UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Representing linguistic knowledge with probabilistic models

Permalink
https://escholarship.org/uc/item/5vp920sn

Author
Meylan, Stephan Charles

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Representing linguistic knowledge with probabilistic models

By

Stephan C. Meylan

A Dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas L. Griffiths, Chair
Professor Terry Regier
Professor Michael C. Frank

Summer 2018

©2018 – Stephan C. Meylan
all rights reserved.

Abstract

Representing linguistic knowledge with probabilistic models

by

Stephan C. Meylan

Doctor of Philosophy in Psychology

University of California, Berkeley

Thomas L. Griffiths, Chair

The use of language is one of the defining features of human cognition. Focusing here on two key features of language, *productivity* and *robustness*, I examine how basic questions regarding linguistic representation can be approached with the help of probabilistic generative language models, or PGLMs. These statistical models, which capture aspects of linguistic structure in terms of distributions over events, can serve as both the product of language learning and as prior knowledge in real-time language processing. In the first two chapters, I show how PGLMs can be used to make inferences about the nature of people's linguistic representations. In Chapter 1, I look at the representations of language learners, tracing the earliest evidence for a noun category in large developmental corpora. In Chapter 2, I evaluate broad-coverage language models reflecting contrasting assumptions about the information sources and abstractions used for in-context spoken word recognition in their ability to capture people's behavior in a large online game of "Telephone." In Chapter 3, I show how these models can be used to examine the properties of lexicons. I use a measure derived from a probabilistic generative model of word structure to provide a novel interpretation of a longstanding linguistic universal, motivating it in terms of cognitive pressures that arise from communication. I conclude by considering the prospects for a unified, expectations-oriented account of language processing and first language learning.

This dissertation is dedicated to the first ape that dared to speak.

Contents

Acknowledgments

I would first like to acknowledge the formative role of many teachers over the years, insofar as writing a dissertation calls heavily on skills and interests cultivated decades earlier: Mal Ellenberg, Emerson Littlefield, Richard Beaton, Charla Gaglio, Richard Moore, David Fields, and Theresa Andrés. For my early language education at Brown, I would like to thank Drs. Jim Morgan, Phil Lieberman, Mark Johnson, Katherine Demuth, and Julie Sedivy. I remember a particular episode from undergrad where I asked Jim a question in class about language acquisition, and he responded by cursing and saying I should go and try to find out the answer. I can't recall what the question was, but at least now I've tried for the better part of a decade.

Many thanks to graduate students and postdocs at Berkeley for being interlocutors, coauthors, co-conspirators, and friends especially David Bourgin, Ruthe Foushee, Josh Peterson, Jess Hamrick, Thomas Langlois, Rachel Jansen, Josh Abbott, Aida Nematzadeh, Luke Maurits, Anna Rafferty, Mike Pacer, and Jordan Suchow. I would like to express additional appreciation to the final two individuals for creating Dissertate, the typesetting system used for this dissertation. For my research assistants, Sissi Wang, Teeranan Pokaprakarn, Neha Dabke, Naomi Jing, and Sathvik Nair, I have infinite gratitude that you all helped implement and conduct experiments and cheerfully received half-baked first-pass theoretical "ideas."

If any scientist can see over mountains, it is because they stand on a mountain of software written by people they have never met. Not to claim that this volume entails seeing over much more than foothills, I would nonetheless like to thank Hadley Wickham (`ggplot2`, used to generate graphs in all chapters), the Jupyter Notebook team (the analyses in all chapters), and Graham Dumpleton (`mod_wsgi`, key to running the Telephone experiment at scale). To think that this dissertation is one of tens of thousands that they contributed to is truly humbling.

Special thanks to my advisor, Tom Griffiths, and my dissertation committee, Mahesh Srinivasan, Terry Regier, and Michael Frank, for helping me develop these ideas, conduct these experiments, analyze data, and repeat for the course of many years.

A final thanks to my family (and my adoptive college-and-since family, Sam Tarakajian) for support and firm insistence that I return to the real world when I think a little too hard about language. And an especially strong thanks to my partner, Lindsay Berkowitz, for regarding my "other girlfriend," language, with minimal jealousy. Lindsay, I love your vowels.

Language is the House of Being.

In its home human beings dwell.

Martin Heidegger

0

Introduction

The ability to use natural language is one of the hallmarks of human cognition. Shared—at least in its totality—by no other species, our language abilities have played a central role in enabling new vectors of cultural transmission, supporting tool making, writing, trade, and science as we know it. Language serves as a crucial medium of thought (Whorf, 1940) and vital means of artistic expression. And beyond the obvious downstream applications to health and technology, to understand how

humans understand and use language is to understand one of the most remarkable evolutionary systems in existence.

Beyond the long-standing philosophical and literary inquiries into the nature of language, the 19th and 20th centuries saw the emergence of a discrete, explicitly scientific endeavor to formally characterize the structure of languages (Grimm, 1819; von Humboldt, 1836; Gabelentz, 1901), as well as to understand the psychological mechanisms that support its use (Kantor, 1936). This includes extensive treatment in the fields of linguistics, psychology, cognitive science, linguistic anthropology, neuroscience, and, increasingly, computer science. These endeavors can be separated into two broad methodological approaches, focusing either on understanding (and in many cases imitating) the *mechanisms* of language use at the level of individuals, or on characterizing the structural regularities that emerge at scale in such communicative systems. These approaches roughly correspond to the fields of neuroscience and psychology (including psycholinguistics and child language acquisition) on the one hand, and the field of linguistics on the other. These twin modes of inquiry meet at the locus of *linguistic representations*, or the knowledge that individuals store about the language(s) they encounter. These representations are simultaneously the output of the process of language learning, and serve as the input for real-time language processing (Bresnan, 1986; Bybee, 2006). In this dissertation, I focus on the form, function, and ontogeny of these representations.

To constrain an otherwise overwhelming space of questions about the nature of these linguistic representations, I focus specifically on how they relate to two critical properties of natural language. The first of these properties is *productivity*, or the combinatorial potential of language structure that provides for the "infinite use of finite means" characteristic of human languages (von Humboldt,

1836). While all people observe a finite quantity of linguistic input, most infer a rich system of linguistic regularities (phonological, morphological, syntactic, semantic, and pragmatic) that allow them to express novel messages that they have never themselves heard previously, and to interpret novel utterances, often from new speakers. The second of these properties is *robustness*: language users show a remarkable ability to communicate despite high levels of perceptual uncertainty and environmental noise (Shannon, 1948, 1951). That spoken communication occurs under noisy conditions is increasingly recognized as the rule rather than a special case (Gibson et al., 2013).

The two properties above highlight the central role of inference in language use: Any language user must infer a complex system of abstractions across many levels of linguistic structure to be able to express new ideas (productivity) and develop expectations regarding what other speakers are likely to say (robustness). Inference thus occurs at multiple timescales: listeners use what they know about language structure to infer a speaker's message in the moment, then on the basis of such experiences iteratively revise their knowledge. While the composition of this ever-changing store of knowledge depends in part on the specific linguistic experience of the individual, it also implicates representations, or the manner in which this knowledge is encoded. The nature and ontogeny of these representations—what is inferred, how it is inferred, when it is inferred, and how it is subsequently used—is thus of utmost importance in understanding the human language faculty, the regularities attested in the world's languages, and the complex relationship between the two.

The effort to uncover the representations implicated in human language use is by no means a new endeavor; indeed, many of the central debates in language research revolve around theoretical disagreements over the nature of representations. Generative accounts of language structure (e.g.,

Chomsky, 1957) emerged in opposition to preceding behaviorist accounts (Skinner, 1957), with the criticism that the latter attributed insufficient representations to people, such that the theories could not account for the richness of the structure observed in human languages. Subsequent usage-based theories (Tomasello, 1992) responded in turn by criticizing the latter for unlearnable representational complexity.

Recent advances have introduced a possible synthesis: learning occurs over a hypothesis space of sophisticated, generative representations. Of particular interest for characterizing this kind of learned knowledge are *probabilistic generative models*, or statistical models that characterize how the data observable to an agent may be generated (Pearl, 1986; Jordan, 1998; Tenenbaum et al., 2011). In many cases these models posit probability distributions over latent, unobservable variables in addition to observable data, for example positing a syntactic parse tree (or distribution over such parse trees) to characterize a sentence heard by a listener.

Probabilistic generative models have proven especially useful for language-related engineering applications such as speech recognition and speech synthesis (Manning & Schütze, 1999). In these cases, their success can be attributed to the fact that they can be trained or fit with huge quantities of linguistic data, both annotated and unannotated. These probabilistic generative language models (henceforth PGLMs) have found increasing purchase in psycholinguistics, where they can be used to derive quantitative predictions regarding processing difficulty; conversely, measures of processing difficulty can be used to evaluate evidence for the representational commitments of these models. Here I concern myself with neither neural plausibility, nor more broadly the ways in which these representations may by learned by people. Instead, following the argument stressing the importance

of understanding cognition at various levels of analysis in (Marr, 1982), I seek explanations at the *computational* level, focusing on the general computational problems that need to be solved and the information required to solve them. In the case of language, it is sufficiently challenging to rigorously define the problems and find appropriate representations in the intersection of learnable and useful.

This dissertation consists of three sections. First, in *The Emergence of Syntactic Productivity*, I seek to answer the question of *when* children first use a key syntactic representation in early language development. In this section I present a model-based method to evaluate support for the existence of a noun category in English-learning children's early use of determiners ("a" and "the"). This model includes at its core a probabilistic generative model of language structure.

Second, in *Evaluating Models of Robust Word Recognition with Serial Reproduction*, I develop a method to evaluate PGLMs in their ability to capture human behavior in a spoken word recognition task. The structure of this task, in which each of a set of utterances is reproduced by a chain of participants as in the game of "Telephone," yields a dataset that is increasingly representative of participants' linguistic expectations. I focus particularly on two outstanding questions regarding information sources in spoken word recognition: whether people form expectations on the basis of preceding linguistic context, and if so, whether people form abstract representations of that context.

Finally, in *Wordforms—Not Just Their Lengths—Are Optimized for Efficient Communication* I examine how the phonological form, or specific sound sequence, of a word is shaped by the need for efficient production and robust word recognition. In this section, wordforms are characterized in terms of their probability under a PGLM that captures a language's phonotactics. I conclude by

reflecting on the prospects of collapsing the traditional distinctions between language processing and language acquisition in favor of an inference / expectations-oriented account, with an emphasis on the explanatory virtues of PGLMs.

James IV of Scotland was said to have sent two children
to be raised by a mute woman isolated on the island
of Inchkeith, to determine if language was learned or
innate.
The children were reported to have spoken good Hebrew.

Wikipedia, "Language Deprivation Experiments"

1

The Emergence of an Abstract Grammatical Category in Children's Early Speech

One of the distinguishing features of natural languages is the use of rich hierarchical structures comprised of words—syntax—to express ideas. By using a complex system of categories at a level of abstraction higher than words, language users are able to comprehend and produce combinations

of words they have never heard or seen before. While treatments of these categories and their relations vary by the particular grammatical theory, the consensus is that the adult grammar consists of groups of words that have similar behaviors, such as nouns, verbs, adjectives, and prepositions, as well as super-ordinate groups of words like noun phrases and verb phrases. But when are these grammatical categories, vital for the productive use of language, first available? In this chapter, I examine how children's earliest use of indefinite and definite determiners—'a' and 'the'—can be used to evaluate the availability of an abstract grammatical category, and hence productivity.

The timeline of the availability of these representations—and by extension their ontogeny—has long been contested in language research. Full productivity accounts assert that these categories are available to children from the beginning of language development, and may guide their earliest inferences about language (Valian, 1986; Valian et al., 2009; Yang, 2010, 2013). Item-based accounts, by contrast, posit that children have notably limited grammatical productivity and are instead largely restricted to reproducing the exact constructions they hear from caregivers in the earliest phases of language production (Pine & Lieven, 1997; Tomasello, 1992). Much of this research has focused on the case of a subset of determiners, the English articles 'a', 'an', and 'the,' whose use in novel contexts provides gold-standard evidence of the existence of a productive noun category. Experimental evidence suggests that children can produce truly novel determiner-noun combinations as young as 2;5. (Tomasello & Olguin, 1993). However, knowledge of this structural regularity may be available earlier yet: in-lab experimental methods may underestimate children's knowledge in that they entail a higher cognitive burden than everyday language use in the home. Corpus-based research in the preceding period (the onset of multiword speech until 2;5) has produced contrasting results regard-

ing early productivity: Valian et al. (2009) and Yang (2013) found evidence of full productivity while (Pine et al., 2013) found evidence of the opposite.

In this section, my coauthors and I use a probabilistic generative model to evaluate the degree of information sharing across nouns implicit in the observed combinatorial behavior in children's article+noun productions. This model includes as components two simple generative language models that reflect the determiner+article usage preferences of a child and their corresponding caregiver. While these models are presented here as components of a larger model for data analysis, the core beta-binomial model (p. 17) can be interpreted as the special case of a fragment of another well-known PGML, a lexicalized probabilistic context-free grammar (PCFG). In fact, generalizing from the two-article case to a multi-article case, and using a Dirichlet prior (i.e. generalizing the model from a beta-binomial to a Dirichlet-multinomial model) is equivalent to a shared Dirichlet prior over the re-write rules for noun phrases in a lexicalized PCFG, a strategy similar to one that has been previously explored for grammar induction in computational lingusitics (Johnson et al., 2007). For a PCFG, this would mean that the expansion rules for a nonterminal noun phrase specific to its noun head would depend on a mixture of specific experience with that noun, as well as shared experience with other nouns. In this sense, the model presented here is similar to the probabilistic generative models of language used elsewhere in this dissertation.

The larger data-analytic model embeds two instances of this probabilistic generative model of language into a model of a larger stochastic process in which only a portion of utterances made by children and caregivers in the time period in question are observed by researchers. This sponsors a notable advance in that the analytic model can make use of both parental input and child produc-

tions from large corpora of child-available speech (several corpora from CHILDES and the Speechome corpus) in order to infer key parameters of interest. This Bayesian model-based approach provides a continuous metric of grammatical productivity appropriate for characterizing children's earliest productions, and provides a principled means of quantifying uncertainty in the estimate of the child's level of productivity. The use of parental language usage as input allows the model to partial out the direct contributions of parental input, or what the child might be copying directly. One of the key parameters of the fit model then maps to a gradient that includes the existing, contrasting theoretical positions as quantitative endpoints: idealized full-productivity and zero-productivity item-based learners are the two endpoints for this continuum. This allows for an evaluation of the two hypotheses—as well as the space of all intermediate ones— for the speech speech samples from different developmental intervals, both to evaluate the absolute levels and to characterize the change over time. This provides a promising new means of tracing the change in linguistic representations over the course of language development using naturalistic corpora.

This chapter originally appeared as an article in *Psychological Science* in 2017 (Meylan et al., 2017). The use of "we" in this chapter refers to myself and co-authors: Brandon C. Roy (MIT Media Lab), Michael C. Frank (Stanford University), and Roger P. Levy (MIT). I would also like to thank Charles Yang for discussion of his model and data preparation, Steven Piantadosi (University of Rochester) for initial discussions, and to the members of the Language and Cognition Lab at Stanford and the Computational Cognitive Science Lab at U.C. Berkeley for valuable feedback. This work was supported by a National Science Foundation Graduate Research Fellowship to S.C.M. under Grant No. DGE-1106400. R.L. was also supported by Alfred P. Sloan Research Fellowship

FG-BR2012-30 and from a fellowship at the Center for Advanced Study in the Behavioral Sciences.

1.1 Introduction

One of the most astonishing parts of children's language acquisition is the emergence of the ability to say and understand things that they have never heard before. This ability, known as *productivity*, is a hallmark of human language (von Humboldt, 1836; Hockett, 1959). Indeed, adults' linguistic representations are almost universally described in terms of syntactic abstractions such as "determiner," "verb," and "noun phrase" (e.g., Chomsky, 1981; Sag et al., 1999). But do these same adult-like abstractions guide how children produce and comprehend language?

Some researchers have suggested a *generativist* view of syntactic acquisition: adult-like abstractions guide children's comprehension and production from as early as it can be measured (Pinker, 1984; Valian, 1986; Yang, 2013). Others have argued that adult-like syntactic categories—or at least their guiding role in behavior—emerge gradually, with the accumulation of experience. On such *constructivist* views, children's representations progress over time from memorized multi-word expressions to specific item-based constructions and eventually generalize to abstract combinatorial rules (Braine, 1976; Pine & Martindale, 1996; Pine & Lieven, 1997; Tomasello, 2003).

Here we focus on a key case study for this debate: the emergence of the capacity in English to produce a noun phrase (NP) by combining a determiner (Det, such as "the" or "a") with a noun (N). This capacity is exemplified for adult English by the context-free rule

$$\text{NP} \rightarrow \text{Det N}$$

Using this knowledge, when adult native English speakers hear a novel count noun with "a," e.g. "a blicket," they know that combining the novel noun with "the" will also produce a permissible

noun phrase, e.g. "the blicket." Adult-like knowledge and use of this part of English syntax requires the category ***Det***, the category ***N***, and the rule specifying how they are combined.

In recent years, this case study—noun phrase productivity, with a focus on the use of determiners—has played an increasingly prominent role in the generativist–constructivist child language acquisition debate (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009; Yang, 2013; Pine et al., 2013). Whereas nouns often have referents in the child's environment, the semantic contribution of determiners to utterance meaning is more subtle (Fenson et al., 1994; Tardif et al., 2008). Thus one might expect determiners to be learned late (Valian et al., 2009). Yet children produce them relatively early, and their uses are overwhelmingly correct by the standards of adult grammar. Is this because children deploy adult-like syntactic knowledge, or because they memorize and reuse specific noun phrases, creating the illusion of full productivity?

Experimental methods have been of limited utility in resolving this question. Tomasello & Olguin (1993) found evidence for the presence of a noun-like productive object word category in children between 20 and 26 months, presenting objects with nonce labels and eliciting reuse in novel syntactic contexts and morphological forms (using a "wug" test; Berko, 1958). But these data do not resolve the extent to which syntactic abstractions guide children's everyday speech. Instead, most work on early syntactic productivity has relied on observational language samples (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009; Yang, 2013; Pine et al., 2013).

Making inferences about children's knowledge from observational evidence is difficult for a number of reasons, however. First, individual child language corpora have typically been small—consisting of weekly or monthly recordings of only a couple of hours. Second, nouns (like other

words) follow a Zipfian frequency distribution (Zipf, 1935), in which a small number of words are heard often, but most are heard only a handful of times. As a result, evidence regarding the range of syntactic contexts in which a given child uses an individual noun is weak for most nouns (Yang, 2013). These inferential challenges are sufficiently severe that within the past several years, researchers on opposing sides of the productivity debate have drawn opposite conclusions from similar datasets (Pine et al., 2013; Yang, 2013). Making progress on children's syntactic productivity requires overcoming these challenges.

Here we present a new, model-based framework for drawing inferences about syntactic productivity, differing from previous work in two critical respects. First, previous approaches assessed productivity via a summary statistic, the overlap score, computed from a child language sample. This statistic is difficult to interpret because it may be biased by the size and composition of the sample (discussed below). Here, in contrast, we model productivity as one feature of a model of child language whose parameters can be estimated from a sample and whose overall fit to the data can be assessed. Second, we explicitly model item-based memorization and reuse of specific determiner–noun pairs from caregiver speech in the child's environment as an additional contributor to child language production alongside syntactic productivity. Our framework encodes a continuum of hypotheses ranging between fully productive and fully item-based, and allows us to assess how a child at any given point in development balances these two knowledge sources in their production of determiner–noun combinations.

We apply this model to a wide range of longitudinal corpora of child speech, including the Speechome Corpus (Roy et al., 2015), a new high-density set of recordings of one child's early input and

productions. Our model reveals that many of the conventional corpora analyzed in previous research (Valian, 1986; Pine & Martindale, 1996; Pine & Lieven, 1997; Valian et al., 2009; Pine et al., 2013) are too small to draw high-confidence inferences. An exploratory analysis of the Speechome data, both denser and from earlier in development, provides evidence for low initial levels of productivity followed by a rapid increase starting around 24 months. Several other datasets provide corroboratory evidence. Contra full-productivity accounts, syntactic productivity is very low in the first months of determiner use in these datasets. At the same time, the current work constrains the timeline of constructivist accounts. We find a rapid early increase in productivity—in Speechome this increase occurs within a few months of the onset of combinatorial speech, prior to the beginning of many of the datasets that have been used previously to address this question. We conclude by discussing the need for denser datasets to provide conclusive evidence on questions about the roots of syntactic abstraction.

1.2 Background

Previous investigations have focused on the overlap score, a summary statistic of productivity (Pine & Martindale, 1996; Pine & Lieven, 1997; Pine et al., 2013). Overlap is calculated from the distribution of determiner–noun pairings in a sample, as the proportion of nouns that appear with both "a" and "the" out of the total number of nouns used with either. While initial investigations suggested that young children use comparatively fewer nouns with both determiners (Pine & Martindale, 1996; Pine & Lieven, 1997), overlap scores are highly dependent on sample size due to the Zipfian dis-

tribution of noun frequencies (Valian et al., 2009; Yang, 2013). In addition, this statistic is not well-suited for distinguishing increases in direct experience—greater exposure to the relevant words (e.g., hearing both "the dog" and "a dog" independently and subsequently repeating these, even without abstraction)—from true changes in grammatical productivity (Valian et al., 2009; Yang, 2013).

Two recent investigations have used more sophisticated techniques to address issues of sample size. Yang (2013) constructed a null-hypothesis "full productivity" model in which each noun has the same distribution over determiner pairings (no item-specific preferences) and showed that it predicted overlap score well for six children in the CHILDES database. Pine et al. (2013), in contrast, developed a noun-controlled method for comparing adult and child productivity scores in a given sample, and rejected a full-productivity null hypothesis. Neither of these methods, however, is well suited to tracking developmental changes in productivity, because of their focus on the overlap score. If item-based knowledge plays a role in children's productions, overlap might increase over time even without any changes in productivity, simply because children have heard more determiner+noun pairs.

Here we take a fundamentally different approach from previous work, to address the challenge of decoupling genuinely productive behavior from what might be expected on the basis of experience. We proceed from the observation that there are two sources of information by which a speaker could know that a particular determiner–noun pair belongs to English, and thus potentially produce it: (1) direct experience with that specific determiner–noun pair and (2) a productive inference using knowledge abstracted from experience with different determiner–noun pairs (and perhaps other input as well). Measuring a given speaker's productivity from corpus data requires assessing

the extent to which the speaker's language use reflects productivity above and beyond what can be attributed to direct experience.

We define a probabilistic model of determiner+noun production that considers both knowledge sources. In our model, the contribution of productive knowledge can range along a continuum from none (a child capable only of imitating caregiver input, like an idealized version of an "island" learner as described in Tomasello, 1992) to complete (a "total generalizer" equivalent to the null-hypothesis model in Yang, 2013). Specific model parameters correspond to the contributions of these two information sources, and we use Bayesian inference to infer likely values of these parameters for a corpus sample given both the child's determiner–noun productions and caregiver input. By comparing temporally successive samples for a given child, we can use this model to estimate the child's change in syntactic productivity over time. Because our model is fully Bayesian, we are also able to estimate the level of certainty in our estimates, critically allowing us to avoid overly confident inferences when data are too sparse.

1.3 Methods

1.3.1 Model

We model the use of each noun token with a specific determiner as the output of a probabilistic generative process. We assume that each noun has its own determiner preference ranging from 0 (a noun used only with "a") to 1 (a noun used only with "the"). We then explicitly model cross-noun variability by assuming some underlying distribution of determiner preferences across all nouns.

Lower cross-noun variability indicates that nouns behave in a more class-like fashion, while higher variability indicates little generalization of determiner use across nouns.

Formally, each noun type can be thought of as a coin whose weight corresponds to its determiner preference. Each use of that noun type with a determiner is thus analogous to the flip of that weighted coin, where heads indicate the use of the definite determiner and tails correspond to the indefinite. A sequence of noun uses are thus draws from a binomial distribution with success parameter μ corresponding to the determiner preference. The determiner preference for each noun is drawn from a beta distribution with mean μ_0 (the underlying "average" preference across all nouns) and scale ν, giving us a *hierarchical beta-binomial* model (Gelman et al., 2004).*

Under this model, a child's determiner productions for each noun she uses are guided by a combination of the two information sources mentioned above—(1) direct experience, and (2) productive knowledge—and the strength of each information source's contribution to the child's productions is determined by a weighting parameter. For (1), a parameter η determines how effectively the child learns from noun-specific determiner productions in its linguistic input; for (2), a parameter ν determines how strongly the child applies productive knowledge of determiner use across *all* nouns. These parameters η and ν do not trade off against each other, but rather play complementary roles in accounting for a child's productions: As η increases, the variability across nouns in a child's determiner productions can more closely match the variability in her input, while as ν increases, the child is increasingly able to produce determiner–noun pairs for which she has not received sufficient

*Many readers may be more familiar with the more common parameterization of the beta distribution in terms of shape parameters $\alpha = \mu\nu$ and $\beta = (1 - \mu)\nu$.

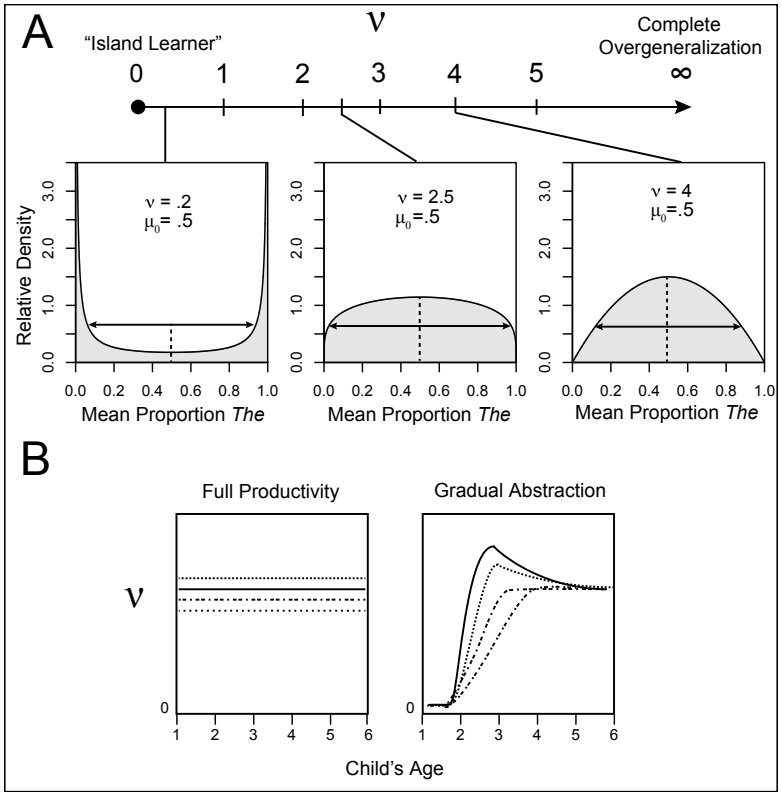


Figure 1.1: A: Interpretation of the ν parameter, a concise metric of grammatical productivity. At low values of ν, little or no information is shared between nouns. At higher ν values, nouns exhibit more consistent usage as a class, indicating the existence of a productive rule governing the combination of determiners and nouns. μ_o represents the mean proportion of definite determiner usage across nouns, set here at .5 in all three panels. **B**: Schematized trajectories for the development of grammatical productivity under two competing theories.

evidence from caregiver input.

At the heart of the model are the contributions of direct experience and productive knowledge. Both contribute to the rate at which a child uses "the" as opposed to "a" for each noun. This rate, μ_i, is taken to be beta-distributed and corresponds to a beta-binomial Bayesian update of a prior of mean μ_o and concentration ν with count data corresponding to the caregiver input, weighted by a factor of η. Thus, larger ν indicates stronger influence from the child's productive knowledge, while

larger η indicates that the child learns the noun-specific nuances of caregiver input more effectively. For more details see Appendix A; Our complete hierarchical Bayesian model and variable definitions are presented in Fig. A1.

Since we lack exhaustive recordings of caregiver input, we treat unrecorded caregiver input as a latent variable drawn from the same distribution as aggregated caregiver input and infer it jointly with model parameters (see Appendix A: Details of the Imputation). The theoretically critical target of inference is ν, the strength of the child's productive knowledge of determiner–noun combinatorial potential, which can range from $\nu = 0$ (an extreme "island" learner whose determiner preference for a given noun is guided exclusively by its direct experience with that noun, and whose noun-specific determiner preferences are likely to be skewed toward 0 or 1) to ν approaching infinity (an extreme over-generalizer who has identical determiner preference for all nouns, Fig. 1.1A).

We use Markov chain Monte Carlo to infer confidence intervals over η, μ_0, and ν from a child's recorded productions and linguistic input. But a single recording of a child typically does not yield high confidence in these estimates because of the relatively low numbers of productions for individual nouns. To overcome this issue, we use two different methods for constructing sufficiently large samples of child and caregiver tokens to evaluate the developmental trajectory of the ν parameter: split-half and sliding window analyses.

First, in the *split-half* analysis, we divide the data for each child into distinct early and late time windows with an equal number of tokens, denoted with the subscripts t_1 and t_2. Separate parameter sets (μ, ν, η) are maintained for the first and second windows; for a given sample from the joint posterior, the changes in parameters from the first window to the second can be calculated as:

$$\Delta\nu = \nu_{t_2} - \nu_{t_1}, \qquad \Delta\mu = \mu_{t_2} - \mu_{t_1}, \qquad \Delta\eta = \eta_{t_2} - \eta_{t_1}. \tag{1.1}$$

These variables may be treated as targets of inference, over which highest posterior density (HPD) intervals may be computed. Our principal target of inference is $\Delta\nu$, the change in the contribution of productive knowledge to the child's determiner use. This two-window approach maximizes statistical power, but does so at the expense of a detailed time-related trajectory: for those children with longer periods of coverage, this estimate may group together several distinct developmental time periods.

Second, as an exploratory technique, we also use our model to measure finer-grained changes in parameter estimates across development via a *sliding window* approach in which the model is fit to successively later subsections of the corpus of child productions. Each window also includes the corresponding adult productions that occur prior to or during that subsection. In this case we fit the model to successive 1024 token windows of the child's speech, advancing by 256 tokens for each sample. This method yields a higher resolution timecourse than the split-half analysis, though at the expense of less-constrained parameter estimates, especially for the smaller corpora. For more details on inferring model parameters, see Appendix A: Model Fitting Procedure.

Our approach is an example of "Bayesian data analysis" (Gelman et al., 2004). We create a cognitively interpretable model that captures the spectrum of different hypotheses, from item-based learning to full productivity. We can then infer, for a particular dataset, where on the spectrum the

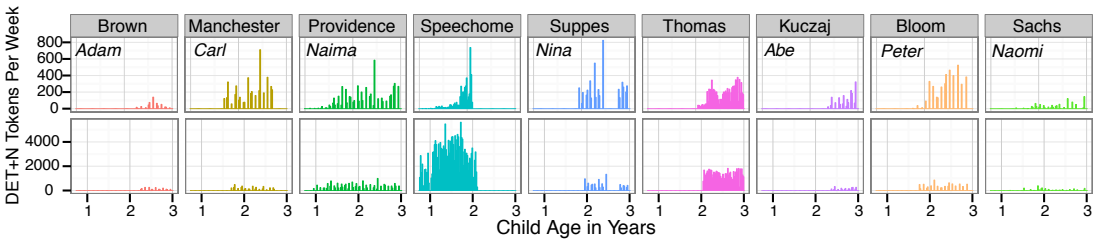


Figure 1.2: Number of recorded determiner+noun uses per week for children (top) and corresponding caregivers (bottom) before three years of age for the child with the most determiner+noun pairs for each of the nine corpora analyzed here. Note that the number of weekly observations from children and caregivers are presented on differing scales (0-800 and 0-6,000 respectively).

data fall. In a classic predictive model, parameters are fit–or overfit—to some external performance standard. In contrast, our model summarizes a particular aspect of the dataset and gives an estimate of the relative certainty we have in this summary measurement.

1.3.2 Data

We used a large set of publicly-available longitudinal developmental corpora of recordings of children and their caregivers from the CHILDES archive (MacWhinney, 2000). Four of these corpora have been examined previously for early evidence of grammatical productivity: the Providence Corpus (Demuth et al., 2006), the Manchester Corpus (Theakston et al., 2001), the Brown Corpus (Brown, 1973), and the Sachs Corpus (Sachs, 1983). We additionally analyze four single-child corpora: Bloom (Bloom et al., 1974), Kuczaj (Kuczaj, 1977), Suppes (Suppes, 1974), and Thomas (Lieven et al., 2009). These eight corpora yield usable data for a total of 26 children. While high-density data with rich annotations exist for all of these corpora, coverage starts in most cases well after the onset of combinatorial speech and is sparse under two years of age, the time interval necessary for characteriz-

ing initial levels of grammatical productivity.

To address these shortcomings, we additionally analyze the densest longitudinal developmental corpus in existence, the Speechome Corpus (Roy et al., 2015). The Speechome Corpus covers the period 9 through 25 months in the life of a single child, and contains video and audio recordings of nearly 70% of the child's waking hours, with transcripts for a substantial portion of these (Vosoughi & Roy, 2012). While transcription of the Speechome Corpus is a work in progress, the version used here contains approximately 4,300 noun phrases with articles produced by the child before 25 months of age and includes dense coverage of child-accessible caregiver speech, with some 196,300 noun phrases in the same time period. The Speechome Corpus supports more detailed inferences about developmental timecourse in the second year of life. The Speechome Corpus is also distinguished in the quantity of child-available adult speech, with nearly 80% more caregiver tokens than the next best-represented child, Thomas. Fig. 1.2 shows comparative densities for adult and child determiner+noun pairs for the child with the most data in each corpus.[†]

We assess our model using seven different methods for extracting determiner+noun data from each corpus. These data treatments reflect a range of assumptions regarding the availability of phrase structure for identifying which noun corresponds to each determiner, whether information can be shared between morphologically inflected forms, and whether the child is considering only singular forms in the language. In the absence of reliable morphological tags, the Thomas and Speechome

[†]Datasets that capture large amounts of an individual family's experience like Speechome pose unique privacy risks. In order to share reproducible data while maintaining privacy, we are distributing determiner+noun count data from the Speechome corpus while obfuscating the identities of the specific nouns the child produced.

corpora were assessed on four data treatments each. For additional technical details refer to Appendix A: Data Preparation. We have made available model code, noun-anonymized Speechome data, and auxiliary code necessary to reproduce our research in a public GitHub repository accessible at ***https://github.com/smeylan/determiner_learning***.

1.4 Results

The two hypotheses represented in the literature—full productivity or gradual abstraction over item-based knowledge (Fig. 1.1B)—make contrasting predictions regarding initial productivity and the effects of developmental change. Full productivity predicts a nonzero initial level combined with a negligible effect of developmental time—productivity does not increase with exposure to more data. Gradual abstraction over item-based knowledge, in contrast, predicts near-zero initial productivity indicating the absence of syntactic category knowledge in the earliest productions, and a positive relationship with developmental time corresponding to the gradual induction of abstract categories throughout childhood.

1.4.1 Split-Half Method

To test for changes in productivity, we assess the null hypothesis that 0 (no change) is within the 99.9% HPD interval for the posterior estimate of $\Delta\nu$, the difference in ν estimates between the first and second half of tokens each child. (We use the 99.9% criterion because of the large number of independent comparisons implied by this analysis—one for each of the 27 children.) By this standard, only one child (Speechome, in 3 of 4 data treatments) shows a significant increase in produc-

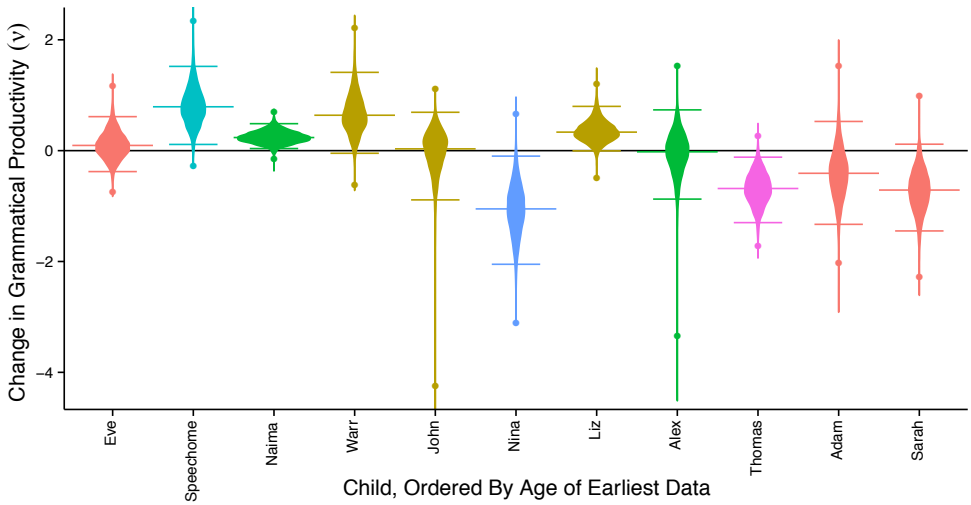


Figure 1.3: Posterior estimates for change in productivity between first and second half of children's corpora ($\Delta\nu$) for each child (n = 11). Longest horizontal lines indicate the median of the posterior, and shorter horizontal lines the 95% HPD. Points indicate the 99.9% HPD. The remainder of the children (n=16) are not displayed on the basis of poorly constrained posteriors (99.9% HPD for ν outside [0,3] for either time period).

tivity (Fig. 1.3 and Fig. 1.4). The remaining data treatment for Speechome is strictly positive within the 95% HPD interval. Three other children—Liz (in 3 of 7 data treatments), Naima (1 of 7 treatments) and Warr (1 of 7 treatments)—have at least one data treatment where change is strictly positive within the 95% HPD. These findings suggest some early increases in productivity. Results across all seven data preparations are presented in Figure S4.

We also found apparent *decreases* in grammatical productivity for several of the older children. Thomas (2 of 4 treatments within the 99.9% HPD interval, 1 in the 95% HPD interval), Sarah (1 in the 99.9% and 4 in the 95% HPD interval), and Nina (2 in the 99.9% and 2 in the 95 % HPD interval) show strictly negative changes. The timing of these decreases are consistent with a phase of overregularization, during which they are more willing to use determiner noun-combinations that are rare

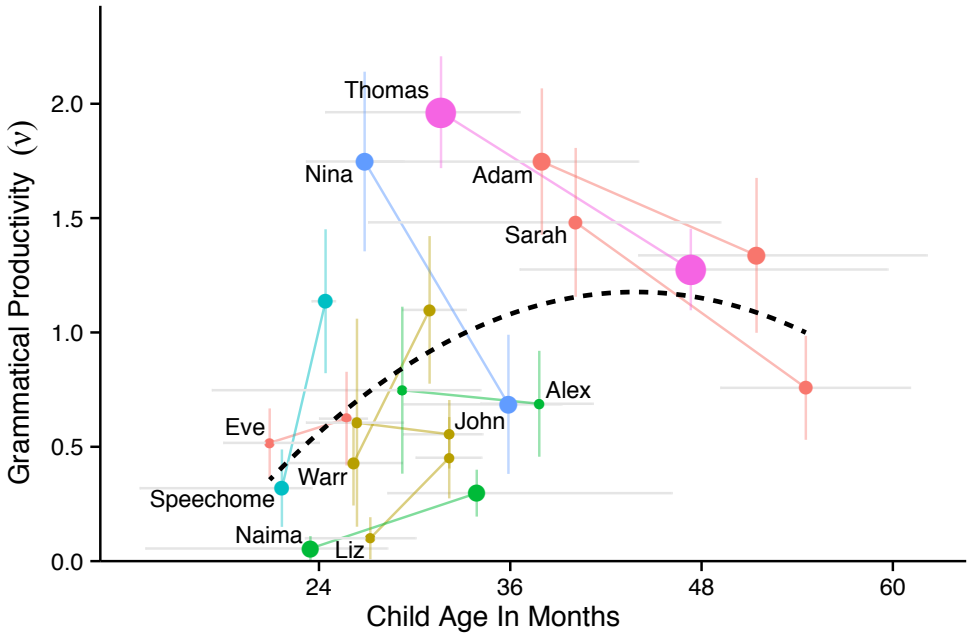


Figure 1.4: The inferred developmental trajectory for determiner productivity, γ, across children (n = 11). Each line shows a two-point productivity trajectory for a single child, plotted by age in months. Marker size corresponds to the number of child tokens used for each child. Gray horizontal lines indicate the temporal extent of the tokens used to parameterize the model at each point; vertical lines indicate the SD of the posterior. The best fitting quadratic trend is shown as a dashed black line.

or unattested in adult speech like *a sky*, followed by a decrease towards adult-like levels. Consistent with this hypothesis, increases in ν tended to occur in datasets from younger children (p=0.009 by rank sum test on the children in Fig. 1.3).

Together these results are broadly consistent with constructivist hypotheses, in that we find minimal evidence of productivity in the earliest multiword utterance coupled with a development-related increase in productivity soon thereafter. However, our results deviate slightly from the proposal of gradual emergence of abstract schema from item-specific exemplars, as set forth in (Abbot-Smith & Tomasello, 2006). The possibility of a decrease in determiner productivity later in development suggests that while children may construct abstract generalizations from their input, they may also use input later in development to constrain overly general abstract schema (along the lines schematized in Figure 1.1B, right, top two trajectories).

Our model is defined independently from overlap score, the primary measure of productivity used in previous literature. We can take advantage of this independence to use overlap as a model validation method. Although a simple overlap measure is not useful for characterizing productivity and comparing *across* children, we can use it to validate our model *within* individuals. We do this by sampling new simulated determiner productions from the fitted model's distribution on child determiners for each time window, computing overlap, and then comparing the results to the empirical values from that same child. Empirical overlap falls within the 95% range of simulated overlap scores for all children, validating the model's overall fit to the data. For additional details see Appendix A: Results.

1.4.2 Sliding Window Method

The higher temporal resolution sliding window method reveals changes in grammatical productivity consistent with the split-half analysis, with major increases in productivity for Speechome and Warr and major decreases for Thomas, Adam, and Sarah (Figure 1.5, column 1). The sliding window models also reveal significant variability in ν not related to age (e.g., Naima from the Providence Corpus). In addition, using the same validation technique described above, simulated overlap from sliding window estimates was strongly correlated with empirical values (Pearson's r of 0.940 – 0.951 across data treatments).

1.5 Discussion

The model-based statistical approach presented here for analyzing child language is the first method that allows the respective contributions of productivity and item-based knowledge to be teased apart. Our analysis reveals two key findings. First, children's syntactic productivity changes over development. Several of the youngest children show increases in productivity, with evidence strongest in the largest dataset, Speechome. In addition, some older children show decreases in productivity. This trend might suggest a period of particularly strong generalization followed by a retreat, similar to the pattern observed in morphological domains (e.g., Rumelhart & McClelland, 1985; Pinker, 1991), as well as verb argument structure (Bowerman, 1988; Ambridge et al., 2011).

Second, for the majority of children, our model placed wide confidence intervals on productivity estimates, indicating that the available data were likely not sufficient to draw precise developmen-

tal conclusions. The data for these children typically included a maximum of one hour per week of transcripts; furthermore, most of the child productions in these datasets were collected after the child's second birthday. If adult-like categories are constructed early—soon after the onset of word combination—many of these datasets begin too late to provide decisive evidence regarding the trajectory of early development. The trend line obtained in Figure 1.4 is suggestive rather than conclusive; additional datasets would be required to test whether the pattern is robust within the developmental trajectory of a single child. These results underscore the critical importance of dense, naturalistic data for understanding the development of linguistic knowledge in early childhood.

1.6 Conclusion

Debates about the emergence of syntactic productivity have typically oscillated between two poles: Immediate, full productivity early in development, or accumulation of item-specific knowledge with gradually increasing levels of productivity. Our approach parameterizes the space of models between these poles. In the future it can be adapted to characterize productivity in other simple morphosyntactic phenomena and in other languages. In the key case study of English determiner productivity, applying our model to new, dense data yields support for constructivist accounts and further constrains the developmental timeline within these accounts. While children's earliest multiword utterances may be island-like, grammatical productivity emerges rapidly thereafter.

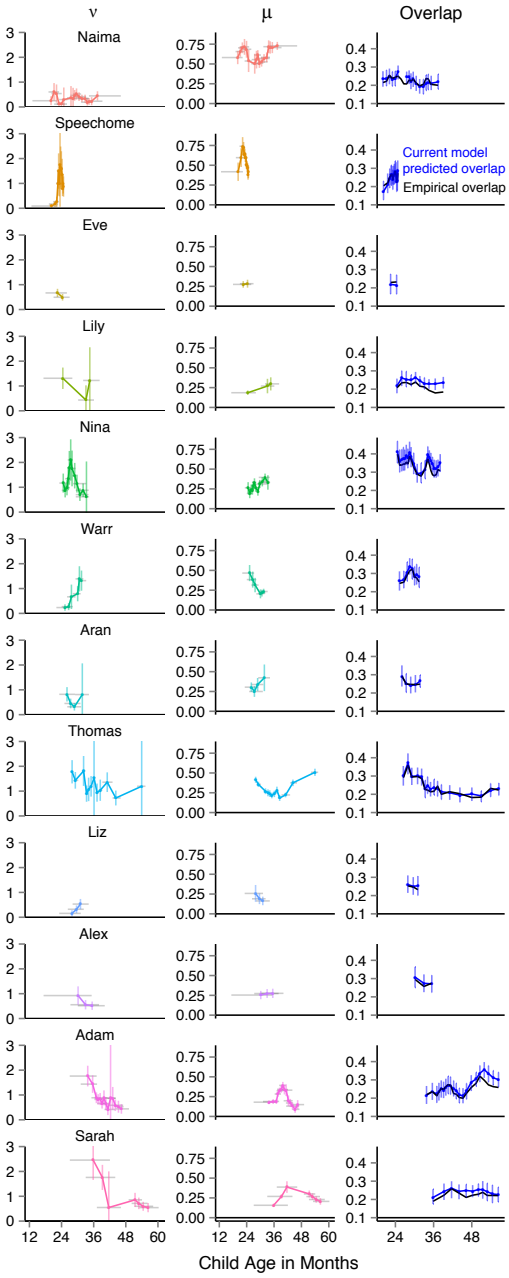


Figure 1.5: Child determiner productivity ν, child mean determiner preference μ, and predicted and empirical overlap scores for the 11 children presented in the split half analysis. Vertical lines show the 99% HPD for ν, μ, and overlap predicted by the current model. Horizontal lines indicate the temporal extent of the tokens used to fit the model at each point.

Question: What's the probability of encountering a dinosaur on the streets of Moscow?

Answer: 50-50: Either you do — or you don't.

Russian proverb

2

Evaluating Models of Robust Word Recognition with Serial Reproduction

Interpreting speech entails dealing with environmental noise, speaker variability, and other factors that make the acoustic signal alone insufficient for robust communication. Rather, listeners' expectations about what others are likely to say make communication possible: people flexibly com-

bine the perceived linguistic signal with their own knowledge regarding plausible—or probable—utterances they should expect from a speaker. Indeed, shortcomings in automatic speech recognition systems highlight deficiencies in machines' abilities to leverage these same knowledge sources. But what form should these linguistic expectations take, and in particular what sort of representations should be attributed to listeners? In this chapter, I evaluate several probabilistic generative models of language (PGLMs) in their ability to capture human linguistic expectations for the task of in-context spoken word recognition.

To move forward on this issue, I use a novel method for eliciting relevant experimental data: a large-scale online game of "Telephone." This sequential experimental design has an extremely useful property: serial reproduction, formally equivalent to iterated learning, reveals the inductive biases that people use to interpret the utterances that they hear. The set of utterances yielded by serial reproduction (Bartlett, 1932; Xu & Griffiths, 2010) can then be used to evaluate a set of PGLMs. This set includes models with architectures that vary according to key theoretical questions regarding the representations that people use for in-context word recognition and sentence processing. An evaluation of model performance shows that models that use abstract representations of preceding context best predict the pattern of changes made by people in the Telephone game, though large *n*-gram models lacking abstract representations perform only slightly worse. I interpret these findings in light of recent work highlighting the interaction of computational constraints and representations in human language processing. A mixed-effects regression model predicting *which* words in a linguistic utterance are most likely to be lost or changed in the course of spoken transmission corroborates these results, and replicates and extends previous results from isolated spoken word recognition and

eye-tracking research for writing linguistic material.

The contents of this chapter are as of yet unpublished. It was co-authored by Sathvik Nair (U.C. Berkeley) and Thomas L. Griffiths (U.C. Berkeley). Special thanks to the Computational Cognitive Science Lab at U.C. Berkeley for valuable feedback, Andrew Silverman (Gracenote) for help with implementation of the browser-based recording interface, and Samuel Tarakajian for recordings of audio stimuli. This material is based on work supported by a National Science Foundation Graduate Research Fellowship to S.C.M. under Grant No. DGE-1106400.

2.1 Introduction

Spoken communication occurs in a noisy channel (Shannon, 1948, 1951). To combat rampant environmental noise, variation within and between speakers, and lexical and syntactic ambiguity, listeners must make extensive use of well-developed linguistic expectations, or knowledge of what speakers are likely to say (Gibson et al., 2013). This requires the integration of many sources of information, including knowledge of word frequencies (Howes, 1957), probable word sequences (Miller et al., 1951), sub-word phonotactic regularities (Vitevitch & Luce, 1999), the plausibility of syntactic relationships (Altmann & Kamide, 1999), and pragmatic expectations (Rohde & Ettlinger, 2012). More broadly, people are able to adjudicate between various candidate interpretations of an acoustic signal in light of the specific discourse context, and bring considerable "general world knowledge" (e.g., knowledge of intuitive physics, properties of people and objects) to the task of natural language understanding (Levesque et al., 2011).

While establishing *which* information sources people use is an important first step in understanding language processing, the next critical challenge is to develop a theory of how they learn, represent, and use this information. Levy (2008) posited an explicit link between knowledge *encoded as probabilistic linguistic expectations* and sentence processing difficulty as revealed by various behavioral measures. This proposal identified the critical role of a "causal bottleneck:" the totality of a person's linguistic expectations—reflecting any combination of the above knowledge sources relevant to language processing—are reflected in a listener's (or reader's) expectations for which word will be encountered next. The negative log probability of a word under a listener's expectations, or

surprisal, provides a succinct measure of a word's expectedness. Levy demonstrated that surprisal estimates derived from a probabilistic generative model of language can be used to approximate human knowledge of language structure, and can provide significant explanatory power above and beyond theories of processing difficulty that focus on memory constraints alone.

Though Levy (2008) used a specific probabilistic generative model (the Earley parser of Hale, 2001), he asserted a broader relationship between surprisal and processing difficulty, such that more sophisticated models capable of capturing additional information sources should be expected to produce surprisal estimates more strongly correlated with observed processing difficulty. Subsequent work in psycholinguistics and cognitive science has demonstrated the utility of a variety of generative language models in understanding sentence processing (Demberg & Keller, 2008, 2009; Frank & Bod, 2011; Fossum & Levy, 2012; Smith & Levy, 2013; Fine et al., 2013; Futrell & Levy, 2017) and shed considerable light on communicative constraints on lexicons (Piantadosi et al., 2011; Mahowald et al., 2013). However, the question of *which* models best capture human linguistic expectations — which information sources and via which representations — remains notably underexplored.

In addition to the utility of increasingly accurate models of people's linguistic expectations for psychology and cognitive science, these same models are of utmost importance for a very broad set of speech-related engineering applications. Using wider linguistic context has been crucial for the development of automatic speech recognition systems, where noise, ambiguity, and speaker variation have been well-known challenges since the 1940s (e.g., Shannon, 1948). As such, the development of generative models of utterance structure has received extensive treatment in the fields of Computational Linguistics, Natural Language Processing, and Automatic Speech Recognition, where such

models are collectively known as "language models." While these models entail strong simplifying assumptions regarding the knowledge of language structure available to people (let alone knowledge of language structure available to linguists), they can nonetheless be used to derive expectations from large corpora or other large-scale datasets. Furthermore, their probabilistic form allows them to be combined in principled ways with acoustic information using Bayesian techniques. A growing interest in commercial applications in recent years has resulted in a profusion of model architectures, especially ones taking advantage of deep artificial neural network (e.g., Hannun et al., 2014; Zaremba et al., 2014; Jozefowicz et al., 2015).

Research in natural language processing typically evaluates generative language models in terms of the probability that they assign to a held-out dataset: the model that assigns a higher probability to a held-out dataset is the better model in the absence of confounding factors (Jurafsky & Martin, 2009). This methodology identifies models that are optimized to reflect the statistics of the corpora they are tested on, which may or may not be representative of people's linguistic expectations in sentence processing. Even the use of naturalistic corpora like Switchboard (Godfrey et al., 1992) or Santa Barbara (Du Bois et al., 2000) fails to address this problem, in that participants' expectations for the purposes of comprehension may deviate significantly from that of any known corpora.

In the current work, we develop a method to approach the linguistic expectations used by people in a naturalistic language task using the technique of *serial reproduction* (Bartlett, 1932; Xu & Griffiths, 2010). Similar to the related technique of *iterated learning* (Kirby, 2001), serial reproduction can be shown to converge to participants' inductive biases (or prior, under a Bayesian interpretation) over a sufficient number of iterations and as long as certain conditions are met (Griffiths & Kalish,

2007). Even with fewer iterations than required for convergence, the yielded samples are still useful in that they reflect an approach to the distribution of interest. The behavioral experiment here specifically targets the gap in systematic, language-wide, and ecologically-valid tests of generative language models by using this method in the form of a large-scale web-based game of "Telephone". In addition to evaluating specific language models, we focus on performance differences between models that arise from broad architectural differences that have received significant theoretical attention in the language processing literature. We focus here on two contested aspects of linguistic knowledge in word recognition and sentence processing: 1) to what degree people use the preceding words in an utterance for prediction of upcoming material and 2) to what degree people use higher-order, abstract representations of that preceding context, like grammatical phrase structure or parts of speech, to inform their expectations.

To preview our results: the changes people make to sentences in the Telephone game are better explained by probabilistic generative models of language that track preceding context. Among those models that track preceding context, the changes are better explained by models that make use of abstract representations of that preceding context. A token-level analysis of which individual words are successfully transmitted from one participant to another provides converging evidence for both results. More broadly, we introduce a method to elicit samples that approach people's expectations for receptive language tasks in the audio modality, and show how that method can be used to evaluate probabilistic generative models of language in their ability to explain key human linguistic behaviors.

"I bought a bear/pear at the farmers' market"

Candidate	Prior	Likelihood	Posterior
bear	Highly improbable (.1)	Fully consistent (1)	.1 /(.1 + .72), p = .12
pear	Probable (.9)	Slightly inconsistent (.8)	.72 /(.1 +.72), p = .88

Figure 2.1: Example of the contributions of the likelihood and prior in spoken word recognition in a sentential context. The phonemes /b/ and /p/ are largely distinguished by a single dimension of difference, their voice onset time, and are relatively easily confused. Using broader language knowledge, a listener can recognize a speaker's likely intended meaning when data is ambiguous or noisy. In this toy example, the prior probability of *pear* overwhelms the evidence for *bear* provided by the data.

2.2 Theoretical Background

In a landmark study, Luce & Pisoni (1998) showed that the recognition of isolated words embedded in noise could be modeled as a competitive process, combining evidence from the received audio signal with each word's probability (relative frequency) in a corpus. Subsequent work has formalized this process of competition among words within an explicitly Bayesian framework and extended it to the recognition of words in sentential contexts (Norris & McQueen, 2008). It is now widely acknowledged that linguistic expectations—probabilistic knowledge regarding what is more or less likely to be said—make the process of word recognition far more robust than it would be as a purely data-driven, bottom-up process (Norris et al., 2016). Figure 2.1 provides an example of how a listener might overcome perceptual noise to arrive at a speaker's intended message — that they bought a *pear* and not a *bear* at a farmers' market — by relying on their prior knowledge.

A critical question at Marr's computational level of analysis (Marr, 1982) is what sort of information sources might be combined to accomplish this task, independent of the precise timecourse or

computational tractability. Luce & Pisoni (1998) use word probability—normalized frequency—as an example of basic word-level knowledge that listeners have access to before encountering a linguistic signal. This would, however, lead to incorrect predictions for the bear/pear example in Figure 2.1, in that most corpora have more instances of the word *bear* than *pear*. Rather, the plausibility of the two candidates reflects more detailed world knowledge, for example what sort of things might be bought and sold at a farmer's market. In the absence of models that encode this sort of world information directly, increasingly sophisticated models of linguistic structure, or probabilistic generative models of language (PGLMs), have been used as a proxy for approximating people's expectations.

The PGLMs under study here vary in structural complexity from none (a unigram model) to distributions over full parse trees identifying fine-grained syntactic relations between all words in a sentence. Language models with higher structural complexity bring the task of in-context spoken word recognition into contact with the task of *sentence processing*: if people think that the utterances they hear will adhere to grammatical rules, they can use that information to constrain their predictions about upcoming words. While people certainly use detailed representations of relationships between words in sentence processing more broadly, it remains an open question whether this same information is used inferring a speaker's meaning for ambiguous acoustic signals in the course of word recognition.

Use of Preceding Linguistic Context

We focus our investigation on two key dimensions of probabilistic generative models: whether they make use of preceding context, and if so, whether they encode abstract structural regularities.

There exist opposing theoretical views on whether sentence processing difficulties reflect unexpected linguistic material or difficulties in integrating newly received words into a listener's semantic representation of a sentence. In principle, processing difficulty on encountering an unexpected noun (as in "I bought a *bear*") could reflect a violation of linguistic expectations, difficulty in constructing a scene representation ("a mixture of memory retrieval and semantic integration processes instigated by the noun itself" per Nieuwland et al., 2018), or both. Thus measures of processing difficulty for low probability, semantically-rich words cannot adjudicate between these two positions.

DeLong et al. (2005) conducted an experiment to evaluate evidence of processing difficulty for words with strong linguistic—but not semantic—expectations using the variants of the English indefinite articles "a" and "an." They measured an N400 response to contextually-predictable versus unpredictable articles, for example, "The day was breezy so the boy went outside to fly...", where a separate norming study established cloze probabilities of 86% for 'a' and 89% for 'kite'. They obtained higher magnitude N400 responses for both nouns and articles, the critical test case, in inverse proportion to their cloze probability (i.e, stronger N400 responses for less probable continuations). Nieuwland et al. (2018) challenge this conclusion with the results of a multi-lab replication study, which failed to obtain a significant result in the case of articles. They interpret this null result as evidence that phonological forms are not pre-activated on the basis of preceding context.

Here we use an alternative means of testing for evidence of prediction, by evaluating evidence for context-reliant inductive biases in spoken word recognition. While transmission errors between a speaker and a listener are expected under both of the above accounts for low probability words, only the prediction-oriented account suggests that the replacements — the misrecognitions of listeners

— reflect a distribution associated with word prediction. In the current experiment, we test whether the set of utterances yielded by serial reproduction increase in probability faster under models that use preceding context versus one that does not. We conduct this analysis first on all utterances, then in the Discussion look at the specific case of indefinite articles, which comprise more than 2% of the collected data.

One advantage of the current approach is that it makes no strong assertions regarding the timeline of the integration of received data with prior expectations, as is required in the case of time-locked stimuli for examining N400 effects. Following Kuperberg & Jaeger (2016) we take a Bayesian interpretation of the term *prediction*, as distinguishing those information sources that are independent of the received acoustic signal from the acoustic signal itself. Thus even if the inference happens well after the receipt of the audio signal, inference still implicates prediction so long as it calls upon independent information sources. A second advantage is that the "Telephone" task used here has no visual queues, one criticism raised by Nieuwland et al. (2018) of Altmann & Kamide (1999).

Another well-cited study, Piantadosi et al. (2011), found that average in-context predictability for a word is a stronger predictor of word length than frequency alone in a sample of 11 European languages, providing indirect cross-linguistic evidence of the importance of preceding context. In that the robustness and cross-linguistic generality of this result is challenged elsewhere in this volume, we stress the importance of this question regarding the role of preceding linguistic context in spoken word recognition.

Abstract Structure in Preceding Context

The second question we address—predicated on an affirmative result in favor of *any* use of context in the above analysis—is whether there is evidence that people use abstract representations of that preceding context. Contrary to previous work finding evidence of extensive use of abstract representations in sentence processing, (e.g., Gibson et al., 2005), Frank & Bod (2011) found no additional predictive value for models that use hierarchically-structured representations of preceding sentence context for predicting reading times, a common measure of processing difficulty. The models under comparison in their study included *n*-gram models, the Roark parser (Roark, 2001), and echo state networks (Jaeger, 2001), a kind of recurrent neural network architecture. By contrast, a replication and extension of that study by Fossum & Levy (2012) found that the use of better lexical *n*-gram controls derived from a larger model and using a more sophisticated smoothing technique eliminated the performance differences between unlexicalized sequential and hierarchical models. Further, they showed that a state-of-the-art lexicalized hierarchical model from Petrov & Klein (2007)—a model that tracks more granular relationships between words and grammatical categories—predicts reading times better yet.

In the current work, we investigate the importance of abstract representations in the auditory domain, using a larger sample of language models. We adopt a different typology of models than that used in the above works: we group recurrent neural network models—of which we include two more recent architectures—with models that infer parse trees. While Fossum & Levy (2012) and Frank & Bod (2011) separate those models that make full hierarchical representations of the phrase

structure ("latent context in the form of syntactic trees") vs. others, we note that recurrent neural networks capture significant higher-order linguistic regularities of a qualitatively similar type to the nonterminals in a phrase structure grammar, and unlike those captured from an *n*-gram model. An examination of the word embeddings from Elman (1990), an extremely simple recurrent neural network, show that the average activation pattern for a word reflects a combination of semantic and syntactic similarity, in that both of these are reflected in the lexical distributions in the surrounding context. While this is admittedly far from a full hierarchical parse tree, the use of *any* sort of abstract context may have a stronger effect on the sort of expectations that are encoded in a model, rather than whether the representation is fully hierarchical. We evaluate support for this partition of models, as well as the performance of the language models in predicting the pattern of changes, using our collected serial reproduction data.

On the Evaluation of Language Models

As noted in the introduction, probabilistic generative models of language are generally evaluated by how likely they are to generate a test sample of language (in natural language processing) or in their ability to predict experimental observables pertaining to processing difficulty (in psycholinguistics). Each of these evaluation methods has notable shortcomings.

Regarding maximizing test sample probability, there is no guarantee that a corpus-derived test set has a high probability under human linguistic expectations. As rational agents, people should be expected to develop expectations that match the task of interpreting linguistic material under noisy conditions, but in the case of language it is very hard to know what constitutes "typical" linguistic

experience. The proportions of written and auditory sources, topics, and speech registers (i.e., levels of formality) is undoubtedly subject to significant individual variation. Furthermore, even with a perfect estimate of a person's experience with language thus far, there is no guarantee that linguistic material encountered in the future will have the same composition as that encountered in the past. In such a case, a rational agent might be expected to shift probability mass to as-of-yet unobserved linguistic events to avoid overfitting on previous experience. As such, we argue that *no* corpus collected from natural sources should be expected to reflect people's linguistic expectations for the task of in-context word recognition.

Psycholinguistic studies, by contrast, have evaluated language models in their ability to predict observables pertaining to processing difficulty, such as reaction times, looking time or regressions in eye tracking, and event-related potentials. We note the presence of a confound in such measures, in that high surprisal linguistic events may cause increased processing difficulty, but they may also simply result in communication errors. This complex trade-off between expectedness, processing difficulty, and the prevalence of errors in communications—which may vary further according to task demands—means that processing difficulty is only a partial record of expectations. Models must thus be evaluated with respect to both processing difficulty and the kinds of errors they induce.

In the present study we use the technique of *serial reproduction* to create a dataset that sidesteps the above issues. Intuitively, we start with a small test corpus, similar to the first evaluation method noted above, and use serial reproduction to gradually change the properties of that corpus so that it better reflects people's linguistic expectations. This transition in properties arises from the fact that participants' expectations are reflected in the changes they make between the utterance they hear and

the utterance they produce at each iteration.

Serial Reproduction

Information transmission by serial reproduction was first studied by Sir Frederic Bartlett, who tracked the evolution of stories and pictures recreated from memory after rapid presentation (Bartlett, 1932). More recent work has identified that the technique, like iterated learning (Kirby, 2001), can be used to experimentally reveal inductive biases, or reasons that people would favor one hypothesis over another independent of observed data in inferential tasks (Mitchell, 1997). If participants (reproducers for serial reproduction, learners for iterated learning) use Bayesian inference to infer the posterior distribution over hypotheses, and then draw from that distribution in production, then their output at each sequential generation reflects a combination of observed data and participants' inductive biases. Over time, the distribution implicit in the output data comes to reflect participants' inductive biases. For a trial in the Telephone game, participants' hypotheses are the set of possible interpretations that they might provide for an utterance, in that the message they infer reflects both the acoustic data and their expectations regarding language. We model these expectations regarding language—the prior a listener uses when inferring a speaker's message–with a suite of probabilistic generative models of language (PGLMs), which are described in greater detail below.

Under certain assumptions, both kinds of transmission chains (iterated learning and serial reproduction) can be interpreted as a form of Gibbs sampling, a common technique in Markov chain Monte Carlo based parameter estimation (Geman & Geman, 1984). The transmission of a linguistic utterance across a succession of participants can be interpreted as a Markov process; given certain as-

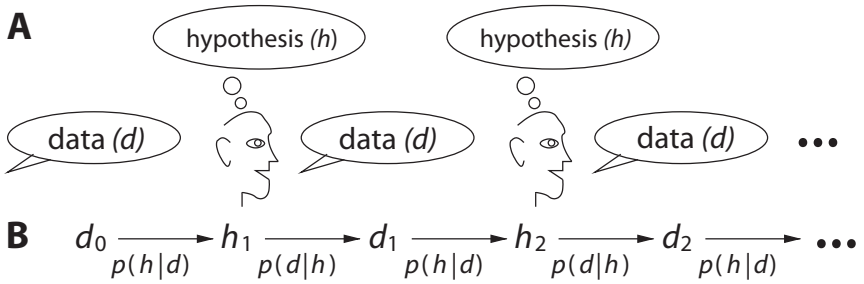


Figure 2.2: Figure 1: A: In serial reproduction, each participant chooses a hypothesis regarding what they heard, and on the basis of that hypothesis produces data for the next listener. B: Under the Bayesian model of the telephone task presented here, participants assign posterior probability $p(h|d)$ to hypotheses regarding what they heard proportional to the product of the data-dependent likelihood $p(d|h)$ and the data-independent prior $p(h)$.

sumptions, the resulting Markov chain is guaranteed to converge to its stationary distribution if run for enough iterations. This means that the probability that a participant selects a hypothesis h converges to their prior distribution $p(h)$. This approach has been profitably used to reveal inductive biases in memory (Xu & Griffiths, 2010), category structure (Griffiths et al., 2008a; Sanborn et al., 2010), and function learning (Griffiths et al., 2008b). In the domain of language specifically, iterated learning has been used to reveal biases relevant to language evolution (Griffiths & Kalish, 2007).

We motivate the use of serial reproduction with language in the audio modality by analogy to the function learning experiments of Griffiths et al. (2008b), which can be thought of as a game of "Telephone" in the space of mathematical functions in two dimensions. In these experiments, participants saw a selection of points drawn from a function in a two-dimensional space. They were then asked to reproduce that function with a small number of points. While different transmission chains started with radically different functions, including some of considerable complexity, all are reduced to simple linear functions in the course of reproduction (Fig. 2.3). In the face of limited, noisy data people revert to a preference for simple linear functions. In the present case we are inter-

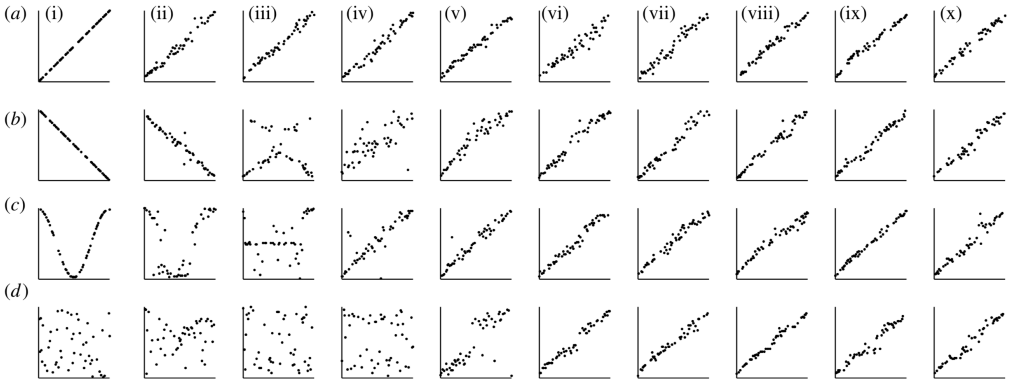


Figure 2.3: Example transmission chains from a function learning experiment. Columns (i-x) correspond to successive participants reproducing different initial functions, (a_i-d_i). The resulting positive, linear functions provide evidence of an inductive bias for function learning. Figure reproduced from Griffiths et al. (2008b).

ested in a much more nuanced set of expectations — what people expect other people to say.*

The process of serial reproduction can consequently be thought of as gradually changing a corpus so that it better reflects what people expect others to say. To the degree that the initial set of utterances in that corpus is not representative of what people expect to hear, then the process of serial reproduction will introduce edits that change them to increasingly reflect the broader linguistic expectations of participants. If, for example, the initial corpus contained an excessive number of low-frequency words pertaining to finance, participants should be expected to misinterpret these and replace them with words prototypical of normal conversational registers. To our knowledge, no previous work has used this technique of serial reproduction to investigate language processing. However, we note that the task of serial reproduction of isolated speech sounds was previously used to investigate whether repeated imitation of environmental noises can give rise to word-like forms

*An intuitive interpretation of this process is that repetition entails reversion to the mean. In the case of linguistic expectations, this "mean" is of unknown character and the subject of strong theoretical interest.

(Edmiston et al., 2017).

2.3 Language Models

The principal objective of this contribution is to evaluate a variety of broad-coverage language models in their ability to capture the linguistic expectations implicit in participants' behavior in a game of Telephone. By "language models," we follow the convention of the natural language processing literature to refer to probabilistic models that encode information regarding the distribution over possible words before encountering acoustic data regarding its identity. We thus exclude many "language models" in the broader sense that are concerned with the online dynamics of word recognition like the cohort model (Marslen-Wilson, 1987a) or logogen model (Morton, 1969) of word recognition. In that modeling human performance in the Telephone game requires large vocabularies and a means of handling out-of-vocabulary items, all models here track expectations over 10,000 or more word types and have a method for handling newly-encountered out-of-vocabulary words. We now provide a high-level overview of the major probabilistic language models examined in this work. Insofar as each model is the subject of many dissertations, conference papers, and journal articles in its own right, the overviews provided here are neither exhaustive nor formally rigorous; we refer readers to the original publications for more details regarding each model.

N-gram models

The simplest PGLMs that we evaluate here are *n*-gram models. First used to model natural language by Shannon (1948), *n*-gram models typically track forward transitional probabilities for words by

conditioning on preceding words. These models make the strong simplifying assumption that the sequence of words constitutes in an utterance or text is a *Markov chain*, in which the probability of each event (i.e., word) depends only on immediately preceding events. These models can condition on a larger or smaller preceding context: an n-gram model tracks conditional probabilities given n-1 words. Bigram models condition on just the preceding word, while trigram models condition on the two previous words. The probability of an utterance is the product of the conditional probabilities of the constituent words,

$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{-1}), \tag{2.1}$$

where $w_1, \ldots, w_m$ is an utterance of comprised of m words and n is the order of the n-gram model. By convention, word sequences are augmented with a start symbol (of probability 1) and an end symbol (the probability of which is tracked by the model, the same as any other word type). These models track statistics over sequences of words, and do not include any explicit encoding of statistical expectations for higher-order abstractions like parts of speech, super-ordinate grammatical categories, or semantic categories. n-gram models may appear to encode expectations related to these abstractions because lexical statistics capture these regularities implicitly: a trigram model will assign almost all of the probability mass following "near a" to nouns and adjectives. However, without support for abstract representations, a trigram model cannot assign higher probabilities to nouns as a class (i.e., nouns not observed following "near a") in predicting the next words in that context. As

such, comparing the probability of the serial transmission chains under the *n*-gram models against the probability under the other models—which all posit and track regularities at a level higher than words—allows us to evaluate evidence for the degree to which people may use abstract representations of context to inform their expectations.

A special case of the *n*-gram model of particular theoretical interest is the unigram, or 1-gram model, where the probability of a word is not conditioned on preceding context. Unigram probability estimates thus largely reflect normalized word frequencies, but may assign nonzero probability mass to out-of-vocabulary words depending on the choice of smoothing technique, described below. Comparing the predictions of the unigram model with those of the trigram and higher-order *n*-gram models allows us to evaluate evidence for the degree to which participants' linguistic expectations are updated to reflect any amount of preceding context, independent of the degree of abstraction.

Besides the length of the context on which they condition, a second dimension of variation in the architecture of *n*-gram models is the choice of *smoothing* technique, or how probability mass is allocated to unobserved word sequences when the model is fit. Using probabilities directly derived from counts, i.e., the maximum likelihood estimate, of an *n*-gram model is generally ill-advised because of sparsity: many word sequences observed in a new dataset may not have been observed in the dataset used to fit the corpus. This reallocation of probability mass among sequences often reflects simple assumptions about the statistical structure of languages. Smoothing complements the practice of assigning some proportion of the unigram probability mass to newly-encountered tokens (known as "Out-Of-Vocabulary" words), such that the model assigns new material nonzero probabilities.

Here we use two smoothing schemes when estimating transition probabilities for higher-order

n-gram models. For larger datasets, we use modified Kneser-Ney smoothing (Chen & Goodman, 1998). This smoothing scheme shifts probability mass to unobserved trigrams that are expected on the basis of the prevalence of constituent loewr-order *n*-grams, e.g., assigning "near San Antonio" a small non-zero probability because "near San" (as in "near San Francisco") and "San Antonio" are both relatively frequent bigrams. For smaller datasets, we use Good-Turing smoothing (Gale & Sampson, 1995), which shifts probability mass from *n*-grams seen once to ones that are not encountered in fitting.

We build several new *n*-gram models using the SRI Language Modeling Toolkit, or SRILM (Stolcke, 2002). We also use a proprietary 5-gram model from the DeepSpeech project (Hannun et al., 2014) which was estimated using KenLM (Heafield, 2011), another language modeling package which was specifically designed with backwards compatibility with SRILM and produces estimates that differ only as a function of small numerical differences in implementation. The datasets used to derive parameter estimates are treated in greater detail below.

PCFGs

The remainder of the probabilistic generative language models under consideration here use some sort of abstract representation above the level of the word to derive word-level expectations. The first class of model, probabilistic context-free grammars (PCFGs) posit that utterances reflect an abstract hierarchical structure comprised of grammatical categories like nouns, verbs, noun phrases, and verb phrases. The task of predicting the next word is thus not just dependent on the previously-observed words as in the *n*-gram models, but also depends on a listener's beliefs about the hierar-

chical grammatical structure implied in the previous words, and how that grammatical structure narrows the set of possible continuations. For example, having heard words that likely form a verb phrase for a transitive verb, a listener should expect a noun phrase containing the object of that verb.

Each hypothesis about the abstract hierarchical structure of an utterance is captured in terms of a parse tree, which describes how an utterance could be generated from a context-free grammar. A context-free grammar is a linguistic formalism that describes sentences as a hierarchy of constituents, starting with a root "sentence" node. This sentence node is composed of grammatical categories such a nouns, verbs, noun phrases, verb phrases, each of which in turn consists of other grammatical categories or words. All terminals, or leaf nodes in the hierarchy, are words. A CFG captures the intuition that the same linear sequence of words may reflect several possible interpretations of the relationship between constituents, for example that "the girl saw the boy with the telescope" has two high probability interpretations pertaining to the attachment point of the prepositional phrase — whether "with the telescope" modifies *boy* or how the girl *saw*. While CFGs cannot capture certain human linguistic phenomena (Stabler, 2004), probabilistic CFGs, or PCFGs, are commonly used as generative probabilistic models of language given their principled account of higher-order structure. Unlike typed dependency parsers (e.g. De Marneffe et al., 2006), which track typed pairwise relationships between words, PCFGs provide a probability distribution over hierarchical parses for an utterance.

The parameters of PCFGs can be learned in an unsupervised fashion from linear word representations alone, or fit using corpora that have been annotated by linguists with gold-standard most-probable parses. Because of their stronger performance, we focus here on the predictions of PCFGs

with supervised training. Given the size of the hypothesis space of possible grammars for a language, as well as parses for a sentence, PCFGs vary in the techniques that they use to find the highest probability parses. Here we use two PCFGs, the Roark (2001) parser and the BLLIP parser (Charniak & Johnson, 2005), which approach the problem of inference in this large hypothesis space in very different ways.

Roark Parser

The Roark parser (Roark, 2001) is a widely used *top-down* parser, meaning that it maintains and updates a set of candidate parse trees as it moves from left to right in an utterance. In combination with other architectural decisions, this allows the Roark parser to calculate probabilities for each sequential word conditioned on the preceding words. This incremental property makes it especially well-suited for modeling online sentence processing, where people update their interpretation of the utterance as they receive new acoustic data (Roark et al., 2009).

BLLIP Parser

Unlike the top-down approach of the Roark Parser, the BLLIP (or Johnson-Charniak) parser uses bottom-up inference in combination with a secondary scoring function on the yielded highest probability trees (Charniak & Johnson, 2005). The bottom-up approach means that the parser starts by positing grammatical categories directly above the level of words, then iteratively identifies possible higher-level grammatical categories up to the sentence root. The BLLIP parser uses bottom-up parsing to generate a set of high probability parses which are then re-ranked using a separate discrim-

inative model with a researcher-specified set of ad-hoc features. Though bottom-up parsing violates the constraints of online auditory language processing in that it requires that the complete signal be received previous to parsing, we include the parser in analyses tracking utterance-level changes because of its strong performance.

For both PCFG language models, the probability of an utterance reflects a marginalization over the possible parse trees that generate that utterance. Because of computational constraints, these parsers typically generate a relatively small number of the highest probability parse trees (in this case, 50 for each model). While this represents a truncation of the true set of possible parse trees, the first few (i.e., the first one to five) are overwhelmingly more probable than the remainder, and thus provide a reliable estimate of utterance probability. In the case of the Roark parser, surprisal estimates for individual words can be derived by querying the probability of the next word given the restricted set of highest probability provisional parse trees (Roark et al., 2009).

Recurrent Neural Network Language Model (RNN LM)

The expectations derived from PCFGs reflect hierarchical parses of sentences, but listeners could also develop expectations by noting abstract commonalities in the usage of words, for example that the words "five" and "six" tend to appear in very similar lexical contexts. We include here two recurrent neural network models (RNNs) that can infer partially syntactic, partially semantic higher-level regularities in word usage, and can use these regularities in the service of prediction. As with PCFGs, RNN language models can track long distance dependencies (Linzen et al., 2016). While models lacking abstract representations of context could in principle capture long-distance dependencies

(e.g., a 9-gram model), the combinatorial richness of language means evidence is far too sparse to infer these regularities when tracking sequences of words alone.

Recurrent neural networks are able to capture and use these higher-level regularities because they use the state of the hidden layer of the network at the previous timestep, often referred to as a "memory," in addition to newly-received data to make predictions. This hidden layer from the previous timestep is a lossy—and thus generalizable—representation of the preceding context. By projecting words in the preceding context into a lower-dimensional space, the observation of a word sponsors the activation of words with similar lexical distributions in the context, imbuing the model with robustness in prediction even if a particular sequence of words has not been seen. Since early demonstrations of their utility for language prediction with small-scale linguistic corpora (Elman, 1990), an extensive literature has scaled up these architectures to deal with web-scale natural language prediction tasks, particularly developing techniques to prevent overfitting given the massive number of parameters, and to manage their computational complexity (Mikolov et al., 2011).

The first RNN we use here, henceforth RNN LM, was first described in Mikolov et al. (2010), while additional performance optimizations for training can be found in Mikolov et al. (2011). We train a network with 40 hidden nodes that uses backpropagation through time (BPTT, Werbos, 1990) for the two preceding timesteps. The vocabulary is limited to the 9,999 highest-frequency word types, with the remainder of types assigned to an <unk> type.

Big Language Model (Big LM)

One of the major challenges with recurrent neural networks is the problem of *vanishing gradients* where the error signal necessary to update connection weights in the network becomes too small to use standard gradient descent techniques. This problem becomes especially prevalent when recurrent neural networks track longer histories of preceding events. The RNN language model described above, for example, can only update weights using the model state at two preceding timesteps. One solution that has received significant attention is to replace nodes in the neural network's hidden layer with *long short-term memory* (LSTM) cells (Hochreiter & Schmidhuber, 1997). While non-LSTM RNNs directly copy the previous state of the hidden layer at each timestep, the LSTM uses these special cells that regulate how information is propagated from one timestep to the next. These cells have input, output, and forgetting "gates" that regulate how the cell's state changes, such that it can maintain state for longer intervals than a typical RNN. Such networks have been widely useful in sequence prediction tasks (Gers & Schmidhuber, 2001), with particular successes in language modeling (Sundermeyer et al., 2012; Zaremba et al., 2014).

One challenge for RNNs, left unaddressed by the use of LSTMs, is the difficulty of scaling to larger vocabularies, a necessity for modeling large naturalistic language corpora. Evaluating the network's loss function, which requires generating a probability distribution over all words, becomes prohibitively computationally expensive owing to the normalizing term in a softmax function. This has lead researchers to limit vocabularies to sizes much smaller than those typical of n-gram models or PCFGs, often in the range of 5,000 to 30,000 word types.

Jozefowicz et al. (2015) address this limitation in their "Big LM" architecture by representing words as points in a lower-dimensional continuous space using a convolutional neural network (CNN). These models can represent words as embeddings (real-valued vectors) that capture perceptual similarity between words on the basis of shared structure, such as word lemmas or part of speech markings like *-ing* (Ling et al., 2015). The LSTM is then used to make predictions in this lower dimensional space, reducing the number of computations in the softmax function from the size of the vocabulary to the dimensionality of the CNN-derived embeddings.

Given the significant technical resources necessary to train such a model (24-36 graphics processing units) and that model hyperparameters have not been made publicly available, we use the model provided by Jozefowicz et al. (2015). This model uses 4096-dimensional character embeddings to represent words, makes use of backpropagation through time for the previous 20 timesteps, and has two layers with 8192-dimensional recurrent state in each layer.

Training Data

All of the above language models are fit or trained using corpus data comprised of text. Ideally, we would evaluate all combinations of language models and corpora to identify how model performance reflects an interaction of model architecture and training data. However computational limitations, licensing restrictions, and limited public-release codebases make this goal unfeasible at present. Instead, wherever possible we evaluate each language model trained on two datasets: one on a large corpus for which it is known to produce competitive or state-of-the-art results, and the other on the contents of the Penn TreeBank (Marcus et al., 1993).

The Penn TreeBank is a standard training, validation, and test dataset in the public domain, consisting of about 930k words in the training set. For those models that take advantage of validation for setting hyperparameters, these models have additional access to 74k tokens. All sentences are annotated with gold-standard parse trees, making this dataset appropriate for training the two supervised PCFG models. While the thematic content and register of the Wall Street Journal are not representative of conversational English, all language models should be equally disadvantaged by this shortcoming.

For the instance of each model architecture trained on a larger dataset, we use a variety of datasets with known strong performance in the psycholinguistics or NLP literature. For *n*-gram models we use the British National Corpus (BNC; BNC, 2007), an approximately 200m word dataset commonly used in psycholinguistics. The BNC consists of material from newspapers, academic and popular books, college essays, and transcriptions of informal conversations. The BLLIP parser is trained on the Google TreeBank (Bies et al., 2012) in addition to the Penn TreeBank. The Google TreeBank contains approximately 255k word tokens in 16,600 sentences with gold-standard parse trees, taken from blogs, newsgroups, emails, and other internet-based sources. Big LM is trained on the One Billion Word Benchmark of Chelba et al. (2013). The DeepSpeech project (Hannun et al., 2014) provides a smooth 5-gram model trained on a proprietary dataset including Librispeech (Panayotov et al., 2015) and Switchboard (Godfrey et al., 1992). No larger model was used in the cases of the Roark parser and the RNN LM.

Computation of Lexical Surprisal

Estimating surprisal — the unexpectedness of an event, in the form of its negative log probability under a predictive model — for each individual word token requires that a model produce conditional word probabilities, also known as prefix probabilities for PCFGs. In that *n*-gram models encode these continuation probabilities directly, these probabilities are straightforwardly accessible.

The Roark parser produces an estimate of lexical continuation probability that marginalizes across the beam of parse trees (highest probability candidate parses given the material seen so far). The bottom-up inferential procedure used by the BLLIP means that it assigns probabilities to parse trees rather than individual words. For this reason, we analyze changes in sentence probability under the BLLIP, but do not use it as a predictor in analyses requiring word-level estimates of surprisal.

The RNN LM yields a set of activations over the vocabulary which is translated with a softmax function into a probability distribution. Big LM produces a probability distribution over continuations, though the softmax function is computed over the lower-dimensional representations of words. In both cases the prediction by the neural network is conditioned on the state of the model at preceding timesteps. We treat the probabilities output by these models in the same way as the conditional probabilities from the *n*-gram models.

2.3.1 Computation of Sentence Probabilities

For all models we treat omit the end of sentence marker in the computation of probability. We adopt this strategy because of variation in the training datasets regarding punctuation, such that

those models trained on datasets with punctuation assign high surprisal to end-of-sentence markers without preceding punctuation, which was not collected in the Telephone game. In the case of *n*-gram models, sentence probability is simply the product of conditional word probabilities because of the Markov property. We make the same simplifying assumption with the two neural language models.

Evaluating the probability of a sentence for the two PCFG models, the Roark and BLLIP parsers, requires marginalizing over the set of possible parse trees. For these two models, we sum over the probabilities of the top 50 parse trees yielded by each model. While there may be many more trees that would yield the same string of words, the first few (1-5) highest ranking parses typically account for nearly all of the aggregate probability mass for that utterance.

2.4 Methods

We use a web experiment to gather chains of audio recordings and transcriptions appropriate for answering our primary research questions. This web experiment lets participants listen to and record audio, coordinates the flow of stimuli such that recordings from one participant can be used as stimuli for later participants, and ensures that the succession of recordings remains interpretable linguistic material, while avoiding explicit judgments of appropriateness by the experimenters. We first describe experience of a participant in the task; then we describe the flow of data in the experiment, focusing on how the successive contributions of participants are used as stimuli for future participants.

Experimental Interface

Upon accepting the experiment on Amazon Mechanical Turk, a participant is provided with a link to a web application that allows them to listen to and record responses to audio recordings of utterances. The participant first sees a screen in which they are encouraged to adjust their speakers or headphones to a comfortable level to hear a recording of an acoustic piano. The participant then proceeds through a sequence of four practice trials. The first practice trial tests whether the participant's speakers or headphones are working properly by having them transcribe a ten-word sentence. The second practice trial tests whether a participant's microphone is working by having them listen to a recording of a sentence and repeating its content. At the beginning of this trial, the participant grants permission in the web browser for the use of their microphone. Participants are prompted to repeat the second practice trial until their recording is similar to the gold-standard transcription of that sentence, as evaluated by a normalized Levenshtein-Damerau distance of .2 between the gold-standard transcription and the output of an automatic speech recognition system run on their recording (described in greater detail below).

After completing the second practice trial, a participant is then introduced to the full trial format, where 1) they listen to a recording 2) they choose whether or not to flag the recording they heard as appropriate and interpretable 3) they record a response (i.e., their best guess of the content of the utterance they heard) 4) they decide whether to flag their *own* recording (e.g., in case of speech errors or excessive ambient noise) 5) they provide a written transcription of the utterance they recorded. After the participant submits a trial, the new recording and transcription are pro grammatically eval-

uated with a number of filters confirming that the audio recording is not blank, that it is of similar length to the previous utterance, and that it passes basic tests of consistency between recording and transcription using an automated speech recognition system. The participant does two practice trials in this full trial format; we provide further details of these five steps and automated filters below.

After completing the practice trials, the participant begins a series of 47 test trials, including 40 target trials and 7 randomly interspersed fillers. The 40 target trials consist of either from a pool of initial recordings (described below) or the most recently-contributed recording of another participant. The seven filler utterances are the complement of the four filler utterances used for the practice trials, and are drawn from an inventory of nine audio recordings with similar properties to the initial stimuli recordings, and two standard test sentences from the TIMIT corpus (Garofolo et al., 1993). If the participant flags the utterance that they heard as inappropriate in (2) or their own recording as compromised in some way (4),the trial ends early, and they progress to the next trial. Otherwise, if the recording passes the set of automated tests, the state of the experiment is updated after each target trial to point to that newest recording as the appropriate stimulus for the next participant.

An example screenshot is presented in Fig. 2.4. We now present in greater detail the five steps an experimental trial and the automated filters outlined above.

1. Listening to a recording

The participant is provided with a button entitled "Click to play next audio recording." Clicking this button starts a three minute audio timer, which remains visible to the user through Step 5, be-

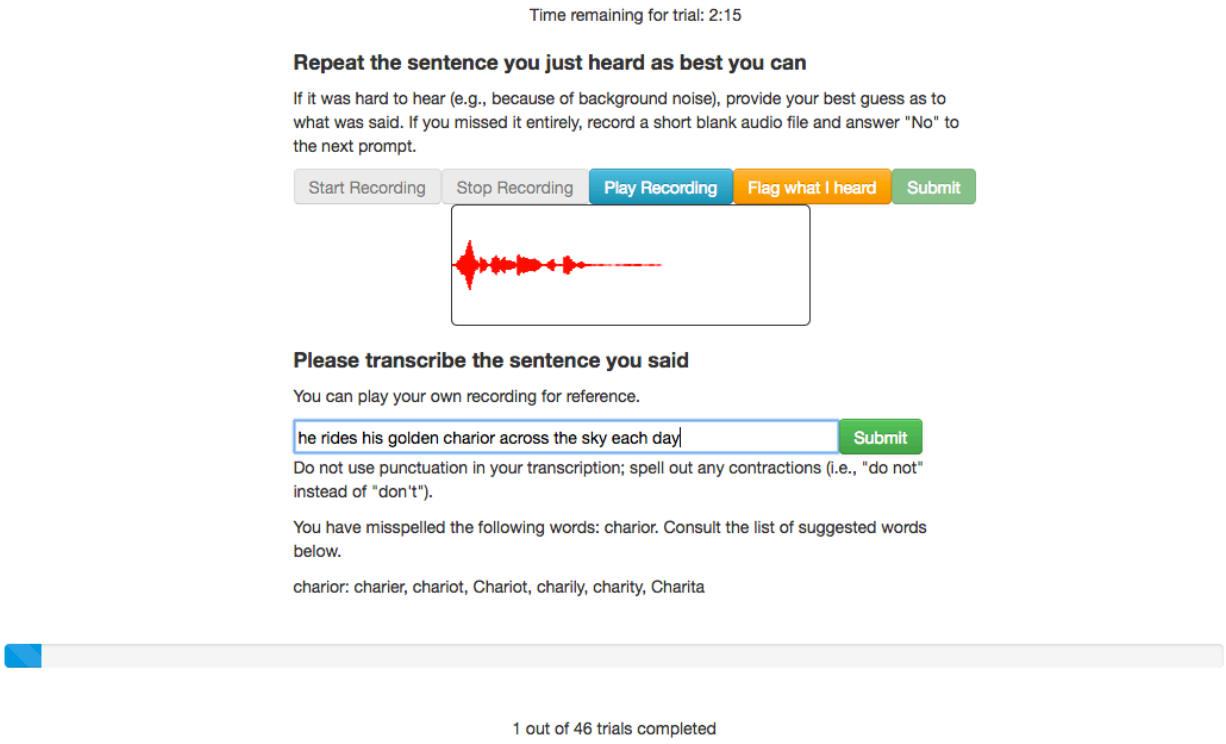


Figure 2.4: The browser-based audio recording and transcription interface used by participants. This screenshot depicts the state of the application after the participant has listened to a recording by a previous participant, recorded a response, and is in the process of submitting a transcription for their contributed recording. In this case the participant has been flagged for misspelling a word.

low. This three minute timer ensures that the web experiment is not blocked by an inactive participant; if time expires, another participant receives this stimulus. Participants receive a mildly-worded warning to try to complete trials more quickly if this timer expires. Besides the button to start playing the audio recording, there are no controls to pause, repeat, or move location in the recording. The audio is presented embedded in noise recorded from a coffee shop (average SNR_{dB} across recordings = -6.8). Pilot experiments established that this naturalistic source of noise introduced a high rate of edits without introducing significant participant attrition, as was observed with similar amplitude white noise. There are between 500 and 1000 ms of noise-only padding preceding and following the target utterance in each initial and participant-contributed recording.

2. Flagging the upstream recording

Though the participant cannot pause or move within the recording that they hear, they may flag the recording at any time. If the participant chooses to flag a recording, they are then asked to choose one of a set of provided reasons: Contains speech errors, Speech starts or stops abruptly, Contains obscenities, and Other. Upon choosing Other, the participant could provide a free-form text response. The trial ends early after choosing a reason.

3. Recording best guess of what was heard

If the participant does not flag the audio recording, they are immediately prompted to record their best guess as to what was said, with the specific prompt of "Repeat the sentence you just heard as best you can." After clicking the Record button, the waveform for the recording is drawn in real

time. The participant may listen to and re-record their response as many times as desired. If a recording is more than eight seconds long, the participant is prompted to record again. When finished, the participant clicks "Submit." Because the HTML5 audio recording specification leaves the choice of sampling rate and bit depth to the client, all recordings are normalized upon receipt by the server to 16 kHz, 16 bit PCM WAV files. The acoustic properties of the recording environments vary between participants in that each participant records their responses in an uncontrolled environment.

4. Self-flagging the new recording

After submitting the recording, the participant is given the opportunity to self-flag the recording, in case of a speech error or other problems such as unexpected background noise. If the participant chooses to self-flag the recording, they are prompted to provide a reason, with the same set of candidate reasons as the upstream flagging procedure in Step 2, above. After providing a reason for self-flagging, the trial ends early.

5. Transcribing the new recording

Finally, if the participant has submitted a recording and has attested that it is of good quality, they are then prompted to provide a written transcription thereof. The inclusion of punctuation or of a misspelled word (as determined with the Linux utility Aspell) returns an error message to the participant, who may then edit the transcription and resubmit. After submitting, the trial ends. The participant is then provided with a button entitled "Click to play next audio recording"; because there is no timer on this screen, they may pause between each trial.

Automated filters

After the participant submits the written transcription, they advance to the start button for the next trial. The web application asynchronously applies a set of automated filters to the audio file and the transcription on the server. We employ these filters to determine if a recording is of sufficient quality to be used as input to other downstream participants. In principle, these filters could be implemented with human participants, but using automated tools allows us to direct participants' effort (and commensurate compensation) towards data collection. Of note from a design and data analysis perspective, these filters must be applied at the time of collection: because future participants hear responses from earlier participants, responses must be filtered in real time to maintain the continuity and integrity of the chain of recordings. These filters include:

- Is the file silent?
- Is the transcription provided by the new participant between 20% longer and 20% shorter (in terms of the number of non-space characters) than the transcription of the input sentence they received? If utterances become too short, they become difficult to characterize with language models.
- Is the transcription provided by the new user more than 2 words longer or 2 words shorter than the transcription of the input sentence? This follows the same logic as above.

A second set of tests pertains to the audio quality and the legibility of the recording. For this we use an automatic speech recognition system to generate a transcription of the received audio file. Specifically, we use one of the most advanced publicly-available automatic speech recognition systems, DeepSpeech (Hannun et al., 2014), to check:

- Is the DeepSpeech-generated transcription of the participant's audio file similar to the transcription they provided? This guards against the possibility that a user would provide an acceptable transcription, but an unrelated audio file (e.g., one filled with obscenities).
- Is the DeepSpeech-generated transcription of the participant's audio file similar to the transcription provided by the upstream participant? This prevents the introduction of material like "I didn't hear the last sentence."

The above checks are operationalized by testing whether the normalized Levenshtein-Damerau distance (Navarro, 2001) between the strings in question is less than some threshold. We use the threshold of .58, arrived at by computing the normalized Levenshtein distance between each sentence in a large corpus and a number of candidate transcriptions, including the correct one and several unrelated foils. On this test corpus with highly dissimilar sentences, this threshold yields a negligible false positive rate; in practice, this threshold is extremely permissive and flags only highly deviant recordings.

If a recording fails any of the above tests, the participant receives feedback at the end of the following trial. This asynchronous evaluation allows for efficient speech recognition (which is very computationally costly and requires the use of a graphics processing unit on the server) and prevents participants from having to wait for the web application to recognize and validate their responses. Utterances that are rejected by any of these filters are retained in that they are potentially relevant to a number of research questions outside of the scope of the current study. If a recording passes the above tests, the recording is combined with a randomly selected interval of cafe background noise (with the same acoustic properties as above) so that it may be used as a stimulus for later participants. The implications of flagging the upstream stimulus, self-flagging, and automated filtering for the

stimuli heard by later participants are described in greater detail below.

After submitting a response for the final stimulus, participants are prompted to take an optional demographic survey detailing age, gender, level of education, current geographical location, proficiency with English, and information regarding previous residence for the purpose of dialectal analyses (outside of the scope of this work).

Stimuli and Experimental Design

Next we describe the initial stimuli and experimental design.

Initial Stimuli

We select a set of 40 initial target sentences and 9 filler sentences from the TASA corpus (Zeno et al., 1995) and the Brown Corpus (Kucera & Francis, 1967). These sentences are chosen to provide maximal variation in probability under different language models for sentences of the same length. First we determine which sentence length (in terms of words and characters) is the most common (10 words, 42 nonspace characters + 9 spaces). All evaluated as grammatical by the experimenters. For this cohort of length-matched sentences, we then obtain their probabilities under unigram and trigram language models trained on the British National Corpus. For the trigram, we used modified Kneser-Ney smoothing (Chen & Goodman, 1998) on transitions of order three (for further details regarding this smoothing scheme, see Language Models below). For both unigram and trigram probabilities, we used the empirical distribution of probabilities for the yielded sentences to generate 20 5-percentile tranches. Then for each tranch, we iterate through sentences, rejecting all

Table 2.1: Example transcriptions of initial stimuli, their unigram and trigram probabilities, percentile ranks, and their character length.

	Unigram		Trigram		
Sentence	Log Prob.	Percentile	Log Prob.	Percentile	Characters
"they found that they had many of the same interests"	-37.28	94	-27.09	92	51
"the molecules that make up the matter do not change"	-38.71	79	-28.51	85	51
"they went on a short hike one warm autumn afternoon"	-42.72	18	-35.89	25	51
"each nonfiction book has a call number on its spine"	-43.56	11	-41.01	5	51

sentences phrased as questions, interpretable as inappropriate, or containing numbers, hyphenated words, or contractions. This yields a single sentence for each of the 20 unigram and 20 trigram tranches. Two additional sentences were chosen from the TIMIT corpus (Garofolo et al., 1993). Initial audio stimuli were read by a male speaker at a normal conversational pace in a soundproof environment.

All chains are initialized with a grammatical, semantically interpretable sentence, but we make no explicit effort to maintain either property over the course of serial reproduction. Because later recordings may not be grammatical sentences, we refer to them as utterances, though they are sentences in a high proportion of cases (see Discussion).

Serial Transmission

The succession of stimuli and responses (the latter constituting stimuli for later participants) can be conceived of in terms of a directed acyclic graph, or DAG. Considering the succession of recordings

for each sentence as a graph—a collection of nodes representing recordings and edges representing the order in which they were collected—allows us to concisely represent the data collected in the course of the experiment and to operationalize the logic for both automated and participant-based flagging of recordings.

Per the specification of the interface above, a recording takes one of five possible states: accepted by fiat because it is an initial stimulus ("protected"), provisionally accepted ("accepted"), flagged by a downstream participant ("downstream-flagged"), flagged by the participant who recorded the sentence ("self-flagged"), or flagged by one of several automated methods ("auto-flagged"). When a participant starts a test trial indexed by a particular initial sentence, they are provided with either the most recent accepted recording or the initial recording itself (if no previous recordings have been accepted). In other words, if a participant flags the input recording they heard, the following participant will then hear the previously accepted recording in the graph. The sequence of recordings appropriate for analysis, or *recording chain*, is then the initial sentence recording and the subsequent succession of accepted recordings. Flagged and self-flagged recordings comprise the complement of the nodes in the graph. We follow the convention established in the iterated learning literature of referring to the sequential position of a recording within a chain as its *generation*, though note that a participant may contribute recordings for different stimuli at different generations, unlike most iterated learning experiments.

The process by which participants are assigned to stimuli can then be considered in terms of *threading*. Each participant must provide recordings for each of the 40 test sentences. Because of the need for strictly successive recordings, only one participant at a time may listen to and record a

response to the same recording. We implement these constraints by implementing a *mutex*, a data structure that limits concurrent access to a resource, in this case recording chains. In combination with multiple independent chains for each stimulus, this setup allows participants to listen and record sentences continuously while maintaining a strictly sequential relationship between the contributed recordings. The three-minute timer on each trial means that the controller will re-assign a stimulus to another participant if a response is not recorded within three minutes. This setup also means that the order in which a participant contributes to each of the recording chains is randomized across participants.

These dynamics mean that any recording may be flagged and removed for the duration of the experiment, even if another participant records a downstream utterance. For example, it is possible that sentence s_1 recorded by participant p_1 is provisionally accepted, and that a succeeding participant p_2 records a downstream repetition, s_2. If, however, p_3 flags s_2, then p_4 will hear s_1, and may in principle flag that recording. We find that such cases of retroactive flagging are relatively rare, but that this mechanism provides an automated method to produce chains of interpretable utterances appropriate for analysis.

This architecture also means that if two participants p_1 and p_2 are progressing through the experiment at the same time, then p_1 may provide the stimulus recording heard by p_2 for some sentences and p_2 may provide the stimulus recording heard by p_1 in others.

Participants and Protection of Human Subjects

n = 266 participants were recruited using Amazon Mechanical Turk. Participants were limited to those living in the United States, with internet access equal to or faster than a symmetrical 512 Kbps connection, and with a microphone attached to their computer. Each IP address and worker identifier (stored locally in hashed form) was allowed to participate only once. The quality of each participant's internet connection was screened at the beginning of the experiment to avoid possible issues with downloading and uploading relatively bandwidth-intensive audio. Data collection methods, including the audio data retention policy, were reviewed and approved by the U.C. Berkeley Committee for Protection of Human Subjects. In addition to providing informed consent, participants also completed a media release allowing their submitted audio recordings (which constitute personally-identifiable information) to be used in publicly-available corpora. With the exception of the audio recordings and the Mechanical Turk worker IDs and IP addresses (stored in a hashed format and discarded after the completion of the experiment), no other personally-identifiable information was collected. Participants were not explicitly told that the recordings that they heard might come from other participants, though the media release states that their recordings could be used as experimental stimuli. The introduction to the experiment stated that the task was designed to gather data on how people recognize words in conditions with high levels of background noise.

2.5 Results

In this section, we evaluate probabilistic generative models of language described above in their ability to predict the changes made by human participants in a large web-based game of "Telephone." We begin by evaluating whether the recording chains—independent sequences of recordings from the game like that shown in Table 2.2—indicate movement towards a consistent set of linguistic expectations. We then evaluate which language models increase the most in probability over the course of the experiment. These analyses focus in particular on whether the degree to which models use preceding context, as well as the way in which they represent that preceding structure, affects their ability to predict the changes made by people. Finally, we analyze which features of word tokens — derived from these language models or from other properties of words — are predictive of their successful transmission from speaker to listener.

Evaluating Movement Towards Convergence

We first evaluate whether the recording chains change in a way that suggests that they are headed towards a single distribution. Convergence among sampling chains in Markov chain Monte Carlo is often evaluated in terms of a *potential scale reduction factor*, or PSRF (Gelman & Rubin, 1992). The PSRF measures the degree to which between-chain variance in parameter estimates reduces with respect to within-chain variance over the course of sampling. In this case, we cannot directly access quantitative estimates of the parameters of the "true" latent model — the linguistic expectations of human participants. Instead we use a proxy measure: whether the probability estimates for in-

Table 2.2: A representative recording chain yielded by serial reproduction. The first transcription is the initial stimulus. Each subsequent transcription is that of a participant who heard the preceding sentence presented in naturalistic background noise. All responses are collected as audio recordings first, then participants are prompted to provide a written transcription.

your teeth begin breaking up the food by chewing it
your teeth begin breaking up the food by chewing it
your teeth end up breaking up the food by chewing
Your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing
your teeth end up breaking up the food by chewing it
her teeth ended up breaking to the food back to you
her teeth ended up breaking as the food got hard
her key ended up breaking off into her car
her key ended up breaking in to her car
our key ended up breaking into the car
Our key ended up breaking into the door
Berkie ended up breaking in to the door
Our key ended up breaking into the door
Our key ended up breaking in to the door
Her key ended up breaking into the door
her key ended up breaking into the door
Her key ended up breaking in the door
her key ended up breaking in the lock
The key ended up breaking in the lock
the key ended up breaking in the lock
The key ended up opening the lock
the key ended up opening the lock
the key ended up opening the lock
The key ended up opening the lock
the key to it is upholding the law

dependent chains become more similar within a model over the course of the serial reproduction experiment. While the probability measures could potentially become progressively more similar for other reasons, this pattern is a necessary condition that the chains are approaching the distribution of interest.

Because the model-based proxy measures above are highly noisy, we aggregate utterance chains into groups. Chains are stratified into four groups on the basis of the probability quartile of the initial sentence as computed by each language model. Initial conditions of the recording chains are known to vary significantly on this dimension, in that initial stimuli reflect a stratified sample based on unigram and trigram probability measures. We take the mean probability within each quartile at each generation (sequential position within the recording chain), and from that compute the inter-quartile variance.

The asymptotic decrease in inter-quartile variance seen in Figure 2.5 suggests that the recording chains are increasingly similar over the course of the experiment under all language models. Figure 2.5 shows that variance drops to less than a quarter of the initial value for all models (with one exception for Big LM at the 25th generation), and around a tenth of the initial value for several of the models trained on large datasets (*n*-gram models trained on the BNC and the DeepSpeech datasets). However, we caution against the stronger assertion that the chains are sampling *directly* from people's expectations because utterances continue to increase in probability at the end of the experiment (see the next section), suggesting that the utterances need to undergo further changes to converge to human expectations. We leave the possibility of sampling directly from the participants' revealed expectations after convergence to future research, and for now use samples generated as participants

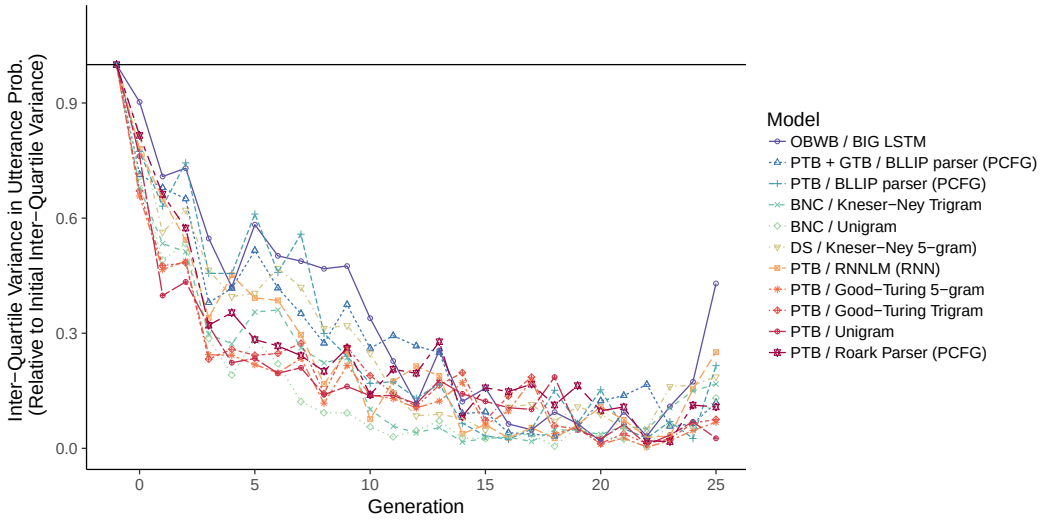


Figure 2.5: Variance in language probability estimates at each generation of the serial reproduction experiment relative to initial variance. Variance estimates are computed with respect to means for four groups of chains, defined by initial sentence probability quartile. The reduction in variance suggests that the utterances in chains starting with high probability sentences and low probability sentences becomes much more similar over the course of the experiment.

approach the distribution of interest.

Evaluating Language Models

Ideally, the language models here could be evaluated using the probability each assigns to a large set of utterances sampled from participants after strong evidence that sampling chains have fully converged with participants' expectations—both that inter-quartile variance approaches zero and that the probability of utterances is no longer increasing. That utterance probabilities continue to increase (treated in further detail below) suggests that these chains are not yet directly sampling from participants' expectations by the end of the experiment.

Instead, we use the guarantee from serial reproduction that the yielded sentences are approaching human expectations in the task, even if they have not yet converged. As such, the pattern of changes

towards the target distribution can be used to evaluate models: more representative models should have a greater magnitude increase in probability (or decrease in average per-word surprisal) over the course of the experiment. In theory, a model reflecting the true expectations of participants would exhibit the largest possible decrease in surprisal over the course of serial transmission. Though we lack access to those expectations or the relevant change in surprisal, the magnitude of the decrease in surprisal for each language model—as an approximation of those human expectations– is sufficient to rank it with respect to others.

As first noted by Bartlett (1932), messages decrease in length over the course of serial reproduction. To eliminate the effect of shorter utterances, which would trivially be assigned higher probabilities under all language models, we divide each utterance's negative log probability by the number of words in that utterance. The resulting measure has an intuitive interpretation as the average surprisal, measurable in bits, for each utterance under each model. The average of this measure across chains over the course of the experiment is shown in Figure 2.6.

Models vary in their surprisal estimates for reasons outside of the scope of the current analysis, especially the choice of smoothing scheme. For example, consider two unigram models that withhold different amounts of probability mass for word tokens not encountered during training, e.g., .05 and .01. If these two models were used to produce probability estimates for a set of sentences comprised of exclusively known tokens, the first model would assign a lower probability estimate compared to the second model, even though the probability estimates would be perfectly correlated. We thus focus not on the intercept for each model but the slope of the change in surprisal, taking advantage of the fact that the change in the number of bits ($\log_2 p(w)$) corresponds to a constant

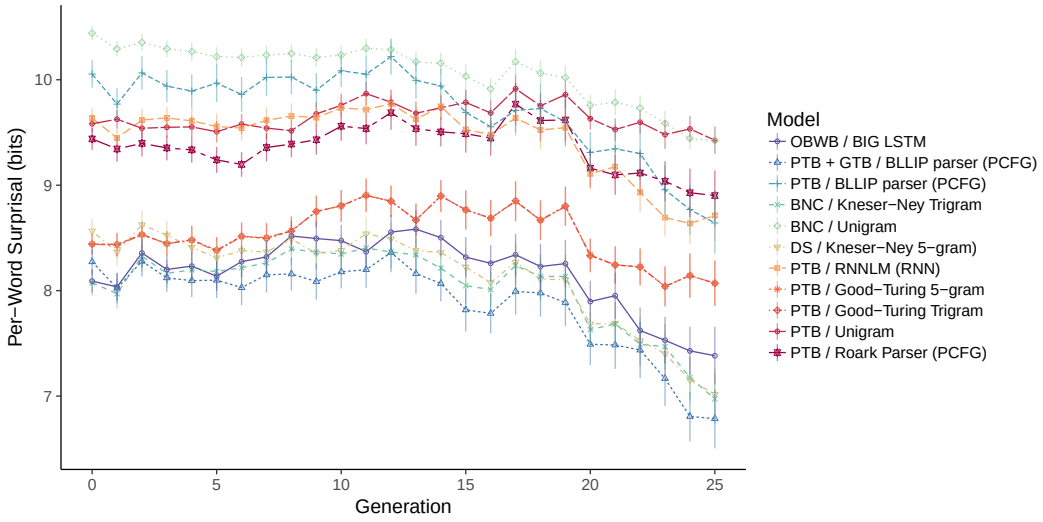


Figure 2.6: Average per-word surprisal in bits ($-\log_2(p(w|c))$) under each language model over the course of serial reproduction. Error bars indicate standard error of the mean.

multiplicative factor in probabilities (an increase of one bit, whether it is the third or the seventh, means that the words in the set of utterances are, on average, half as probable).

Average per-word surprisal under each model for recording chains through 25 sequential transmissions is plotted in Figure 2.6. The change in average per-word surprisal (in bits) is plotted in Figure 2.7. This latter graph reveals that higher-order *n*-gram models trained on large datasets and the BLLIP PCFG parser trained on the Penn Treebank best capture the changes observed in the course of serial reproduction. The observed pattern of increases in probability corresponds with the absolute model probabilities, in that these models are the same ones that assign the highest probability to the final generations in the sampling chains. *n*-gram models and the Roark parser (trained on the same TreeBank dataset) show statistically significant — though more modest — increases in probability. Big LM, arguably the most sophisticated publicly-available deep neural network-based

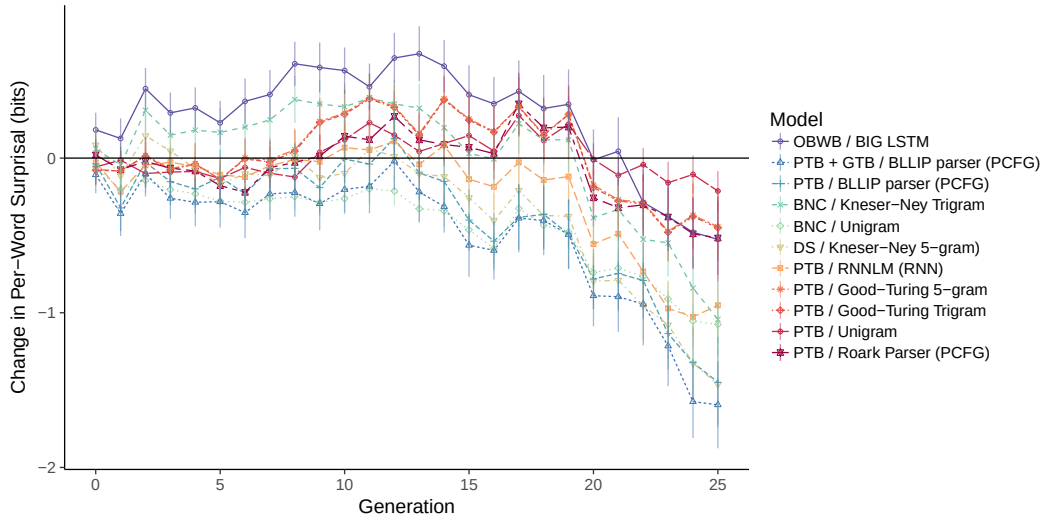


Figure 2.7: Change in the average per-word surprisal for utterances over the course of the telephone game under 11 language models. A decrease in surprisal of one bit means that a word is twice as probabile under a model.

language model, exhibits higher surprisal estimates for much of the experiment, a point we return to in the Discussion.

We next turn to the question of whether there are differences in performance between broad classes of models as a result of theoretically-relevant architectural distinctions, in particular whether models that use preceding context are more representative of the revealed linguistic expectations than the unigram model. Among the models that use preceding context, we evaluate whether those models that use higher-order abstract representations more accurately reflect the changes made in the course of the serial reproduction experiment.

To confirm that the identified theoretical contrasts (usage of context and usage of abstract higher-order representations) are indeed reflected in the probability estimates produced by the models, we first measure the similarity between language models on the basis of the probability estimates they

provide for all $n = 3{,}193$ utterance transcriptions yielded by the experiment. If the model performance is sensitive to these features, correlations should be strongest within the partitions identified above. Because the form of the relationship may not be linear, we use Spearman's rank correlation coefficient to evaluate the pairwise similarity. We limit this analysis to models trained on the Penn TreeBank to eliminate the effect of training dataset on model performance.

The results of this analysis yield two major clusters, the *n*-gram models, which do not use higher-order representations for prediction, and the RNN-LM and the two PCFG models which do (Fig. 2.8). Among the *n*-gram models, the unigram model is distinguished from the higher-order *n*-gram models; these longer-context *n*-gram models are more similar than the unigram model to the models with higher-order abstract structure. The yielded similarities show that the predictions derived from these models reflect the key architectural distinctions of theoretical interest regarding the use of context. Further, these results substantiate our claim that recurrent neural network language models are better grouped with the PCFG models than the *n*-gram models given their representational capatities, in contrast to the classification used in Frank & Bod (2011) and Fossum & Levy (2012).

Having demonstrated that the model-based probability estimates are sensitive to the architectural differences of interest, we then examine whether the tracked utterances see a more pronounced increase in probability—a more pronounced decrease in average per-word surprisal—under those models that use preceding context. This distinction pits the unigram *n*-gram models (which do not use preceding context to inform expectations) against the remainder of the models (which do). We construct a mixed-effects linear regression model with the following predictor variables: dataset (Penn TreeBank vs. large representative dataset), context (used vs. not used), and the interaction

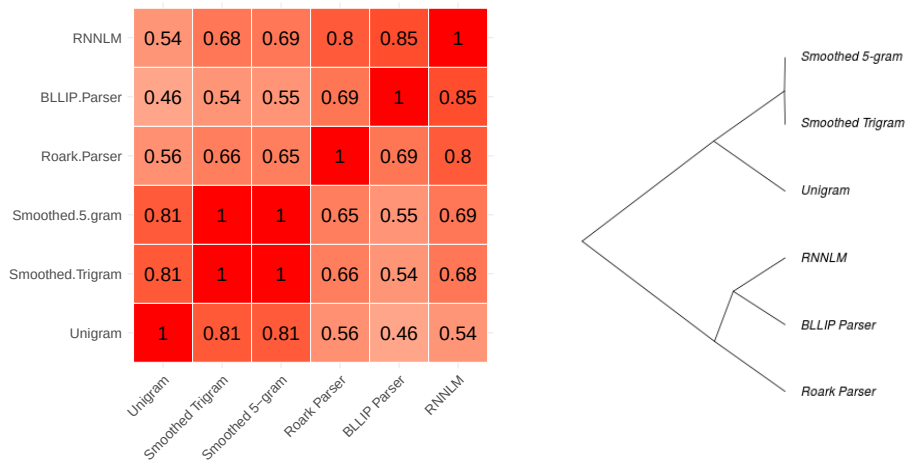


Figure 2.8: *A.* Spearman's rank correlation for pairwise combinations of sentence probabilities across the 3,193 sentences from the experiment. *B.* The resulting dendrogram, derived from the rank correlations using Ward's method, shows that the model-based probability estimates reflect the architectural differences of interest.

of both preceding terms with generation (sequential participant in the recording chain). For the outcome, we use average per-word surprisal. In that the unigram model is the sole model that does not use context, we do not use a more specific designation of model architecture (e.g., Trigram vs. 5-gram vs. Roark) as a predictor. Unique recording chains (independent sequences of utterances from the Telephone game) are treated as random intercepts, and we account for a ***recording chain*** × ***generation*** random slope. The model is fit with 11 model estimates of per-word surprisal for each of 2,864 utterances produced between the 1st and 26th participant (thus omitting the initial sentences).

First, we evaluate whether accounting for each language model's use of preceding context as a factor improves the overall fit of the mixed-effects model by comparing the above full mixed-effects model with a nested one lacking the use of preceding context as a predictor. This reveals that the

Table 2.3: A mixed-effects linear regression examining average per-word surprisal estimates (negative log probability under a model) across 11 language models as a function of 1) whether the model uses preceding context to inform expectations and 2) the dataset used to fit the model. Significance of fixed-effects is computed following Satterthwaite (1946).

Fixed Effects				
	Coef β	SE(β)	t value	p
(Intercept)	2.9076	0.025	116.42	<.0001
generation	-0.001	0.0023	-0.43	0.7
context: used	-0.3381	0.0104	-32.59	<.0001
dataset: PTB	0.1619	0.008	20.14	<.0001
generation x context: used	-0.0026	8e-04	-3.16	<.01
generation x dataset: PTB	0.0059	6e-04	9.16	<.0001
Random Effects				
	Std. Dev			
(Intercept) | Recording Chain	0.32			
generation | Recording Chain	0.03			

model that includes use of context and a use of context x generation interaction exhibits significantly better fit (χ^2 = 3792.2, p <.0001) For the model including the use of context as a factor (Tab. 2.3) there is a statistically significant negative coefficient for ***generation*** × ***context: used*** (β = -0.0026, t value = -3.16, $p < .01$), indicating a greater magnitude decrease in surprisal estimates over the course of the experiment for those language models that take preceding context into consideration. A positive coefficient for ***generation*** × ***dataset: PTB*** indicates that language models that were trained on the Penn TreeBank, a relatively small news corpus, are less representative of people's expectations than the larger language models (β = 0.0059, t value = 9.16, $p < .0001$).

The second architectural question we address is whether models that represent higher-order regularities, such as phrase structure or abstract semantic representations of the preceding context, better

reflect the revealed linguistic expectations of participants than models that track only specific preceding words. The n-gram models (where $n > 1$) use only lexical representation of preceding context, and do not posit any sort of higher-order abstractions such as verb phrases or noun phrases. In contrast, the PCFGs explicitly represent phrase structure, and recurrent neural network language models are sensitive to higher-order syntactic and semantic regularities insofar as these are captured in the lower-dimensional embedding of preceding context. Similar to the approach taken for evaluating the utility of preceding context, we construct a mixed effects linear regression model with the following predictors: dataset (Penn TreeBank vs. large representative dataset), abstract representation of context (used vs. not used), and the interaction of both preceding terms with generation. As the outcome measure, we again use average per-word surprisal. The mixed-effects model uses the same random effects structure and is fit with the same set of utterances as above, but exclude measures from the two unigram language models, which do not track preceding context.

We first test whether accounting for each language model's type of representation (abstract vs. not) improves the overall fit of the mixed-effects model by comparing the above full model specification with a nested model lacking type of context representation as a predictor. This reveals that the model that includes the type of context representation as well as its interaction with generation exhibits significantly better fit (χ^2 = 1401.1, p <.0001) For this full model (Tab. 2.4), there is a statistically significant negative coefficient for the interaction for ***generation*** * ***abstract structure: used*** (β = -0.0044, t value = -6, $p < .0001$). This greater magnitude decrease in surprisal estimates suggest that the changes made by people are better reflected by the language models that use abstract representations of preceding context. As with the previous analysis, we find that models that are trained

Table 2.4: A mixed-effects linear regression examining average per-word surprisal estimates (negative log probability under a model) across 9 language models as a function of 1) whether the model uses an abstract representation of the preceding context to inform expectations and 2) the dataset used to fit the model. Significance of fixed-effects is computed following Satterthwaite (1946).

Fixed Effects				
	Coef β	SE(β)	t value	p
(Intercept)	2.3362	0.0272	85.81	<.0001
generation	0.001	0.0025	0.38	0.7
abstract structure: used	0.2382	0.0092	25.8	<.0001
dataset: PTB	0.2923	0.0088	33.38	<.0001
generation x abstract structure: used	-0.0044	7e-04	-6	<.0001
generation x dataset: PTB	0.0042	7e-04	6.03	<.0001
Random Effects				
	Std. Dev			
(Intercept) | Recording Chain	0.35			
generation | Recording Chain	0.03			

on the Penn TreeBank are less representative of people's linguistic expectations than models trained on larger datasets ($\beta = 0.0042$, t value $= 6.03$, $p < .0001$).

Predicting Word-Level Errors

Finally, we evaluate whether model-derived probability estimates, in combination with other features of words, are sufficient for predicting which words are misheard: altered or deleted in transmission.[†] Instances of deletions and substitutions are identified using dynamic programming, using the edit operations corresponding to the Levenshtein distance, or minimum edit distance between

[†]While we also collect data regarding insertions in the course of serial reproduction, it is harder to identify the relevant properties of the preceding recording that prompt the insertion of material.

utterances. As is commonly done to estimate Word Error Rate (WER) in natural language processing (Popović & Ney, 2007), this computation is applied to word sequences rather than character strings (Fig. 2.5). We then construct a mixed-effects logistic regression model using features of words including model-based estimates of surprisal, position in sentence, age of acquisition ratings (Kuperman et al., 2012a), concreteness (Brysbaert et al., 2014a), number of phonemes, number of syllables (Brysbaert & New, 2009), and phonological neighborhood density (Yarkoni et al., 2008) as fixed effects to predict whether the word changes. Because of the high level of correlation between language models, we use a residualization scheme to identify their respective contributions. Unigram surprisal (negative log probability) is used directly as a predictor. We then predict trigram probability from unigram probability, and use the residuals as a predictor representative of the contribution of a language model which considers the two words preceding the words in question as the preceding context. For each of the remaining models with word-level surprisal estimates (5-gram on the DeepSpeech dataset, BigLM on the One Billion Word Benchmark, the Roark Parser trained on the Penn TreeBank) we take the residuals after predicting its surprisal values from both unigram and trigram models.

The identities of the listener and the identity of the speaker are treated as random intercepts to account for variability in comprehension performance and speaker intelligibility. The model is fit with 27,290 instances where a participant heard a word and either 1) reproduced it faithfully in their own recording (22,482 cases) 2) produced an identifiable substitute (substitution) or 3) did not produce an identifiable substitute (deletion). Cases 2) and 3) were collapsed into a single category of transmission failure. We fit the full model and conduct no pruning of predictors.

Table 2.5: Example edit string from input sentence to output sentence. M, D, I, and S indicate Match, Deletion, Insertion, and Substitution respectively.

M	M	M	M	D	I	I	M	D	M	S	M
you	may	not	notice	yourself			grow	from	day	to	day
you	may	not	notice		as	you	grow		day	by	day

This model shows that words with higher unigram surprisal are more likely to change (β = 0.3827, z value = 17.25, $p<$ .0001), as are words with higher trigram surprisal with unigram surprisal partialed out (β = 0.3729, z value = 19.37, $p<$ .0001). For all models with abstract structure (with trigram surprisal partialed out), higher residualized surprisal estimates are predictive of a transmission failure. Words are more likely to change if they are appear later in the utterance (β = 0.0989, z value = 15.57, $p<$ .0001)or if that word is acquired later in development (β = 0.0544, z value = 3.96, $p<$ .0001). By contrast, words with more syllables or more phonemes are *less* likely to change (β = -0.249, z value = -4.91, $p<$ .0001), as are words rated as highly concrete (β = -0.0856, z value = -4.46, $p<$ .0001). Finally, words in sparse phonological neighborhoods — words with high average phonological Levenshtein distance to the 20 most similar competitors (or PLD-20, per (Yarkoni et al., 2008)) are less likely to change (β = -0.0948, z value = -2.22, $p<$ .05). We return to these results in greater detail in the Discussion.

The fit model can be used to estimate the probability of change for each word in the dataset. The area under the ROC curve for the above model, .728, indicates that for a large sample of randomly chosen pairs of words in which one word was successfully recovered by the listener and the other was not, this model assigns a higher probability of change to the word that changed 72.8% of the time. Examples of utterances with each word colorized by the probability of successful recovery by

Table 2.6: Mixed-effects logistic regression predicting whether a word will be transmitted successfully on the basis of its surprisal under various language models as well as other word properties. Significance of fixed-effects is computed following Satterthwaite (1946).

Fixed Effects				
	Coef β	SE(β)	z value	$Pr(>\|z\|)$
(Intercept)	-2.6554	0.0933	-28.45	<.0001
BNC unigram surprisal	0.3827	0.0222	17.25	<.0001
Residualized BNC trigram surprisal	0.3729	0.0193	19.37	<.0001
Residualized Roark PCFG syntactic surprisal	0.1159	0.0266	4.36	<.0001
Residualized Big LM surprisal	0.1959	0.0181	10.82	<.0001
Residualized DS 5-gram surprisal	0.1593	0.0262	6.07	<.0001
Position in sentence	0.0989	0.0064	15.57	<.0001
Age of acquisition	0.0544	0.0137	3.96	<.0001
Number of phonemes	-0.0609	0.0231	-2.64	<.01
Number of syllables	-0.249	0.0507	-4.91	<.0001
Concreteness	-0.0856	0.0192	-4.46	<.0001
Phonological Neighborhood Density (PLD20)	-0.0948	0.0427	-2.22	<.05
Random Effects				
	Std. Dev.			
Listener ID	0.62			
Speaker ID	0.3			

the listener are shown in Fig. 2.7.

2.6 Discussion

In this work, we investigate several probabilistic generative models of linguistic structure in their ability to capture people's linguistic expectations. A serial reproduction task — the game of Telephone, where participants reproduce recordings made by other participants in sequence— reveals

Table 2.7: Probability that each word is successfully recovered by an average listener under a mixed effects logistic regression model. Red indicates that a word is likely to be misheard (resulting in a substitution or deletion), while green suggests that a listener is likely to reproduce the word successfully. Sentence are from the set of initial stimuli.

a dietitian goes to college for at least four years
the iris absorbs all of the light waves except blue
some acids you may know are vinegar and lemon juice
the chase leads across a field toward a nearby farm
your teeth begin breaking up the food by chewing it
the raspberry leaves are not very tasty to a rabbit
a fly buzzed over the oilcloth on the kitchen table
the brain helps all parts of the body work together
often the village was burned to the ground by fires
goods are exchanged in the market place of an oasis
you may not notice yourself growing from day to day
the discovery of oil has caused many cities to grow
they can read the label and use the medicine safely
a county may have several towns or cities within it
now the plane was going one thousand miles an hour
the molecules that make up the matter do not change
the captain closed the door behind us and bolted it
the third and fourth waves seemed to be the highest
how do you know the difference between hot and cold
meadow mice and gophers eat the roots of some weeds

participants' prior expectations in a naturalistic spoken word recognition task. We find evidence that people use preceding linguistic context to inform word recognition. Further, we find evidence that people use abstract representations of that context to inform their expectations, in line with the results of Fossum & Levy (2012) for reading.

The Role of Preceding Context

The role of prediction in sentence processing has inspired significant debate, with recent work (Nieuwland et al., 2018) challenging long-standing experimental evidence of people's use of prediction in word recognition. The serial reproduction experiment conducted here can be used to evaluate evidence for prediction in word recognition, insofar as we understand prediction as the use of data independent of the received acoustic signal for a word (Kuperberg & Jaeger, 2016). While both accounts predict that transmission failures (deletions and substitutions) are more common for improbable words (insofar as they are both unpredictable and implicate less prototypical world states), the accounts are distinguished in what material they posit might *replace* words in the case of communicative failures. Prediction-centric accounts suggest that people will replace words that are not transmitted successfully with words that have a higher probability of occurring under people's linguistic expectations. Integrationist accounts, by contrast, make no clear prediction as to what should replace these mis-recognized words. We use frequency (in the form of unigram surprisal) as a baseline for integrationist accounts.[‡] Our results showing that utterances increase in probability faster over the course of the experiment under models that condition on preceding linguistic context are consistent with accounts that posit a key role for prediction in word recognition, and are not well-explained by integration-centric accounts.

Indefinite articles comprise a substantial portion of the words in the serial reproduction experi-

[‡]One might imagine that an integrationist account would assert that people tend towards easy-to-integrate words in such situations. However, we argue that this would be isomorphic to the prediction-based account, in that this would implicate the same sort of prior (data-independent) knowledge.

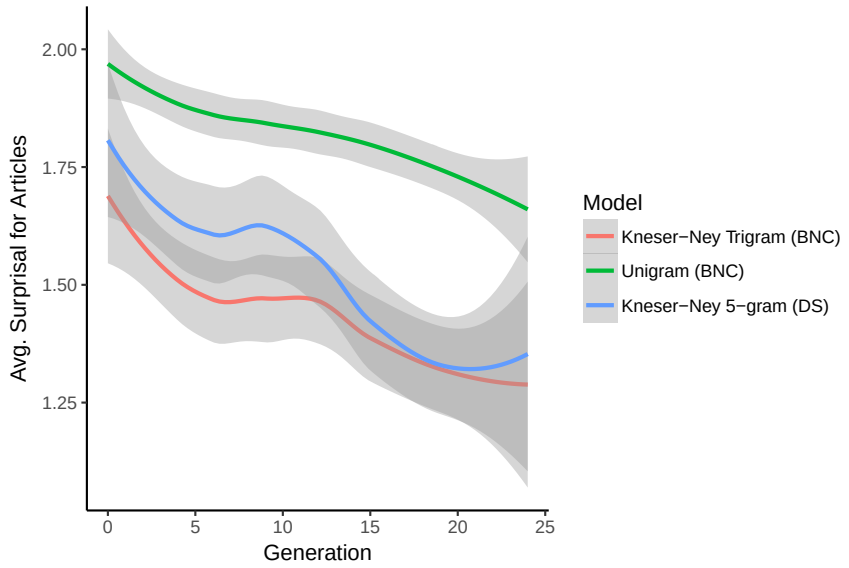


Figure 2.9: Log likelihood ratio of the set of indefinite articles under three language models, reflecting the probability at each generation relative to their probability at convergence. The higher rate of increase for the models that use preceding context suggest that the composition of articles is sensitive to prediction on the basis of preceding context.

ment (2.6%), permitting an analysis of items in the test case highlighted by DeLong et al. (2005) and Nieuwland et al. (2018). We limit this analysis to the largest *n*-gram models, in that this provides for a direct comparison of the utility of preceding context, independent of the degree of abstraction. We follow a similar analytical approach as the analysis of context in the Results section, but limit the scope of the analysis to just individual indefinite articles rather than whole sentences. As above, the specific model is not represented in addition to context use; recording chains are treated as random intercepts, with a ***recording chain*** × ***generation*** random slope.

This analysis (Fig 2.9; Table 2.8) shows that the per-word surprisal (negative log probability) estimates for articles alone decreases faster over the course of the experiment for language models that condition on preceding context (β = -0.0068, t value = -2.33, $p < .05$). This finding is consistent

Table 2.8: Mixed-effects linear regression predicting average surprisal among articles at each generation as a function of whether the models track preceding context.

Fixed Effects				
	Coef β	SE(β)	t value	p
(Intercept)	1.9404	0.0646	30.04	<.0001
context: used	-0.2613	0.0359	-7.27	<.0001
generation	-0.0045	0.0048	-0.92	0.4
context: used x generation	-0.0068	0.0029	-2.33	<.05
Random Effects				
	Std. Dev			
(Intercept) | Recording Chain	0.46			
generation | Recording Chain	0.02			

with the analysis of all words presented above, and suggests that people are changing articles as a function of their in-context predictability.

We note, however, that there is no reason to believe in the mutual exclusivity of integration-based and prediction-based difficulty in word recognition and sentence processing beyond the usual desire for parsimony. The analysis presented here is only capable of evaluating evidence for prediction, and cannot dismiss the possibility of additional integration-related difficulty, either in word recognition or when listeners construct fine-grained representations of meaning in the course of sentence processing.

The Role of Abstract Structure

Before further interpretation of the results regarding the utility of abstract structure, we clarify the exact nature of our claims and note several methodological challenges that temper the strength and generality of these null results.

First a top-level theoretical clarification is in order: Higher-level abstractions are thoroughly implicated in the processes of sentence comprehension and production (e.g., verb argument structure). Rather, our objective in this study is to investigate whether such abstractions may inform the lower-level process of word recognition.

Second, we highlight an inherent brittleness in using a small cohort of probabilistic generative models to characterize human knowledge more generally: the encoded expectations reflect a complex interaction of architecture, training data, and fitting procedure. Limited explanatory power for human performance may reflect any of one of the above aspects, or a complex interaction in between them. Further, each language model is drawn from a larger space of possible language models, such that the generalizations we make about model architectures on the basis of a few examples may not be robust.

On the matter of fitting, the encoded expectations may reflect local minima or maxima. While this is not a concern with *n*-gram models fit with count-based methods, this problem increases in severity for language models which have large numbers of randomly-initialized parameters, especially the neural network language models. For these models, continued research focuses on how to improve the speed and robustness of the training procedure. All models may suffer from the

problem of overfitting (Geman et al., 1992), in which learned distributions reflect properties of the training data at the expense of their generality in extending to new data. The relationship between corpus size and performance is modulated by architecture: Large models trained on small datasets may be more prone to overfiting. More abstract structure may allow models to perform better on smaller datasets because they can fall back on expectations for higher-level abstractions in the absence of experience with specific words or sequences. But smaller training corpora may also make it harder for models to arrive at useful abstractions in the course of fitting.

We took several steps to address these pitfalls regarding model performance. First, we confirmed that all models exhibit characteristic levels of performance on standard test datasets, suggesting that the fitting procedures used here are congruent with previous benchmarks. Though we note above the broader shortcomings of evaluating language models in terms of the perplexity on a held-out dataset, perplexity is nonetheless useful insofar as it allows us to check whether the fitting procedure yields models comparable or equivalent to those used in other studies. Second, we treat the dataset on which each model is trained as a separate predictor in our analyses (binarized into those trained on the Penn TreeBank and those trained on large datasets known to yield highly performant models). This helps isolate the contribution of the model architecture apart from the dataset on which the model was trained. Third, we note that deficiencies in model fitting would most likely affect the more sophisticated neural network models or PCFG parsers, which would yield a null result rather than the positive one in favor of structure obtained here.

Another caveat is that the set of models investigated here represents only a small sample of possible models. Sampling a larger set of models may reveal that the differences in fit to human behavior

that we find for model architectures are not robust, or that other distinctions in model architecture are predictive of larger differences in performance. We acknowledge this limitation of the current study, and note the possibility of evaluating a large set of models on the basis of a larger space of architectural features in future work.

Though the analysis yielded a statistically negative coefficient for ***generation*** × ***abstract structure: used***, we note that this is a relatively minor quantitative effect. Several of the *n*-gram models, such as the DeepSpeech Kneser-Ney 5-gram and the BNC trigram model, exhibit a pattern of surprisal estimates that is very similar to models with considerably more sophisticated representations of abstract structure (BLLIP and Big LM). This pattern of near-parity between certain large, higher-order *n*-gram models and more sophisticated generative models can be explained by a recent theoretical proposal regarding processing difficulty that posits graded use of detailed structural representations in people's linguistic expectations. *Noisy-context surprisal* suggests that while people may use structured, abstract representations of the preceding linguistic context, that such representations are imperfect and in particular tend to degrade the longer they are kept active in memory (Futrell & Levy, 2017).

Futrell & Levy (2017) highlight the case of *structural forgetting effects*, where people do not effectively use preceding grammatical structures to predict the remainder of the sentence. For example, among the two utterances,

1. *The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_1$ well-decorated.
2. The apartment$_1$ that the maid$_2$ who the cleaning service$_3$ had$_3$ sent over was$_2$ cleaning every week was$_1$ well-decorated.

the first is grammatically ill-formed in that no verb corresponds with noun phrase with the head "maid$_2$", yet people give it consistently higher grammaticality ratings than the second sentence (originally presented in Gibson & Thomas, 1999). This effect, they argue, arises from an "information locality" effect, whereby structural expectations—which should give infinitely more probability mass to the grammatically correct sentence than the incorrect one—are attenuated for material with longer intervening intervals (in terms of words or time). Under this account, people have fleeting access to the parse trees that they form in sentence processing. *n*-gram models can be thought of as an approximation of as an abstract, structured generative model, but one with a particularly sharp memory decay function such that an extremely limited sample of the preceding context is used to predict the identity of the next word.

The memory-based effect identified by Futrell & Levy (2017) may be further exacerbated by high levels of background noise in the current experiment, such that we see relatively small advantages of abstract structural representations. While imperfect memory imposes noise on the representation of context even under optimal acoustic conditions, this decay may be yet stronger if participants lack peaked estimates regarding what actually constitutes the preceding context.

An Error Model for In-context Spoken Word Recognition

Data on word-level changes in the course of serial reproduction permits an analysis of which features of individual word tokens make them more or less likely to be recovered successfully by a listener. To our knowledge, this is the first study to collect this data using recordings from other participants in an experiment as stimuli. Our analysis shows a predictive utility for the unigram model, as well as

the information uniquely encoded by each model with progressively longer or more abstract representations of context.

The finding that the age of acquisition of a word is predictive of successful transmission—earlier-acquired words are more likely to be recognized than later-acquired ones—extends previous results from isolated visual word recognition (Morrison & Ellis, 2000) into the auditory realm. This is consistent with the hypothesis that such words enjoy a privileged status above and beyond their frequency, perhaps relating the problem of retrieving semantic representations (Austerweil et al., 2012) to the compositional structure of the lexicon (Vincent-Lamarre et al., 2016).

A second interesting finding is that once the relationship between word length and unigram surprisal is accounted for, longer words are *more* likely to be recognized. This can be interpreted as evidence that people are better able to recognize words that are more perceptually distinctive, in that words with more syllables have have fewer perceptually similar competitors, above and beyond edit-distance-based measures of neighborhood density.

Limitations of Serial Reproduction

A potential caveat to the generality of the results is that the expectations implicit in the utterances obtained from the Telephone game may be task-dependent, and of limited utility for characterizing linguistic expectations more generally. For example, participants could infer that the task they are performing is qualitatively unlike "normal" language use, and make use of a different set of expectations such that the collected data is not representative of the distributions of interest for language processing. It would be particularly concerning, for example, if the colelcted utterances demon-

strated marked decreases in grammatical acceptability or semantic interpretability over the course of serial reproduction. To evaluate whether participants produced grammatical and semantically-interpretable recordings for the duration of the experiment, authors S.M. and S.N. conducted a follow-up analysis in which they independently coded all utterances from generations 23-25 with binary judgments of grammaticality and semantic interpretability. The latter category was used to distinguish sentences that are structurally well-formed but not interpretable, e.g., "the bus and bus driver were opening the door." Utterances in this set were judged as 87% and 88% grammatical (Cohen's $\kappa = 0.94$) and 78% and 79% semantically interpretable (Cohen's $\kappa = 0.89$). The results of this analysis suggest that participants largely maintain the grammaticality and semantic well-formedness of their responses through the course of the serial reproduction experiment, and suggest that participants are tapping into a similar set of expectations as normal language use.

Another important possibility is that participants could be modulating how they use preceding context in word recognition based on the level of noise in the experiment. Because they may be unsure of the preceding context for a particular word, they may prefer shorter, less-structured representations of context in the current experiment, whereas they might rely on that context more heavily in a noise-free environment. This basic logic of noise-modulated expectations is substantiated by the finding of Luce & Pisoni (1998) that participants' reliance on word frequency in isolated spoken word recognition increases as a function of the level of background noise. Audio stimuli here are embedded in relatively high levels of noise, qualitatively dissimilar to the reading tasks of (Frank & Bod, 2011) and (Fossum & Levy, 2012). We highlight the importance of characterizing variation in the use of linguistic expectations as a function of environmental noise as an important next step with this

paradigm.

Finally, we emphasize the limitations in the generality of these results with respect to individual variation among speakers of English, as well as variation across speakers of different languages. This experiment makes the strong simplifying assumption that the process of serial reproduction yields samples from a single, unified set of linguistic expectations that are shared across all English-speaking participants for the purposes of word recognition. Of course, linguistic expectations should vary significantly as a function of linguistic experience, and may vary between speakers for other reasons like population-level variability in working memory. At a higher level, the expectations of English speakers are certainly not representative of the expectations of speakers in other languages. Given the pronounced typological diversity of languages, expectations may take qualitatively different forms. For example, listeners may rely less on sequential word order in languages with more flexible word order. Future work will be needed to characterize the ways and extent to which expectations vary across natural languages.

2.7 Conclusion

In this study, we collect data on how utterances change in the course of a web-based game of "Telephone." We use this data, which better represents people's linguistic expectations for in-context spoken word recognition than existing corpora, to evaluate a broad range of probabilistic generative models of language. Models that use preceding linguistic context to inform expectations regarding upcoming words are more strongly reflect the changes made by people; further, we find that

the changes people make reflect a larger magnitude increase in probability under language models that use abstract representations of preceding context versus those that do not. These results shed light on contemporary theoretical debates in word recognition and sentence processing, while the paradigm offers greater promise in helping to better understand humans' remarkable language processing abilities.

3

Wordforms—Not Just Their Lengths—Are Optimized for Efficient Communication

The question of whether there exist features or properties shared across all languages, or *linguistic universals*, has received significant attention in linguistics, psychology, and cognitive science. Indeed many candidate commonalities have been identified (Greenberg, 1963), but their status as "univer-

sals" remains controversial (Evans & Levinson, 2009). But even if these commonalities fall short of truly universal status, their prevalence still demands explanation. One possibility is that these commonalities arise from inductive biases shared among language learners (Culbertson et al., 2012). Another (mutually-compatible) possibility is that these commonalities reflect "design principles" that languages must adhere to in order to effectively serve the purpose of communication (Evans & Levinson, 2009). Under this latter account, cross-linguistic regularities may emerge from pressures exerted on languages, for example the need to robustly transmit ideas from speakers to listeners, or the impetus to minimize articulatory effort on the part of speakers.

In this section, I investigate the well-known correspondence in natural languages between the frequency of words and the length of the corresponding wordforms (Zipf, 1935, 1949; Bentz & Ferrer-i-Cancho, 2016). Specifically, I endeavor to motivate this broad commonality by linking it to cognitive mechanisms implicated in language processing. I approach this problem using a probabilistic generative language model (PGLM), in this case to characterize the amount of information conveyed to a listener by a wordform. This model measures the degree to which a word's sound sequence deviates from a listener's expectations for words in their language, operationalized as its probability under a simple *n*-gram model of the phoneme sequences in the lexicon. Following Shannon (1948), I consider this probability in terms of the quantity of information it conveys, characterizing words in terms of their *phonological information content* (PIC).

This simple probabilistic model is simultaneously motivated by Bayesian models of spoken word recognition (Luce & Pisoni, 1998; Norris & McQueen, 2008) and measures of phonotactic well-formedness (Jusczyk et al., 1994). I show that this treatment of information content includes Zipf's

"Law of Abbreviation" (Zipf, 1935, 1949) as a special case, but that explicitly accounting for a language's phonotactics (learnable from linguistic corpora) accounts for significant additional variance in word frequency. Though the representation used by the model is a coarse approximation of human linguistic knowledge, the yielded pattern is clear: for words of the same length, the less frequent one is likely to be composed of less prototypical sounds and sound sequences. This pattern proves robust across a broad sample of large linguistic datasets, and significantly surpasses baseline correlations between length and frequency in almost all cases. I also examine how PIC relates to a word's average in-context predictability, which has been demonstrated to correlate more strongly than frequency with word length (Piantadosi et al., 2011).

The kernel of this chapter — language-wide evaluation of the correspondence between phonological information content and lexical surprisal in English — was presented at the CUNY Sentence Processing conference in 2015. Results for other languages and other analyses are as of yet unpublished. It was co-authored by Thomas L. Griffiths (U.C. Berkeley). Special thanks to Steven Piantadosi for sharing materials, helpful commentary on early drafts from Terry Regier, Susanne Gahl, and Keith Johnson, and members of the Computational Cognitive Science Lab at UC Berkeley for valuable discussion. This material is based upon work supported by the US National Science Foundation Graduate Research Fellowship under grant no. DGE-1106400 and NSF grant no. SMA-1228541.

3.1 Introduction

While natural languages are highly diverse in many respects, they display striking structural regularities (Greenberg, 1963; Evans & Levinson, 2009; Futrell et al., 2015). How these structural regularities relate to human cognition — especially whether they shape cognition or vice versa — remains an open question with implications for linguistics, psychology, and neuroscience (Hauser et al., 2002; Evans & Levinson, 2009; Kemp & Regier, 2012; Fedzechkina et al., 2012). One of the most robust statistical laws that describe human languages is the relationship between word length and frequency, often called Zipf's *Law of Abbreviation*: frequently-used words tend to be short (Zipf, 1935). To date, this basic relationship has been demonstrated to hold in all of the approximately one thousand languages that have been tested, with no known counter-examples (Bentz & Ferrer-i-Cancho, 2016).

Despite its ubiquity, critical questions remain regarding the underlying cause of the Law of Abbreviation. Zipf originally posited that this pattern emerges from speakers' desire to minimize articulatory effort to the degree possible by using the shortest form for words that are used most often, following what he later labeled the *Principle of Least Effort* (Zipf, 1949).* While the underlying cause of Zipf's Law (i.e., the correspondence between frequency and frequency rank) has been the subject of extensive debate (Yule, 1944; Miller, 1957; Mandelbrot, 1954; Ferrer-i-Cancho & Solé, 2003; Piantadosi, 2014), the Law of Abbreviation has received less attention (though see Ferrer-i-Cancho

*These competing pressures can be traced back to the late 19th century to the opposition between "striving for ease" (*Bequemlichkeitsstreben*) and "striving for clarity" (*Deutlichkeitsstreben*)" identified by von der Gabelentz (1901; translation in Haspelmath, 1999).

2016). This is surprising given the centrality of frequency effects in language processing (Baayen et al., 2016), and in particular the relevance to key theoretical questions regarding systematic variation in language known broadly as *reduction*, where speakers deviate from standard wordforms (or multi-word constructions) either through omission, shortening, or other variations that reduce articulatory effort (Aylett & Turk, 2006; Bell et al., 2009; Gahl et al., 2012). Jaeger & Buz (2017) summarize evidence that pressures for both articulatory economy (towards reduced forms) and robustness (limiting reduction, to avoid ambiguity and thus potential communicative failure) operate simultaneous in the case of reduction, and point future research towards investigating the complex interactions between these two forces. Precisely this dynamic in reduction — economy vs. robustness — can be shown to produce the empirical signature of the Law of Abbreviation in a lab-based experiment: Kanwal et al. (2017) show that a pattern qualitatively similar to Zipf's law of abbreviation *only* emerges under the simultaneous presence of pressures to both minimize communication time and maintain communicative robustness.

Critical to the development of a causal understanding of Zipf's Law of Abbreviation – either in relation to reduction or to other pressures — is a better understanding of the functional form of the relationship. Zipf's original observation in (1935) that "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" leaves open the question of the precise relationship between frequency and length; Ferrer-i-Cancho (2016) similarly focuses on measures of correlation that do not impose a functional form. Characterizing the functional form of the empirical relationship is a necessary first step for evaluating hypothesized relationships to cognitive and communicative pressures.

More challenging yet is establishing precisely *which* variables are implicated in the most basic causal relationship, and for which the variables the correlation is an artifact. Contra Zipf's longstanding observation regarding the relationship of word length and frequency, Piantadosi et al. (2011) demonstrated an even more robust relationship between average in-context predictability and word length across the lexicons of 11 European languages. The strength of this alternative relationship points to the possibility that even a robust, well-cited law may be an artifact of another relationship (see also Baayen et al., 2016 regarding the complex role of frequency in lexical processing). In principle, the degree to which a listener expects a word (encompassing both frequency and predicability, depending on the availability of context) might be more strongly predictive of *another* property of wordforms besides length. Candidate properties of wordforms could include phonotactic probability (the prototypicality of the sound sequence), duration, number of lexical "neighbors" with similar perceptual forms, or another measure of aggregate perceptual similarity to other words. More generally, the variables involved belong to a (dizzyingly) complex network of correlated properties in the lexicon: previous work has found robust correlations between many pairs of variables among word frequency, in-context predictability (Piantadosi et al., 2011), neighborhood density (number of perceptually similar words), phonotactic probability (Vitevitch et al., 1999), age of acquisition (Kuperman et al., 2012b), number of word senses (Baayen & del Prado Martín, 2005), concreteness (Brysbaert et al., 2014b), centrality in a semantic network (Vincent-Lamarre et al., 2016), longevity in the lexicon, and rate of language change (Pagel et al., 2007), among others; furthermore, these word-level properties are varyingly reflected in psycholinguistic measures such as recognition rates in noise (Luce & Pisoni, 1998), response times in lexical decisions (Balota et al., 2007), eye tracking

behavior, and measurements of neural activity (DeLong et al., 2005). Establishing precisely which correspondences are the most robust across languages will help identify which relationships may be most profitably linked to cognitive theories, and in turn improve our understanding of the remainder (though see Ladd et al., 2015 regarding the limitations of correlational studies).

Here, we demonstrate a novel generalization of Zipf's Law of Abbreviation that bears on these questions about the forces shaping language. We show that high frequency words are not only short, but also typically contain higher probability sound sequences than low frequency words. The measure of *phonological information content*, or *PIC*—the negative log probability of a phoneme sequence under a simple model of phonological sequences in a language (Cohen Priva, 2008)—provides a succinct metric of wordform structure motivated by both articulatory economy and communicative robustness. By building a model of phoneme transition probabilities over unique word types in the lexicon and using only short sequences we avoid the obvious circular relationship between word frequency and token-weighted phonotactic probability (which are definitionally equivalent). Evaluation of this hypothesis across 13 languages drawn from three large-scale corpora reveals that PIC, as computed over machine-generated phonemic transcriptions, accounts for significantly more frequency-related variance, yielding a similar improvement in robustness over the basic form of Zipf's Law of Abbreviation to that found by Piantadosi et al. (2011). We conclude by discussing the implications for theories of linguistic reduction.

3.2 Background

We outline two arguments why the correlation of phonotactic probability with word frequency should exceed that of word length and frequency. Both arguments pertain to "reduction"— the systematic underarticulation, shortening, weakening, or wholesale omission— of frequent or highly predictable linguistic material, for example using "probly" in place of "probably" in conversational English (Aylett & Turk, 2006; Bell et al., 2009; Gahl et al., 2012). On longer timescales, reduced variants may eclipse their long-form predecessors to become the dominant (or indeed only) form in the lexicon, e.g., *bus* superseding *omnibus* (Mahowald et al., 2013). The principal argument in favor of a relationship between frequency and phonotactic probability is that all of the above phenomena may be reflected in the probability of a wordform, whereas length only captures shortening and omission. The change from Middle English *aks* to *ask*, for example, would result in a change in the phonotactic probability for the wordfrom, though both forms have the same number of phonemes. Consequently, phonotactic probability provides a better generalized measure of wordfom magnitude, which should be expected to vary with word frequency given the above communicative pressures.

The first argument is that phonotactic probability is a better measure than length of the articulatory effort required of speakers to produce a word. Vitevitch & Luce (2005) found that common articulatory sequences are faster to produce, an effect which is robust for non-words and in the absence of listeners. Retaining the classic logic of Zipf's Principle of Least Effort, in which speakers prefer languages with lower total articulatory costs, then a more accurate measure of articulatory cost – e.g. one that assigning context-dependent costs to phonemic material – should be expected to

correlate more strongly with frequency.

The second argument emerges with respect to the countervailing force of communicative robustness. In the absence of other pressures, the speaker-oriented optimization of the sort described above would lead speakers to continue shortening *all* forms; in the limit using a single short, ambiguous signal for every word in the language, similar to the situation described by Piantadosi et al. (2012) in motivating context-disambiguated polysemy. But while speakers prefer easier-to-produce, shorter phonological forms, they are constrained by listeners' need for *sufficiently* distinctive wordforms such that each word can be recognized in spoken word recognition – that is, differentiated from competitors. Phonotactic probability is also useful in this regard as a measure of distinctiveness, as an aggregate measure of the degree to which other words in the lexicon are consistent with a given speech signal. Given that listeners are known to rely on prior probabilities of linguistic events to infer speaker's intended meanings (Gibson et al., 2013), one potential explanation for the observed pattern is that a less common/predictable word needs, other factors held equal, a more distinctive wordform—reflected in a lower phonotactic probability—to have the same probability of successful recognition on the part of the listener. In other words a high frequency word has sufficient support for its identity outside of the wordform, whereas a low frequency word is highly reliant on the distinctiveness of its wordform for a listener to successfully distinguish it from competitors.

Phonotactic probability is closely related to metrics of lexical neighborhood density used in research on both visual and spoken word recognition. Neighborhood density for a word reflects how many other words have a similar wordform; while proposals vary on how to best operationalize similarity, proposals share the intuition that words with more similar wordforms ("neighbors") are

harder to recognize because there are more competitors consistent with a given received audio signal (or visual cue, in the case of reading). The most common definition is words within Levenshtein distance of one (one substitution, deletion, or insertion) (Coltheart et al., 1977). Previous work, e.g., Vitevitch et al. (1999), has demonstrated a strong correlation between phonotactic probability and neighborhood density computed in this way. There are nonetheless important differences between neighborhood density and phonotactic probability. First, it is possible to find words with high phonotactic probability but no neighbors within a single edit for words of moderate length. Previous work has used this distinction to investigate differences between listeners's phonological and word-level expectations in word recognition (Storkel et al., 2006). Second, cross-linguistic work suggests another dissociation in that neighborhood densities may be on average higher in natural languages than expected under a lexicon-wide phonotactic model (Dautriche et al., 2017). We investigate this correspondence in greater detail below.

3.3 Model

We employ a model-based estimate of the probability of phoneme sequence to produce a fine-grained estimate of the phonological typicality of wordforms. First, we define the phonological information content (PIC) of a wordform as the surprisal—negative log probability—of its phoneme sequence. Estimating the probability of a sequence $P(s^w)$ then depends critically on 1) the choice of probabilistic generative model of wordform structure and 2) the dataset used to parameterize that model.

We begin by demonstrating that word length can be used as a coarse estimate of phoneme sequence probability, insofar as length is the determining factor of string probability under a generative model with extremely strong simplifying assumptions. We then detail a method to capture statistical nonindependence in wordforms and discuss the possiblity of capturing more sophisticated types of hierarchical organization with more elaborate language models. To simplify exposition, we treat wordforms as sequences of phonemes; note however that the logic presented here also applies to characters in languages with phonemic writing systems (where characters approximate phonemic material, with obvious exceptions of digraphs like English *ch*). This robustness of the correspondence between phoneme- and character-based models is evaluated below.

3.3.1 Word Length and String Probability

The length of a word can be seen as a measure of probability under an extremely simple generative model of wordform structure, specifically one that treats phoneme string generation as the result of a memoryless, uniform random process. This kind of random process has long been used as a null hypothesis in statistical language research, often colorfully characterized as the random typing of "monkeys on typewriters" (Mandelbrot, 1954; Miller, 1957). While notably deficient in capturing the key aspects of human-generated linguistic samples (Ferrer-i-Cancho & Solé, 2002), these models capture the key correspondence that longer wordforms are less probable in a language: as long as there is more than one phoneme in the language–such that the probability any phoneme is less than 1– then a wordform that is one phoneme longer is necessarily less probable.

This model yields an estimate of phoneme sequence probability — and hence phonological in-

formation content — that is strictly proportional to the length of the wordform. Specifically, if a wordform s^w is comprised of a string of symbols s_1^w, ..., s_n^w that are equiprobable and independent— $P(s_i^w) = 1/|v|$, where $|v|$ is the number of items in the relevant (phonemic or orthographic) inventory— then PIC is strictly proportional to the length of the wordform:

$$-\log P(s^w) = -\log P(s_1^w) \times \ldots \times P(s_n^w) \tag{3.1}$$

$$-\log P(s^w) = -\log P(1/|v|)^n \tag{3.2}$$

$$-\log P(s^w) = n \times -\log P(1/|v|) \tag{3.3}$$

Finally because $-\log P(1/|v|)$ can be factored out as a constant scaling factor,

$$-\log P(s^w) \propto n \tag{3.4}$$

$$PIC(s^w) \propto n \tag{3.5}$$

This result establishes a correspondence between phoneme sequence probability and length, and supports further investigation into the possibility that Zipf's original formulation may be a special case of a more general relationship between frequency and phonotactic probability. We now consider an elaborated model that posits additional structure in the generative model for wordforms, yielding more precise, graded predictions regarding phoneme sequence probability, and hence better estimates of phonological information content.

3.3.2 *N*-gram / Markov Models

While the above length-only models capture a relationship between word length and phonotactic probability, they nonetheless fail to capture two key regularities in the lexical substructure observed in natural languages. First, phonemes are not equiprobable: taking English as an example, /w/ is substantially less common than /e/. Second, phonemes are not statistically independent: the phoneme /t/ in English is followed more frequently within a word by /i/ or /e/ and very rarely—if ever—by /b/ or /g/. Just as they have rich knowledge of which words follow others, people have rich knowledge of the relative prominence of these sequences (Shannon, 1951; Vitevitch & Luce, 1999; Luce & Large, 2001). Other lines of work suggest that people are capable of using sub-word information incrementally, for example using sounds from the beginning of a word to restrict the set of consistent continuations (Marslen-Wilson & Welsh, 1978; Marslen-Wilson, 1987b; Zwitserlood, 1989; Eberhard et al., 1995).

An *n*-gram model of phoneme sequences adds these two key features, by introducing a statistical structure that uses the probability of a phoneme by conditioning on preceding content, and replaces the equiprobability assumption by estimating transitional probabilities from corpus data. Under

these models, the phonological information content $PIC(s^w)$ of a wordform s^w is defined as:

$$PIC(s^w) = -\log P(s^w) \tag{3.6}$$

$$= -\log P(l_1, \ldots, l_{|s^w|}) \text{ for } l \in s^w \tag{3.7}$$

$$= -\sum_{i=1}^{|s^w|} \log P(l_i | l_{i-(n-1)}, \ldots, l_{i-1}). \tag{3.8}$$

where l are the phonemes that comprise the sequence s^w, $|s^w|$ is the length (in phones) of s_w, and n is the sequence length, or order, of the model (e.g., 1 = unigram, 2= bigram, 3 = trigram).

PIC computed under an n-gram model produces predictions contrary to the length-only model. Depth /dɛpθ/ (*depth*) contains fewer phonemes yet has a higher PIC (28.7 bits) than /gɹɑʊnd/ (ground, 12.89 bits) because the latter is comprised of significantly more common subsequences. A comparison of PIC estimates from the monkeys-on-typewriters model and a more sophisticated trigram model for the word *motorcycle* is presented in Table 3.1. In that n-gram models do not posit higher-order structure, they do not explicitly account for the morphological structure of a word. Rather, the relative prevalence of morphemes in the language is reflected in the phoneme-to-phoneme transition statistics estimated from corpora. In the name of brevity, we henceforth use "PIC" to refer to PIC computed under the n-gram phonological transition model.

We note three key dimensions of variation among n-gram models that modulate their appropriateness for a particular task, especially for modeling phoneme sequences. The first is that n-gram models vary in the number of preceding events (in this case phonemes) used to predict the next one: in unigram (1-gram) models phonemes are drawn independently, whereas in bigram (2-gram)

Table 3.1: Estimates of the phonological information content (in bits) under a naive model vs. under a probabilistic phonological model for English

"M"	"O"	"T"	"O"	...	Σ Bits
Uniform Character Probabilities					
$-\log_2 P(M)$	$-\log_2 P(O)$	$-\log_2 P(T)$	$-\log_2 P(O)$		
4.700	4.700	4.700	4.700	...	47.004
3-Character Phonological Information Content Model					
$-\log_2 P(M \mid \triangleright)$	$-\log_2 P(O \mid \triangleright M)$	$-\log_2 P(T \mid MO)$	$-\log_2 P(O \mid OT)$		
4.400	2.184	3.217	3.672	...	35.602

Note The negative log probability of a wordform under the uniform character probability model is computed assuming 27 characters in the inventory and an end symbol for the wordform. The probabilistic language model takes into account sequential dependencies up to length 3, obtained from 25,000 most frequent words in the English Google Books (2012) corpus.

models the continuation probabilities are conditioned on the preceding phoneme. Tracking longer sequences may provide a better fit to data by tracking more granular events, but may also introduce overfitting in that longer sequences are less likely to be observed. We consider models of order 3 (i.e. predicting each phoneme with up to two preceding phonemes) to capture some among of phonotactic and morphological structure, but without risking overfitting. In the case of a type-weighted model, we note that if we track the full history available for every word then all wordforms have a phonotactic probability of $\frac{1}{|\mathbb{W}|}$, where $|\mathbb{W}|$ is the size of the lexicon. Intuitively, if transitions are tracked up to the length of the longest word, then only a single word type displays each exact phoneme sequence and novel sequences are assigned zero probability.

A second consideration is how to apportion probability mass from observed phoneme sequences to unobserved ones, or *smoothing*. In that phonological inventories are much smaller than lexical

vocabularies, sparsity is less problematic for phoneme-level language models than for word-level ones. Nonetheless, we adopt Good-Turing smoothing (Gale & Sampson, 1995) to ensure that the model assigns nonzero probability to a larger class of possible transitions than those observed in the dataset. Finally, some amount of probability mass must be assigned to unseen phonemes, in case they are encountered in the test set. While this is not a concern for the core phonemes that comprise a phonological inventory, loanwords may contain singleton phonemes (e.g. the final vowel in "rapproachment"). Here we map these out-of-vocabulary phonemes to an unknown token, which is assigned a small probability mass; this unknown token is then treated as any other by the smoothing scheme.

3.3.3 More Complex Models of Wordform Structure

We briefly note two more complex probablistic generative models of language that have been used in linguistics and psycholinguistics to characterize wordform structure that are potentially appropriate for producing yet better estimates of phonological information content: hidden Markov Models (HMMs) and probabilistic context-free grammars (PCFGs). In principle, both of these model classes can produce better estimates of phonological information content because they assign lower probability mass to words that have internal structure unlike attested words. An HMM does this by adding an additional inventory of unobserved states (e.g., vowels vs. consonants or onsets vs. codas), and conditions observed data on the unobserved state. Besides their ubiquitous use in Natural Language Processing and Automatic Speech Recognition, HMMs have recently been used to examine the extent of gradient lexical competition effects (Strand & Liben-Nowell, 2016). A PCFG posits

that the observed data (*terminals*) are generated from a set of latent states (*nonterminals*) following a set of probabilistic re-write rules; unlike an HMM, a PCFG is capable of capturing recursion. Futrell et al. (2017) show a modest improvement over *n*-gram models in predicting phonotactic structure in 14 languages in the WOLEX corpus (Futrell et al., 2017).

While useful for illustrating how more sophisticated generative models of wordforms may provide better characterizations of the magnitude of words, we adopt smoothed *n*-gram models for the analyses presented here. Futrell et al. (2017) show relatively small advantages for a sophisticated feature-interaction PCFG and Dautriche et al. (2016) show slightly worse performance than higher-order *n*-gram models. Further, supervised PCFG induction requires that wordform data be annotated with nonterminal categories, which are not available for many of the languages in the sample examined here. The unsupervised induction of useful PCFGs — which requires learning in a very large hypothesis space — remains an open problem for research.

3.4 Methods

To approach this problem empirically, we examine the strength of the relationship between word frequency and phonological information content (PIC) across a wide range of languages and datasets. First, we produce new frequency estimates for web-scale corpora in 13 languages and three datasets. For each corpus, we take the top 25,000 most frequent words and construct a type-weighted phonotactic model, which we then use to produce model-based information content estimates — negative log probability under the model – for those 25,000 highest-frequency wordforms. We then evaluate

the correspondence between frequency and phonological information content across the words and corresponding wordforms in each corpus. The null hypothesis is that the correspondence between frequency and PIC, as evaluated with Spearman's rank correlation coefficient, is no stronger than the correspondence between frequency and word length. We additionally test the strength of the correspondence between a word's average in-context predictability and its phonological information context, and compare that to the correspondence between average in-context predictability and word length.

3.4.1 Datasets for Frequency and Average Surprisal Estimates

The Google Web 1T datasets were downloaded from the Linguistic Data Consortium (Brants & Franz, 2006, 2009); the Google Books 2012 datasets were downloaded from storage.googleapis.com/books/ngrams/books/da (Michel et al., 2011), and OPUS (2013) from opensubtitles.org (Tiedemann, 2012). All punctuation-only word tokens were discarded, and punctuation marks appearing with other text, with the exception of apostrophes, were removed. We make the simplifying assumption that the tokenized written forms correspond to lexical items used by speakers; while this assumption may not hold for all forms (e.g., German compound nouns, French contractions), it holds for the vast majority of word forms in the analysis (see (Baayen et al., 2016) for further discussion of the importance of variation in orthographic segmentation conventions for frequency analyses). For the purposes of computing frequency, all tokens were converted to lowercase using the relevant POSIX locale; US English and European Portuguese were used for English and Portuguese, respectively. In the case of Google Books 2012, part-of-speech tags were discarded, and instances from earlier than 1800 removed from

the analysis. UTF-8 encoding was maintained throughout for all languages and datasets; Hebrew strings were represented with right-normalized forms. Counts were stored using ZS, a specialized file format for efficient retrieval of *n*-gram counts (Smith, tted).

3.4.2 Estimating Average Lexical Information Content

In addition to frequency, we estimate the overall predictability of each word, operationalized as its average lexical information content. Following Piantadosi et al. (2011), we compute the negative mean log trigram probability across contexts:

$$-\frac{1}{N}\sum_{i=1}^{N}\log P(W = w|C = c_i). \tag{3.9}$$

where c_i is the context for the ith occurrence of w and N is the frequency of w in the dataset. Because estimates of mean information content are highly biased for small datasets, we do not compute these values for datasets in the OPUS corpus.

3.4.3 Estimating Phonological Information Content

For each language and dataset, a three-character transition model was estimated using the 25,000 most frequent in-dictionary words also appearing in the corresponding OPUS subtitle corpus. Diphthongs (vowel sequences) were treated as sequential instances of discrete vowels. We also produced analogous three-phone transition models (excluding the Hebrew OPUS and Hebrew Google

Books 2012 datasets) using the IPA transcriptions from an automatic speech synthesizer, eSpeak. While these broad phonological transcripts are imperfect, using IPA representations for words accounts for language-specific variation in orthographic conventions. For example, written Spanish includes accents only when the placement of prosodic stress cannot be deduced from more general rules in the language. Using an IPA transcription avoids the need for developing language-specific processing rules, for example deciding whether 'a' vs. 'á' should be merged or kept as separate orthographic variants in Spanish.

Loan words and acronyms can greatly affect the obtained transition probabilities for the phonotactic model, especially because types contribute equally to the transition weighting (e.g., the transitions in "Okeechobee," "mañana," and "ACLU" would be as heavily weighted in a phonotactic model of English as the transitions in "they" and "will"). To minimize these effects, we used only non-capitalized word types present in the relevant Aspell dictionary to build sound and character transition models for each language (with the exception of German, in which nouns, which are capitalized by convention, were retained).

To avoid overfitting among higher order sequences, phone and character transition probabilities were computed with Witten-Bell smoothing (Chen & Goodman, 1999) with interpolation on transitions of order 3 using the SRILM toolkit (Stolcke, 2002), as is commonly used for character-level language models. Each word's phonotactic probability was calculated as the product of the probabilities of each symbol given the preceding symbol string up to two symbols, including a start symbol $\triangleright$ and an end symbol $\triangleleft$, e.g., $P(the) = P(t|\triangleright) \times P(h|\triangleright t) \times P(e|th) \times P(\triangleleft|he)$. We convert this sequence probability to phonological information content—which keeps the same directionality

and approximate range of word length— by taking the negative log probability.

3.4.4 Evaluating Correlations

Following the basic methodology adopted in Piantadosi et al. (2011), we examine the correlation between word-level predictors (log frequency and average information content) and each of two quantitative measures of structural form (either word length or PIC) for the 25,000 most frequent types in each language. Unlike that work, we limit our analysis to in-dictionary types, thereby excluding person names, place names, acronyms, and loan words from the analysis.

We evaluate the strength of each of these correlations using Spearman's rank correlation coefficient, which evaluates the degree of monotonicity of the function rather than linear correspondence. Correlations are computed for all pairs of variables spanning an inventory of lexical variables and wordform measures. Lexical variables include frequency (raw number of occurrences) and average trigram information content, as described above. Wordform measures included the length in phonemes, the length in characters, the phonological information content as estimated under an n-gram model of order 3 built on phoneme transitions, and the approximation of phonological information content as estimated under a Markov model of order 3 built on character transitions. The statistical significance of the difference between correlations is evaluated in each case using bootstrapped estimation of the difference scores and comparing the resulting distribution to 0.

To evaluate the relationship of PIC and frequency in the absence of word length, we partial out word length (i.e., use the residuals from predicting PIC from word length with a linear model) and compute the correlation with frequency. We perform the analogous operation on length, obtaining

the residuals from predicting length from PIC and computing the correlation with frequency. This provides a strong test of the unique predictive value of these variables.

Finally, we compute correlations between frequency and PIC for three random baselines for each language in Google 1T to confirm that the obtained correlations are nontrivial. In the first model, the assignment of wordform to frequency is randomly permuted. In the second model, word forms are drawn from the collected phonemic material of the language (without replacement), maintaining each word form's length in the source language but not other properties. In the third model, the order of phonemes or characters for each word are permuted, the phonotactic model refit, and PIC recomputed. This control maintains the unigram phoneme statistics, but perturbs the higher-order phoneme transitions that may exist in a language. Under the third model PIC correlations from the natural languages are expected to significantly exceed all three of these random baselines.

3.5 Results

We investigate the correlation between word length and a measure of the phonotactic probability and word frequency in large corpus samples (43m to 266b words) in 13 languages, across three large-scale datasets from different linguistic sources. If a word's phonotactic probability is indeed a stronger correlate of frequency, then we may conclude that communicative pressures for robustness and economy of articulation effort are better reflected in the probability of wordforms.

Across languages, we obtain a systematically stronger negative correlation between log frequency and PIC than log frequency and word length (Figure 3.1). Building the model from phonemic tran-

scriptions, this pattern holds in 11 of 11 languages in the Google 1T datasets, 4 of 6 languages from Google Books 2012, and 12 of 12 languages from the 2013 OPUS corpus (phonemic transcriptions were not available for Hebrew). Building the model from characters—as an approximation of phonological forms—this pattern holds in all languages from Google 1T, 4 of 7 (results are consistent in 2 of the remaining languages, but fail to reach significance), and all languages from the 2013 OPUS corpus. Partial correlations reveal a significant length-related contribution to PIC and vice versa, though in some cases PIC with length partialed out is a stronger correlate of frequency than length is, e.g., English 1T when PIC is computed over IPA representations. In other words, among words of the same length in a given language, PIC explains substantial additional variance in word frequency (Figure 3.2): high frequency words have higher probability (lower PIC) sound sequences. Dutch, English, and German show the same pattern of results for the set of words in CELEX (Baayen et al., 1995); this pattern holds among monomorphemic, multimorphemic, and and an aggregate analysis of all words.

A similar pattern of results emerges regardless of whether PIC is computed over characters or phonemic representations. The Russian Google Books 2012 dataset is the only dataset showing consistent evidence in favor of a stronger relationship between length and frequency—however, this is contrary to the results of the Russian OPUS results, which exhibit the prevailing dominance of PIC. Russian shows the lowest correlation between frequencies obtained from Google Books and OPUS (Pearson's $r = .48$), as well as the lowest correlation between PIC estimates derived from Google Books and those derived from OPUS (Pearson's $r = .63$). Across languages, models built over phonemes and character transitions provide similar estimates of PIC (Pearson's r between .789

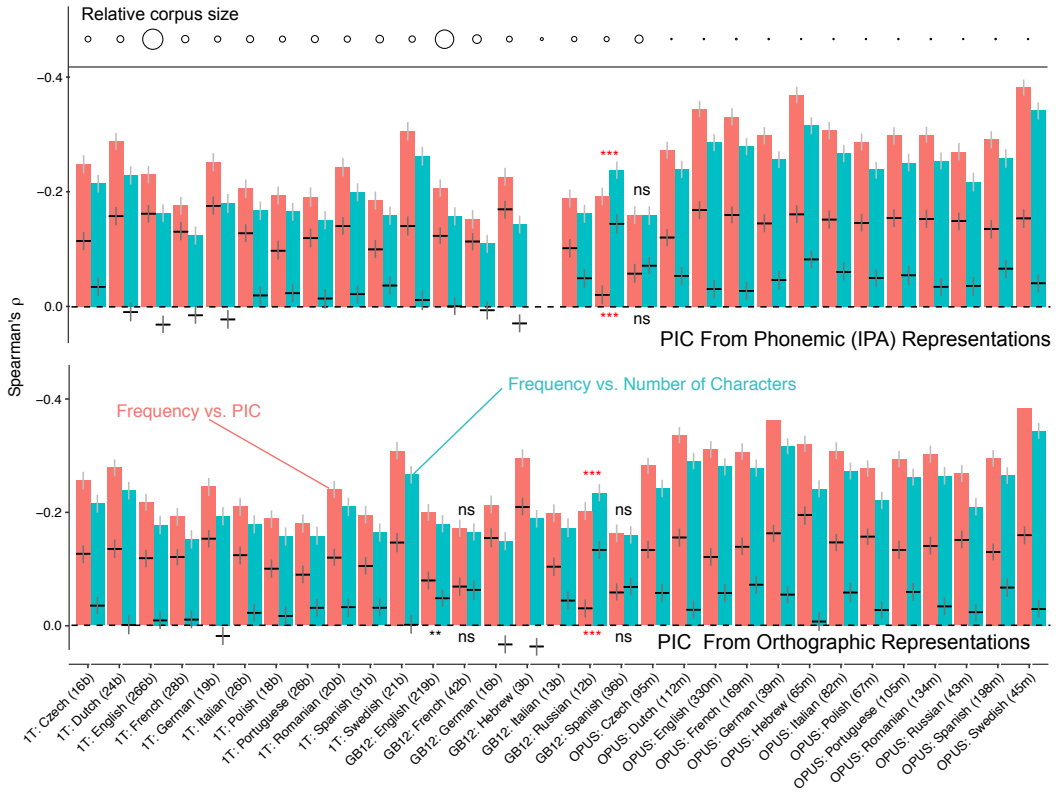


Figure 3.1: Frequency exhibits a stronger negative correlation with PIC than length. Bars indicate Spearman's ς for the two variables for the $n = 25000$ most frequent words in each dataset. Gray lines indicate the 99% bootstrapped confidence interval. Black lines indicate the correlation with the other measure of word difficulty partialed out.

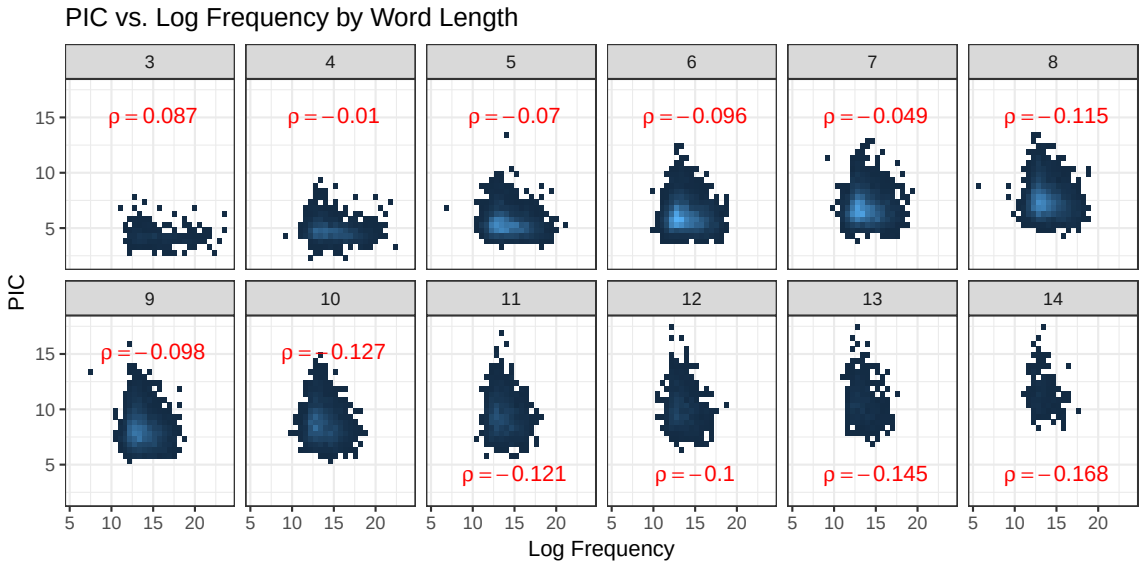


Figure 3.2: Among words with the same number sounds, more frequent words have higher phonotactic probability as measured with PIC. This figure shows this relationship in the English Google Books (2012) dataset. Correlations display Spearman's ϱ among words of the same length in phonemes.

and .919 across languages, median = .874). While more research is required to extend these findings beyond Germanic, Romance, and Slavic languages, Hebrew provides an important test of whether this relationship holds in languages with extensive nonconcatenative morphology.

PIC computed from three-phone and three-character transition models from natural languages substantially exceed the correlations observed for PIC computed under three random baseline languages (Figure 3.3). While these correlations are not statistically significantly higher for natural languages when PIC is computed using a token-weighted model, we find that the correlation obtained for natural language are larger than all same-language baselines for every language ($p < .001$, by bootstrapped tests of the difference of correlations between each <baseline, natural language> pair).

Whereas Piantadosi et al. (2011) found that taking into account contextual predictability (in the

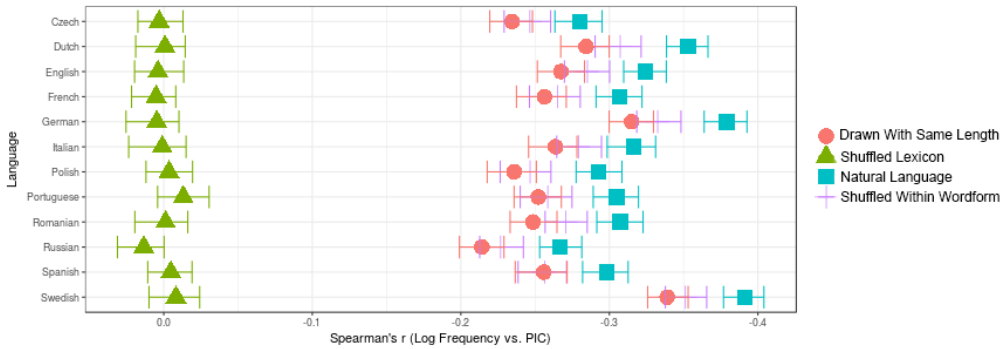


Figure 3.3: PIC computed from three-character transition models on the lexicons of natural langua ges ("Natural Language") have a significantly higher correlation with frequency than three baselines: 1) when PIC is computed for randomly shuffled wordforms ("Shuffled Lexicon") 2) when the wordform is drawn using single-character probabilities and 3) when the order of characters is shuffled within each wordform ("Shuffled Within Wordform").

form of mean trigram surprisal) better predicts word length than using frequency (negative log probability), we find a qualitatively different pattern of results for phonological information content. Examining the same set of words as above, we find that the correlation between frequency and PIC is *higher* than the correlation between mean trigram surprisal and PIC in all cases (Figure 3.4, B). The correlation between frequency (negative log unigram probability) and PIC is greater than the correlation between mean trigram surprisal and word length in most datasets, the principle exceptions being English Google Books, English Google 1T and the German Google Books (Figure 3.4, C). We find substantially attenuated support for the principal claim in Piantadosi et al. (2011), in that we find unigram frequencies better predict word length than does mean trigram surprisal (Figure 3.4, D). This discrepancy may reflect refinements to the list of lexical items analyzed, improvements in the analysis methodology in the current work (especially, maintaining proper character encoding), or issues in computing mean trigram surprisal in languages with richer inflectional morphology.

3.6 Discussion

The relationship between word length and frequency is one of the most robust empirical findings regarding the structure of lexicons. However, this relationship may be an artifact of an even broader relationship, that of word frequency and the probability of the sound sequences of wordforms. Analyses of large-scale corpora from 13 languages across three datasets substantiate this proposal, and provide evidence that the countervailing pressures that govern linguistic reduction—towards articulatory economy on the one hand and away from ambiguity on the other—are strongly reflected in the probability of wordforms. We reflect further on the possible mechanisms by which this pattern might arise, the correspondence between the measure of PIC and recent measures of lexical information content, and note how the measure could be used to characterize variation between tokens.

3.6.1 Possbile Mechanisms

Our motivation for a stronger correspondence between phonotactic probability and word frequency drew from both speaker- and listener- oriented accounts of reduction, and was not intended to directly evaluate the relevance of these two accounts, unlike Gahl et al. (2012) or Kanwal et al. (2017). However, the obtained results regarding the relationship between both wordform measures and frequency reveals an intriguing asymmetry: while high frequency words are necessarily short, low frequency words may also be relatively short and phonotactically probable, e.g. English *ewe*, *gut*, *whey* (of note, these tend to be highly preserved forms in the language). This heteroskedastic rela-

tionship suggests that the pressure from communicative robustness does not result in an augmentation of words to maintain a correspondence between frequency and wordform. This may be due to the fact that speakers have a range of options for reducing a wordform, but have limited options for augmenting it, such as epenthesis, or varying the duration (Gahl, 2008). Another possibility is that these forms are highly predictable in the contexts in which they appear, such that no alteration in the wordform is necessary, though this contextual predictability is not captured by average trigram information content.

3.6.2 Neighborhood Density

How strong is the correlation between neighborhood density and phonological information content? We compared the obtained phonological information content estimates with two standard measures of neighborhood density, orthographic Levenshtein distance-20 (OLD20) and phonological Levenshtein distance-20 (PLD20) (Yarkoni et al., 2008). Both take the average Levenshtein-Damerau distance (or edit distance, with tranposition counted as a single operation) to the twenty closest neighbors for a word. This yields Spearman's *rho* of .930 and .895, respectively. This relationship remains robust when frequency (in the form of unigram surprisal) is partialed out of both predictors (Spearman's *rho* = .906 and .851, respectively). Though neighborhood density and phonotactic probability correspond to distinct theoretical constructs (the lexicon vs. the phonological inventory), their empirical signatures are extremely similar.

3.6.3 Token-Level Variability

While the corpus studies in the main analysis here use the citation forms for words, PIC may be used to characterize within-word variation, in that articulatory reduction results in different phonemes and phoneme sequences for realization of the same word. PIC can in principle capture such differences that are otherwise the same length, such as instances of metathesis or vowel substitution. In future research we will examine the relationship between the frequency and in-context predictability of particular word tokens and the phonological information content of individual tokens in a large naturalistic corpus such as the Buckeye Corpus (Pitt et al., 2005).

3.7 Conclusion

The canonical inverse relationship between the length of wordforms and their frequency is a special case of an even broader relationship between phonotactic probability and frequency. Phonotactic probability may serve as a better index of articulatory costs; alternatively, it can be interpreted as an index of wordform distinctiveness, critical to successful word recognition. While speakers prefer to simplify and shorten words to reduce articulatory costs, they are limited by listeners' requirements for sufficiently distinctive wordforms for successful recognition. The observed correspondence between frequency and PIC provides preliminary evidence that the psycholinguistic processes at work in producing and perceiving speech may help to shape human languages at the broadest scales.

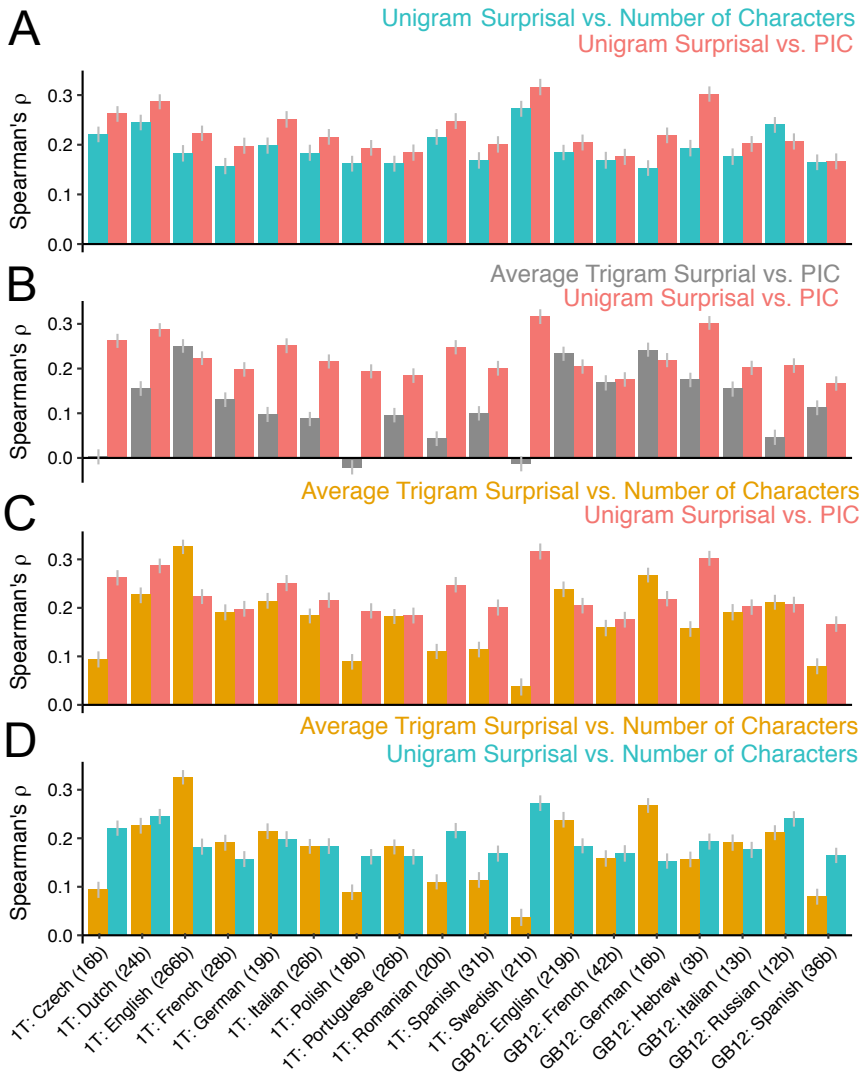


Figure 3.4: Comparison of several key correlations between measures of word expectedness (frequency, predictability) and measures of wordform (length, PIC) in the Google 1T (1T) and Google Books (GB) datasets. **A.** Correlations obtained for PIC and frequency (here treated as unigram surprisal) vs. length and frequency, from the main analysis. **B.** Correlations obtained for PIC and unigram surprisal vs. PIC and in-context predictability, treated as average trigram surprisal following Piantadosi et al. (2011). **C.** Correlations obtained for word length (number of characters) and average trigram surprisal vs. PIC and unigram surprisal. **D.** Correlations obtained for word length (number of characters) and average trigram surprisal vs. word length and unigram surprisal.

"[...] all models are wrong [...]."

Box (1976)

4

Conclusion

In this dissertation, I show how probabilistic generative models of language (PGLMs) can be used to further our understanding of people's linguistic knowledge. The first chapter demonstrates how these models can be used to infer children's knowledge of abstract structural regularities in their language. The second chapter presents a new method for evaluating these PGLMs in their fit to people's linguistic expectations in an in-context spoken word recognition task. The third chapter

shows how even a simple PGLM can help characterize language users' expectations in a way that may explain cross-linguistic regularities in language structure. At a broader level, these chapters demonstrate the utility of PGLMs in understanding language learning (Chapter 1), language processing (Chapter 2), and language structure (Chapter 3). While Box's dictum in the epigraph holds true as ever—PGLMs are not the same as human linguistic knowledge but an approximation thereof—these models nonetheless demonstrate a remarkable level of utility in furthering our understanding the human linguistic faculty.

A fitting conclusion to this dissertation is to then consider the broader prospects of adopting an expectations-oriented framework for understanding the complex iterative relationship between language acquisition and language processing. In particular, I argue that such a framework provides for a principled and fruitful way to relate these mutually constructive processes, and consider the theoretical commitments and implications. To clarify the scope of this endeavor: language researchers are well aware of the shortcomings and inaccuracies of the common-sense shorthand that "children learn language" (though perhaps to varying degrees); my objective here is to consider an alternative shorthand (learners "revise linguistic expectations" throughout the lifespan) that entails fewer compromises in fidelity to the phenomena under study.

Expectations bridge processing and learning

The widespread use of expectations generated from PGLMs in natural language processing makes their utility eminently clear for real-world linguistic tasks like word recognition and sentence parsing. Further, behavioral experiments such as those presented in Hale (2001) and Levy (2008) demon-

strate that these probabilistic expectations derived from PGLMs are strongly predictive of sentence processing difficulty for people. But in addition to their role as *inputs* for probabilistic inference in episodic language processing, expectations can be seen as the *output* of the inferential processes of language learning on longer timescales, including first language learning (Bannard et al., 2009; Perfors et al., 2011). While PGLMs have been used to model language processing and language learning separately, few if any attempts have been made to model both roles together (though see McMurray et al., 2012 for a related endeavor to understand the relationship of online referent selection and word learning). PGLMs are ideally suited as a formalism capable of serving both inferential processes; as such, these two processes can be linked through a common knowledge store.

As a simultaneous store of linguistic knowledge and the product of learning, a PGLM is closely related to the concept of a *linguistic grammar*, or a language user's internalized knowledge of the set of rules governing the composition of their language. As with a grammar, a PGLM can be updated to reflect new data. However, a PGLM further refines this concept by explicitly positing that knowledge of language includes fine-grained probabilistic expectations: rather than judgments of what constitutes "acceptable" or "unacceptable" strings in a language, this knowledge encodes expectations of what people are more or less likely to say. This latter information is far more useful for the inferential requirements of language processing (see Chapter 2).

PGLMs clarify the role of episodic linguistic events, including context

As a consequence of their status as simultaneously *usable* and *updatable* stores of language knowledge, PGLMs provide a principled way to think about individual episodic linguistic events, or what

happens when a listener hears an individual utterance in the real world. Hearing an individual utterance simultaneously presents a language user with the options of 1) deploying their existing language knowledge and/or 2) revising that knowledge. If, for example, a listener finds that a speaker has used a novel pronunciation of a word, then the listener should be able to use their expectations to recover the word intended by the speaker. PGLMs readily handle this task (Luce & Pisoni, 1998; Norris & McQueen, 2008). At the same time, encountering a novel pronunciation may provide data that prompts a listener to revise the model that generates their expectations—updating the model to reflect the proclivities of this speaker, of the dialect group the speaker belongs to, of the existence of multiple phonological realizations of the same word, or many other possible causal pathways for this new pronunciation (see Kleinschmidt & Jaeger, 2015 for further examples in updating phonological expectations).

A distinguishing feature of episodic language use is the availability of non-linguistic context. The non-linguistic context of an utterance makes critical contributions to the inferential tasks of language processing and language learning. For processing, the task of recognizing a familiar noun from a speech stream may be made significantly easier when a limited set of referents are on hand; in the same vein, social cues provide an additional source of information. For language learning, the task of inferring the referent of a novel noun can be made significantly easier given these same information sources available in the specific environmental context (Frank et al., 2009).

PGLMs support representational fluidity

Linguistic inquiry has long focused on the specific form of the adult grammar in a homogeneous and stable speech community (Chomsky, 1957). The expectations-oriented framework shifts the emphasis from the specific form of the representations (and especially commonalities at the level of the population) to the utility of representations for language processing and production for individuals. Under this view, more abstract representations of linguistic structure are useful to a person insofar as they help them to understand others, and relatedly the degree to which they help others understand them. As such, early language learners should be expected to pass through a wide variety of different representations as they revise their expectations to better explain the linguistic data they encounter. Further, language users in the same speech community may arrive at a wide variety of hypotheses about the latent structure of "their language." Individuals may posit widely varying latent structure for the language they speak, as long as the derived expectations of interlocutors are sufficiently similar to allow communication in a noisy channel.

PGLMs are ideally suited for modeling this representational fluidity. Some PGLMs update directly as they process new data sequentially (e.g., LSTMs). In other cases, variation in representations can be modeled by comparing PGLMs fit on sequentially larger datasets (Bannard et al., 2009). Either can be seen as finding a generative model that maximizes the posterior probability given the data seen so far, or using an approximation to that posterior. PGLMs can thus be used to model the process by which learners revise their structural assumptions about language. Perfors et al. (2011) for example show how a learner might transition to increasingly abstract structural representations of

language to best explain the data that they encounter. Explorations of shifts in the representations in PGLMs have so far focused on children's language production (Chapter 1 of this volume, Bannard et al., 2009, Perfors et al., 2011); how these models yielded from acquisition can account for shifts in online language processing is a potentially fruitful avenue of further research.

PGLMs reflect continuity and change over the lifespan

A related feature of this framework is that it provides a principled way to handle simultaneous change and continuity in language knowledge over a person's lifespan. The parameters in a PGLMs are data-dependent, such that expectations may change as a function of exposure to additional data. Nonetheless, the process of inference remains the same for the entirety of the lifespan: newly received data is interpreted in light of existing linguistic expectations, and linguistic expectations are revised in light of the newly received data.

Of course, changes in language knowledge tend to be less drastic for older language users. This reflects a combination of factors. First, older individuals may haveincreasingly strong expectations in episodic language processing such that "novel" events may be interpreted as instances of familiar ones. Second, newly encountered counter-examples are less likely to prompt a large change in language knowledge in light of the overwhelming weight of past experience: a learner is likely to posit a new grammatical category at 2;0, but not at 20;0 or 80;0. PGLMs, and the expectations they yield, naturally exhibit this pattern: changes in expectations decrease in magnitude as as a function of seeing more data, so long as newly-observed data is consistent with that previously observed.

PGLMs can be used for multiple levels of linguistic structure

Re-usable, compositional structure is a defining property of language: sentences, words, and morphemes are composed of smaller units, sequentially arranged (Hockett, 1959). Many PGLMs, including *n*-gram models, LSTMs, hidden Markov models (HMMs), and PCFGs have been shown to induce useful representations of structure when trained on a variety of domains, including sequences of phonemes, morphemes, or words. While parsimony is rarely a strong argument on its own, in this case it suggests that similar pattern extraction mechanisms—however they are implemented at the neurological level in people or in circuits for machines—*could* account for structure at multiple resolutions in language.

PGLMs explicitly encode uncertainty

One of the most challenging aspects of studying cognition more broadly is a dual role of uncertainty: cognitive agents have uncertainty about the world they inhabit, and researchers (insofar as they are cognitive agents themselves) must deal with uncertainty about the inputs, representations, and behavior of cognitive agents. In effect, a researcher is inferring the way in which a cognitive agent does inference; the researcher has access to noisy data often relies on prior knowledge. PGLMs, thankfully, can provide some amount of help, in that they provide a way to explicitly represent uncertainty about language on the part of a cognitive agent. This uncertainty plays a critical role in both language processing and learning. For language processing, how peaked is the support for a particular word given expectations and data? For language learning, how peaked is the support for a par-

ticular model? This quantitative representation of a cognitive agent's uncertainty can be embedded in a larger probabilistic generative model of the data available to a researcher (see Chapter 1 and Appendix 1 for an example). This makes it easier for researchers to determine the degree of uncertainty about their own hypotheses; in other words, we as researchers can be quite certain that a cognitive agent is uncertain.

These features together recommend a joint account of learning and processing focusing on linguistic expectations, and using PGLMs as a consistent, testable computational modeling framework.

References

(2007). The British National Corpus. Version 3, XML Edition. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Abbot-Smith, K. & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, 23, 275–290.

Altmann, G. T. & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73, 247–264.

Ambridge, B., Pine, J., & Rowland, C. (2011). Children use verb semantics to retreat from over-generalization errors: A novel verb grammaticality judgment study. *Cognitive Linguistics*, 22(2), 303–323.

Austerweil, J., Abbott, J., & Griffiths, T. (2012). Human memory search as a random walk in a semantic network. In *Advances in Neural Information Processing Systems* (pp. 3041–3049).

Aylett, M. & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, 119(5), 3048–3058.

Baayen, H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database. Release 2 (CD-ROM).

Baayen, R. & del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, 81(3), 666–698.

Baayen, R., Milin, P., & Ramscar, M. (2016). Frequency in lexical processing. *Aphasiology*, 30(11), 1174–1220.

Balota, D., Yap, M., Hutchison, K., Cortese, M., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39(3), 445–459.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*.

Bartlett, F. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, 60(1), 92–111.

Bentz, C. & Ferrer-i-Cancho, R. (2016). Zipf's Law of Abbreviation as a Language Universal. In *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*: Universitätsbibliothek Tübingen.

Berko, J. (1958). The Child's Learning of English Morphology. *Word*, (pp. 150–177).

Bickel, B. & Nichols, J. (2005). Inflectional morphology. In T. Shopen (Ed.), *Language Typology and Syntactic Description*. Cambridge, UK: Cambridge University Press.

Bielec, D. (1998). *Polish: An Essential Grammar*. London: Routledge.

Bies, A., Mott, J., Warner, C., & Kulick, S. (2012). English Web Treebank. *Linguistic Data Consortium, Philadelphia, PA*.

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

Bowerman, . (1988). The "no negative evidence" problem: How do children avoid constructing an overly general grammar? In *Explaining language universals* (pp. 73–101). Basil Blackwell.

Box, G. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356), 791–799.

Braine, M. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development*, 41(1).

Brants, T. & Franz, A. (2006). Web 1T 5-gram Version 1 LDC2006T13.

Brants, T. & Franz, A. (2009). Web 1T 5-gram, 10 European Languages Version 1 LDC2009T25.

Bresnan, J. (1986). The mental representation of grammatical relations. *Computational Linguistics*, 12(2).

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.

Brysbaert, M., Warriner, A., & Kuperman, V. (2014a). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911.

Brysbaert, M., Warriner, A., & Kuperman, V. (2014b). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–911.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, (pp. 711–733).

Charniak, E. & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting pf the Association for Computational Linguistics* (pp. 173–180).: Association for Computational Linguistics.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Chen, S. & Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language*, 13(4), 359–393.

Chen, S. & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computing, Speech & Language*, 13(4), 359–393.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton and Co.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris Publishers.

Cohen Priva, U. (2008). Using information content to predict phone deletion. In *Proceedings of the 27th West Coast Conference on Formal Linguistics* (pp. 90–98). Somerville, MA: Cascadilla Proceedings Project.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Lawrence Erlbaum Associates.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3), 306–329.

Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163, 128–145.

Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. (2016). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, (pp. 1–21).

De Marneffe, M., MacCartney, B., & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6 (pp. 449–454).

DeLong, K., Urbach, T., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117.

Demberg, V. & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2), 193–210.

Demberg, V. & Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Demuth, K., Culbertson, J., & Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language & Speech*, 49(2), 137–173.

Du Bois, J., Chafe, W., Meyer, C., Thompson, S., & Martey, N. (2000). Santa Barbara corpus of spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Eberhard, K., Spivey-Knowlton, M., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6), 409–436.

Edmiston, P., Perlman, M., & Lupyan, G. (2017). Creating words from iterated vocal imitation. In *The 39th Annual Conference of the Cognitive Science Society* (pp. 331–336).: Cognitive Science Society.

Elman, J. (1990). Finding Structure in Time. *Cognitive Science*, 14, 179–211.

Evans, N. & Levinson, S. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav Brain Sci*, 32(5), 429–448.

Fedzechkina, M., Jaeger, T., & Newport, E. (2012). Language learners restructure their input to facilitate efficient communication. *Proc. Natl. Acad. Sci. U.S.A.*, 109(44), 17897–17902.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, (pp. 1–185).

Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf's law for word frequencies. *Complexity*, 21(S2), 409–411.

Ferrer-i-Cancho, R. & Solé, R. (2002). Zipf's law and random texts. *Advances in Complex Systems*, 5(01), 1–6.

Ferrer-i-Cancho, R. & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788–791.

Fine, A., Jaeger, T., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10), e77661.

Fossum, V. & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–69).: Association for Computational Linguistics.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5), 578–85.

Frank, S. & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6), 829–834.

Futrell, R., Albright, A., Graff, P., & O'Donnell, T. (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics*, 5, 73–86.

Futrell, R. & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th C2onference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1 (pp. 688–698).

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences U.S.A.*, 112(33), 10336–10341.

Gabelentz, G. v. d. (1901). *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Weigel.

Gahl, S. (2008). Time and thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language*, 84(3), 474–496.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806.

Gale, W. & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3), 217–237.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., & Pallett, D. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus. *NASA STI/Recon Technical Report*, 93.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A. & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1), 1–58.

Geman, S. & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721–741.

Gers, F. & Schmidhuber, J. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6), 1333–1340.

Gibson, E., Bergen, L., & Piantadosi, S. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proc. Natl. Acad. Sci. U.S.A.*, 110(20), 8051–8056.

Gibson, E., Desmet, T., Grodner, D., & Watson, D. Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics*, 16(2).

Gibson, E. & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3), 225–248.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP* (pp. 517–520).

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of Human Language* (pp. 73–113). Cambridge, MA: MIT Press.

Griffiths, T., Christian, B., & Kalish, M. (2008a). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science*, 32, 68–107.

Griffiths, T. & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31(3), 441–480.

Griffiths, T., Kalish, M., & Lewandowsky, S. (2008b). Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 363(1509), 3503–3514.

Grimm, J. (1819). *Deutsche Grammatik*. Gottingen: Bei Dieterich.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, (pp. 1–8).

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep Speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hart, B. & Risley, T. (1995). *Meaningful Differences in the Everyday Experience of Young American Children*. Brookes Publishing Company.

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, 18(2), 180–205.

Hauser, M., Chomsky, N., & Fitch, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation* (pp. 187–197).: Association for Computational Linguistics.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hockett, C. (1959). Animal languages and human languages. *Human Biology*, 31(1), 32–39.

Howes, D. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *The Journal of the Acoustical Society of America*, 29(2), 296–305.

Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, 148(34), 13.

Jaeger, T. F. & Buz, E. (2017). Signal reduction and linguistic encoding. *The Handbook of Psycholinguistics*, (pp. 38–81).

Johnson, M., Griffiths, T., & Goldwater, S. (2007). Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (pp. 139–146).

Jordan, M. I. (1998). *Learning in graphical models*, volume 89. Springer Science & Business Media.

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning* (pp. 2342–2350).

Jurafsky, D. & Martin, J. (2009). *Speech & Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, 33(5), 630.

Kantor, J. (1936). An objective psychology of grammar.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165, 45–52.

Kemp, C. & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084), 1049–1054.

Key, M. & Comrie, B., Eds. (2015). *Intercontinental Dictionary Series (IDS).* Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2), 102–110.

Kleinschmidt, D. & Jaeger, T. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148.

Kucera, H. & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuczaj, S. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16, 589–600.

Kuperberg, G. & Jaeger, T. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1), 32–59.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012a). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012b). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990.

Ladd, D., Roberts, S. G., & Dediu, D. (2015). Correlational studies in typological and historical linguistics. *Annu. Rev. Linguist.*, 1, 221–41.

Levesque, H., Davis, E., & Morgenstern, L. (2011). The Winograd schema challenge. In *AAAI Spring Symposium: Logical formalizations of commonsense reasoning*, volume 46 (pp.47).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–77.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, 20(3), 481–508.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the Google Books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations* (pp. 169–174).: Association for Computational Linguistics.

Ling, W., Luís, T., Marujo, L., Astudillo, R., Amir, S., Dyer, C., Black, A., & Trancoso, I. (2015). Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

Luce, P. & Large, N. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processing*, 16(5-6), 565–581.

Luce, P. & Pisoni, D. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hear*, 19(1), 1–36.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Mahowald, K., Fedorenko, E., Piantadosi, S., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318.

Mandelbrot, B. (1954). Simple games of strategy occurring in communication through natural languages. *Transaction of the IRE Professional Group on Information Theory PGIT*, 3(3), 124–137.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press.

Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

Marslen-Wilson, W. (1987a). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71–102.

Marslen-Wilson, W. (1987b). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71–102.

Marslen-Wilson, W. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1), 29 – 63.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, 119(4), 831.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R., & Pennebaker, J. (2007). Are women really more talkative than men? *Science*, 317, 82.

Meylan, S. C., Frank, M. C., Roy, B. C., & Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. *Psychological Science*, 28(2), 181–192.

Michel, J., Shen, Y., Aiden, A. P., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. a., & Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science (New York, N.Y.)*, 331(6014), 176–182.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, T., Kombrink, S., Burget, L., Černockỳ, J., & Khudanpur, S. (2011). Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (pp. 5528–5531).: IEEE.

Miller, G. (1957). Some effects of intermittent silence. *American Journal of Psychology*, 70, 311–314.

Miller, G., Heise, G., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, 41(5), 329.

Mitchell, T. (1997). *Machine Learning*. New York, NY, USA: McGraw-Hill, Inc.

Morrison, C. & Ellis, A. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, 91(2), 167–180.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76(2), 165.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, 33(1), 31–88.

Nieuwland, M., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Zu W., Von Grebmer, S., Bartolozzi, F., Kogan, V., Ito, A., et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, 7, e33468.

Norris, D. & McQueen, J. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, 115(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, 31(1), 4–18.

Pagel, M., Atkinson, Q., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449(7163), 717.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an ASR corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 5206–5210).: IEEE.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3), 241–288.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118(3), 306–338.

Petrov, S. & Klein, D. (2007). Learning and inference for hierarchically split pcfgs. In *Proceedings of the 22nd National Conference on Artificial Intelligence* (pp. 1663).: Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112–1130.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–9.

Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280–291.

Pine, J., Freudenthal, D., Krajewski, G., & Gobet, F. (2013). Do young children have adult-like syntactic categories? Zipf's Law and the case of the determiner. *Cognit*, 127, 345–360.

Pine, J. & Martindale, H. (1996). Syntactic categories in the speech of young children: The Case of the Determiner. *Journal of Child Language*, 23(2), 369–395.

Pine, J. M. & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Appl Psycholing*, 18(2), 123–138.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge University Press.

Pinker, S. (1991). Rules of language. *Science*, 253, 530–535.

Pitt, M., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1), 89–95.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Popović, M. & Ney, H. (2007). Word error rates: Decomposition over pos classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 48–55).: Association for Computational Linguistics.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational linguistics*, 27(2), 249–276.

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333).: Association for Computational Linguistics.

Rohde, H. & Ettlinger, M. (2012). Integration of pragmatic and phonetic cues in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4), 967–983.

Roy, B., Frank, M., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41), 12663–12668.

Roy, B., Frank, M., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Rumelhart, D. & McClelland, J. (1985). On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2. Cambridge, MA: MIT Press.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent–child discourse. In *Children's Language*, volume 4. Lawrence Erlbaum Associates.

Sag, I., Wasow, T., & Bender, E. (1999). *Syntactic theory: A formal introduction*. Stanford, CA: CSLI Press.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2010). Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3), 705–729.

Sanborn, A., Griffiths, T., & Shiffrin, R. (2010). Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology*, 60(2), 63–106.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110–114.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(July 1928), 379–423.

Shannon, C. (1951). Prediction and Entropy of Printed English. *Bell Systems Technical Journal*, 30, 50–64.

Skinner, B. (1957). *Verbal Behavior*. New York: Appleton-Century-Crofts.

Smith, N. (Submitted). ZS: A file format for efficiently distributing, using, and archiving record-oriented data sets of any size.

Smith, N. & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–19.

Stabler, E. (2004). Varieties of crossing dependencies: structure dependence and mild context sensitivity. *Cognitive Science*, 28(5), 699–720.

Stolcke, A. (2002). SRILM – An Extensible Language Modeling Toolkit. In *Proc. Intl. Conf. on Spoken Language Processing*, volume 2 (pp. 901–904).

Storkel, H., Armbrüster, J., & Hogan, T. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49(6), 1175–1192.

Strand, J. & Liben-Nowell, D. (2016). Making long-distance relationships work: Quantifying lexical competition with hidden markov models. *Journal of Memory and Language*, 90, 88 – 102.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, 29, 103–114.

Swadesh, M. (1971). *The Origin and Diversification of Language*. Chicago: Aldine.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. (2008). Baby's first 10 words. *Developmental Psychology*, 44, 929.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022), 1279–1285.

Theakston, A., Lieven, E., Pine, J., & Rowland, C. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28, 127–152.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, 8(4), 451–464.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (pp. 252–259).

Valian, V. (1986). Syntactic categories in the speech of young children. *Dev Psych*, 2, 562–579.

Valian, V., Solt, S., & Stewart, J. (2009). Abstract categories or limited-scope formulae? The case of children's determiners. *J Child Lang*, 36, 743–778.

Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., & Harnad, S. (2016). The latent structure of dictionaries. *Topics in Cognitive Science*, 8(3), 625–659.

Vitevitch, M. & Luce, P. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, 40(3), 374–408.

Vitevitch, M. & Luce, P. (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language*, 52(2), 193–204.

Vitevitch, M., Luce, P., Pisoni, D., & Auer, E. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, 68(1-2), 306–311.

von Humboldt, W. (1970/1836). *On Language: On the diversity of human language construction and its influence on the mental development of the human species*. Cambridge University Press.

Vosoughi, S. & Roy, D. (2012). An automatic child-directed speech detector for the study of child language development. In *Proceedings of Interspeech*.

Wade, T. (1992). *A Comprehensive Russian Grammar*. Oxford: Blackwell.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10), 1550–1560.

Whorf, B. L. (1940). *Science and linguistics*. Bobbs-Merrill Indianapolis, IN.

Xu, J. & Griffiths, T. (2010). A rational analysis of the effects of memory biases on serial reproduction. *Cognitive Psychology*, 60(2), 107–126.

Yang, C. (2010). Who's afraid of George Kingsley Zipf? Unpublished manuscript.

Yang, C. (2013). Ontogeny and phylogeny of language. *Proceedings of the National Academy of Sciences*.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: a new measure of orthographic similarity. *Psychonomic Bulletin and Review*, 15(5), 971–979.

Yule, G. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Touchstone Applied Science Associates (TASA), Inc.

Zipf, G. (1935). *The Psychobiology of Language*. Houghton-Mifflin.

Zipf, G. (1949). *Human Behaviour and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley.

Zwitserlood, P. (1989). The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32(1), 25–64.

A

A.1 Supplementary Material for Chapter 1

A.1.1 Model

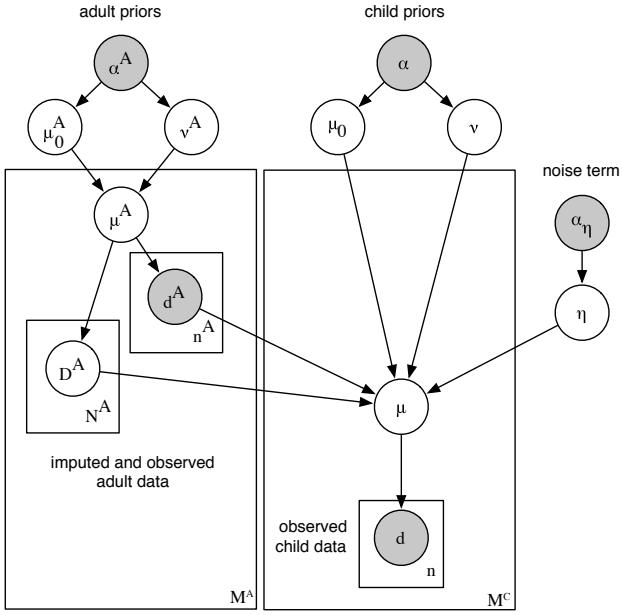
Parameters of the Beta-Binomial Model

The rate at which a child uses "the" rather than "a" for each noun i, is treated as a beta-distributed random variable, μ_i. μ_i has mean $\frac{\mu_o \nu + \eta\left(r_i^A + R_i^A\right)}{\nu + \eta\left(n_i^A + N_i^A\right)}$ and concentration $\nu + \eta\left(n_i^A + N_i^A\right)$, where the child has experienced r_i^A researcher-observed and R_i^A researcher-unobserved uses of noun i with

"the" and, respectively, $n_i^A - r_i^A$ and $N_i^A - R_i^A$ with "a." μ_o and ν describe the prior over determiner preferences across all nouns. Specifically, μ_o indicates the mean determiner preference and ν indicates the concentration (higher values imply that μ_i values are closer to μ_o). η mediates how effectively the child learns from caregiver input the noun-specific determiner preference for each noun. See Figure SA.1 for the complete graphical model.

Details of the Imputation

In our model the child learns from the totality of the linguistic input in his or her lifetime, of which the caregiver speech in our datasets represents only a sample. A side effect of Bayesian inference in our model is the imputation of unobserved caregiver input—D^A in Fig. A.1. For a window starting at time t and ending at time t', we estimate the child's total lifetime number of {*a*,*the*}+noun input tokens from birth through t' based on a rate of 15 million total words of input per year Hart & Risley (1995); Mehl et al. (2007); Roy et al. (2009) and 20 determiner–noun pairs per 1,000 words Godfrey et al. (1992), and assume that nouns occur in the same relative frequencies in the observed and unobserved portions of this total lifetime input. As can be seen in the graphical model in Fig. A.1, inferences about the distribution of determiners for noun i in unobserved caregiver input is constrained by three information sources: Observed caregiver utterances involving noun i, observed caregiver utterances involving *other* nouns, which carry information about the "top-level" caregiver determiner preference (modeled as a beta prior with mean μ_o^A and concentration ν^A on noun-specific caregiver determiner preferences), and observed child utterances, which are in part guided by caregiver input.



Model equations for noun i:

$$\mu_i^A \sim \text{Beta}\left(\mu_o^A, \nu^A\right)$$

$$\mu_i \sim \text{Beta}\left(\frac{\mu_o \nu + \eta\left(r_i^A + R_i^A\right)}{\nu + \eta\left(n_i^A + N_i^A\right)}, \nu + \eta\left(n_i^A + N_i^A\right)\right)$$

$$r_i^A \sim \text{Binom}\left(n_i^A, \mu_i^A\right)$$

$$r_i \sim \text{Binom}\left(n_i, \mu_i\right)$$

$$R_i^A \sim \text{Binom}\left(N_i^A, \mu_i^A\right)$$

Variable definitions:

- ν Strength of child's generalized knowledge regarding determiner preference
- μ_o Child's generalized determiner preference
- μ Child's noun-specific determiner preferences
- η Noise parameter indicating child's effectiveness at learning noun-specific determiner preferences from input
- ν^A Dispersion of caregivers' noun-specific determiner preferences
- μ_o^A Caregivers' generalized determiner preference
- μ^A Caregivers' noun-specific determiner preferences
- α Uninformative prior over μ_o, ν
- α_η Uninformative prior over η
- α^A Uninformative prior over μ_o^A, ν^A
- d Child-produced determiner-noun pairs observed in dataset (comprised of r_i "the" instances and $n_i - r_i$ "a" instances for noun i)
- d^A Caregiver-produced determiner-noun pairs observed in dataset (comprised of r_i^A "the" instances and $n_i^A - r_i^A$ "a" instances for noun i)
- D^A Caregiver-produced determiner-noun pairs *not* observed in dataset (comprised of R_i^A "the" instances and $N_i^A - R_i^A$ "a" instances for noun i)

Figure A.1: Graphical representation of our model. Variables with A superscripts (e.g., μ^A) are "adult" (caregiver) parameters; unsuperscripted variables are child parameters. Shaded nodes indicate observed data (adult and child determiner+noun productions d^A and d) or uninformative priors set by the researcher (α^A, α, and α_η).The M^A and M^C plates correspond to noun types used by the caregiver(s) and the child, respectively; the N^A plate corresponds to adult imputed uses of a given noun, n^A to observed adult uses, and n to observed child uses.

Model Fitting Procedure

We implemented this model using JAGS for Markov chain Monte Carlo based Bayesian inference Plummer (2003). For each model, we took 5 chains of 5000 samples after a burn-in of 2000 adaptive samples and 2000 updates, with thinning of 5 samples (yielding 1000 samples per chain, and 5000 samples total). If the Gelman and Rubin Diagnostic—that the 99th percentile of the potential scale reduction factor, $\hat{R}$ was below 1.1, we considered the model to have converged Gelman et al. (2004), otherwise we ran the chains until convergence in 1000 sample increments. If the model did not meet these convergence criteria by 20,000 samples (100,000 without thinning), we report it as non-converging. Low autocorrelation and good mixing were confirmed through spot visual inspection.

To determine the expectation and distribution of overlap scores predicted by our fitted model for a given child's productions in some time window where each noun i is observed N_i times with either *a* or *the*, we first draw a sample vector of noun-specific child determiner preferences $\{\hat{\mu}\}$ from our MCMC-chain approximation to the posterior over $\{\mu\}$, and then draw for each noun i a new binomially distributed sample of size N_i with mean $\hat{\mu}_i$. The proportion of such samples with at least one instance of both *a* and *the* constitutes a single predicted overlap score for that window. By repeating this process over many sample vectors from the chain, we approximate the posterior predictive distribution on the overlap score for that window, and use it to compute expectations and corresponding HPD intervals.

A.1.2 Data Extraction and Preparation

Corpus work at the scale we describe here is necessarily noisy: poor audio quality, annotator idiosyncrasies, and probabilistic methods for extracting hundreds of thousands of tokens mean that the input to our model inevitably deviates from an ideal data source. Our strategy was thus to test our model across a variety of data preparations to confirm that deviations are of acceptably small magnitude to provide reliable input to the model; indeed, none of the analyses provide us with evidence of systemic problems that might compromise the integrity of our results.

Data Sources

Transcripts for eight developmental corpora Brown (1973); Suppes (1974); Bloom et al. (1974); Kuczaj (1977); Sachs (1983); Theakston et al. (2001); Demuth et al. (2006); Lieven et al. (2009) were downloaded from the CHILDES project at ***childes.psy.cmu.edu***. Utterances from these children ($n = 26$) and their respective caregivers—typically mothers, but also including fathers—were extracted from CHAT-formatted transcripts MacWhinney (2000). These specific corpora were selected because they provide longitudinal coverage within the developmental time period of interest, contain annotated samples of both child and caregiver speech, and in many cases have been used extensively in previous research on grammatical productivity.

To test the model on higher-density data than the corpora available in the CHILDES database, we additionally extracted noun phrases from a ninth corpus, the Speechome Corpus Roy et al. (2015). This annotated corpus spans the 9 through 24 month age range of a child's life (n=1). Em-

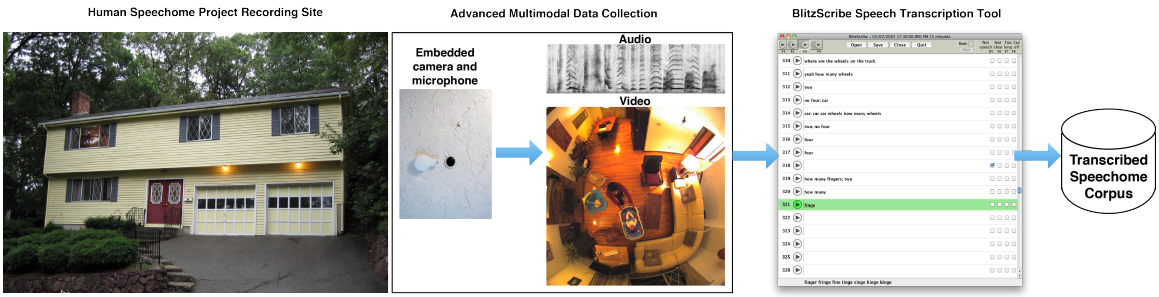


Figure A.2: The Human Speechome Project consists of dense, longitudinal data collection from cameras and microphones embedded in each room of a single child's house. These recordings were transcribed using the custom BlitzScribe transcription tool. The corpus consists of more than 8 million words of transcribed speech and 200,000 hours of audio and video, comprising more than 200 terabytes of media.

bedded cameras and microphones located throughout the child's house were used to achieve an unprecedented level of coverage of language learning in a naturalistic context (Figure SA.2). Annotating the Speechome Corpus was accomplished using new, semi-automatic tools designed for speed and efficiency. Speech from approximately 10 hours per day of raw audio were preprocessed using BlitzScribe, an automated system that uses machine learning techniques to detect and segment speech and assign speaker identities. These samples were then manually transcribed. An estimated 72% (3,618 of 5,000 utterances) of caregiver speech from a balanced sample across time is child-directed, while the remainder is spoken in the presence of the child but not to the child Vosoughi & Roy (2012).

Data Preparation

Determine-noun pairs were extracted from the corpora using three alternative processing pipelines. In the first pipeline ("CLAN"), we extracted determiner and noun pairs from all corpora with CHILDES-compliant annotations using either manually-annotated or, more commonly, machine-

generated dependency parses Sagae et al. (2010). While CLAN is a simple rule-based dependency parser, it incorporates significant domain knowledge and uses special annotations available in CHILDES-formatted files in generating parses. As such, it avoids some of the pitfalls that undermine statistical part-of-speech taggers, often trained on adult speech, when run on samples of early child language.

The two largest datasets, Thomas and Speechome, lack canonical CHILDES annotation, and can only be processed using a statistical part of speech tagger. For this reason we employed two alternative pipelines for extracting determiner+noun pairs, both using a state-of-the-art statistical part-of-speech tagger Toutanova et al. (2003). Determiners that appear without nouns because of interruptions in conversational turn-taking or speech errors were discarded. When the POS tagger identified a series of nouns, we took the first noun as the head of the phrase (the "FN" pipeline) or the last noun as the head of the phrase (the "LN" pipeline).

For all three data extraction pipelines, unrecognizable nouns ("xxx" and "yyy" in CHILDES-formatted files), proper names,* and types shorter than three characters were discarded. Both extraction methods accommodate words intervening between the determiner and noun (e.g. an adjective).

The correct treatment of grammatical variants of similar nouns is not immediately obvious. For example, should a model of determiner productivity track separate counts for "dog" and "dogs," or should these be merged into counts for a single noun? For the CLAN extraction pipeline, we produced three variants of the determiner+noun pairs for each CHILDES dataset. The "Complete" morphology treatment maintained separate counts for all variants; for example, "dog," "doggy," and

*While proper names are generally unlikely to prepended by a determiner, there are many exceptions, including family names ("The Johnsons"), toponyms ("The Gambia","The Hamptons"), historical eras ('The Great Depression'), and publications ("The New York Times").

"dogs" were treated as separate nouns, and their counts were tracked separately. In the "Lemmatized" morphology treatment, records were merged by the lemmatized stem—tokens for any of the three above noun types would be counted as the noun "dog". In the "Singulars" treatment, only singular, unmarked nouns were kept (i.e. counts for "dogs" and "doggy" were discarded). Because the Lemmatized morphology treatment requires morphological parses of the nouns from the CLAN-parsed files, only the Complete and Singulars morphology treatments were available for the LN and FN pipelines.

The combination extraction pipelines and morphology treatments produced seven datasets for each child with fully compliant CHILDES-annotated data, and four datasets for the remaining datasets (Speechome and Thomas). These include 1: Complete-FN, 2: Complete-LN, 3: Complete-CLAN, 4: Lemmatized-CLAN, 5: Singulars-FN, 6: Singulars-LN (the data preparation presented in the main text), and 7: Singulars-CLAN. We conduct our model-based analysis on all available variants for each child, but stress in the main text the results of the model run on singular nouns from the LN extraction pipeline for both consistency with previous work Yang (2013) and high accuracy and precision when compared with gold-standard manual annotation (described below). Descriptive properties for all datasets (LN-Singulars treatment) are provided in Table A.1.

Extraction Procedure Validation

To test the accuracy of the automated extraction pipelines, we compared the lists of identified determiner+noun tokens (before filtering by morphological criteria) with a gold-standard set identified by human annotators. Three paid annotators on Amazon Mechanical Turk found deter-

Corpus	Child	Age Range Yr;Mo	Distinct Days	Interval in Days	Child Tokens (Types)	Caregiver Tokens (Types)	Child % After Filter
Bloom	Peter	1;9–3;1	20	492	4,357 (540)	7,824 (731)	82.2
Brown	Adam*	2;3–5;2	53	1,070	6,370 (911)	5,852 (1,005)	78.4
Brown	Eve*	1;6–2;3	10	275	1,304 (332)	2,890 (483)	70.6
Brown	Sarah*	2;3–5;1	131	1,037	3,958 (773)	8,454 (1,181)	82.1
Kuczaj	Abe	2;4–5;0	190	972	6,360 (1,070)	4,935 (1,070)	81.0
Manch.	Anne	1;10–2;9	31	336	1,515 (369)	6,514 (711)	79.5
Manch.	Aran	1;11–2;10	33	340	2,194 (419)	8,168 (996)	77.3
Manch.	Becky	2;0–2;11	33	338	1,787 (424)	4,335 (638)	75.8
Manch.	Carl	1;8–2;8	33	364	4,392 (410)	4,206 (516)	71.0
Manch.	Domin	1;10–2;10	35	363	467 (147)	4,752 (532)	81.9
Manch.	Gail	1;11–2;11	34	362	1,145 (386)	4,404 (870)	79.1
Manch.	Joel	1;11–2;10	35	339	1,429 (402)	4,694 (846)	78.1
Manch.	John*	1;11–2;10	32	338	2,081 (363)	4,561 (753)	71.1
Manch.	Liz*	1;11–2;10	34	338	1,632 (348)	3,716 (624)	70.8
Manch.	Nic	2;0–3;0	33	362	936 (279)	5,312 (850)	71.9
Manch.	Ruth	1;11–2;11	33	367	928 (226)	5,377 (696)	81.5
Manch.	Warr*	1;10–2;9	33	340	2,901 (438)	6,748 (833)	73.0
Prov.	Alex*	1;4–3;5	51	759	1,706 (367)	6,618 (1,063)	77.7
Prov.	Ethan	0;11–2;11	50	731	1,750 (570)	10,299 (1,225)	79.3
Prov.	Lily	1;1–4;0	80	1,067	3,425 (864)	19,077 (2,287)	80.7
Prov.	Naima*	0;11–3;10	85	1,062	5,710 (1,030)	18,478 (1,880)	76.9
Prov.	Violet	1;2–3;11	51	1,014	1,325 (428)	6,562 (1,315)	76.0
Prov.	William	1;4–3;4	44	733	1,332 (355)	6,164 (952)	76.1
Sachs	Naomi	1;2–4;9	65	1,304	1,472 (438)	2,784 (634)	71.9
Speech.	Speech.*	0;9–2;1	419	488	4,281 (448)	196,331 (6,212)	71.0
Suppes	Nina*	1;11–3;3	48	489	6,367 (704)	11,830 (878)	70.1
Thomas	Thomas*	2;0–5;0	376	1,076	18,989 (1,870)	110,720 (3,958)	85.5

Table A.1: Age range, type and token counts and other properties of corpora analyzed. Counts reflect a data preparation in which only singular nouns are retained and the last noun of any automatically-identified sequence of nouns is assumed to be the head ("Singulars-LN"). Starred children meet the model's convergence criterion in the main analysis (n=11). Child % After Filter indicates the proportion of tokens retained after the application of repetition and imitation filters similar to those used in Yang (2013).

			CLAN		LN		FN	
Transcript	Child	Speaker	Precision	Recall	Precision	Recall	Precision	Recall
First	Alex	Child	—	—	—	—	—	—
		Caregiver	0.93	0.95	0.95	0.90	0.93	0.88
	Eve	Child	0.80	1.00	1.00	1.00	1.00	1.00
		Caregiver	1.00	0.97	0.91	0.91	0.91	0.91
	Warr	Child	1.00	1.00	1.00	1.00	0.92	0.92
		Caregiver	1.00	1.00	0.96	0.96	0.93	0.93
Last	Alex	Child	0.95	0.95	0.76	0.94	0.67	0.70
		Caregiver	0.94	0.93	0.84	0.88	0.75	0.79
	Eve	Child	0.94	0.94	0.93	0.93	0.93	0.87
		Caregiver	0.85	0.92	0.92	0.96	0.92	0.88
	Warr	Child	0.59	0.78	0.92	0.94	0.81	0.83
		Caregiver	0.80	0.89	0.94	0.94	0.84	0.84

Table A.2: Performance of the three automated extraction pipelines compared to gold-standard human annotations for six corpus samples. Recall, the proportion of determiner+noun pairs found by the extraction scripts out of those found by human annotators, reflects the completeness of the extraction method. Precision, the proportion of determiner+noun pairs that were found by human annotators out of those found by the extraction script, reflects the number of false positives. Alex (the child) had no determiner+noun pairs in his first transcript.

miner+noun pairs in the first 1000 lines in the first and last corpora for three children: Alex from the Providence corpus, Eve from the Brown Corpus, and Warr from the Manchester Corpus. Discrepancies between annotators were resolved by majority rule.

The three automated extraction pipelines generally provide similar lists of determiner+noun pairs compared to the manual annotations (Table A.2). Both the CLAN and LN extraction pipelines outperform the FN extraction stack in terms of recall on the twelve transcripts ($p = .012$ and $p = .011$ respectively, per one-tailed Wilcoxon signed rank tests[†]). The LN extraction pipeline outperforms the FN extraction stack on precision as well ($p = .005$ following the same test).

[†]Normal approximations of p values were computed using a continuity correction; zero differences were discarded before ranking absolute differences.

Imputing Determiner+Noun Counts In Adult Speech

For the imputation procedure in the model, unigram probabilities for nouns in the CHILDES American English datasets were obtained by counting all nouns used with a definite or indefinite determiner by maternal or paternal caregivers in all CHILDES American English corpora as of December 2013. Imputation data for the Manchester datasets from the CLAN pipeline are from Manchester alone; for the LN and FN pipeline both British English datasets (Manchester and Thomas) were used. Dialectal differences and conflicting orthographic conventions motivated this decision to maintain separate counts for the imputation. Counts used in the Speechome dataset are from that dataset alone. The imputed caregiver count for each noun is defined as $\lfloor p(n)rd \rfloor$, where $p(n)$ the probability of that noun in the relevant dataset (normalized by the total number of nouns), r is the daily rate of caregiver determiner+noun tokens (here 822), and d is the child age in days.

The coverage provided by the Speechome Corpus allows for an evaluation of the estimated daily rate of determiner+noun pairs used in the imputation step. Given a rate of 15 million total words of input per year Hart & Risley (1995); Mehl et al. (2007) and 20 determiner+noun pairs per 1,000 words in the Switchboard Corpus Godfrey et al. (1992), we estimated that a child hears 822 determiner+noun pairs per day. Daily totals of caregiver tokens from Speechome are higher than this estimate (Figure SA.3). Given that the Speechome corpus is thought to contain approximately 50% of the daily experiences of the target child (∼70% captured, of which ∼70% of the determiner+noun tokens have been annotated), an average of 480 recorded tokens per day corresponds to approximately 960 total determiner+noun tokens per day. We retain the 822 tokens per day as a more con-

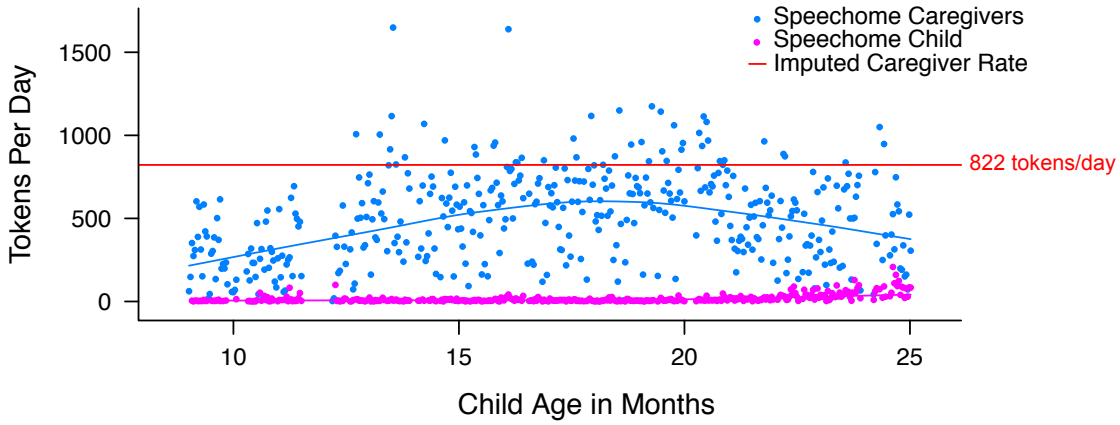


Figure A.3: The observed daily rate of caregiver determiner+noun tokens from the Speechome corpus (blue) is slightly lower than the rate of 822 tokens per day used in the imputation of adult data (marked in red). Loess lines for child and caregiver speech have a span of .67.

servative estimate.

Additional Analyses of the Speechome Corpus

One potential issue in this particular source data is a bias in the assignment of determiner labels on the annotation process. Portions of the audio from the Speechome dataset were periodically assigned and transcribed by multiple annotators, providing a way to assess the quality of the annotations in this dataset. Each speech segment has a primary transcript, but may also have a list of alternate transcripts. These alternates can be used to assess quality by computing inter-annotator agreement, the degree to which multiple annotators independently produce the same transcript for a speech segment.

Our analyses are based on a probabilistic model of determiner choice and are thus robust to some

level of annotation error. However, we wished to determine if there are biases in annotation errors in that a strong bias in determiner classification toward one of the two determiners could artificially inflate ν. Determiner classification can be technically challenging for automated methods and human annotators alike because it involves distinguishing between highly similar, phonetically reduced segments in fluent speech. Both human annotators and automated methods can take advantage of high-level cues to infer a determiner identity different than that present in the audio signal. Since our main concern here is the child's use of determiners, we selected the subset of child speech segments used in our analyses where alternates were available.

Each alternate transcript is first coded as having either ***none***, "a", "the", or ***both*** determiners present. The latter "***both***" category is required, since in some cases a transcript contains both determiners and it is not always possible to align the determiner to the same target noun used in the primary transcript. The primary transcript, on the other hand, is labeled with the determiner that was linked to the target noun in our analysis (but note that a primary transcript containing multiple determiner+noun pairs may enter into this accuracy calculation multiple times with both "a" and "the" labels.) For a speech segment with $k > 0$ alternates, the counts across the above four categories are accumulated, including the primary transcript category, and normalized by the total number of transcripts $k + 1$. These count vectors are grouped by the primary determiner label, accumulated, and again normalized to yield a confusion matrix shown in Table A.3. The vast majority of alternates agree with the determiner label on the primary transcription; the discrepancies are largely cases where the determiner is dropped. Crucially, there are very few confusions between "the" and "a", and there is no evidence of bias either to switch "the" labels to "a" labels or to switch labels in the

Primary Label	None	"the"	"a"	Both	Total Segments
"the"	.22	.74	.03	.00	353
"a"	.28	.03	.67	.03	137

Table A.3: Determiner annotation agreement scores for speech segments with multiple transcripts in the Speechome dataset.

opposite direction. Misclassifications remain symmetric over time while the monthly rate of misclassifications decreases with age. This lends support to the analyses and conclusions based on this data.

A second potential issue—in this case also specific to the Speechome dataset given its reliance on automated speaker identification—is the erroneous assignment of caregiver determiner+noun tokens to the child and vice versa. Low precision in automated speaker identification, corresponding to the attribution of caregiver determiner+noun tokens to the child, would inflate the child's ν estimate. To address this concern, two of the co-authors (MCF and BCR) assessed the accuracy of the speaker identification of all child determiner+noun tokens from the Speechome dataset using clips of the original audio data. Of 9,898 machine-identified determiners attributed to the child, 6,875 were confirmed by manual review (Cohen's $\kappa = 0.979$; discrepancies were resolved by discussion). Of these 6,875 tokens, 2,594 were excluded in the LN treatment, and 2,664 in the FN treatment because of determiners without corresponding nouns (i.e. from reformulations), fragmented words, or out-of-vocabulary words. For Speechome, utterances from all adult speakers were aggregated into a single "caregiver" speaker (14% from father, 18% from the mother, 14% from the nanny, 39% attributed to multiple adult speakers, and 13% unsure).

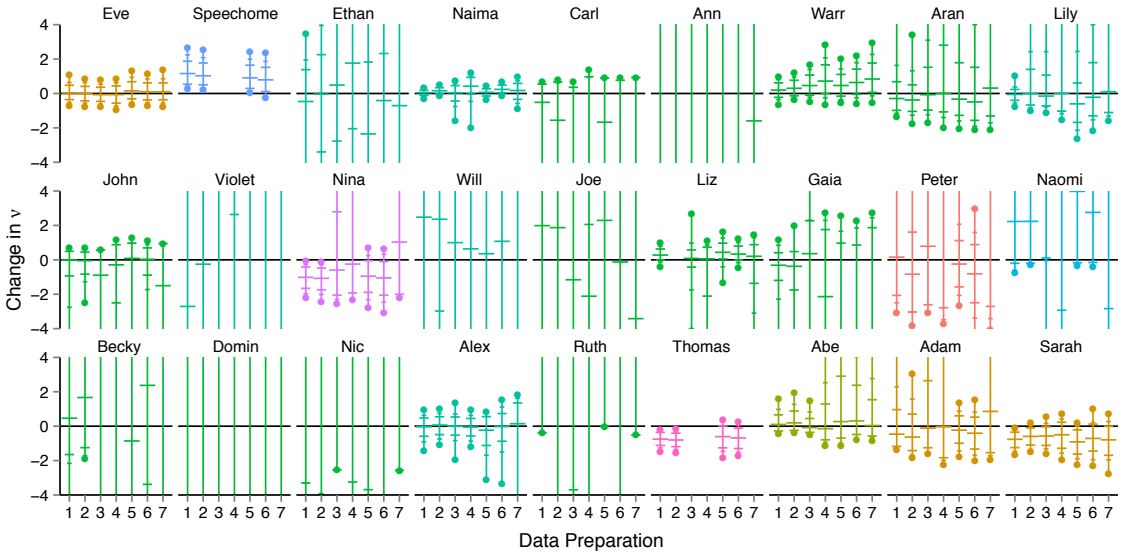


Figure A.4: Posterior distribution of $\Delta\nu$ for all children under all data preparations, 1: Complete-FN, 2: Complete-LN, 3: Complete-CLAN, 4: Lemmatized-CLAN, 5: Singulars-FN, 6: Singulars-LN (the data preparation presented in the main text), 7: Singulars-CLAN . Children are ordered by the mean age of their first interval under the singulars-LN treatment; color indicates the corpus. Vertical lines, from longest to shortest indicate the median of the posterior, the 95% HPD, and the 99% HPD. Points indicate the 99.9% HPD.

A.1.3 Results

Model Convergence

All split-half models converged. Most sliding window models converged (minimum of 267 out of 279 models, in the Complete-LN data preparation).

Predicted vs. Empirical Overlap

Overlap predicted by forward sampling from our model is presented in Table SA.4. For each data preparation, we performed a Wilcoxon rank sum test comparing the empirical overlap with the overlap computed over forward-sampled det+noun tokens. In no condition did the rank sum test reach significance.

Imitation and Repetition Filters for Data

Yang (2013) excluded from analysis child determiner+noun tokens if they were tagged as imitations of the parental speech, as well as within-utterance repetitions by the child. For example, the second instance of "a puzzle" would be discarded if the child said "a puzzle, a puzzle;" if the parent had said "a puzzle" in the preceding utterances *both* would be discarded. A high proportion of repetition and imitation of parental speech on the part of the child could mask initial productivity. On the other hand, such behavior can also be interpreted as genuinely *reflecting* the child's knowledge at that point, in which case excluding such instances from the analysis constitutes an artificial thinning of the data. Constructivist positions assert that the prevalence of rote repetition is itself an impor-

	Current Model			
Data Preparation	r	*RMSE*	$<$ 30 mo.	$>$ 30 mo.
(1) Complete-FN	0.947	0.024	0.023	0.025
(2) Complete-LN	0.941	0.025	0.027	0.024
(3) Complete-CLAN	0.957	0.027	0.027	0.027
(4) Lemmatized-CLAN	0.958	0.032	0.032	0.032
(5) Singulars-FN	0.960	0.028	0.028	0.027
(6) Singulars-LN	0.954	0.029	0.032	0.026
(7) Singulars-CLAN	0.961	0.034	0.036	0.032

Table A.4: Pearson's r and root mean squared error for the current model on the split-half data.

tant characteristic of children's early speech, rather than noise that must be filtered out to discover some underlying knowledge state Lieven et al. (2009); Pine & Lieven (1997). Additionally, imitative and non-imitative uses are hard to distinguish in the real world. Conventions of joint reference in English often lead to cases where two adults use the definite determiner with a noun to refer to some salient discourse referent; to say that one adult speaker imitates the other in such cases is notably problematic.

We chose not to apply this same filter in our primary analysis in that we consider it to be overly conservative for the reasons outlined above, but we report here the results following an approximation to the data preparation in Yang (2013). Because some CHILDES datasets are not annotated with imitation tags and others may follow different classification convention for imitative vs. non-

imitative speech, we applied a uniform filter based on repetition of identical tokens in successive utterances. For CHILDES datasets, a child determiner+noun token was omitted from the analysis if it was used by a caregiver in one of the three immediately preceding caregiver utterances in the same file. CHILDES datasets generally lack timestamps, so this method may erroneously exclude child tokens that follow long intervals without annotated material. The Speechome dataset includes high-resolution temporal information that allows for the application of a more fine-grained filter, in which a token was omitted if it occurred within 15 seconds of a caregiver use or another instance of a child use. The proportion of tokens omitted through these filters ranges from 15-30%, with a strong inverse relationship between mean age and the proportion of tokens omitted (see the rightmost column in Table A.1).

Crucially, the results of our analysis with these filters are consistent with those presented in the main analysis, though confidence intervals for the estimates are substantially wider (compare Figures 3 and SA.5). For the Singulars+LN data preparation, only Naima from the Providence corpus reaches the convergence criteria used in the main analysis of 99.9% HPDs for ν in the interval [0,3] in both the first and second half of tokens. Only five children reach convergence when the criteria are weakened to include children with 99.9% HPDs for ν in the interval [0,9]

These children (Speechome, Eve and Adam from the Brown Corpus, Naima from the Providence Corpus, and Thomas) exhibit similar changes in ν from the first to the second period as in the primary analysis, revealing an overall similar pattern of change (Figure SA.6). In that HPD intervals are significantly wider, in no case can we reject the null hypothesis of no change between developmental time periods (the decrease for Thomas is marginally significant, however, with no change

outside of the 90% HPD). For all children, ν estimates are higher than those reported in the main analysis, suggesting that including repetitions and imitations does indeed produce lower productivity estimates; however, the time-related trends remain robust. We obtain similarly high correlations between predicted and observed overlap (.961–.976 across data preparations), suggesting that this model is equally appropriate for imitation- and repetition- filtered datasets as for unfiltered datasets.

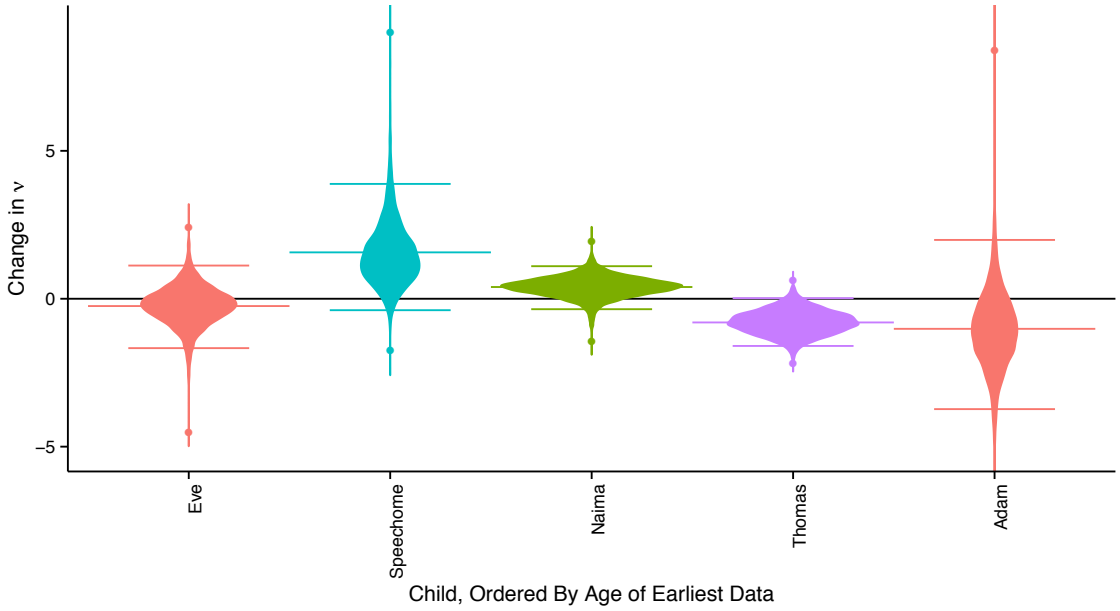


Figure A.5: Posterior estimates for change in productivity between first and second half of children's corpora ($\Delta\nu$) after applying repetition and imitation filters similar to those used in Yang (2013). Longest horizontal lines indicate the median of the posterior, and shorter horizontal lines the 95% HPD. Points indicate the 99.9% HPD. The remainder of the children (n=22) are omitted on the basis of poorly constrained posteriors relating to small sample sizes (99.9% HPD for ν exceeding [0,9] for either time period).

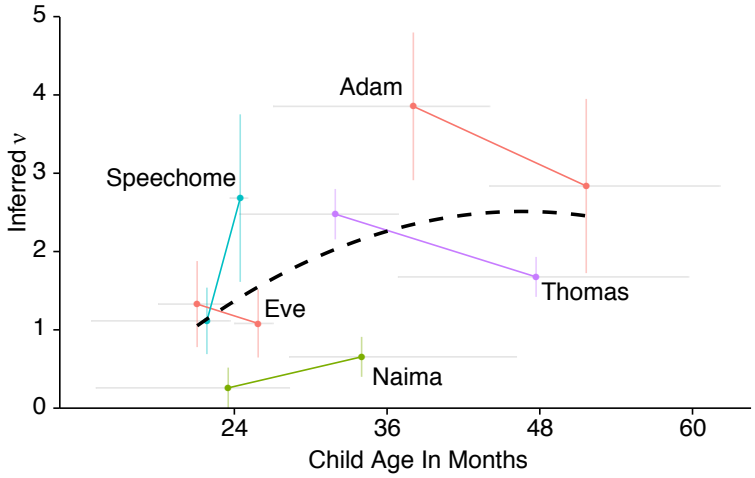


Figure A.6: The inferred developmental trajectory for determiner productivity, ν, across children reaching the convergence criterion after imitations and repetitions are filtered out ($n = 5$). Each line shows a two-point productivity trajectory for a single child, plotted by age in months. Marker size corresponds to the number of child tokens used for each child. Gray horizontal lines indicate the temporal extent of the tokens used to parameterize the model at each point; vertical lines indicate the SD of the posterior. The best fitting quadratic trend is shown as a dashed black line.

B

B.1 Supplementary Material for Chapter 3

In Chapter 2 of this volume I report substantially attenuated support for the finding in Piantadosi et al. (2011) that a word's average information content, or average log probability under a trigram model, is a better predictor than word frequency of word length. In this Appendix, I explore this finding in greater detail, taking into consideration the role of pre-processing and data analysis methods. We first present critiques of the data-processing and analysis choices in Piantadosi et al. (2011), then conduct the analysis with our proposed improved methods. While the results are robust for

English, these analyses suggest that the stronger cross-linguistic conclusion is premature given the data currently available and the method used to estimate average information content.

B.1.1 Text Encoding

Piantadosi et al. (2011) converted word tokens in each language to the closest ASCII transliteration. For English this is a clever data processing choice: almost all English orthographic words can be encoded using ASCII character representations, and data processing operations with ASCII representations are significantly faster—often an order of magnitude so—than those with UTF-8 encoded text. But while there are minimal implications in English, this decision has much more significant implications in other languages. Transliteration incorrectly merges distinct forms, e.g. Spanish *si* ("if") and *sí* ("yes"); consequently, the statistical profiles of multiple wordforms that are distinct in the language's orthography may be erroneously combined. A comparison of lexical information content estimates for Czech, one set computed over UTF-8 text and another from closest ASCII representations, reveals that the downsampling process may significantly perturb information content estimates across the lexicon (Figure B.1)

B.1.2 Words vs. Strings

Piantadosi et al. (2011) test their hypothesis on the 25,000 most frequent strings in the Google 1T datasets, filtering each language by the criterion that a word must appear one or more times in the corresponding OPUS corpus (Tiedemann, 2012). They motivate these inclusion criterion as appropriate for evaluating a broad linguistic claim regarding the correspondence between string length

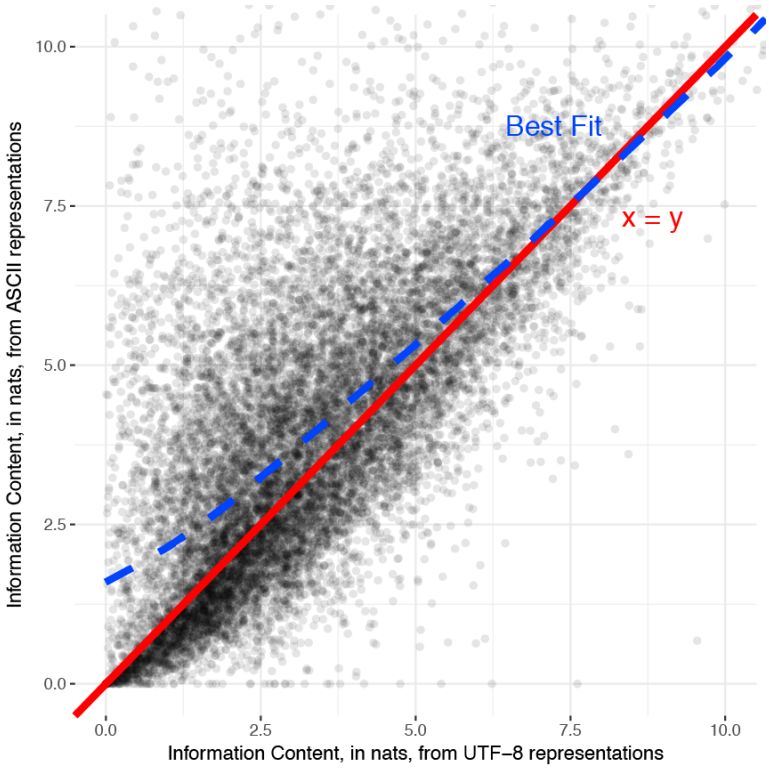


Figure B.1: Correlation in average information content estimates for n = 25000 words under a model computed over ASCII representations and a model computed over UTF-8 representations of the Czech Google 1T corpus. Deviations of the LOESS regression (blue) from the line of identity (red) near the origin indicate that ASCII representations lead to overestimates of information content for highly predictable words.

and predictability in context, not limited to any one type of linguistic material (speech, books, the content of web pages, etc.). However, as an unfortunate consequence of forgoing additional filtering steps, a relatively high proportion of strings in their analysis are of questionable linguistic status. This includes extensive linguistic content from languages other than the target language, with an especially strong presence of English words in many of the other corpora. While language contact and exchange, including the gradual process of loanword adaptation, are standard features of language evolution, we argue that the Google 1T dataset—web scrapes where the source language was identified by a relatively unsophisticated statistical identification—and OPUS—crowd-generated movie subtitles—may have excessive cross-linguistic data pollution that may drive the observed results.

To investigate the composition of the words analyzed in Piantadosi et al. (2011), we used Aspell dictionaries to sort strings into *in dictionary*, *out of dictionary* and *English* categories (tokens from English were sorted only into the first two categories). Aspell is a classic UNIX command line utility for language-specific spell checking; as the backend for system-wide spell checking in other applications, the vocabularies are up to date and extremely large by comparison to traditional dictionaries. By means of a simple combinatorial grammar, Aspell can also evaluate words in languages with complex affixal morphology. Words in a language's dictionary that are also present in English were marked as in-dictionary, e.g., Spanish *pan* (bread). For several of the languages, the resulting word lists were spot-checked by native speakers or proficient L2 speakers to confirm that Aspell-based classifications were appropriate.

Labeling the strings in the analysis with these three categories reveals that a substantial proportion of strings are not commonly accepted word types in the language (Figure B.2). More than 1 in

10 words was found in an English dictionary but not the relevant dictionary for the wordlists for Czech, French, Portuguese, Spanish, and Swedish. While borrowings are to be expected, these rates are significantly higher than those found for Google Books 2012, which range from 0 to 3% of types.

The hypothesis that length is more strongly correlated with average in-context information content than with frequency can then be evaluated *within* within each of these categories: if the relationship is robust, we expect it to obtain among prototypical, in-dictionary types in a language as well. We find that the relationship for the in-dictionary subsets across languages (Figure B.3, column 2) does not reproduce the global pattern for each language found in (Piantadosi et al., 2011) (column 1). While a stronger correlation between length and in-context information content is obtained for within-dictionary types for Dutch, English, French, and Portuguese, we find the opposite preference among the remainder of languages under analysis. Out-of-dictionary types show no strong prevailing pattern (column 3) In languages besides English, types from English show a stronger relationship between frequency (in the target language) and word length (column 4).

Using these this tripartite categorization, we can also investigate how the frequency and length distributions relate between these three subsets across the languages in the sample. This analysis is revealing: the high global correlation between mean in-context information content and word length emerges from the inclusion of all three of these word categories: out-of-dictionary items and items from English are shorter than in-dictionary words, and have lower average information content (Figure B.4, right). On the other hand, words from English have a similar distribution of negative log frequencies, and those not found in either dictionary tend to have a higher negative log frequency that those from the language (Figure B.4, left). In other words the peculiar profile of

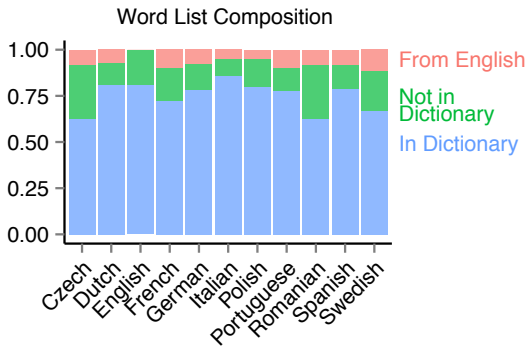


Figure B.2: A significant proportion of the word types in Piantadosi et al. (2011) are not found in the relevant language's dictionary. Word types are especially common in the samples from other languages.

words from English and those found in neither dictionary—short and highly predictable, yet relatively infrequent—depresses the correlation between frequency and character length and inflates the correlation between trigram information content and character length. Taken together, these analyses suggest that the obtained global correlations are highly sensitive to the set of words in the analysis, and that in-context information content is no more predictive than frequency of word length when the correlation is evaluated on words in the respective dictionaries.

B.1.3 Morphology and Orthographic Conventions

Languages vary in the degree of morphological complexity of wordforms. Among those present in the sample in Piantadosi et al. (2011), there is variation in the richness of case-marking systems for nouns, e.g., 6+ cases in Polish (Bielec, 1998), vs. 2 in English (Quirk et al., 1985); degree of inflectional synthesis of verb forms (Bickel & Nichols, 2005), and propensity for pronominal forms to attach to verb forms (pronominal clitics; e.g., Spanish). This variation can have profound consequences for

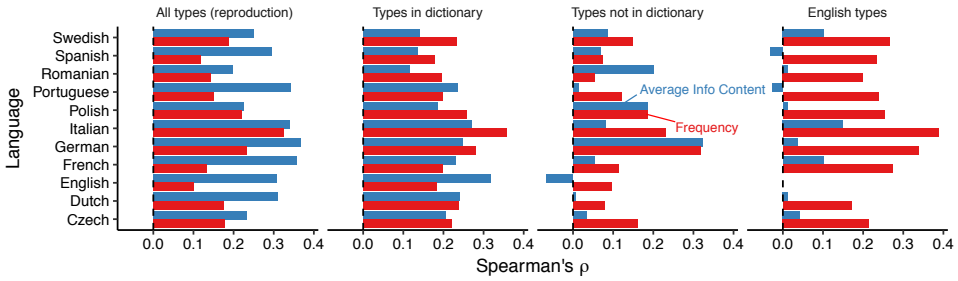


Figure B.3: Piantadosi et al. (2011) found higher global correlations between word length and in-context predictability (information content) as measured by mean trigram surprisal (blue bars in panel 1) than between word length and frequency (unigram surprisal), red bars in panel 1). The pattern is substantially weakened or reversed within each of three sub-groups of word types: those that are found in the relevant dictionary (panel 2), those found in English (panel 4), and those found in neither dictionary (panel 3).

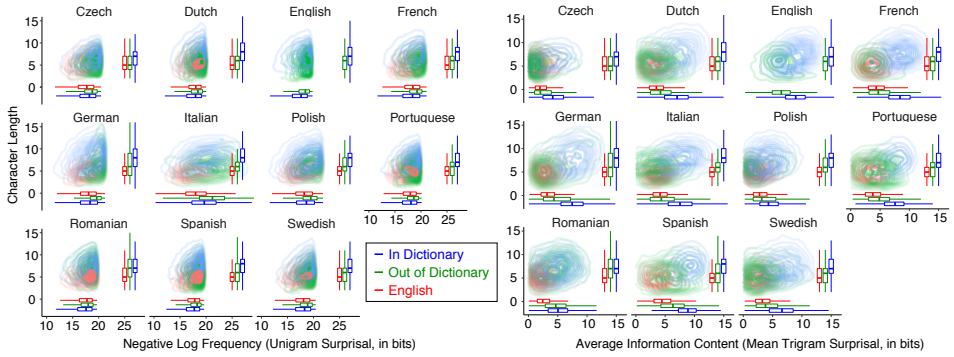


Figure B.4: 2-D density plots depicting the relationship between sentence-level predictor (negative log frequency and mean information content under a trigram model) and word length for the 25,000 most frequent words in 11 languages used in Piantadosi et al. (2011). Words are stratified into three categories: those in each language's spelling dictionary of the target language (blue), those in a spelling dictionary for English (red), and those not in either (green). Marginal boxplots show the median, inter-quartile range (IQR), and 1.5 * IQR for each of the groups. Densities are normalized per category.

the composition of the set of word types under analysis. For example, whereas English would have entries for a handful of forms for the verb *sell* (e.g., *sell, sells, sold, selling*), Spanish, a language with much richer tense system of for verbs, needs to have many more entries for the corresponding verb *vender* owing to the combinatorial space of possible conjugations and object clitics, approximately 160 of which are attested in Google Books 2012. Because the lemma frequency is approximately Zipfian and the use frequency of *sell* is relatively high, 20 variants of *vender* enter the top 25,000 words in the analysis. Depending on what parts of speech have high morphological complexity in each language, substantive differences may emerge in the composition of the word list under analysis across languages: Spanish may have a preponderance of verb forms, while a language with extensive nominal case marking (Russian per Wade 1992). Again, it is unclear what bias may be introduced by this sort of variation in the wordlists under analysis: the implications may vary by language and interact in complex ways with other factors.

While a set of lemmas would be preferable as the target of analysis, appropriate lexical resources are not available across languages. Instead, we propose a method for controlling the wordforms in the analysis. To do this, we conduct the analysis over a subset of word forms from each language intended to match semantic content to the degree possible; specifically we use elicited labels for a matched set of concepts from the Intercontinental Dictionary Series (Key & Comrie, 2015). Similar to Swadesh lists (Swadesh, 1971), IDS datasets contain sets of synonyms matched on conceptual content across a broad sample of languages. Unlike Swadesh lists, IDS datasets include a larger set of approximately 1300 concepts, of which a subset (800-1200) are present in a given language. The dictionary datasets generally contain the unmarked form of the concept, e.g., the singular in the case

of an English count noun, or the infinitive in the case of a Spanish verb. Conducting the analysis on the IDS subset thus both controls for the set of lemmas in the analysis (and excluding lemmas that may be more frequent in a web scrape) and for the number of variants for a given lemma (generally only one).

There are two disadvantage of using the IDS that are worth noting. Because the concepts are present cross-linguistically, the word forms tend to be relatively short and frequent with respect to the broader set of words under analysis in Piantadosi et al. (2011). Second, the total number of items in each language is relatively low, ranging between 800 and 1200 entries. As such, this analysis may overlook a pattern that arises over the totality of the lexicon. However, analogous to the logic above regarding conducting the analysis for in-dictionary words, we should expect to see the same pattern of results to hold for this subset of words if indeed length correlates more strongly with information content than frequency.

Given these refinements and extensions on the methodology in Piantadosi et al. (2011), we conduct a new analysis to evaluate support for the hypothesis that word length are more strongly correlated with information content than frequency. We retain UTF-8 encoded word representations throughout all analyses. In Study 1, the most similar analysis to the original study, we limit words under analysis to the 25,000 most frequent strings that are also seen at lest once in the OPUS corpus. In Study 2, we analyze these two relationships of interest among the 25,000 most frequent words that are also in the relevant language's Aspell dictionary. In Study 3, we take sets of matched IDS wordlists to account for the conceptual and morphological variation under analysis. In all cases, we extend these analyses to an additional large-scale cross-linguistic dataset that has since been made

available: Google Books 2012 (Lin et al., 2012).

B.1.4 Study 1: All Word Types in OPUS Subtitle Corpora

We first reproduce the analysis from Piantadosi et al. (2011), examining all words appearing one or more times in OPUS but recomputing frequencies and trigram surprisal estimates over UTF-8 representations. In addition to the Google 1T dataset used previously we also evaluate the correlation for datasets in Google Book 2012. In this case, the only distinction is that information content estimates are computed over the UTF-8 version of the dataset. Even with this relatively minor data processing manipulation, we find somewhat attenuated support for the pattern of results found by Piantadosi et al. (2011) (Figure B.5). Among the Google 1T corpora, Czech and Polish fail to reach significance. Among the Google Books 2012 corpora, Hebrew, Russian and Spanish exhibit the opposite pattern, with frequency as the stronger correlate of word length, reaching significance in the first case.

B.1.5 Study 2: In-Dictionary Word Types

In the second study, we enforce a stronger constraint on word types entering the analysis: we limit word types to those whose lowercase form can be found in the relevant Aspell dictionary [*]. This typically excludes proper nouns including person and place names, acronyms, and loanwords from other languages. Information content estimates for words within this list reflect all word types.

Enforcing this stronger constraint on the word types under analysis results in a substantive

[*]We allowed uppercase forms for German, which capitalizes all nouns by convention.

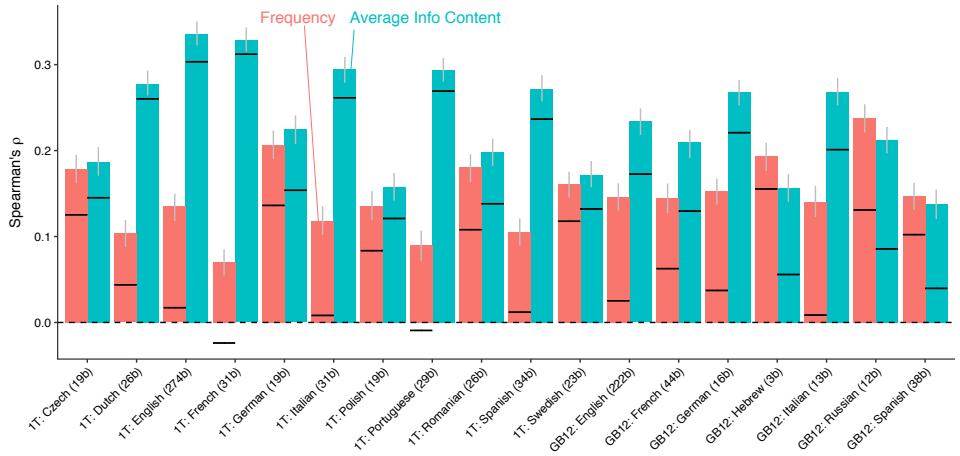


Figure B.5: Correlations between information content and word length (green) and frequency and word length (red) on Google 1T and Books 2012 corpora for the 25,000 most frequent word types in each language that are also present in OPUS. Error bars indicate bootstrapped 95% confidence intervals.

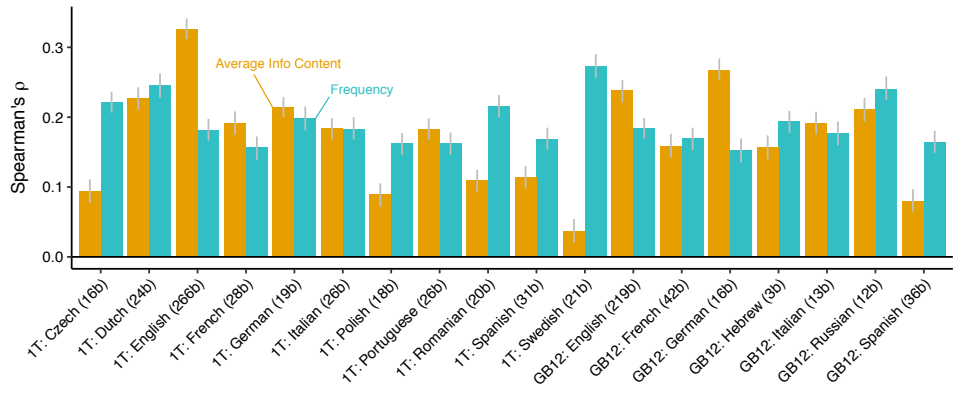


Figure B.6: Correlations between information content and word length (brown) and frequency (negative log unigram probability) and word length (green) on eleven languages in the Google 1T corpus for the 25,000 most frequent in-dictionary word types in each language. Error bars indicate bootstrapped 95% confidence intervals.

change in the pattern of results (Figure B.6). Among the eleven languages from the Google 1T corpora, only two (English and French show significantly higher correlations for information content and word length than frequency (treated here as unigram surprisal, or negative log probability) and word length. 5 of 11 languages (Czech, Polish, Romanian, Spanish, and Swedish) show a significantly higher correlation for word length and frequency. Neither predictor is significantly stronger among the remaining 4 languages (Dutch, German, Italian, and Portuguese).

B.1.6 Study 3: Types in the International Dictionary Series

In our third study, we limit the set of types under analysis to those in the International Dictionary Series, or IDS. This limits the number of word types in the analysis associated with any one lemma, in that only morphologically unmarked forms are used to construct the dictionary for each language.

Matching semantic content to the items in the IDS means that the identity of concepts can be used as a control variable. Here we compare two mixed-effects regression models that predict word length, one using unigram surprisal and the other mean trigram surprisal (average information content). Unigram surprisal, trigram surprisal, and word length were all standardized within each language. Language and concept identity were both treated as random intercepts. This corresponds to the intuition that different concepts have different average word lengths across languages, and that some languages may have longer orthographic representations than others.

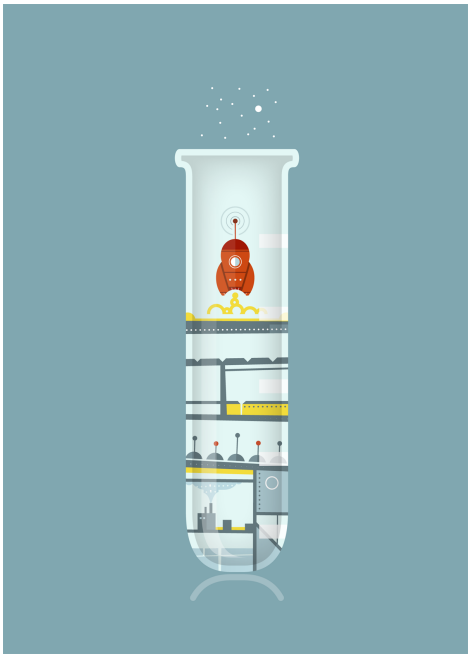
Results from the model comparison are presented in Table B.1. Model 2, where negative log probability is included as the sole fixed effect, demonstrates a better fit than Model 1, where average in-

Table B.1: Comparison of Mixed Effects Models for Word Types in IDS

	Model 1	Model 2
Average information content	0.248***	
	(0.009)	
Negative log probability		0.391***
		(0.010)
Constant	0.025	0.011
	(0.020)	(0.021)
N	13248	13248
Log Likelihood	−16858.080	−16485.720
AIC	33726.170	32981.450
BIC	33763.620	33018.900
	***p < .01; **p < .05; *p < .1	
Random Effects		
# of IDS Concepts	1299	1299
IDS Concepts Standard Deviation	0.546	0.561
# of Languages	11	11
Languages Standard Deviation	0.034	0.039

formation content is the sole fixed effect, according to model log likelihood, AIC and BIC. Thus in a small (n=1200) sample of morphologically unmarked forms, we find that frequency is a better predictor than average predictability of word length.

The results of these three analyses challenge the conclusion that word length is driven by in-context predictability across languages. Positive results for certain larger datasets (e.g., English in Study 2) suggest that this pattern is robust in some languages, and leave open the possibility that larger datasets and better means of estimating information content may reveal that the pattern originally claimed by Piantadosi et al. (2011) is indeed robust. For now, we conclude that this relationship demands further study, with careful consideration of data processing, the set of lexical items under analysis, and the model with which in-context predictability is measured.



This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

The above illustration, *Science Experiment 02*, was created by Ben Schlitter and released under CC BY-NC-ND 3.0. A template that can be used to format a PhD dissertation with this look & feel has been released under the permissive AGPL license, and can be found online at github.com/suchow/Dissertate or from its lead author, Jordan Suchow, at suchow@post.harvard.edu.