication and extension. Language Learning and Development, 11(4):369–380.

Berglund, E., Eriksson, M., and Westerlund, M. (2005). Communicative skills in relation to gender, birth order, childcare and socioeconomic status in 18-month-old children. Scandinavian journal of psychology, 46(6):485–491.

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

Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., and Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. Cognition, 140:122–134.

Bleses, D., Basbøll, H., and Vach, W. (2011). Is danish difficult to acquire? evidence from nordic past-tense studies. Language and Cognitive Processes, 26(8):1193–1231.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., and Basbøll, H. (2008). Early vocabulary development in danish and other languages: A cdi-based comparison. Journal of child language, 35(3):619–650.

Bloom, L., Tinker, E., and Margulis, C. (1993). The words children learn: Evidence against a noun bias in early vocabularies. Cognitive Development, 8(4):431–450.

Bloom, L., Tinker, E., and Scholnick, E. (2001). The intentionality model and language acquisition: Engagement, effort, and the essential tension in development. Monographs of the Society for Research in Child Development, 66(4).

Bloom, P. (2000). How children learn the meanings of words. Number Sirsi) i9780262523295. MIT Press, Cambridge, MA.

Bloom, P. (2002). How children learn the meanings of words. MIT Press, Cambridge, MA.

Bornstein, M. H. (1985). On the development of color naming in young children: Data and theory. Brain and Language, 26(1):72–93.

Bornstein, M. H. (2013). Cultural approaches to parenting. Psychology Press.

Bornstein, M. H. and Cote, L. R. (2005). Expressive vocabulary in language learners from two ecological settings in three language communities. Infancy, 7(3):299–316.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., and Vyt, A. (2004a). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. Child development, 75(4):1115–1139.

Bornstein, M. H., Hahn, C.-S., and Haynes, O. M. (2004b). Specific and general language performance across early childhood: Stability and gender considerations. First Language, 24(3):267–304.

Bornstein, M. H. and Haynes, O. M. (1998). Vocabulary competence in early childhood: Measurement, latent construct, and predictive validity. Child development, 69(3):654–671.

Bornstein, M. H., Hendricks, C., Hahn, C.-S., Haynes, O. M., Painter, K. M., and Tamis-LeMonda, C. S. (2003). Contributors to self-perceived competence, satisfaction, investment, and role balance in maternal parenting: A multivariate ecological analysis. Parenting: Science and Practice, 3(4):285–326.

Bornstein, M. H. and Putnick, D. L. (2012). Stability of language in childhood: A multiage, multidomain, multimeasure, and multisource study. Developmental psychology, 48(2):477.

Bowerman, M. (1996). The origins of children's spatial semantic categories: Cognitive versus linguistic determinants. In Gumperz, J. J. and Levinson, S. C., editors, Rethinking Linguistic Relativity, pages 145–176. Cambridge University Press.

Bradley, R. H. and Corwyn, R. F. (2002). Socioeconomic status and child development. Annual review of psychology, 53(1):371–399.

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (2015). Developmental changes in the relationship between grammar and the lexicon. In Proceedings of the 37th Annual Meeting of the Cognitive Science Society.

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (2016). From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In Proceedings of the 38th Annual Meeting of the Cognitive Science Society.

Braginsky, M., Yurovsky, D., Marchman, V. A., and Frank, M. C. (under revision). Consistency and variability in word learning across languages.

Brinchmann, E. I., Braeken, J., and Lyster, S.-A. H. (2018). Is there a direct relation between the development of vocabulary and grammar? Developmental science, page e12709.

Brooks, R. and Meltzoff, A. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: a longitudinal, growth curve modeling study. Journal of Child Language, 35(01):207–220.

Brown, R. (1973). A first language: The early stages. Harvard Universirty Press, Cambridge, MA.

Bruner, J. (1985). Child's talk: Learning to use language. Child Language Teaching and Therapy, 1(1):111–114.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. Behav Res Meth, 46(3):904–911.

Busby, J. and Suddendorf, T. (2005). Recalling yesterday and predicting tomorrow. Cognitive Development, 20(3):362–372.

Butterworth, G. (2003). Pointing is the royal road to language for babies. In Pointing, pages 17–42. Psychology Press.

Carey, S. (1978). The child as word learner. In Linguistic theory and psychological reality, pages 264–293. MIT Press, Cambridge, MA.

Carey, S. (2009). The Origin of Concepts. Oxford University Press, Oxford, England.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., and Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monographs of the society for research in child development, 63(4).

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., and Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. Proceedings of the National Academy of Sciences, 110(28):11278–11283.

Caselli, C., Casadio, P., and Bates, E. (1999). A comparison of the transition from first words to grammar in english and italian. Journal of child language, 26(01):69–111.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., and Weir, J. (1995). A cross-linguistic study of early lexical development. Cog Dev, 10(2):159–199.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. Journal of Statistical Software, 48(6):1–29.

Chalmers, R. P. et al. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. Journal of Statistical Software, 71(5):1–39.

Charman, T., Drew, A., Baird, C., and Baird, G. (2003). Measuring early language development in preschool children with autism spectrum disorder using the macarthur communicative development inventory (infant form). Journal of Child Language, 30(1):213–236.

Choi, S. and Gopnik, A. (1995). Early acquisition of verbs in korean: A cross-linguistic study. Journal of child language, 22(03):497–529.

Chomsky, N. (1957). Syntactic Structures. Mouton, The Hague.

Chomsky, N. (1981). Principles and parameters in syntactic theory. Explanation in linguistics: The logical problem of language acquisition, pages 32–75.

Chomsky, N. (2014). The minimalist program. MIT press.

Clark, E. (2003). First language acquisition. Cambridge University Press, Cambridge, UK.

Clark, E. V. (1977). From gesture to word: on the natural history of deixis in language acquisition. In Bruner, J. S. and Garton, A., editors, Human growth and development: Wolfson College Lectures, pages 85–120. Oxford University Press., Oxford.

Clark, E. V. and Hecht, B. F. (1983). Comprehension, production, and language acquisition. Annual review of psychology, 34(1):325–349.

Colombo, J. (2001). The development of visual attention in infancy. Annual review of psychology, 52(1):337–367.

Conwell, E. and Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. Cognition, 103(2):163–179.

Crain, S. and Thornton, R. (2000). Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics. MIT Press.

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., and Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. Child development, 85(4):1330–1345.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological bulletin, 52(4):281.

Curtiss, S. (1977). Genie: A psychological study of a modern-day wild child. Perspectives in Neurolinguistics and Psycholinguistics. Academic Press, Boston, MA.

Dale, P. S. (2015). Adaptations, Not Translations!

Dale, P. S., Dionne, G., Eley, T. C., and Plomin, R. (2000). Lexical and grammatical development: A behavioural genetic perspective. Journal of child language, 27(3):619–642.

Dale, P. S. and Fenson, L. (1996). Lexical development norms for young children. Behavior Research Methods, Instruments, & Computers, 28(1):125–127.

Dale, P. S. and Penfold, M. (2011). Adaptations of the MacArthur-Bates CDI Into Non-US English Languages.

Dale, P. S., Price, T. S., Bishop, D. V., and Plomin, R. (2003). Outcomes of early language delay: I. predicting persistent and transient language difficulties at 3 and 4 years. Journal of Speech, Language, and Hearing Research, 46(3):544–560.

Dapretto, M. and Bjork, E. L. (2000). The development of word retrieval abilities in the second year and its relation to early vocabulary growth. Child Development, 71(3):635–648.

Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. Neuropsychologia, 44(11):2037–2078.

De Houwer, A., Bornstein, M. H., and Leach, D. B. (2005). Assessing early communicative ability: A cross-reporter cumulative score for the macarthur cdi. Journal of Child Language, 32(4):735–758.

Decker, D. M. et al. (1999). Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press.

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., and Bates, E. (2005). A crosslinguistic study of the relationship between grammar and lexical development. Journal of Child Language, 32(4):759–786.

Diamond, A. (2013). Executive functions. Annual review of psychology, 64:135–168.

Diamond, A. and Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. Science, 333(6045):959–964.

Dickinson, D. K. and Tabors, P. O. (2001). Beginning literacy with language: Young children learning at home and school. Paul H Brookes Publishing.

Dionne, G., Dale, P. S., Boivin, M., and Plomin, R. (2003). Genetic evidence for bidirectional effects of early lexical and grammatical development. Child development, 74(2):394–412.

Dixon, J. A. and Marchman, V. A. (2007). Grammar and the lexicon: Developmental ordering in language acquisition. Child Development, 78(1):190–212.

Dore, J. (1974). A pragmatic description of early language development. Journal of psycholinguistic Research, 3(4):343–350.

Duddington, J. (2012). espeak text to speech.

Dunn, L. M. and Dunn, L. M. (2007). Peabody picture vocabulary test. AGS Publishing/Pearson Assessments, Parsippany, NJ, 4th edition edition.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. Cognition, 173:43–59.

Eigsti, I.-M., de Marchena, A. B., Schuh, J. M., and Kelley, E. (2011). Language acquisition in autism spectrum disorders: A developmental review. Research in Autism Spectrum Disorders, 5(2):681–691.

Elman, J. L., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. MIT Press, Cambridge, MA.

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pérez Pereira, M., Wehberg, S., Marjanovič-Umek, L., Gayraud, F., Kovacevic, M., and Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. British Journal of Developmental Psychology, 30(2):326–343.

Evans, G. W. (2004). The environment of childhood poverty. American psychologist, 59(2):77.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and brain sciences, 32(5):429–448.

Farkas, G. and Beron, K. (2004). The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. Social Science Research, 33(3):464–497.

Fazly, A., Alishahi, A., and Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. Cognitive Science, 34(6):1017–1063.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs-Lasky, M., Rockette, H. E., and Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. Child Development, 76(4):856–868.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., and Paradise, J. L. (2000). Measurement properties of the macarthur communicative development inventories at ages one and two years. Child development, 71(2):310–322.

Fenson, L., Bates, E., Dale, P., Goodman, J. C., Reznick, J. S., and Thal, D. (2000). Reply: Measuring variability in early child language: Don't shoot the messenger. Child development, 71(2):323–328.

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., and Thal, D. J. (2007). MacArthur-Bates Communicative Development Inventories. Brookes Publishing Company.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C., and Stiles, J. (1994). Variability in early communicative development. Monogr Soc Res Child Dev, 59(5).

Fernald, A. and Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. Child development, 83(1):203–222.

Fernald, A., Marchman, V. A., and Weisleder, A. (2013). Ses differences in language processing skill and vocabulary are evident at 18 months. Developmental Science, 16(2):234–248.

Fernald, A. and Morikawa, H. (1993). Common themes and cultural variations in japanese and american mothers' speech to infants. Child Development, 64:637–56.

Fernald, A., Perfors, A., and Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. Developmental psychology, 42(1):98.

Fernald, L. C., Kariger, P., Hidrobo, M., and Gertler, P. J. (2012). Socioeconomic gradients in child development in very young children: Evidence from india, indonesia, peru, and senegal. Proceedings of the National Academy of Sciences, page 201121241.

Fiser, J. and Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. Proceedings of the National Academy of Sciences, 99(24):15822–15826.

Fisher, C., Gertner, Y., Scott, R. M., and Yuan, S. (2010). Syntactic bootstrapping. Wiley Interdisciplinary Reviews: Cognitive Science, 1(2):143–149.

Floccia, C., Sambrook, T. D., Delle Luche, C., Kwok, R., Goslin, J., White, L., Cattani, A., Sullivan, E., Abbot-Smith, K., Krott, A., et al. (2018). Vocabulary of 2-year-olds learning english and an additional language: Norms and effects of linguistic distance. Monographs of the Society for Research in Child Development, 83(1):1–135.

Flynn, J. R. (1987). Massive iq gains in 14 nations: What iq tests really measure. Psychological bulletin, 101(2):171.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Infancy, 22(4):421–435.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2016a). Wordbank: An open repository for developmental vocabulary data. Journal of child language.

Frank, M. C., Goldwater, S., Griffiths, T., and Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. Cognition, 117(2):107–125.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. Psychological Science, 20:578–585.

Frank, M. C., Lewis, M., and MacDonald, K. (2016b). A performance model for early word learning. In Proceedings of the 38th annual conference of the cognitive science society.

Frank, M. C., Tenenbaum, J. B., and Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. Language, Learning, and Development.

Frank, S. L., Bod, R., and Christiansen, M. H. (2012). How hierarchical is language use? Proceedings of the Royal Society of London B: Biological Sciences, page rspb20121741.

Freudenthal, D., Pine, J., and Gobet, F. (2010). Explaining quantitative variation in the rate of optional infinitive errors across languages: a comparison of mosaic and the variational learning model. Journal of child language, 37(3):643–669.

Friend, M. and Keplinger, M. (2008). Reliability and validity of the computerized comprehension task (cct): data from american english and mexican spanish infants. Journal of child language, 35(1):77–98.

Ganger, J. and Brent, M. R. (2004). Reexamining the vocabulary spurt. Developmental psychology, 40(4):621.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., and Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. Memory, 2(2):103–127.

Gathercole, V. C. M., Thomas, E. M., Roberts, E., Hughes, C., and Hughes, E. K. (2013). Why assessment needs to take exposure into account: Vocabulary and grammatical abilities in bilingual children. Issues in the Assessment of Bilinguals, pages 20–55.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. science, 332(6033):1100–1104.

Gelman, A., Jakulin, A., Pittau, M. G., Su, Y.-S., et al. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2(4):1360–1383.

Gentner, D. (1978). On relational meaning: The acquisition of verb meaning. Child development, pages 988–998.

Gentner, D. and Boroditsky, L. (2001). Individuation, relativity, and early word learning. In Lang Acquis and concept dev. Cambridge University Press.

Gertner, Y., Fisher, C., and Eisengart, J. (2006). Learning words and rules. Psychological Science, 17(8):684.

Gillette, J., Gleitman, H., Gleitman, L., and Lederer, A. (1999). Human simulations of vocabulary learning. Cognition, 73(2):135–176.

Gleitman, L. (1990). The structural sources of verb meanings. Lang Acquis, 1(1):3–55.

Goldberg, A. E. (2006). Constructions at work: The nature of generalization in language. Oxford University Press on Demand.

Goldfield, B. A. (1993). Noun bias in maternal speech to one-year-olds. Journal of child language, 20(1):85–99.

Goldfield, B. A. and Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. Journal of child language, 17(1):171–183.

Goldin-Meadow, S. (1998). The development of gesture and speech as an integrated system. New Directions for Child and Adolescent Development, 1998(79):29–42.

Goldin-Meadow, S. and Mylander, C. (1983). Gestural communication in deaf children: Noneffect of parental input on language development. Science, 221(4608):372–374.

Gómez, R. and Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. Cognition, 70:109–135.

Goodman, J., Dale, P., and Li, P. (2008). Does frequency count? parental input and the acquisition of vocabulary. Journal of child language, 35(3):515.

Gopnik, A., Choi, S., and Baumberger, T. (1996). Cross-linguistic differences in early semantic and cognitive development. Cognitive Development, 11(2):197–225.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. Psychological review, 111(1):3.

Greenberg, J. H. (1963). Universals of language.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). Fundamentals of item response theory, volume 2. Sage.

Hammer, C. S., Farkas, G., and Maczuga, S. (2010). The language and literacy development of head start children: A study using the family and child experiences survey database. Language, Speech, and Hearing Services in Schools, 41(1):70–83.

Hao, M., Shu, H., Xing, A., and Li, P. (2008). Early vocabulary inventory for mandarin chinese. Behavior Research Methods, 40(3):728–733.

Hardwicke, T. E., Mathur, M., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal cognition.

Hart, B. and Risley, T. (1995). Meaningful differences in the everyday experience of young American children. Brookes Publishing Company, Baltimore, MD.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? science, 298(5598):1569–1579.

Hayiou-Thomas, M. E., Dale, P. S., and Plomin, R. (2012). The etiology of variation in language skills changes with development: A longitudinal twin study of language from 2 to 12 years. Developmental science, 15(2):233–249.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3):101–102.

Hidaka, S. (2015). Estimating the latent number of types in growing corpora with reduced cost–accuracy trade-off. Journal of Child Language, pages 1–28.

Hills, T. T., Maouene, J., Riordan, B., and Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63(3):259–273.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., and Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? Psychological Science, 20(6):729–739.

Hirsh-Pasek, K., Adamson, L. B., Bakeman, R., Owen, M. T., Golinkoff, R. M., Pace, A., Yust, P. K., and Suma, K. (2015). The contribution of early communication quality to low-income children's language success. Psychological Science, 26(7):1071–1083.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. Child development, 74(5):1368–1378.

Hoff, E. (2006). How social contexts support and shape language development. Developmental review, 26(1):55–88.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., and Parra, M. (2012). Dual language exposure and early bilingual development. Journal of child language, 39(1):1–27.

Hoff, E. and Naigles, L. (2002). How children use input to acquire a lexicon. Child development, 73(2):418–433.

Hoff, E., Quinn, J. M., and Giguere, D. (2017). What explains the correlation between growth in vocabulary and grammar? new evidence from latent change score analyses of simultaneous bilingual development. Developmental science.

Hoff-Ginsberg, E. (1998). The relation of birth order and socioeconomic status to children's language experience and language development. Applied Psycholinguistics, 19(4):603–629.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., Rocroi, C., and Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. Monogr Soc Res Child Dev, pages i–135.

Holman, E. W., Wichmann, S., Brown, C. H., Folia, V. V., Müller, A., and Bakker, D. (2008). Explorations in automated language classification. Folia Linguistica, 42:331–354.

Horowitz, A. C., Schneider, R. M., and Frank, M. C. (2017). The trouble with quantifiers: Exploring children's deficits in scalar implicature. Child development.

Hurtado, N., Marchman, V., and Fernald, A. (2008). Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children. Developmental Science, 11(6):F31–F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., and Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. Developmental Psychology, 27(2):236–248.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., and Levine, S. (2002). Language input and child syntax. Cognitive psychology, 45(3):337–374.

Hyde, J. S. and Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis.

Iverson, J. M., Capirci, O., and Caselli, M. C. (1994). From communication to language in two modalities. Cognitive development, 9(1):23–43.

Iverson, J. M. and Goldin-Meadow, S. (2005). Gesture paves the way for language development. Psychological science, 16(5):367–371.

Jørgensen, R. N., Dale, P. S., Bleses, D., and Fenson, L. (2010). CLEX: A cross-linguistic lexical norms database. Journal of Child Language, 37(02):419–428.

Kachergis, G., Yu, C., and Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. Psychonomic bulletin & review, 19(2):317–324.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. Psychological bulletin, 109(3):490.

Kamhi, A. G. (1986). The elusive first word: The importance of the naming insight for the development of referential speech. Journal of child language, 13(1):155–161.

Katsos, N., Cummins, C., Ezeizabarrena, M.-J., Gavarró, A., Kraljević, J. K., Hrzica, G., Grohmann, K. K., Skordi, A., De Lopez, K. J., Sundahl, L., et al. (2016). Cross-linguistic patterns in the acquisition of quantifiers. Proceedings of the National Academy of Sciences, 113(33):9244–9249.

Kauschke, C. and Hofmeister, C. (2002). Early lexical development in german: A study on vocabulary growth and vocabulary composition during the second and third year of life. Journal of child language, 29(4):735–757.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., and Cook, R. (2009). The world color survey. CSLI Publications Stanford.

Kim, M., McGregor, K. K., and Thompson, C. K. (2000). Early lexical development in english-and korean-speaking children: Language-general and language-specific patterns. Journal of Child Language, 27(2):225–254.

Kristoffersen, K. E., Simonsen, H. G., Bleses, D., Wehberg, S., Jørgensen, R. N., Eiesland, E. A., and Henriksen, L. Y. (2013). The use of the internet in collecting cdi data–an example from norway. Journal of Child Language, 40(03):567–585.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. Nature reviews neuroscience, 5(11):831–843.

Kuhn, T. S. (1970). The structure of scientific revolutions. University ot Chicago Press, Chicago, pages 84–85.

Landau, B., Gleitman, L. R., and Landau, B. (2009). Language and experience: Evidence from the blind child. Harvard University Press.

Lany, J. and Saffran, J. R. (2010). From statistics to meaning: Infants acquisition of lexical categories. Psychological Science, 21(2):284–291.

Legate, J. A. and Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. The Linguistic Review, 18(1-2):151–162.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10:707–710.

Libertus, M. E., Odic, D., Feigenson, L., and Halberda, J. (2015). A developmental vocabulary assessment for parents (dvap): Validating parental report of vocabulary size in 2-to 7-year-old children. Journal of Cognition and Development, 16(3):442–454.

Lidz, J., Waxman, S., and Freedman, J. (2003). What infants know about syntax but couldn't have learned: experimental evidence for syntactic structure at 18 months. Cognition, 89(3):295–303.

Lieven, E., Salomo, D., and Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. Cognitive Linguistics, 20(3):481–507.

Lieven, E. V., Pine, J. M., and Baldwin, G. (1997). Lexically-based learning and early grammatical development. Journal of child language, 24(1):187–219.

Lieven, E. V., Pine, J. M., and Barnes, H. D. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. Journal of child language, 19(2):287–310.

Liszkowski, U., Carpenter, M., and Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. Developmental science, 10(2):F1–F7.

Lutchmaya, S., Baron-Cohen, S., and Raggatt, P. (2001). Foetal testosterone and vocabulary size in 18-and 24-month-old infants. Infant Behavior and Development, 24(4):418–424.

Luyster, R., Lopez, K., and Lord, C. (2007). Characterizing communicative development in children referred for autism spectrum disorders using the macarthur-bates communicative development inventory (cdi). Journal of Child Language, 34(3):623–654.

Maccoby, E. E. and Jacklin, C. N. (1974). The Psychology of Sex Differences. Stanford University Press.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. Third Edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Maital, S. L., Dromi, E., Sagi, A., and Bornstein, M. H. (2000). The hebrew communicative development inventory: Language specific properties and cross-linguistic generalizations. Journal of Child Language, 27(01):43–67.

Makransky, G., Dale, P. S., Havmose, P., and Bleses, D. (2016). An item response theory–based, computerized adaptive testing version of the macarthur–bates communicative development inventory: Words & sentences (cdi: Ws). Journal of Speech, Language, and Hearing Research, 59(2):281–289.

Marchman, V. A. (1997). Children's productivity in the english past tense: The role of frequency, phonology, and neighborhood structure. Cognitive Science, 21(3):283–304.

Marchman, V. A. and Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. Journal of child language, 21(2):339–366.

Marchman, V. A. and Dale, P. S. (2017). 3 assessing receptive and expressive vocabulary in child language. Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide, page 40.

Marchman, V. A. and Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. Developmental Science, 11(3):F9–F16.

Marchman, V. A. and Martínez-Sussmann, C. (2002). Concurrent validity of caregiver/parent report measures of language for children who are learning both english and spanish. Journal of Speech, Language, and Hearing Research, 45(5):983–997.

Marcus, G. (1995). The acquisition of the english past tense in children and multilayered connectionist networks. Cognition, 56(3):271–279.

Mariscal, S. and Gallego, C. (2012). The relationship between early lexical and grammatical development in spanish: Evidence in children with different linguistic levels. The Spanish journal of psychology, 15(1):112–123.

Marjanovič-Umek, L., Fekonja-Peklaj, U., and Podlesek, A. (2013). Characteristics of early vocabulary and grammar development in slovenian-speaking infants and toddlers: a cdi-adaptation study. Journal of child language, 40(4):779–798.

Markus, H. R. and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. Psychological review, 98(2):224.

Masur, E. F. (1990). Gestural development, dual-directional signaling, and the transition to words. In From gesture to language in hearing and deaf children, pages 18–30. Springer.

Mayor, J. and Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from cdi analysis. Developmental Science, 14(4):769–785.

Mayor, J. and Plunkett, K. (2014). Shared understanding and idiosyncratic expression in early vocabularies. Developmental science, 17(3):412–423.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. Science, 317(5838):631–631.

McMurray, B., Horst, J. S., and Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. Psychological review, 119(4):831.

Medina, T., Snedeker, J., Trueswell, J., and Gleitman, L. (2011). How words can and cannot be learned by observation. Proceedings of the National Academy of Sciences, 108(22):9014.

Meylan, S. C., Frank, M. C., Roy, B. C., and Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. Psychological science, 28(2):181–192.

Miller, D. I. and Halpern, D. F. (2014). The new science of cognitive sex differences. Trends in cognitive sciences, 18(1):37–45.

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

Mollica, F. and Piantadosi, S. T. (2017). How data drive early word learning: A cross-linguistic waiting time analysis. Open Mind, 1(2):67–77.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A.-L., De Schryver, M., De Winne, J., and Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisuisition for 4,300 dutch words. Behav Res Meth, 45(1):169–177.

Moyle, M. J., Weismer, S. E., Evans, J. L., and Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. Journal of Speech, Language, and Hearing Research, 50(2):508–528.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1:0021.

Naigles, L. (1990). Children use syntax to learn verb meanings. Journal of child language, 17(2):357–374.

Nelson, C. A. (2007). A neurobiological perspective on early human deprivation. Child development perspectives, 1(1):13–18.

Nelson, K. (1973). Structure and strategy in learning to talk. Monographs of the Society for Research in Child Development, pages 1–135.

Nisbett, R. E., Peng, K., Choi, I., and Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. Psychological review, 108(2):291.

Nordmeyer, A. E. and Frank, M. C. (2014). The role of context in young children's comprehension of negation. Journal of Memory and Language, 77:25–39.

Nordmeyer, A. E. and Frank, M. C. (2018). Uninformative negation is infelicitous to both adults and children. Language Learning and Development.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., et al. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences, 106(26):10593–10597.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349(6251):aac4716.

Paul, R. (1996). Clinical implications of the natural history of slow expressive language development. American Journal of Speech-Language Pathology, 5(2):5–21.

Pea, R. D. (1982). Origins of verbal logic: Spontaneous denials by two-and three-year olds. Journal of child language, 9(3):597–626.

Pearson, B. Z., Fernandez, S. C., and Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. Language learning, 43(1):93–120.

Perry, L. K., Perlman, M., and Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisuisition. PloS One, 10(9):e0137147.

Piaget, J. (1962). Play, dreams and imitation in childhood. Routledge.

Piantadosi, S. T. and Gibson, E. (2014). Quantitative standards for absolute linguistic universals. Cognitive Science, 38(4):736–756.

Pine, J. M. and Lieven, E. V. (1990). Referential style at thirteen months: why age-defined cross-sectional measures are inappropriate for the study of strategy differences in early language development. Journal of Child Language, 17(3):625–631.

Pine, J. M. and Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. Applied psycholinguistics, 18(2):123–138.

Pinker, S. (1991). Rules of language. Science, 253(5019):530–535.

Pinker, S. and Jackendoff, R. (2005). The faculty of language: what's special about it? Cognition, 95(2):201–236.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Connections and symbols, pages 73–193.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. Cognition, 38(1):43–102.

Plunkett, K. and Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. Cognition, 48(1):21–69.

Plunkett, K. and Marchman, V. (1996). Learning from a connectionist model of the acquisition of the english past tense. Cognition, 61(3):299–308.

Poulin-Dubois, D., Graham, S., and Sippola, L. (1995). Early lexical development: The contribution of parental labelling and infants' categorization abilities. Journal of Child Language, 22(2):325–343.

Pullum, G. K. and Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. The linguistic review, 18(1-2):9–50.

Quine, W. (1960). Word and object. The MIT Press.

Redington, M., Crater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. Cognitive Science: A Multidisciplinary Journal, 22(4):425–469.

Reese, E. and Read, S. (2000). Predictive validity of the new zealand macarthur communicative development inventory: Words and sentences. Journal of Child Language, 27(2):255–266.

Regier, T. and Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. Cognition, 93(2):147–155.

Rescorla, L. (1989). The language development surveya screening tool for delayed language in toddlers. Journal of Speech and Hearing Disorders, 54(4):587–599.

Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay. Journal of Speech, Language, and Hearing Research, 52(1):16–30.

Rescorla, L., Dahlsgaard, K., and Roberts, J. (2000). Late-talking toddlers: Mlu and ipsyn outcomes at 3; 0 and 4; 0. Journal of child language, 27(3):643–664.

Rescorla, L., Roberts, J., and Dahlsgaard, K. (1997). Late talkers at 2: Outcome at age 3. Journal of Speech, Language, and Hearing Research, 40(3):556–566.

Roberts, J. E., Burchinal, M., and Durham, M. (1999). Parents' report of vocabulary and grammatical development of african american preschoolers: Child and environmental associations. Child Development, 70(1):92–106.

Robinson, J. P. and Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. American Educational Research Journal, 48(2):268–302.

Ross-sheehy, S., Oakes, L. M., and Luck, S. J. (2003). The development of visual short-term memory capacity in infants. Child development, 74(6):1807–1822.

Rovee-Collier, C. (1997). Dissociations in infant memory: rethinking the development of implicit and explicit memory. Psychological review, 104(3):467.

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

Rowe, M. L. and Goldin-Meadow, S. (2009). Early gesture selectively predicts later language learning. Developmental science, 12(1):182–187.

Rowe, M. L., Suskind, D. L., and Hoff, E. (2012). Early language gaps: Sources and solutions.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. (2015). Predicting the birth of a spoken word. Proc Natl Acad Sci, 112(41):12663–12668.

Roy, D. and Pentland, A. (2002). Learning words from sights and sounds: a computational model. Cognitive Science, 26:113–146.

Rumelhart, D. E., McClelland, J. L., and the PDP research group (1986). Parallel distributed processing: Explorations in the microstructure of cognition. MIT Press, Cambridge, MA.

Saffran, J. R., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. Science, 274(5294):1926.

Saffran, J. R. and Kirkham, N. Z. (2018). Infant statistical learning. Annual review of psychology, 69.

Schneider, R., Yurovsky, D., and Frank, M. C. (2015). Large-scale investigations of variability in children's first words. In Proceedings of the Cognitive Science Society.

Schults, A. and Tulviste, T. (2016). Composition of estonian infants expressive lexicon according to the adaptation of cdi/words and gestures. First Language, 36(5):485–504.

Schwartz, R. G. and Terrell, B. Y. (1983). The role of input frequency in lexical acquisuisition. J Child Lang, 10(01):57–64.

Shafto, P., Goodman, N. D., and Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. Perspectives on Psychological Science, 7(4):341–351.

Shneidman, L. A. and Goldin-Meadow, S. (2012). Language input and acquisition in a mayan village: How important is directed speech? Developmental science, 15(5):659–673.

Shukla, M., White, K. S., and Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. Proceedings of the National Academy of Sciences, 108(15):6038–6043.

Singer Harris, N. G., Bellugi, U., Bates, E., Jones, W., and Rossen, M. (1997). Contrasting profiles of language development in children with williams and down syndromes. Developmental Neuropsychology, 13(3):345–370.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. Cognition, 61:39–91.

Slobin, D. I. (1985). The crosslinguistic study of language acquisition: Theoretical issues, volume 2. Psychology Press.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In Gumperz, J. J. and Levinson, S. C., editors, Rethinking Linguistic Relativity, pages 70–96.

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

Snedeker, J. (2009). Word learning. In Squire, L., editor, Encyclopedia of Neuroscience, pages 503–508. Elsevier.

Snedeker, J., Geren, J., and Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. Psychol Sci, 18(1):79–87.

Snedeker, J., Geren, J., and Shafto, C. L. (2012). Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of english in internationally-adopted children. Cognitive Psychology, 65(1):39–76.

Spelke, E. S. and Kinzler, K. D. (2007). Core knowledge. Developmental science, 10(1):89–96.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, 106(26):10587–10592.

Stoet, G. and Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within-and across-nation assessment of 10 years of pisa data. PloS one, 8(3):e57988.

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. J Speech Lang Hear Res, 53(3):670–683.

Stolt, S., Haataja, L., Lapinleimu, H., and Lehtonen, L. (2009). Associations between lexicon and grammar at the end of the second year in finnish children. Journal of Child Language, 36(4):779–806.

Størvold, G. V., Aarethun, K., and Bratberg, G. H. (2013). Age for onset of walking and prewalking strategies. Early human development, 89(9):655–659.

Swingley, D. and Humphrey, C. (2017). Quantitative linguistic predictors of infants' learning of specific English words. Chi Dev.

Szagun, G., Steinbrink, C., Franik, M., and Stumper, B. (2006). Development of vocabulary and grammar in young german-speaking children assessed with a german language development inventory. First Language, 26(3):259–280.

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., and Yoder, P. (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. Journal of Speech, Language, and Hearing Research, 52(3):643–652.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. Developmental Psychology, 32(3):492.

Tardif, T., Fletcher, P., Liang, W., and Kaciroti, N. (2009). Early vocabulary development in mandarin (putonghua) and cantonese. Journal of Child Language, 36(5):1115–1144.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., and Marchman, V. A. (2008). Baby's first 10 words. Developmental Psychology, 44(4):929.

Tardif, T., Gelman, S. A., and Xu, F. (1999). Putting the noun bias in context: A comparison of english and mandarin. Child Development, 70(3):620–635.

Tardif, T., Shatz, M., and Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of english, italian, and mandarin. Journal of Child Language, 24(3):535–565.

Thal, D. and Bates, E. (1988). Language and gesture in late talkers. Journal of Speech, Language, and Hearing Research, 31(1):115–123.

Thal, D., Jackson-Maldonado, D., and Acosta, D. (2000). Validity of a parent-report measure of vocabulary and grammar for spanish-speaking toddlers. Journal of Speech, Language, and Hearing Research, 43(5):1087–1100.

Thal, D., Tobias, S., and Morrison, D. (1991). Language and gesture in late talkers: A 1-year follow-up. Journal of Speech, Language, and Hearing Research, 34(3):604–612.

Thal, D. J., Bates, E., Goodman, J., and Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late-and early-talking toddlers. Developmental Neuropsychology, 13(3):239–273.

Thal, D. J., Bates, E., Zappia, M. J., and Oroz, M. (1996). Ties between lexical and grammatical development: Evidence from early-talkers. Journal of Child Language, 23(2):349–368.

Thal, D. J., O'Hanlon, L., Clemmons, M., and Fralin, L. (1999). Validity of a parent report measure of vocabulary and syntax for preschool children with language impairment. Journal of Speech, Language, and Hearing Research, 42(2):482–496.

Thordardottir, E. T., Weismer, S. E., and Evans, J. L. (2002). Continuity in lexical and morphological development in icelandic and english-speaking 2-year-olds. First Language, 22(1):3–28.

Tillman, K. A. and Barner, D. (2015). Learning the language of time: Childrens acquisition of duration words. Cognitive psychology, 78:57–77.

Tillman, K. A., Marghetis, T., Barner, D., and Srinivasan, M. (2017). Today is tomorrows yesterday: Childrens acquisition of deictic time words. Cognitive psychology, 92:87–100.

Tincoff, R. and Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. Psychological Science, 10(2):172–175.

Tincoff, R. and Jusczyk, P. W. (2012). Six-month-olds comprehend words that refer to parts of the body. Infancy, 17(4):432–444.

Tomasello, M. (2000). Do young children have adult syntactic competence? Cognition, 74(3):209–253.

Tomasello, M. (2003). Constructing a language: A usage-based approach to child language acquisition.

Tomasello, M. (2010). Origins of human communication. MIT press.

Tomasello, M. and Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. Monographs of the Society for Research in Child Development, 59(5):174–179.

Trueswell, J. C., Medina, T. N., Hafri, A., and Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. Cognitive psychology, 66(1):126–156.

Tsao, F.-M., Liu, H.-M., and Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. Child development, 75(4):1067–1084.

Valian, V. (1986). Syntactic categories in the speech of young children. Developmental psychology, 22(4):562.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., and Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. Proceedings of the National Academy of Sciences, 113(23):6454–6459.

Vlach, H. A. and Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. Cognition, 127(3):375–382.

Volterra, V. and Caselli, M. C. (1985). From gestures and vocalizations to signs and words. Sign language research, 83.

Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. Harvard university press.

Wagner, K., Dobkins, K., and Barner, D. (2013). Slow mapping: Color word learning as a gradual inductive process. Cognition, 127(3):307–317.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behav Res Meth, 45(4):1191–1207.

WEBER, A. M., MARCHMAN, V. A., Yatma, D., and FERNALD, A. (2018). Validity of caregiver-report measures of language skill for wolof-learning infants and toddlers living in rural african villages. Journal of child language, pages 1–20.

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

Werner, H. and Kaplan, B. (1963). Symbol formation: An organismic-developmental approach to language and the expression of thought.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. Lingua, 106(1-4):23–79.

Wichmann, S., Holman, E. W., Bakker, D., and Brown, C. H. (2010). Evaluating linguistic distance measures. Physica A, 389(17):3632–3639.

Yang, C. D. (2002). Knowledge and learning in natural language. Oxford University Press on Demand.

Yang, C. D. (2004). Universal Grammar, statistics or both? Trends in Cognitive Sciences, 8(10):451–456.

Yang, C. D. (2013). Ontogeny and phylogeny of language. Proceedings of the National Academy of Sciences, 110(16):6324–6327.

Yang, C. D. (2016). The price of linguistic productivity: How children learn to break the rules of language. MIT Press.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. Neurocomputing, 70(13):2149–2165.

Yu, C. and Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. Psychological Science, 18(5):414–420.

Yurovsky, D. and Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. Cognition, 145:53–62.

Yurovsky, D., Smith, L. B., and Yu, C. (2013). Statistical word learning at scale: The baby's view is better. Developmental science, 16(6):959–966.

Zipf, G. K. (1935). The psycho-biology of language.