Predicting the birth of a spoken word

Article *in* Proceedings of the National Academy of Sciences · September 2015

DOI: 10.1073/pnas.1419773112

Brandon Cain Roy
Massachusetts Institute of Technology

Predicting the birth of a spoken word

Brandon C. Roy[a,b,1], Michael C. Frank[b], Philip DeCamp[a], Matthew Miller[a], and Deb Roy[a]

[a]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Department of Psychology, Stanford University, Stanford, CA 94305

Edited by Richard N. Aslin, University of Rochester, Rochester, NY, and approved August 11, 2015 (received for review October 19, 2014)

Children learn words through an accumulation of interactions grounded in context. Although many factors in the learning environment have been shown to contribute to word learning in individual studies, no empirical synthesis connects across factors. We introduce a new ultradense corpus of audio and video recordings of a single child's life that allows us to measure the child's experience of each word in his vocabulary. This corpus provides the first direct comparison, to our knowledge, between different predictors of the child's production of individual words. We develop a series of new measures of the distinctiveness of the spatial, temporal, and linguistic contexts in which a word appears, and show that these measures are stronger predictors of learning than frequency of use and that, unlike frequency, they play a consistent role across different syntactic categories. Our findings provide a concrete instantiation of classic ideas about the role of coherent activities in word learning and demonstrate the value of multimodal data in understanding children's language acquisition.

word learning | language acquisition | multimodal corpus analysis | diary study

Adults swim effortlessly through a sea of words, recognizing and producing tens of thousands every day. Children are immersed in these waters from birth, gaining expertise in navigating with language over their first years. Their skills grow gradually over millions of small interactions within the context of their daily lives. How do these experiences combine to support the emergence of new knowledge? In our current study, we describe an analysis of how individual interactions enable the child to learn and use words, using a high-density corpus of a single child's experiences and novel analysis methods for characterizing the child's exposure to each word.

Learning words requires children to reason synthetically, putting together their emerging language understanding with their knowledge about both the world and the people in it (1, 2). Many factors contribute to word learning, ranging from social information about speakers' intentions (3, 4) to biases that lead children to extend categories appropriately (5, 6). However, the contribution of individual factors is usually measured either for a single word in the laboratory or else at the level of a child's vocabulary size (4, 6, 7). Although a handful of studies have attempted to predict the acquisition of individual words outside the laboratory, they have typically been limited to analyses of only a single factor: frequency of use in the language the child hears (8, 9). Despite the importance of synthesis, both for theory and for applications like language intervention, virtually no research in this area connects across factors to ask which ones are most predictive of learning.

Creating such a synthesis, our goal here, requires two ingredients: predictor variables measuring features of language input and outcome variables measuring learning. Both of these sets of measurements can be problematic.

Examining predictor variables first, the primary empirical focus has been on the quantity of language the child hears. Word frequencies can easily be calculated from transcripts (7, 8), and overall quantity can even be estimated via automated methods (10). Sheer frequency may not be the best predictor of word learning, however. Although some quantity of speech is a prerequisite for learning, the quality of this speech, and the interactions that support it, is likely to be a better predictor of learning (2, 11, 12). In the laboratory, language that is embedded within coherent and comprehensible social activities gives strong support for meaning learning (3, 13). In addition, the quantity of speech directed toward the child predicts development more effectively than total speech overheard by the child (14).

Presumably, what makes high-quality, child-directed speech valuable is that this kind of talk is grounded in a set of rich activities and interactions that support the child's inferences about meaning (2, 11). Measuring contextually grounded talk of this type is an important goal, yet one that is challenging to achieve at scale. In our analyses, we introduce data-driven measures that quantify whether words are used in distinctive activities and interactions, and we test whether these measures predict the child's development.

Outcome variables regarding overall language uptake are also difficult to measure, especially for young children. Language uptake can refer to both word comprehension and word production, with comprehension typically occurring substantially earlier for any given word (15). In-laboratory procedures using looking time, pointing, or event-related potentials can yield reliable and detailed measures of young children's comprehension, but, typically, only for a handful of words (e.g., refs. 14, 16). For systematic assessment of overall vocabulary size, the only methods standardly used with children younger than the age of 3 y are parent report checklists (15) and assessment of production through vocabulary samples (8). We adopt this second method here. By leveraging an extremely dense dataset, we can make precise and objective estimates of the child's productive vocabulary through

Significance

The emergence of productive language is a critical milestone in a child's life. Laboratory studies have identified many individual factors that contribute to word learning, and larger scale studies show correlations between aspects of the home environment and language outcomes. To date, no study has compared across many factors involved in word learning. We introduce a new ultradense set of recordings that capture a single child's daily experience during the emergence of language. We show that words used in distinctive spatial, temporal, and linguistic contexts are produced earlier, suggesting they are easier to learn. These findings support the importance of multimodal context in word learning for one child and provide new methods for quantifying the quality of children's language input.

Author contributions: D.R. conceived and supervised the Human Speechome Project; P.D. developed the data recording infrastructure; B.C.R., P.D., M.M., and D.R. developed new analytic tools; B.C.R., M.C.F., and D.R. designed research; B.C.R., M.C.F., M.M., and D.R. performed research; B.C.R. and M.C.F. analyzed data; B.C.R., M.C.F., and D.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in GitHub, a web-based repository hosting service, https://github.com/bcroy/HSP_wordbirth.

[1]To whom correspondence should be addressed. Email: bcroy@media.mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1419773112/-/DCSupplemental.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

the identification of the first instance of producing an individual word. Although this method does not yield estimates of comprehension vocabulary, production can be considered a conservative measure: If a child is able to use a word appropriately, he or she typically (although not always) can understand it as well.

In addition to the measurement issues described above, studies that attempt to link input to uptake suffer from another problem. The many intertwined connections between parent and child (genetic, linguistic, and emotional) complicate direct causal interpretations of the relationship between input and learning (17). Some analyses use longitudinal designs or additional measurements to control for these factors (e.g., refs. 7, 14). Here, we take a different approach: We use a classic technique from cognitive (18) and developmental psychology (19), the in-depth case study of a single individual, treating the word as the level of analysis rather than the child. We make distinct predictions about individual words based on the particular input the child receives for that word (holding the child and caregiving environment constant across words).

Using this single-child case study, we conduct two primary analyses. First, we measure the contribution of input frequency in predicting the child's first production of individual words and examine how it compares with other linguistic predictors at a word-by-word level, examining this relationship both within and across syntactic categories. Next, we add to this analysis a set of novel predictors based on the distinctiveness of the contexts in which a word is used; these predictors dominate frequency when both are included in a single model.

The contribution of this work is twofold. First, we develop a set of novel methods for measuring both language uptake and the distinctiveness of the contexts in which words appear and show how these methods can be applied to a dense, multimodal corpus. Second, we provide an empirical proof of concept that these contextual variables are strong predictors of language production, even controlling for other factors. Although the relationship between the contexts of use for a word and its acquisition has been proposed by many theorists (2, 11), it has yet to be shown empirically. Because our empirical findings come from correlational analyses of data from a single child, whose individual environment is, by definition, unique, these findings must be confirmed with much larger, representative samples and experimental interventions to measure causality. Nevertheless, the strength of the relationships we document suggests that such work should be a priority.

Current Study

We conducted a large-scale, longitudinal observation of a single, typically developing male child's daily life. The full dataset consists of audio and video recordings from all rooms of the child's house (Fig. S1) from birth to the age of 3 y, adding up to more than 200,000 h of data. For the current study, we focus on the child's life from 9–24 mo of age, spanning the period from his first words ("mama" at 9 mo) through the emergence of consistent word combinations. From our data, we identified 679 unique words that the child produced. Although it is quite difficult to extrapolate from this production-based measure exactly how the child would have scored on a standardized assessment, 341 of the child's words appear on the MacArthur–Bates Communicative Development Inventory Words and Sentences form. With these words checked, he would have scored in approximately the 50th percentile for vocabulary (15). By the end of the study, when the child was 25 mo old, he was combining words frequently and his mean length of utterance (MLU) was ∼2.5 words.

Recording took place ∼10 h each day during this period, capturing roughly 70% of the child's waking hours. Automatic transcription for such naturalistic, multispeaker audio is beyond the current state of the art, with results below 20% accuracy in our experiments (20); therefore, using newly developed, machine-assisted speech transcription software (21), we manually transcribed nearly 90% of these recordings. We only transcribed speech recorded from rooms within hearing range of the child and during his waking hours. The resulting high-quality corpus consists of ∼8 million words (2 million utterances) of both child speech and child-available speech by caregivers that could contribute to the child's linguistic input. Each utterance was labeled with speaker identity using a fully automatic system (more details of data processing and transcription are provided in *SI Materials and Methods* and Figs. S2 and S3).

Our primary outcome of interest was the child's production of individual words. For each of the words the child produced in the transcripts, we labeled the age of first production (AoFP) as the point at which the child first made use of a phonological form with an identifiable meaning [even though forms often change (e.g., "gaga" for "water"); *SI Materials and Methods*]. These AoFP events were identified automatically from transcripts and then verified manually (Figs. S4–S7). Although the child's abilities to comprehend a word and to generalize it to new situations are also important, these abilities are nearly impossible to assess with confidence from observational data. In contrast, we were able to estimate AoFP with high precision.

Predicting Production

Unlike smaller corpora, our dataset allows us to quantify and compare predictors of word production. In our initial comparison, we focus on three variables: ease of producing a word, complexity of the syntactic contexts in which it appears (22), and amount of exposure to it (7). In each case, we use a very simple metric: length of the target word (in adult phonemes); mean length (in words) of the caregiver utterances in which the target word occurs before the child first produces it (MLU); and logarithm of the average frequency of the target word's occurrence each day, again before the child's first production. Although there are more complex proxies for ease of production (23) or syntactic complexity of the input contexts (24), these simple computations provide robust, theory-neutral measures that can easily be implemented with other corpora.

Each of these three predictors was a significant independent correlate of AoFP ($r_{phones} = 0.25$, $r_{MLU} = 0.19$, and $r_{freq} = -0.18$, all $P < 0.001$). Longer words and words heard in longer sentences tended to be produced later, whereas those words heard more frequently tended to be produced earlier. These relationships remained relatively stable when all three factors were entered into a single linear model (Fig. 1*A*, baseline model), although the effect of frequency was somewhat mitigated.

A notable aspect of this analysis is the role played by predictors across syntactic categories. Frequency of occurrence was most predictive of production for nouns, although it had little effect for predicates or closed-class words (Fig. 1). Higher use frequency may allow children to make more accurate inferences about noun meaning just by virtue of increased contextual co-occurrence (25, 26). In contrast, the complexity of the syntactic contexts in which predicate terms occur appears to be more predictive of the age at which they are acquired (27). Like predicates, closed-class words were also learned later and were better predicted by MLU than by frequency. Those closed-class words appearing in simple sentences (e.g., "here," "more") were learned early, whereas those closed-class words typically found in longer sentences were learned late (e.g., "but," "if"), as would be expected if producing these words depended on inferring their meaning in complex sentences.

Successively incorporating predictors allows us to examine the relationship between individual predictors and particular words through improvements in predicted AoFP [Fig. 2 and online interactive version (wordbirths.stanford.edu/)]. Long words like "breakfast," "motorcycle," or "beautiful" are predicted to be learned later when the number of phonemes is added to the model; words

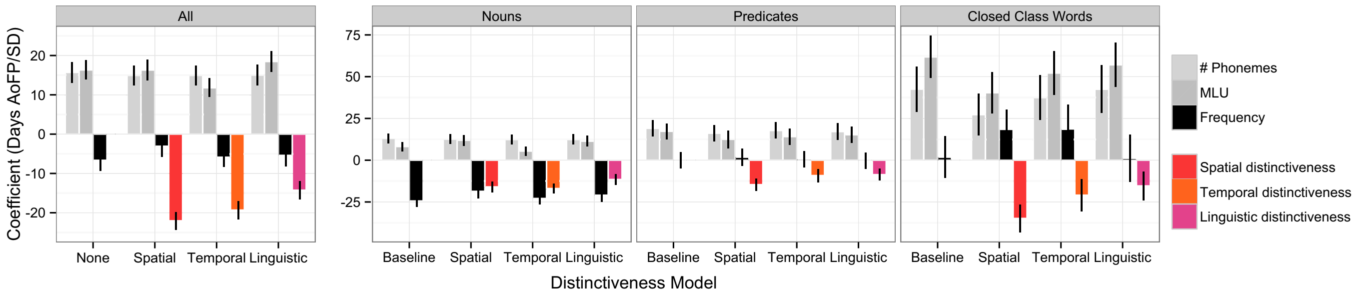

Fig. 1. Regression coefficients (±SE) for each predictor in a linear model predicting AoFP. Each grouping of bars indicates a separate model: a baseline model with only the number of phonemes, MLU, and frequency or a model that includes one of the three distinctiveness predictors. Red/orange/purple bars indicate distinctiveness predictors (spatial/temporal/linguistic). Coefficients represent number of days earlier/later that the child will first produce a word per SD difference on a predictor. (*Right*) Three plots show these models for subsets of the vocabulary.

that often occur alone or in short sentences like "no," "hi," and "bye" are predicted to be learned earlier when MLU is added. Although previous work on vocabulary development has relied on between-child analyses of vocabulary size, our analyses illustrate how these trends play out within the vocabulary of a single child.

Quantifying Distinctive Moments in Acquisition

Jerome Bruner hypothesized the importance of "interaction formats" for children's language learning (11). These formats were repeated patterns that were highly predictable to the child, including contexts like mealtime or games like "peek-a-boo," within which the task of decoding word meaning could be situated. He posited that inside these well-understood, coherent activities, the child could infer word meanings much more effectively. Such activities might therefore play a critical role in learning.

Inspired by this idea, we developed a set of formal methods for measuring the role of such distinctive moments in word learning. We examined three dimensions of the context in which a word appears: the location in physical space where it is spoken, the time of day at which it is spoken, and the other words that appear nearby it in the conversation. We hypothesized that distinctiveness in each of these dimensions would provide a proxy for whether a word was used preferentially in coherent activities.

For each dimension (time, space, and language), we created a baseline distribution of the contexts of language use generally and measured deviations from it. We derived spatial distributions from motion in the videos, capturing the regions in the child's home where there was motion while words were being used. We first clustered the pixels of video in which coherent motion existed and then measured motion in the 487 resulting clusters (most spanned 0.35–0.65 m^2) during 10-s time windows surrounding each word. Automatic motion detection is a robust indicator of both the location and trajectories of human activity. Temporal distributions were created based on the hour of the day in which a word was uttered.

Linguistic context distributions were built by using a latent Dirichlet allocation (LDA) topic model, which produced a set of distinct linguistic topics based on a partition of the corpus into a set of 10-min "documents" (28). At this temporal resolution, language related to everyday activities, such as book reading and mealtime, is identifiable and might span one or a few 10-min episodes, yielding topics that reflect linguistic regularities related to these activities. To map this distribution onto individual words, we computed the distribution of topics for each document within which a word occurred.

Once we had created context distributions for each dimension, we computed the distinctiveness of words along that dimension. We took the Kullback–Leibler (KL) divergence between the distribution for each word and the grand average distribution (e.g., the overall spatial distribution of language use across the child's home) (29). Because KL divergence estimators are biased with respect to frequency (30), we explored a number of methods for correcting this bias, settling on using linear regression to remove frequency information from each predictor (*SI Materials and Methods*). The resulting distinctiveness measures capture the distance between the contextual distribution of the word and the contextual distribution of language more generally. For example,

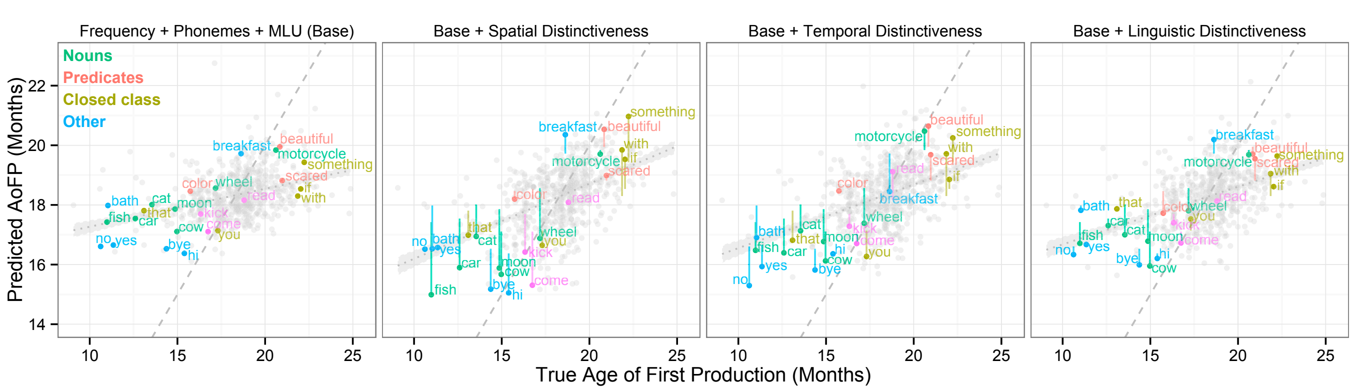

Fig. 2. Predicted AoFP plotted by true AoFP for successive regression models. Each dot represents a single word, with selected words labeled and lines showing the change in prediction due to the additional predictor for those words. Color denotes word category, the dotted line shows the regression trend, and the dashed line shows perfect prediction. (*Left*) Plot shows the baseline model, which includes frequency, phonemes, and utterance length. (*Right*) Subsequent three plots show change due to each distinctiveness predictor when added to the baseline model. An interactive version of this analysis is available at wordbirths.stanford.edu/.

words like "fish" or "kick" have far more distinct spatial, temporal, and linguistic distributions than the word "with" (Fig. 3).

The more tied a word is to particular activities, the more distinctive it should be along all three measures, and the easier it should be to learn. Consistent with this hypothesis, contextual distinctiveness (whether in space, time, or language) was a strong independent predictor of the child's production. Each of the three predictors correlated with the child's production more robustly than frequency, MLU, or word length, with greater contextual distinctiveness leading to earlier production ($r_{spatial} = -0.40$, $r_{temporal} = -0.34$, $r_{linguistic} = -0.28$, all $P < 0.001$).

These relationships were maintained when the distinctiveness predictors were entered into the regression models described above (Fig. 1*A*). Because the distinctiveness predictors were highly correlated with one another ($r = 0.50$–0.57, all $P < 0.001$; Fig. S8), we do not report a single joint analysis [although it is available in our interactive visualization (wordbirths.stanford.edu/)]; models with such collinear predictors are difficult to interpret.

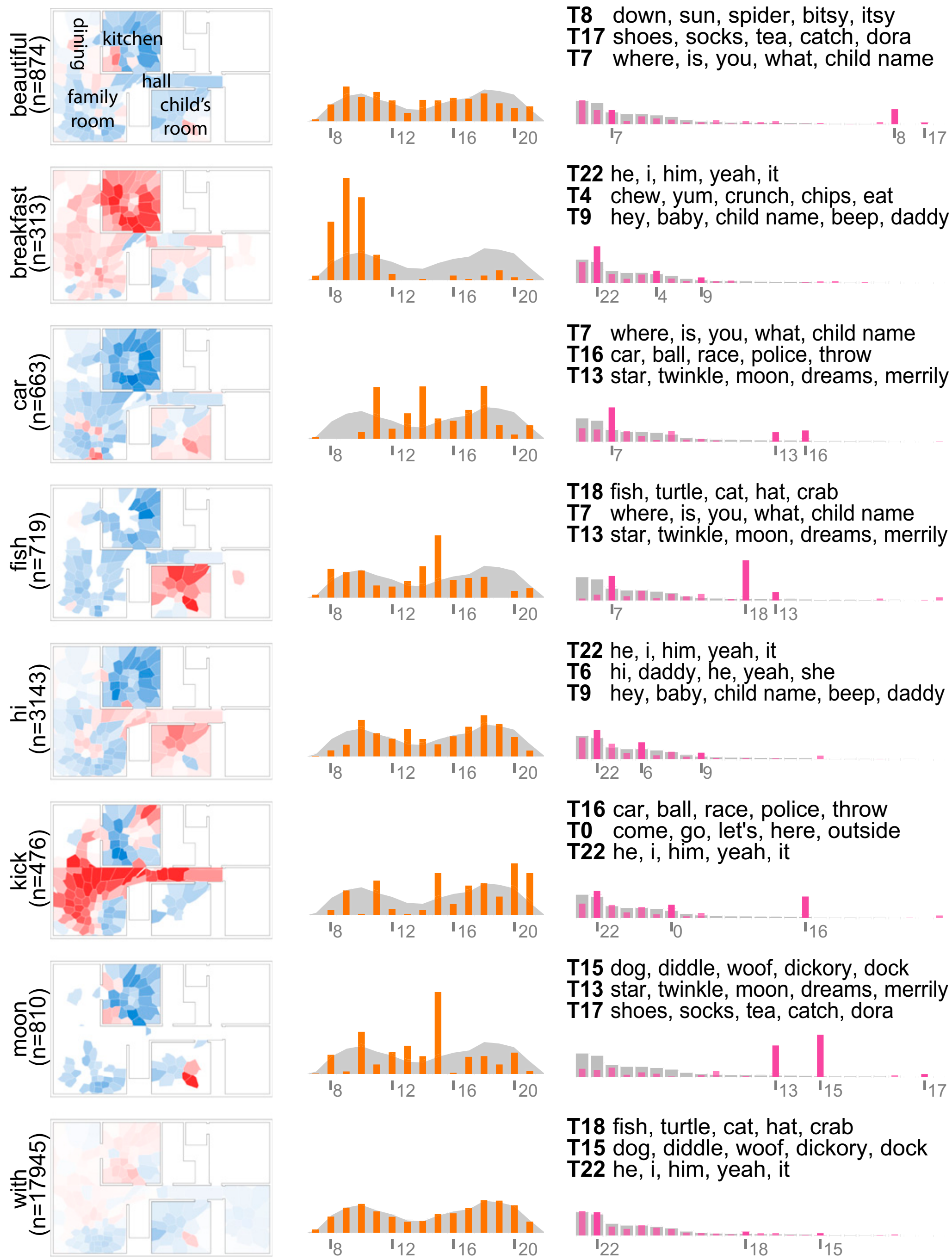

Fig. 3. Examples of eight spatial, temporal, and linguistic context distributions for words. Spatial distributions show the regions of the house where the word was more (red) and less (blue) likely than baseline to be used. Rooms are labeled in the topmost plot. Temporal distributions show the use of the target word throughout the day, grouped into 1-h bins (orange) and compared with baseline (gray). Linguistic distributions show the distribution of the word across topics (purple), compared with the baseline distribution (gray). The top five words from the three topics in which the target word was most active are shown above the topic distribution.

Nevertheless, the distinctiveness predictors did make different predictions for some words. For example, the words "diaper" and "change" were highly concentrated spatially but quite diffuse in time, consistent with their use in a single activity (Table S1).

All three distinctiveness measures were significant predictors of AoFP, with spatial distinctiveness and temporal distinctiveness being the strongest predictors in their respective models. The strength of word frequency was reduced dramatically in all models, despite its very low correlation with the distinctiveness predictors (r values between −0.09 and −0.02; Tables S2–S6).

Our distinctiveness measures did not simply pick out different syntactic categories. Instead, and in contrast to word frequency, they had relatively consistent effects across classes (Fig. 1). For predicates, there was essentially no effect of frequency, but all three distinctiveness predictors still had significant effects. In contrast, frequency was still a strong predictor for nouns even when distinctiveness was included. In some models of closed-class words, frequency was even a positive predictor of AoFP (higher frequency leading to later production), presumably because the most frequent closed-class words are among the most abstract and least grounded in the details of specific contexts (e.g., "the," "and," "of").

The distinctiveness predictors also did not simply recreate psycholinguistic constructs like imageability. We identified the 430 words in the child's vocabulary for which adult psycholinguistic norms were available (31). Within this subset of words, all three distinctiveness factors were still significant predictors when controlling for factors like imageability and concreteness.

In sum, despite the radically different data they were derived from (video of activities, time of day for each utterance, and transcripts themselves), the three distinctiveness variables showed strong correlations with one another and striking consistency as predictors of the age at which words were first produced. This consistency supports the hypothesis that each is a proxy for a single underlying pattern: Some words are used within coherent activities like meals or play time (e.g., breakfast, kick), whereas others are used more broadly across many contexts. These differences may be a powerful driver of word learning.

Conclusions

Children learn words through conversations that are embedded in the context of daily life. Understanding this process is both an important scientific question and a foundational part of building appropriate policies to address inequities in development. To advance this goal, our work here created measures of the grounded context of children's language input, and not just its quantity. We used distributional distinctiveness of words in space, time, and language as proxies for the broader notion of their participation in distinctive activities and contexts. We hypothesized that these activities provide consistent, supportive environments for word learning.

We found support for this hypothesis in dense data from a single child. Across words and word categories, those words that were experienced in more distinctive contexts were produced earlier. Because the distinctiveness measures, especially spatial distinctiveness, were more predictive of learning than quantity of linguistic exposure, our findings support the utility of probing the contexts within which words are used and provide a strong argument for the importance of multimodal datasets.

The causal structure of language acquisition is complex and multifactorial. The greater children's fluency is, the greater is the complexity of their parents' language (32), and the more words children know, the better they can guess the meanings of others (5). In the face of this complexity, about which relatively little is still known, we chose to use simple linear regression, rather than venturing into more sophisticated analyses. This conservative choice may even understate the degree to which our primary predictors of interest affect the child's earliest words, because our models fail to take into account the increasing diversification of the child's learning abilities over his or her second year (1, 2, 6).

Nevertheless, because our data came from a single child, establishing the generality of these techniques will require more evidence. One strength of the methods we present lies in their applicability to other datasets via automated and semiautomated techniques. With the growth of inexpensive computation and increasingly precise speech recognition, which are hallmarks of the era of "big data," datasets that afford such in-depth analyses will become increasingly feasible to collect. In addition to replication of our correlational analyses, a second important direction for future work is to make tighter experimental tests of the causal importance of contextual distinctiveness in word learning.

Theorists of language acquisition have long posited the importance of rich activities and contexts for learning (2, 11, 12). Our contribution here is to show how these ideas can be instantiated using new tools and datasets. We hope this work spurs further innovation aimed at capturing the nature of children's language learning at scale.

Materials and Methods

Video Processing. The spatial distinctiveness analysis first identifies regions of pixels that exhibit motion, yielding a 487-dimensional binary motion vector summarizing the active regions across all cameras. Characterizing motion relative to regions, rather than individual pixels, is robust to pixel-level noise and provides a low-dimensional representation of activity. Region-level activity for any point in time is obtained by measuring pixel value changes in the region for video frames within ±5 s of the target time. This low-dimensional representation is advantageous because it requires no human annotation and is robust to noise while also capturing the locations of activity and a gist of activity trajectories. More detail on these computations, including how regions are defined, is provided in *SI Materials and Methods*.

Extracting Spatial Distinctiveness. A word's spatial distribution summarizes where activity tended to occur when the word was uttered. This distribution is computed from the condensed, region-activity representation of the recorded video described above. First, for any word that the child learns, all child-available caregiver utterances containing that word before the word birth are identified. For each such exposure, the region activity vector is calculated for the utterance time stamp, capturing the immediate situational context of the child's exposure to the target word, including the positions of the participants and their trajectories if they are in motion. These vectors are then summed and normalized to obtain the word's spatial distribution.

A word's spatial distribution may not be particularly revealing about its link to location, because locations will generally have different overall activity levels. Instead, word spatial distributions are compared with a baseline: the background distribution of all caregiver language use. The background distribution is computed in the same manner as word spatial distributions except that the entire corpus is processed for all caregivers, and not just the pre-AoFP utterances. To quantify spatial distinctiveness, we compute the frequency-corrected KL-divergence between the word's spatial distribution and the background. The raw KL-divergence (also known as relative entropy) (29) between discrete distributions p and q is written as $D(p \parallel q) = \sum_i p_i \log \frac{p_i}{q_i}$, and it is 0 if $p = q$; otherwise, it is positive. The caveat in using KL-divergence directly for comparing distinctiveness between different words is that it is a biased estimator and depends on the number of samples used in estimating p and q. To address this issue, we use a word frequency-adjusted KL-divergence measure, which is discussed below.

Extracting Temporal Distinctiveness. A word's temporal distribution reflects the time of day it is used at an hour-level granularity, from 0 (12:00–12:59 AM) to 23 (11:00–11:59 PM). As with the spatial distribution, for each word the child learns, all child-available caregiver utterances containing that word before AoFP are identified. For this set, the hour of the day is extracted from each utterance time stamp and the values are used to estimate the parameters of a multinomial by accumulating the counts and normalizing. The hour of day associated with a word can be viewed as a sample drawn from the word's temporal distribution. As with spatial distinctiveness, we use frequency-adjusted KL-divergence to compare a word's temporal distribution with a background distribution computed over all caregiver utterances in the corpus. Larger KL-divergence values indicate more temporally distinct word distributions, which tend to be more temporally grounded and used at particular times of the day.

Extracting Linguistic Distinctiveness. The child's exposure to a word occurs in the context of other words, which are naturally linked to one another through topical and other relationships. A word's embedding in recurring topics of everyday speech may be helpful in decoding word meaning, and the topics themselves may reflect activities that make up the child's early experience. To identify linguistic topics, we used LDA (28), a probabilistic model over discrete data that is often applied to text. LDA begins with a corpus of documents and returns a set of latent topics. Each topic is a distribution over words, and each document is viewed as a mixture of topics. We used the computed topics to extract the topic distribution for each word that the child produced. More details of LDA analysis are provided in *SI Materials and Methods*. As with both of the previous two distinctiveness measures, we used frequency-adjusted KL-divergence to compare a word's pre-AoFP topic distribution with the background distribution.

Bias Correction for Divergence Estimates. The distinctiveness measures quantify how a word's use by caregivers differs from the overall background language use across spatial, temporal, and linguistic contexts. Within a contextual modality, for a particular word, we wish to compare the pre-AoFP caregiver word conditional distribution against the baseline distribution, where the distributions are modeled as multinomials. Although maximum likelihood estimates of multinomial parameters from count data are unbiased, KL-divergence estimates are not. To address this issue, we empirically examined several approaches to quantifying word distinctiveness. The raw KL-divergence value is strongly correlated with the sample counts used in constructing the word multinomial distribution, as expected, and generally follows a power law with $\log D(p_w \parallel p_{bg}) \sim -\alpha \log n_w$, where p_w is the estimated word distribution, n_w is the number of word samples used, and p_{bg} is the background distribution. The method we adopted was to use the residual log KL-divergence after regressing on log count. The distinctiveness score is calculated as $\text{Score}_w = \log D(p_w \parallel p_{bg}) - (\alpha_0 + \alpha_1 \log n_w)$, where α_0 and α_1 are the regression model parameters. More details are provided in *SI Materials and Methods*.

Variable Transformations. All predictor variables were standardized; frequencies were log-transformed. More details are provided in *SI Materials and Methods*.

Ethics, Privacy, and Data Accessibility. Data collection for this project was approved by the MIT Committee on the Use of Humans as Experimental Subjects. Regular members of the household (family, baby-sitters, or close friends) provided written informed consent for use of the recordings for noncommercial research purposes. Occasional visitors were notified of recording activity and provided verbal consent; otherwise, recording was temporarily suspended or the relevant data were deleted. Datasets such as ours open up new research opportunities but pose new and unknown ethical concerns for researchers. To safeguard the privacy of the child and family being studied here, we are not able to make available the full video and audio dataset. Nevertheless, we make aggregate data about individual words available via the GitHub web-based repository hosting service (github.com/bcroy/HSP_wordbirth), and we encourage interested researchers to investigate these data.

ACKNOWLEDGMENTS. Rupal Patel, Soroush Vosoughi, Michael Fleischman, Rony Kubat, Stefanie Tellex, Alexia Salata, Karina Lundahl, and the Human Speechome Project transcription team helped shape and support this research. Walter Bender and the MIT Media Lab industrial consortium provided funding for this research.

1. Bloom P (2002) *How Children Learn the Meanings of Words* (MIT Press, Cambridge, MA).
2. Clark EV (2009) *First Language Acquisition* (Cambridge Univ Press, Cambridge, UK).
3. Baldwin DA (1991) Infants' contribution to the achievement of joint reference. *Child Dev* 62(5):875–890.
4. Carpenter M, Nagell K, Tomasello M (1998) Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monogr Soc Res Child Dev* 63(4):i–vi, 1–143.
5. Markman EM (1991) *Categorization and Naming in Children: Problems of Induction* (MIT Press, Cambridge, MA).
6. Smith LB, Jones SS, Landau B, Gershkoff-Stowe L, Samuelson L (2002) Object name learning provides on-the-job training for attention. *Psychol Sci* 13(1):13–19.
7. Hart B, Risley TR (1995) *Meaningful Differences in the Everyday Experience of Young American Children* (Brookes Publishing Company, Baltimore).
8. Huttenlocher J, Haight W, Bryk A, Seltzer M, Lyons T (1991) Early vocabulary growth: Relation to language input and gender. *Dev Psychol* 27(2):1236–1248.
9. Goodman JC, Dale PS, Li P (2008) Does frequency count? Parental input and the acquisition of vocabulary. *J Child Lang* 35(3):515–531.
10. Oller DK, et al. (2010) Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proc Natl Acad Sci USA* 107(30):13354–13359.
11. Bruner J (1985) *Child's Talk: Learning to Use Language* (W. W. Norton & Company, New York).
12. Cartmill EA, et al. (2013) Quality of early parent input predicts child vocabulary 3 years later. *Proc Natl Acad Sci USA* 110(28):11278–11283.
13. Akhtar N, Carpenter M, Tomasello M (1996) The role of discourse novelty in early word learning. *Child Dev* 67:635–645.
14. Weisleder A, Fernald A (2013) Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychol Sci* 24(11):2143–2152.
15. Fenson L, et al. (1994) Variability in early communicative development. *Monogr Soc Res Child Dev* 59(5):1–173, discussion 174–185.
16. Friend M, Keplinger M (2008) Reliability and validity of the computerized comprehension task (CCT): Data from American English and Mexican Spanish infants. *J Child Lang* 35(1):77–98.
17. Duncan GJ, Magnuson KA, Ludwig J (2004) The endogeneity problem in developmental studies. *Res Hum Dev* 1(1-2):59–80.
18. Ebbinghaus H (1913) *Memory: A Contribution to Experimental Psychology* (Teachers College, New York).
19. Piaget J (1929) *The Child's Conception of the World* (Routledge, London).
20. Vosoughi S (2010) Interactions of caregiver speech and early word learning in the Speechome corpus: Computational explorations. Master's thesis (Massachusetts Institute of Technology, Cambridge, MA).
21. Roy BC, Roy D (2009) Fast transcription of unstructured audio recordings. *Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009)* (ISCA, Brighton, UK).
22. Brent MR, Siskind JM (2001) The role of exposure to isolated words in early vocabulary development. *Cognition* 81(2):B33–B44.
23. Storkel HL (2001) Learning new words: Phonotactic probability in language development. *J Speech Lang Hear Res* 44(6):1321–1337.
24. Newport EL, Gleitman H, Gleitman LR (1977) Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. *Talking to Children: Language Input and Acquisition*, eds Snow CE, Ferguson CA (Cambridge Univ Press, Cambridge, UK), pp 109–149.
25. Yu C, Smith LB (2007) Rapid word learning under uncertainty via cross-situational statistics. *Psychol Sci* 18(5):414–420.
26. Frank MC, Goodman ND, Tenenbaum JB (2009) Using speakers' referential intentions to model early cross-situational word learning. *Psychol Sci* 20(5):578–585.
27. Gleitman L (1990) The structural sources of verb meanings. *Lang Acquis* 1:3–55.
28. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3:993–1022.
29. Cover TM, Thomas JA (2006) *Elements of Information Theory* (Wiley, New York).
30. Miller GA (1955) *Information Theory in Psychology: Problems and Methods* (Free Press, Glencoe, IL), Vol 2.
31. Coltheart M (1981) The MRC psycholinguistic database. *Q J Exp Psychol* 33(4): 497–505.
32. Ferguson C, Snow C (1978) *Talking to Children* (Cambridge Univ Press, Cambridge, UK).
33. Kubat R, DeCamp P, Roy B, Roy D (2007) TotalRecall: Visualization and semi-automatic annotation of very large audio-visual corpora. *Proceedings of the 9th International Conference on Multimodal Interfaces* (ACM, New York).
34. Fiscus J (1998) Sclite scoring package, version 1.5. US National Institute of Standard Technology (NIST). Available at www.nist.gov/itl/iad/mig/tools.cfm. Accessed August 30, 2015.
35. Jurafsky D, Martin JH, Kehler A (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (MIT Press, Cambridge, MA).
36. Reynolds DA, Quatieri TF, Dunn RB (2000) Speaker verification using adapted gaussian mixture models. *Digital Sig Proc* 10(1):19–41.
37. Dromi E (1987) *Early Lexical Development* (Cambridge Univ Press, Cambridge, UK).
38. Gopnik A, Meltzoff A (1987) The development of categorization in the second year and its relation to other cognitive and linguistic developments. *Child Dev* 58(6): 1523–1531.
39. McMurray B (2007) Defusing the childhood vocabulary explosion. *Science* 317(5838):631.
40. Roy BC (2013) The birth of a word. PhD thesis (Massachusetts Institute of Technology, Cambridge, MA).
41. Chao A, Shen T-J (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ Ecol Stat* 10(4):429–443.
42. Paninski L (2003) Estimation of entropy and mutual information. *Neural Comput* 15(6):1191–1253.
43. Zipf GK (1949) *Human Behavior and the Principle of Least Effort* (Addison–Wesley Press, Cambridge, MA).
44. Piantadosi ST, Tily H, Gibson E (2011) Word lengths are optimized for efficient communication. *Proc Natl Acad Sci USA* 108(9):3526–3529.
45. Weide R (1998) The Carnegie Mellon University Pronouncing Dictionary, release 0.7a. Available at www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed August 30, 2015.
46. Bates E, et al. (1994) Developmental and stylistic variation in the composition of early vocabulary. *J Child Lang* 21(1):85–123.
47. Caselli C, Casadio P, Bates E (1999) A comparison of the transition from first words to grammar in English and Italian. *J Child Lang* 26(1):69–111.
48. Huber PJ (2011) *Robust Statistics* (Springer, Hoboken, NJ).