Predicting the birth of a spoken word

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.

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

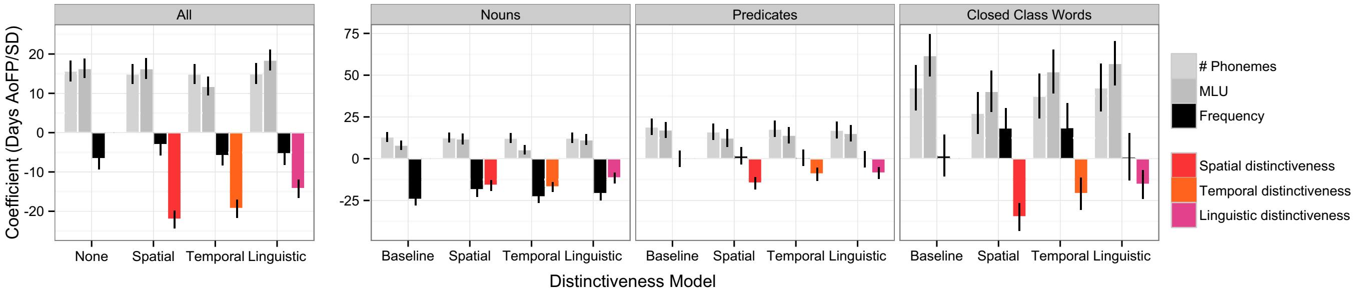


Fig. 1. Regression coefficients (±SE) for each predictor in a linear model predicting AoFP. Each grouping of bars indicates a separate model: a baseline model with only the number of phonemes, MLU, and frequency or a model that includes one of the three distinctiveness predictors. Red/orange/purple bars 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,

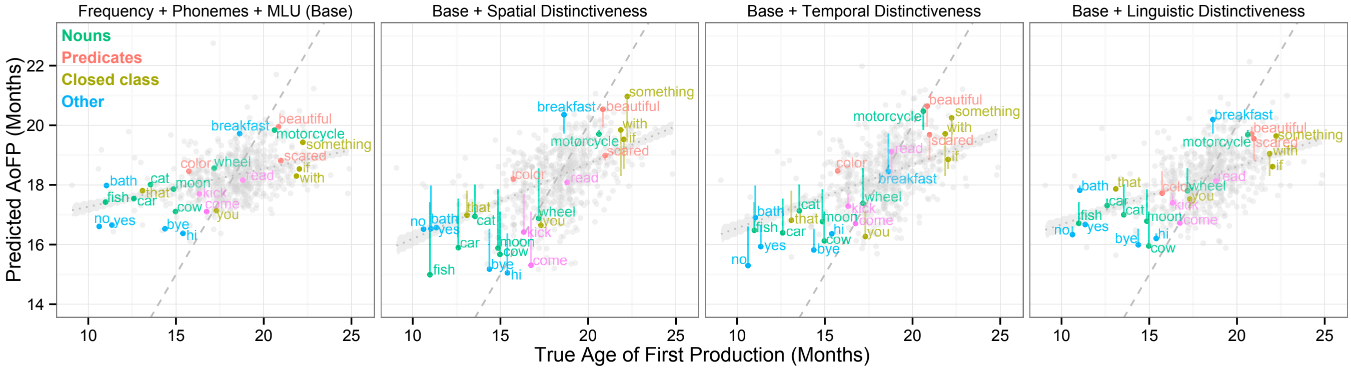


Fig. 2. Predicted AoFP plotted by true AoFP for successive regression models. Each dot represents a single word, with selected words labeled and lines showing the change in prediction due to the additional predictor for those words. Color denotes word category, the dotted line shows the regression trend, 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 ($r_{\text{spatial}} = -0.40$, $r_{\text{temporal}} = -0.34$, $r_{\text{linguistic}} = -0.28$, 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.

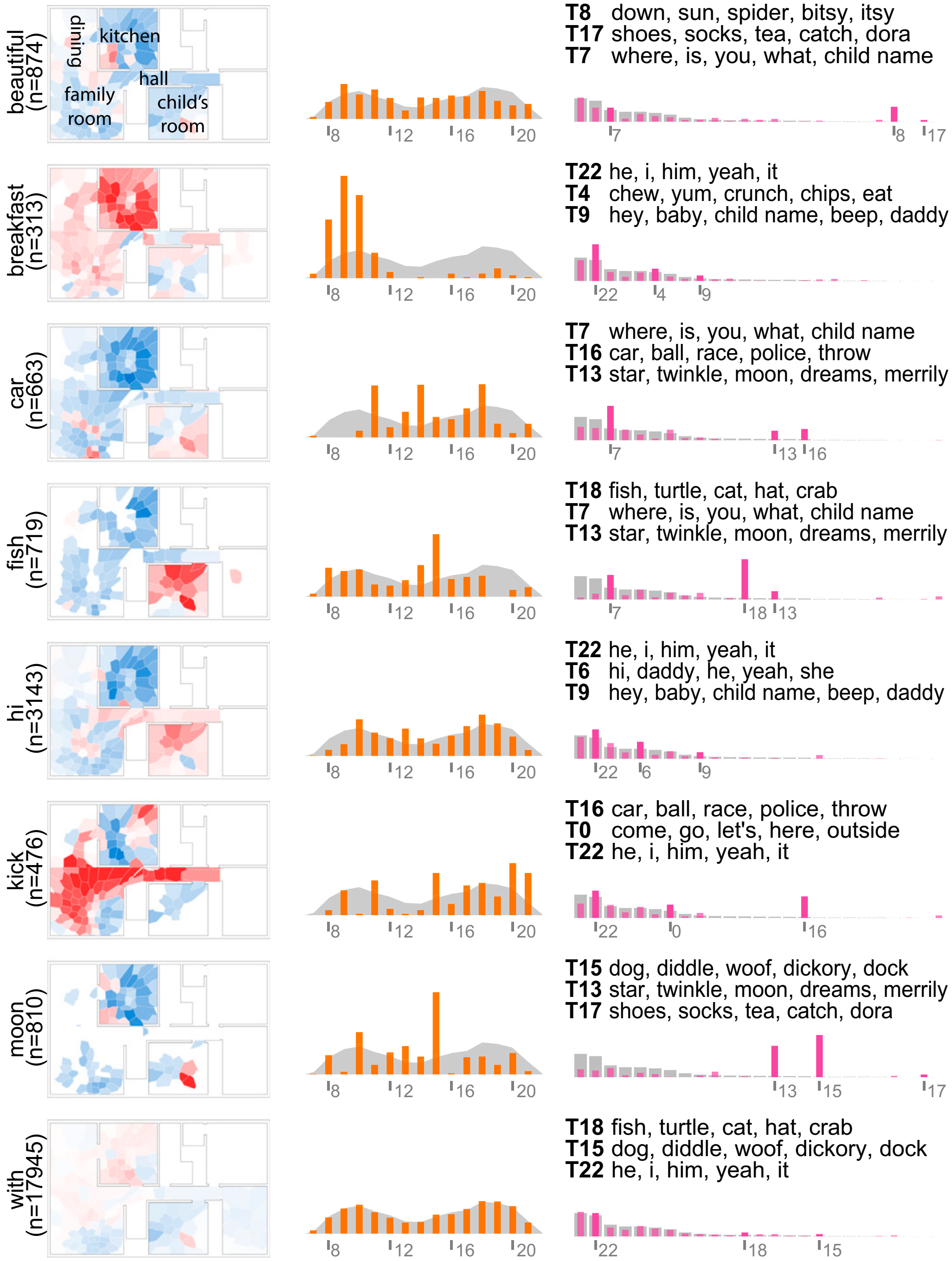


Fig. 3. Examples of eight spatial, temporal, and linguistic context distributions for words. Spatial distributions show the regions of the house where the word was more (red) and less (blue) likely than baseline to be used. Rooms are labeled in the topmost plot. Temporal distributions show the use of 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 $\mathrm{Score}_w = \log D(p_w \parallel p_{bg}) - (\alpha_0 + \alpha_1 \log n_w)$, 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).

Supporting Information

Roy et al. 10.1073/pnas.1419773112

SI Materials and Methods

Data Collection. Data collection spanned the child's first 3 y of life. Audio and video recordings were captured from a custom recording system in the child's home, consisting of 11 cameras and 14 microphones embedded in the ceilings. This system was unobtrusive while achieving full spatial coverage. Cameras were fitted with fisheye lenses to obtain a full view of each room, and recordings were made at ∼15 frames per second and 1-megapixel resolution. Audio was recorded from boundary-layer microphones, which were able to capture whispered speech from any location by using the entire ceiling as a pickup. Audio was digitized at 48 KHz and 16-bit sample resolution. Fig. S1 shows the family's home, a view into the living room, and some components of the recording system. Altogether, roughly 90,000 h of video and 120,000 h of audio were recorded and stored on servers housed at the MIT Media Lab. Fig. S2 shows the full data-processing system used in the current study.

Speech Transcription. The transcribed subset of the data spans the period during which the child was aged 9–24 mo. Recordings are included from 444 of the 488 d in this period (with exclusions due to random subsampling in the transcription process). During this time frame, an average of 10 h of multitrack audio was captured per day.

In general, the audio-video recording system ran all day and captured substantial amounts of silence, nonspeech audio, and adult speech during the child's naps. To minimize the amount of audio to transcribe and to focus on the speech relevant to the child's language learning, we identified a subset of multitrack audio recordings for transcription using a manual preprocessing step. By viewing the video, we first annotated the room the child was in and whether he was awake or asleep across the day's recording. Annotation was performed using TotalRecall (33), a tool we developed for browsing and annotating audio and video. The resultant "where-is-baby" time series of annotations were then used to exclude audio from rooms that were out of the child's hearing range. Furthermore, when the child was asleep, audio from all rooms was excluded. We refer to the nonchild speech contained in this filtered subset as child-available speech, because it can reasonably be considered his linguistic input.

Even after filtering, fully manual transcription at this scale would have been prohibitively time-consuming and expensive, and fully automatic speech recognition would have been too inaccurate. We developed a new speech transcription tool called BlitzScribe (21) that combines automatic and manual processing. BlitzScribe uses automatic audio-processing algorithms to scan through the unstructured audio recordings to find speech and create short, easily transcribable segments. The speech detection algorithm splits audio into short 30-ms frames with a 15-ms overlap, extracts spectral features from each frame, and applies boosted decision trees to classify audio frames as speech or nonspeech. A segmentation algorithm then groups classified frames into short segments of speech and nonspeech.

Automatically identified speech segments were then loaded into a simplified user interface that presented each segment as a blank row in a list where the transcript could be typed. Audio playback was controlled using the keyboard, obviating the need to switch between the keyboard and mouse. Because the speech segments were automatically detected, if nonspeech was incorrectly labeled as speech (false-positive error), the transcriber simply left the segment blank and it was automatically marked as nonspeech. The system was tuned to favor false-positive over false-negative errors, because false-positive errors are easier to correct.

The primary output of BlitzScribe was a sequence of speech transcripts linked to the corresponding audio segments. Transcribed speech segments were generally between 500 ms and 5 s long, tuned to support ease of transcription as well as fine-grained temporal resolution for each transcribed token. In addition to the speech transcripts, the labeled speech and nonspeech segment information could be used to retrain and improve the speech detection algorithms.

Transcription quality was assessed on an ongoing basis by assigning the same 15-min blocks of audio to multiple annotators and evaluating interannotator agreement on these assignments. Our system incorporated the US National Institute of Standards and Technology sclite text alignment algorithm (34) to calculate interannotator agreement. This measure was primarily used to track transcriber performance and identify cases where transcription conventions may have been misunderstood, which was particularly important as nearly 70 annotators contributed to this project over the course of 5 y. We reviewed cases where a transcriber's average pairwise interannotator agreement score against all other annotators dropped below ∼0.85. In some cases, low reliability would lead to greater training for individual transcribers or the establishment of transcription conventions for particular words or phrases. Some assignments were inherently more difficult, however, and had lower average interannotator agreement scores due to background noise or overlapping speech, for example.

Speaker Identification. Speaker identity was labeled using a fully automatic system, although manual annotations were included where available. The automatic system used acoustic features to learn a decision boundary between the four primary speakers: mother, father, nanny, and child. We used mel-frequency cepstral coefficient (MFCC) features, MFCC deltas, and MFCC delta-deltas, which are effective and commonly used in automatic speech-processing algorithms (35). Audio samples in a speech segment were partitioned into a set of 30-ms frames (with 15-ms overlap), and acoustic features were extracted from each frame in the same manner as for speech detection. The frames were classified by comparing the likelihood of these observations under a trained Gaussian mixture model for each speaker. Our system uses a universal background model trained across different speakers as a starting point for speaker-specific mixture models, similar to other approaches (36).

For any speech segment, there are potentially multiple speaker annotations produced either by different versions of the automatic speaker identification system or by different human annotators. The logic for choosing the speaker annotation is always to prefer human annotations to machine annotations, and then to select the most recently produced annotation. Roughly 2.2% (about 51,000) of the speech segments were human-annotated, and the remaining segments were produced automatically. Although manual speaker labeling is expensive in terms of human effort, a small number of segments (∼540) were annotated independently by multiple annotators to assess interannotator agreement. Interannotator agreement on speaker labeling was high, at roughly 96% agreement and $\kappa = 0.94$.

Each automatically generated speaker annotation also provides a confidence score. We used a confidence threshold to tune the tradeoff between data yield and accuracy. In the results reported here, we used a confidence threshold that preserved at least 80%

of the data for each speaker and achieved accuracy in excess of 90%. Details on the relationship between the confidence threshold, accuracy, and yield are provided in Fig. S3. Note that, as described below, AoFP by the child for each word was manually verified to avoid faulty speaker identification leading to errors in this measure.

Child's Productive Vocabulary

The primary outcome variable for our study was the child's AoFP for individual words. Finding these first productions in roughly 8 million tokens of child and adult speech is challenging because the subset of child speech alone consists of hundreds of hours of audio, which is too much to listen to manually. On the other hand, the naive strategy of simply searching the transcripts for the child's first production of a word is also problematic; small annotation error rates for transcripts and speaker identification labels can result in many false-positive errors, which could erroneously lead to attributing adult-produced words to the child. To narrow down candidate words, we followed a two-step process, first filtering annotation errors and then conducting manual review of the filtered set of words.

There are two primary annotation error types that might lead to incorrect identification of a word's first production by the child: errors in transcription and errors in speaker identification. Transcription errors are less common, and because speech transcripts are human-generated, further human review of a speech segment may not yield a better or more authoritative transcript. In contrast, most speaker identification annotations are produced by an automatic system with a higher error rate, and speaker identity is relatively easy to discern for a human annotator. We addressed these issues through a combination of automatic and manual approaches. An automatic inference procedure identified candidate words and word birth dates for the child's vocabulary from the large amount of observed data, and a software tool was developed to enable rapid manual review and annotation.

Automatically Identifying Candidate Word Births. The automatic inference procedure was the first step. We began by modeling the speaker label associated with a particular token in an utterance as a noisy observation. There are two primary error types that could result from the speaker identification system with respect to identifying the child's true vocabulary. A false-negative result is a case in which a child's true production of a word is mislabeled as nonchild speech. Although a single true production of a particular word may be mislabeled as nonchild speech, the chance that all such true productions are mislabeled quickly decreases toward zero with each production. For this reason, and because scouring all nonchild-labeled speech for false-negative results would be extremely costly, we do not directly address false-negative errors. However, we do address false-positive errors, in which a nonchild word production is mislabeled as child speech. False-positive errors can lead to attributing words to the child's productive vocabulary erroneously or to identifying AoFP earlier than the child's first production.

To infer automatically whether and when the child first produced a word in the presence of false-positive errors, we use an hypothesis testing procedure to compare a model of observed word occurrence counts parameterized by word birth month to a null hypothesis model. Under the null hypothesis, the child never produced the word and all observed occurrences are false-positive errors. In the parameterized model, all observed child productions in the preacquisition regime are false-positive errors, whereas those observed child productions in the postacquisition regime are a combination of false-positive errors and true-positive counts. A likelihood ratio test can be used both to test whether the child acquired the word and to determine what the most likely word birth month would be. Fig. S4 shows the occurrence counts of the word "star" by month. Although there are child-labeled occurrences of this word for every month (shown in red), the likelihood ratio test procedure identifies month 16 as the mostly likely word birth month and, furthermore, that the likelihood of the observed data under this model is significantly higher than under the null model ($P < 0.05$).

With this method, we proceeded as follows. First, only child utterances with a speaker identification confidence at or greater than 0.4 were considered. This threshold preserved 90% of the child's true utterances at a false-positive rate of about 0.05. All words in these utterances were then tokenized and normalized via manually generated mapping, reducing alternate spellings, plurals, gerunds, and some common misspellings to a canonical form, resulting in 6,064 word types. Next, words that were uttered two or fewer times by the child and five or fewer times overall were removed. Without a sufficient number of examples of the child using a word, even manual review may be unreliable. A similar criterion for child speech was used by Dromi (37), which required three consistent vocalizations in various contexts for a word to be admitted into the lexicon. We also noted that the long tail of rare words often contained misspellings of more common words. These thresholds were chosen to be permissive and yielded a set of 2,197 candidate words, which is many more than expected for a 2-y-old (15). Reducing the thresholds further would have required additional human review later in the analysis pipeline but with little expected change to the final set of word births. After filtering, we applied the hypothesis testing procedure described above to each of these words, yielding a candidate set of 1,375 word births.

Manual Word Birth Review and Annotation. The final vocabulary growth time line used for our analyses was manually reviewed and verified using the "Word Birth Browser," a tool we designed specifically for this purpose. This tool loads a set of candidate words and their AoFP values, and allows the user to play back the corresponding audio segment. The user is also presented with all other utterances containing the target word, which can be sorted by date and speaker identity so that prior or subsequent candidate occurrences may also be reviewed. Finally, because interpreting the speech in an isolated utterance can be challenging, a contextual window with all utterances in the surrounding few minutes is also available and can be used for playback. This tool is shown in Fig. S5. Several members of our transcription team helped to annotate word births using this tool. After several weeks of effort, 679 words and their AoFP dates were identified and used in the results reported in the main text.

We believe this final set of words is quite accurate, although our results may still be biased in a number of ways. First, we had no method for finding false-negative errors, so we likely understate the child's vocabulary, especially for words learned later (for which there are fewer opportunities for detection). Second, low-frequency words may be more likely to be detected later than their actual first production, because individual instances of production might be missed.

Tracking Lexical and Syntactic Development. The child's productive vocabulary grew slowly at first, consisting of about 10–15 words by 12 mo of age, and then rapidly accelerated over the next 6 mo. Although the child's vocabulary continued to grow, the rate of growth decreased substantially after 18 mo of age. Fig. S6*A* depicts the number of new words added to the child's productive vocabulary over time, illustrating the dynamic nature of the child's lexical growth.

Researchers have noted the rapid growth of many children's early vocabularies, which is sometimes referred to as a "vocabulary spurt." Some have suggested this vocabulary spurt is a byproduct of a new insight children gain about categories (38), and others suggest that it is a mathematical consequence of the natural distribution of word difficulty (39). Furthermore, some children's lexical growth rate may not accelerate but exhibit greater development in other areas, such as combinatorial productivity (2). Less commonly discussed is the decline we observe in growth rate; it has been suggested this decline may signify a transition into a different learning "stage" (37) or a statistical sampling artifact (1), although the scale and density of the Human Speechome Project corpus mitigates sampling issues. Fig. S6*B* shows the MLU (in words) of the child over time, an indicator of the child's grammatical and general productive language development. The transition from single-word utterances to multiword productions seems to begin around 18 mo of age. Notably, the decline in lexical acquisition rate also occurs around this time. This pattern of decreased productive lexical acquisition rate, coinciding with an increase in combinatorial speech, aligns with the findings by Dromi (37), who argued for distinct learning stages. Certainly, grammatical combinatorial speech requires a sufficiently (and syntactically) rich productive vocabulary, supporting a dependence of MLU on vocabulary size. However, it is less clear why the onset of combinatorial speech should coincide with a decrease in the lexical acquisition rate. Although more research is needed, Fig. S6 illustrates that there are multiple strands of communicative development underway that may share important interdependencies.

Fig. S7 shows the overall breakdown of utterances and tokens by speaker, after removing utterances consisting only of nonword vocalizations. The child's role as a communicative participant clearly increases with time. The pattern of engagement roughly tracks vocabulary size and shows a substantial increase around months 17 and 18, roughly tracking the rapid increase in vocabulary size in these months.

Methods for Distinctiveness Measures

Video Processing. Spatial distinctiveness was calculated across spatial regions rather than at the pixel level, which yielded a lower dimensional spatial representation that also provided robustness to pixel noise. To obtain regions that faithfully captured the spatial and activity structures of interest, the raw 960 × 960-pixel video from each camera was first down-sampled to 120 × 120 pixels. Background subtraction was applied to each down-sampled frame to identify "active" pixels that differed significantly from their average "background" value, resulting in streams of binary video. For each stream, pairs of pixels with highly correlated values and within a short spatial distance of each other were clustered together, yielding a total of 487 regions across nine of the 11 cameras (the master bedroom and bathroom were again omitted from this analysis).

Region activities for a point in time were computed as follows. First, background subtraction was applied to all reduced-resolution video frames within a temporal window of ±5 s of the target time. For each region, we calculated the fraction of active pixels in the region for all frames in the temporal window and then thresholded. In this way, the activity at any point in time was summarized as a 487-dimensional binary vector indicating the active regions.

LDA Modeling. We partitioned the entire corpus of speech transcripts into a set of documents by splitting the 9- to 24-mo time range into a nonoverlapping sequence of 10-min windows, and grouped all transcripts that occurred in a 10-min window together into a document. This process resulted in ~18,700 documents, which we referred to as "episodes." We selected 10-min windows through some experimentation, but with an aim toward choosing a time scale that would capture enough natural speech to include one or a small number of identifiable, discrete activities. Shorter (5 min) and longer (15 min) episodes also yielded similar topics and regression results. Note that in the extreme, very short documents consisting of a single word provide no other words of linguistic context. On the other hand, very long documents (e.g., at the day level) would not capture how clusters of co-occurring words and activities shift and change over the course of a day.

In the standard LDA formulation, documents are treated as an unordered set of words. Each document was first processed to identify a common vocabulary shared across all documents. As is common in probabilistic text modeling, where parameters must be estimated for every word, we reduced the vocabulary size by first removing a small set of "stop words" that were expected to contribute little topic information (e.g., and, "or," "not"). We then applied a stemming algorithm to combine morphological variants into a single word type (e.g., mapping "runs," "running," and "run" to a common form). Finally, we removed words that occurred fewer than six times or occurred in fewer than five documents. The resultant vocabulary consisted of 6,731 words. Note that although these thresholds better condition the input data for LDA modeling (because removing rare words reduces the number of parameters to estimate), the downstream distinctiveness analysis is not particularly sensitive to these thresholds. In general, a rare word is less likely to have an impact on a document's topic distribution, and the distinctiveness measure derives from the topic distributions of pre-AoFP documents containing the target word.

We applied LDA to this corpus. LDA takes as input a target number of topics to identify; choosing the appropriate number requires some intuition and experimentation. We settled on 25 topics after a number of early experiments, largely because the resulting topics were fairly coherent and interpretable (but note that distinctiveness results were also fairly robust to different numbers of topics). Some of the topics that emerged seemed to correspond to activities such as mealtime, book reading, bath time, and playing with toys. In addition, 25 topics corresponded approximately to the number of everyday activities that human annotators noted in a separate annotation effort of a subset of the corpus [more details on this manual activity annotation and analysis are provided elsewhere (40)].

As with spatial and temporal context, we computed a topic distribution for each word based on caregiver word use before AoFP. To do so, we identified all 10-min episodes (documents) before AoFP. We apportioned caregiver uses of the target word during the episode to topics according to the episode's topic mixture and then summed and normalized to obtain the topic distribution for the word.

A topic that is strongly associated with a word will thus have a high conditional probability $\Pr(topic_i|w)$, but as with spatial and temporal context, the topic conditional probability distribution must be compared with a background distribution to quantify its distinctiveness. The background topic distribution was computed in the same manner as the per-word topic distribution, except by summing over all episodes in the corpus. It is the weighted average of all of the episode topic distributions, weighted by the number of words in each episode. Linguistic topic distinctiveness is defined as the frequency-adjusted KL-divergence between the word conditional topic distribution and the background topic distribution.

Bias Correction for KL-Divergence Estimates. The distinctiveness measures compare a word's spatial, temporal, or topical distribution against the "background" distribution of language use in the modality. These distributions are modeled as multinomials and estimated from observed data. Although the multinomial parameter estimates are unbiased, the KL-divergence values for these estimated distributions are not; instead, they depend on the number of samples used in estimating the underlying distributions. With fewer samples, the KL-divergence estimates are biased upward, decreasing toward the true KL-divergence as the number of samples increases.

This bias is problematic when comparing KL-divergence values between words whose distributions are derived from different numbers of observations. Because the number of observations for a word depends on both its frequency and AoFP, the raw KL-divergence measure will reflect both true distributional differences in use patterns and frequency-derived bias. Therefore, we explored several bias correction strategies to characterize word distinctiveness properly.

Miller (30) investigated the bias in estimates of entropy, a closely related quantity. He showed that the highest order bias terms depend on k, the number of bins in the multinomial, and n, the number of samples used in estimating the multinomial. The bias decreases toward zero following a $\frac{1}{n}$ relationship. It is straightforward to show that the KL-divergence bias follows the same $\frac{1}{n}$ falloff toward zero. [This bias can be seen by expressing KL-divergence as the cross-entropy minus the entropy, or $D(p \parallel q) = H(p, q) - H(p)$, and recognizing that the cross-entropy estimator is unbiased for multinomial distributions.] Miller (30) suggests a bias correction that can be applied when n is not too small (i.e., when $n \gg k$); unfortunately, this condition is not valid for many words, particularly for spatial distinctiveness, where the number of multinomial bins (i.e., regions) is large.

Chao and Shen (41) present another approach to entropy bias correction for characterizing species diversity from sample counts. Here, the number of species corresponds to the number of multinomial bins, which is unknown. In our scenario, the number of bins is known, although in the case of spatial distinctiveness, some regions may never be active for the set of learned words. A thorough discussion of the bias in information theoretic estimators is presented by Paninski (42).

With these issues in mind, 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. Applying the corrections of Miller (30) and Chao-Shen (41) also generally yielded values negatively correlated with count. This correlation may reflect a real property of word use that more distinctive words are less frequent, but in combined regression models, collinearity with other variables is a concern as a potential confound.

Therefore, we took a conservative approach and decided to remove the effect of count completely in defining distinctiveness: We used the residual log KL-divergence value after regressing on log count. Although this residualization step may diminish the predictive power of KL-divergence, particularly if log sample count correlates with AoFP (although it generally does not), it effectively reduces collinearity with other predictors. Intuitively, the regression line captures the average log KL-divergence by log count, and the residual for a particular word reflects how much more or less contextually distinctive the word is relative to others with the same sample count.

Supporting Data and Analytical Details

In this section, we give additional details on selected analyses; full code to reproduce all reported analyses is available in the linked repository. For interested readers who wish to explore the raw data linked in our GitHub repository (github.com/bcroy/HSP_wordbirth), the measures (and variable names) are as follows: word frequency (sln.freq.pre), MLU (s.uttlen.pre), number of phonemes (s.cmu.phon), spatial distinctiveness (srl.sp.KL), temporal distinctiveness (srl.temp.KL), and linguistic distinctiveness (srl.topic.KL). The variables are named according to the following conventions: standardized variables are prefixed by s, normalized variables are prefixed by n, and logged variables are prefixed by l. The distinctiveness measures are all residualized, denoted with the prefix r.

Correlational Structure Between Variables. Correlations between variables are shown in Fig. S8. The baseline predictors (MLU, number of phonemes, and frequency) were relatively uncorrelated, with one exception. Number of phonemes is a measure of word length, which has been known since Zipf (43) to be correlated with word frequency [perhaps as a consequence of the evolution of vocabulary to facilitate efficient communication (44)].

In contrast, spatial, temporal, and topical distinctiveness was largely uncorrelated with the baseline predictors. We note that correlations between log frequency and the distinctiveness predictors are close to zero but nonzero, despite the fact that the distinctiveness predictors are frequency-controlled, as described above. This effect arises because the counts on which the distinctiveness predictors are residualized are not the same as those counts used to estimate word frequency. There is some small variance in the counts used for each of the distinctiveness predictors relative to frequency, due to both missing video data for a very small subset of transcripts and minor differences in data treatment across approaches (e.g., how multiple uses of a word within the same time window affect distinctiveness distributions).

Finally, we note that there is a high degree of correlation between the distinctiveness predictors (shown by the red dashed line in Fig. S8). For this reason, in the main text, we report models using only one of the predictors, although a model that includes all predictors is shown below.

Differences Between Distinctiveness Variables. Although the primary focus in our analyses is the commonality between the three distinctiveness predictors, we note that they do differ for certain words. We calculated an index of differences between the distinctiveness predictors by calculating the summed squared difference between each prediction and the mean of all three. Table S1 shows the top 10 words on this deviation measure. The results are clearly interpretable. Words like "diaper," "change," and "poop" are very spatially distinctive but are temporally very diffuse, probably because their associated activity is spatially localized (the changing table) but happens at different times throughout the day. In contrast, the word "breakfast" is temporally very distinct but is said throughout the house, probably because the child is being called to eat breakfast at a particular time each morning. These results support the idea that these predictors reveal aspects of the activity structure in which the words are used.

Distinctiveness of Speaker Context. Thanks to the suggestion of the editor and one of the reviewers, we also examined the role of caregiver presence during word use as another measure of a word's contextual distinctiveness. We defined a new variable to capture caregiver context in the same manner as the other distinctiveness measures by first computing a word's pre-AoFP caregiver use distribution (which served as a proxy for caregiver presence, because only speech in the child's vicinity was transcribed.) Thus, words used more frequently in the child's presence by a particular caregiver would have a corresponding peak in the word's speaker distribution. As with the other distinctiveness predictors, we then defined the speaker context distinctiveness as the residualized KL-divergence of the word's speaker distribution relative to the baseline speaker distribution.

By itself, this variable is predictive of AoFP, but when added to the baseline model, it is only significant in predicting the AoFP for nouns and is still weaker than the other three distinctiveness predictors. However, the relationship is directionally the same, indicating that words (or at least nouns) that are more strongly tied to particular caregivers tend to be produced earlier by the child. We tentatively view this analysis as supportive of our hypothesis that linguistic exposure in stable activities, as reflected by distinctive spatial, temporal, linguistic, and caregiver presence measures, contributes to earlier productive acquisition. Table S7 summarizes the relevant distinctiveness values for the combined speaker distinctiveness model, which can be compared with Fig. 1.

Predictor Variables. We used a number of variable transformations in our analysis, as described below. All regression coefficients were standardized (variables prefixed by s) by subtracting the mean and dividing by the SD. This step was taken to create coefficient values whose magnitudes were interpretable as number of days of AoFP per SD of change on a predictor; standardization does not affect the reliability estimates of either individual coefficients or the model as a whole.

Word frequency. We examined a number of ways of including word frequency into our models. From our transcripts, we extracted a count of the number of times a word occurred in the caregivers' speech before AoFP. This count represents a biased estimate of frequency before AoFP, however, because our transcripts omit the first 9 mo of the child's life; for a word learned very early, this count would be artificially low. To remedy this issue, we normalized frequency to a frequency-per-day measure by dividing by the approximate number of days before AoFP for which we had transcripts (variables prefixed by n, denoting normalized). Then, because word frequencies are Zipfian in their distribution (43), we took the natural logarithm of frequency per day (variables prefixed by l, denoting logged). The final predictor we use was thus the standardized, logged, normalized word frequency during the period before AoFP (sln.freq.pre). (We note that this set of variable transformations maximizes the correlation between frequency and age of acquisition relative to other variants.)

MLU. Because morphological analyses were not available for our data, MLU was calculated in words for each sentence in which a target word occurred, again using only those utterances before AoFP. These means were then standardized for the final analysis (s.uttlen.pre).

Number of phonemes. We extracted the number of phonemes in each word by identifying matches in the Carnegie Mellon University Pronouncing Dictionary (45). There were 13 words for which no match was found. We then standardized length in phonemes for the final analysis (s.cmu.phon).

Distinctiveness predictors. The three distinctiveness predictors were also log-transformed, residualized (as noted above), and standardized.

Word category. We first categorized words using the standard MacArthur–Bates Communicative Development Inventory (CDI) categories (small.cat) (15). We then further merged these categories to create syntactic categories, using the category merging scheme of Bates (46) (also ref. 47). Note that this conservative scheme excludes all words marked as "Games and Social Routines" from the nominals category because they may not be true nominals but, instead, words that are used in particular restricted routines.

Regression Models. We note that although we used ordinary least squares regression, all results are qualitatively unchanged via the use of robust regression (48). Results from these analyses are available through our interactive visualization application (wordbirths.stanford.edu/).

In the tables below, we give the full details of the four primary regression models pictured in Fig. 1. Models for subsets of the data can be recomputed easily using the code available in the linked repository. Tables S2–S5 give the baseline model, followed by the three individual distinctiveness predictor models.

Table S6 shows a model including all three distinctiveness predictors. In this model, spatial distinctiveness is assigned the largest predictive weight, whereas temporal distinctiveness remains reliable as well (although considerably smaller than when it is entered separately). Linguistic distinctiveness is not significant in this model, however, suggesting that it did not explain unique variance in AoFP over and above the other distinctiveness predictors. This relatively smaller effect of linguistic distinctiveness is consistent with both its smaller coefficient value in the regression when including it alone (Table S5) and its substantially reduced predictive power when controlling for imageability (discussed below).

Control Analyses for Other Psycholinguistic Variables. To test whether our distinctiveness predictors corresponded to other psycholinguistic variables, we merged the Medical Research Council (MRC) psycholinguistic norms for familiarity, imageability, and concreteness with the child's vocabulary (31). There were 430 words in common between these two sets. Imageability and concreteness were almost indistinguishable ($r = 0.93$), and neither was particularly correlated with any distinctiveness predictor ($r_{max} = 0.35$, $r_{min} = 0.22$), although these correlations were all very reliable, given the large number of words over which they were computed. Familiarity was almost uncorrelated with the distinctiveness predictors ($r_{spatial} = -0.05$, $r_{temporal} = -0.10$, $r_{linguistic} = -0.08$), although it was highly correlated with our frequency measure ($r = 0.55$).

We next examined whether regression coefficients were altered by controlling for variables in the MRC database (within the subset of words for which these variables were available). Intriguingly, the magnitude of spatial distinctiveness for this subset decreased relatively little when controlling for imageability (−25.71 d/SD to −20.77 d/SD), whereas the magnitude of temporal distinctiveness decreased somewhat more (−17.65 d/SD to −12.25 d/SD), and the magnitude of linguistic distinctiveness decreased the most (−13.21 d/SD to −6.47 d/SD). Importantly, in all three models, the distinctiveness predictor was still reliable even when controlling for imageability. The same pattern of results was observed for concreteness.

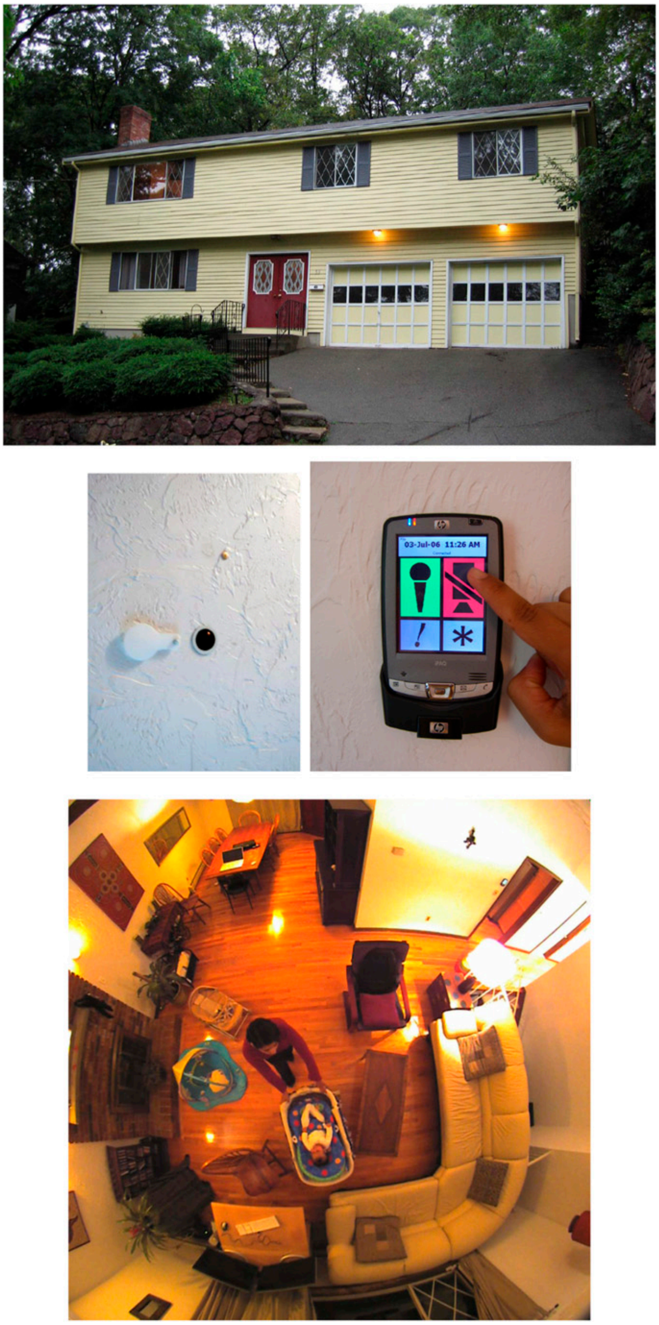


Fig. S1. Site of the Human Speechome Project, where all recording took place. Also shown is the ceiling-mounted camera with an open privacy shutter, the microphone, the recording controller, and a view into the living room.

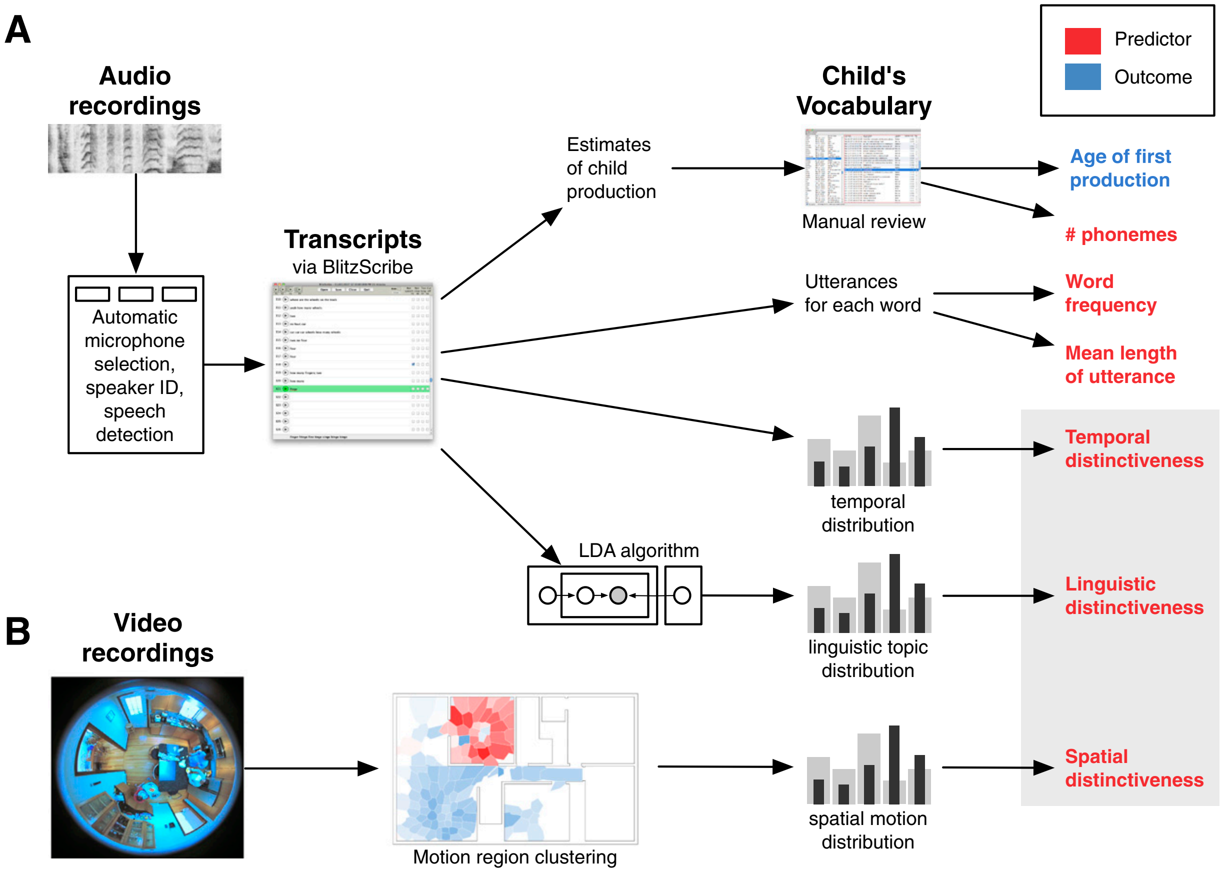


Fig. S2. Schematic of data collection and processing for our dataset, leading to our outcome (blue) and predictor (red) variables. (*A*) Audio recordings are filtered automatically for speech and speaker identity and then transcribed. Transcripts are used for the identification of the child's productions, extraction of frequency, MLU, and temporal distinctiveness predictors, as well as for clustering via topic models (LDA) to extract the linguistic distinctiveness measure. (*B*) Video recordings are processed via motion-based clustering. Region-of-motion distributions for each word are then compared with a base motion distribution for all linguistic events, yielding the spatial distinctiveness predictor.

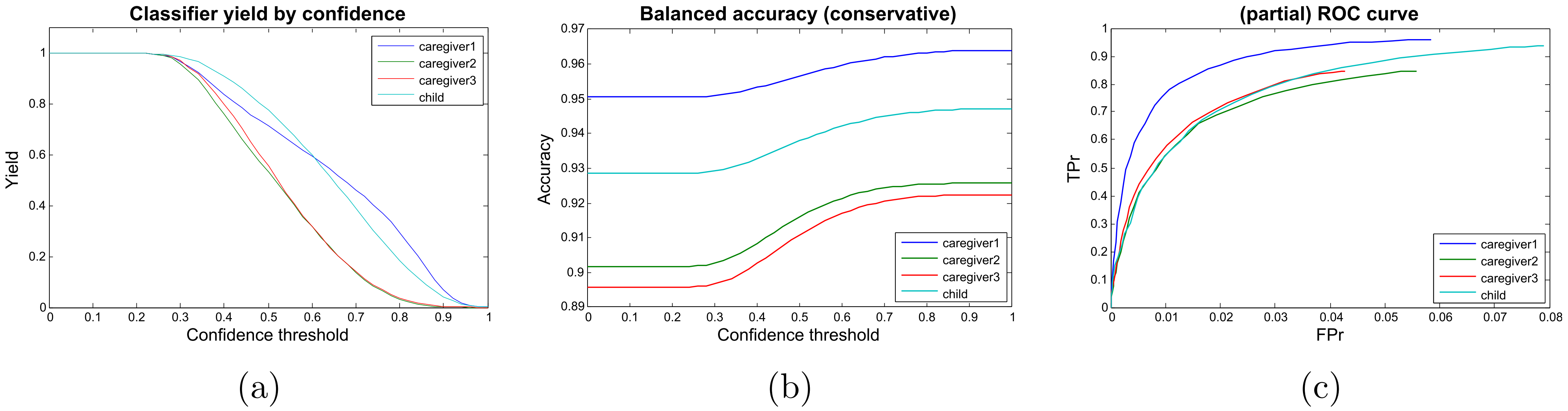


Fig. S3. Speaker identification performance curves. Classifier yield is the fraction of the speaker classifications above a confidence threshold (*A*), and accuracy is the fraction of above-threshold classifications correct for each speaker (*B*). (*C*) Receiver operating characteristic (ROC) curve displays the relationship between true-positive (TPr) and false-positive (FPr) rates for each speaker as the confidence threshold is varied.

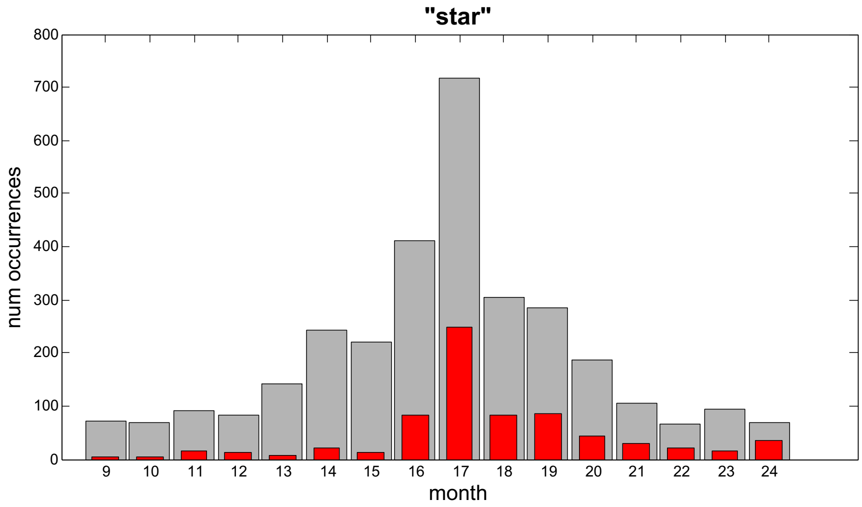
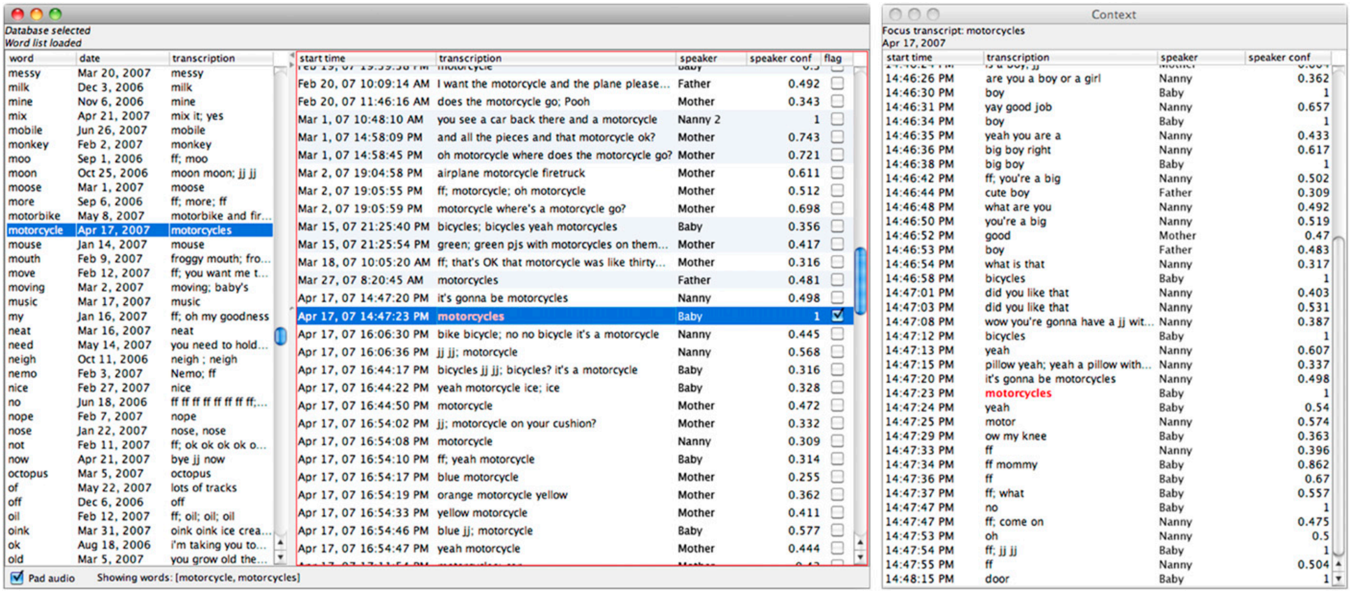


Fig. S4. Counts for the word "star" by month. Child-labeled counts are shown in red, whereas total counts across all speakers are shown in gray.



Database selected
Word list loaded

word	date	transcription
messy	Mar 20, 2007	messy
milk	Dec 3, 2006	milk
mine	Nov 6, 2006	mine
mix	Apr 21, 2007	mix it; yes
mobile	Jun 26, 2007	mobile
monkey	Feb 2, 2007	monkey
moo	Sep 1, 2006	ff; moo
moon	Oct 25, 2006	moon moon; jj jj
moose	Mar 1, 2007	moose
more	Sep 6, 2006	ff; more; ff
motorbike	May 8, 2007	motorbike and fir...
motorcycle	Apr 17, 2007	motorcycles
mouse	Jan 14, 2007	mouse
mouth	Feb 9, 2007	froggy mouth; fro...
move	Feb 12, 2007	ff; you want me t...
moving	Mar 2, 2007	moving; baby's
music	Mar 17, 2007	music
my	Jan 16, 2007	ff; oh my goodness
neat	Mar 16, 2007	neat
need	May 14, 2007	you need to hold...
neigh	Oct 11, 2006	neigh ; neigh
nemo	Feb 3, 2007	Nemo; ff
nice	Feb 27, 2007	nice
no	Jun 18, 2006	ff ff ff ff ff ff ff ff;...
nope	Feb 7, 2007	nope
nose	Jan 22, 2007	nose, nose
not	Feb 11, 2007	ff; ok ok ok ok o...
now	Apr 21, 2007	bye jj now
octopus	Mar 5, 2007	octopus
of	May 22, 2007	lots of tracks
off	Dec 6, 2006	off
oil	Feb 12, 2007	ff; oil; oil; oil
oink	Mar 31, 2007	oink oink ice crea...
ok	Aug 18, 2006	i'm taking you to...
old	Mar 5, 2007	you grow old the...

start time	transcription	speaker	speaker conf	flag
Feb 20, 07 10:09:14 AM	I want the motorcycle and the plane please...	Father	0.492	□
Feb 20, 07 11:46:16 AM	does the motorcycle go; Pooh	Mother	0.343	□
Mar 1, 07 10:48:10 AM	you see a car back there and a motorcycle	Nanny 2	1	□
Mar 1, 07 14:58:09 PM	and all the pieces and that motorcycle ok?	Mother	0.743	□
Mar 1, 07 14:58:45 PM	oh motorcycle where does the motorcycle go?	Mother	0.721	□
Mar 2, 07 19:04:58 PM	airplane motorcycle firetruck	Mother	0.611	□
Mar 2, 07 19:05:55 PM	ff; motorcycle; oh motorcycle	Mother	0.512	□
Mar 2, 07 19:05:59 PM	motorcycle where's a motorcycle go?	Mother	0.698	□
Mar 15, 07 21:25:40 PM	bicycles; bicycles yeah motorcycles	Baby	0.356	□
Mar 15, 07 21:25:54 PM	green; green pjs with motorcycles on them...	Mother	0.417	□
Mar 18, 07 10:05:20 AM	ff; that's OK that motorcycle was like thirty...	Mother	0.316	□
Mar 27, 07 8:20:45 AM	motorcycles	Father	0.481	□
Apr 17, 07 14:47:20 PM	it's gonna be motorcycles	Nanny	0.498	□
Apr 17, 07 14:47:23 PM	motorcycles	Baby	1	☑
Apr 17, 07 16:06:30 PM	bike bicycle; no no bicycle it's a motorcycle	Nanny	0.445	□
Apr 17, 07 16:06:36 PM	jj jj; motorcycle	Nanny	0.568	□
Apr 17, 07 16:44:17 PM	bicycles jj jj; bicycles? it's a motorcycle	Baby	0.316	□
Apr 17, 07 16:44:22 PM	yeah motorcycle ice; ice	Baby	0.328	□
Apr 17, 07 16:44:50 PM	motorcycle	Mother	0.472	□
Apr 17, 07 16:54:02 PM	jj; motorcycle on your cushion?	Mother	0.332	□
Apr 17, 07 16:54:08 PM	motorcycle	Nanny	0.309	□
Apr 17, 07 16:54:10 PM	ff; yeah motorcycle	Baby	0.314	□
Apr 17, 07 16:54:17 PM	blue motorcycle	Mother	0.255	□
Apr 17, 07 16:54:19 PM	orange motorcycle yellow	Mother	0.362	□
Apr 17, 07 16:54:33 PM	yellow motorcycle	Mother	0.411	□
Apr 17, 07 16:54:46 PM	blue jj; motorcycle	Baby	0.577	□
Apr 17, 07 16:54:47 PM	yeah motorcycle	Mother	0.444	□

☑ Pad audio   Showing words: [motorcycle, motorcycles]

Context

Focus transcript: motorcycles
Apr 17, 2007

start time	transcription	speaker	speaker conf
14:46:26 PM	are you a boy or a girl	Nanny	0.362
14:46:30 PM	boy	Baby	1
14:46:31 PM	yay good job	Nanny	0.657
14:46:34 PM	boy	Baby	1
14:46:35 PM	yeah you are a	Nanny	0.433
14:46:36 PM	big boy right	Nanny	0.617
14:46:38 PM	big boy	Baby	1
14:46:42 PM	ff; you're a big	Nanny	0.502
14:46:44 PM	cute boy	Father	0.309
14:46:48 PM	what are you	Nanny	0.492
14:46:50 PM	you're a big	Nanny	0.519
14:46:52 PM	good	Mother	0.47
14:46:53 PM	boy	Father	0.483
14:46:54 PM	what is that	Nanny	0.317
14:46:58 PM	bicycles	Baby	1
14:47:01 PM	did you like that	Nanny	0.403
14:47:03 PM	did you like that	Nanny	0.531
14:47:08 PM	wow you're gonna have a jj wit...	Nanny	0.387
14:47:12 PM	bicycles	Baby	1
14:47:13 PM	yeah	Nanny	0.607
14:47:15 PM	pillow yeah; yeah a pillow with...	Nanny	0.337
14:47:20 PM	it's gonna be motorcycles	Nanny	0.498
14:47:23 PM	motorcycles	Baby	1
14:47:24 PM	yeah	Baby	0.54
14:47:25 PM	motor	Nanny	0.574
14:47:29 PM	ow my knee	Baby	0.363
14:47:33 PM	ff	Nanny	0.396
14:47:34 PM	ff mommy	Baby	0.862
14:47:36 PM	ff	Baby	0.67
14:47:37 PM	ff; what	Baby	0.557
14:47:47 PM	no	Baby	1
14:47:47 PM	ff; come on	Nanny	0.475
14:47:53 PM	oh	Nanny	0.5
14:47:54 PM	ff; jj jj	Baby	1
14:47:55 PM	ff	Nanny	0.504
14:48:15 PM	door	Baby	1

Fig. S5. Screen shot of the Word Birth Browser tool showing the main window (*Left*) and context window (*Right*). In the main window, the left pane is used to select a word to review and the right pane presents all utterances containing the target word, which can be sorted by different attributes. The context window presents the utterances that surround the selected utterance within a temporal window of 1–2 min.

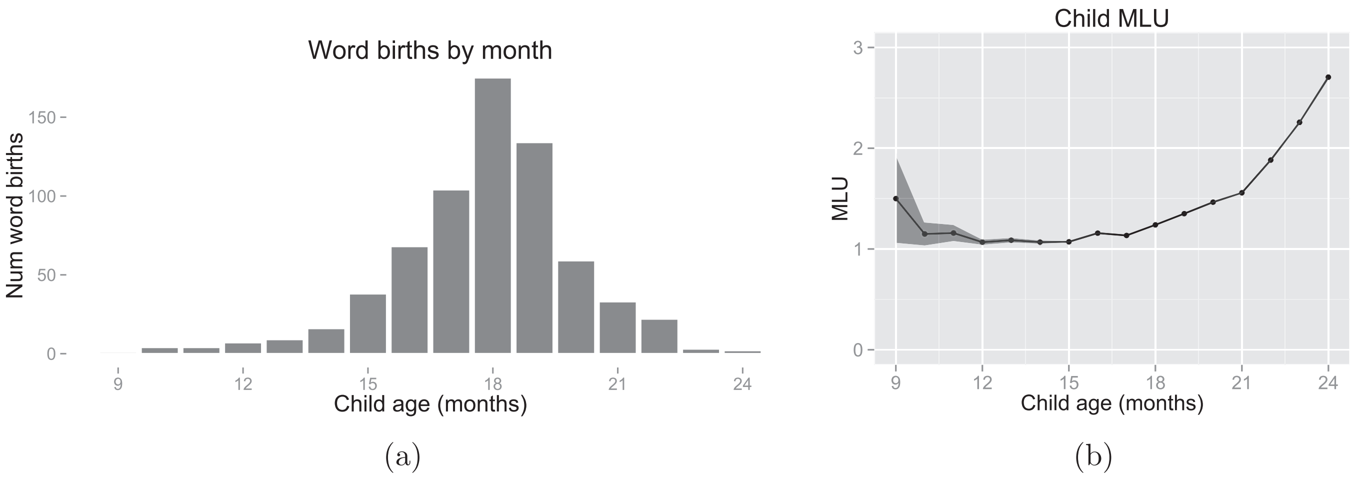


Fig. S6. Child's word birth count (*A*) and MLU (*B*) by month (95% confidence interval shaded). The child's total vocabulary is increasing across the full 9- to 24-mo age range, but the growth rate exhibits an increase up to 18 mo of age, followed by a decline. However, MLU remains relatively flat (at ~ 1) until 18 mo. Num, number.

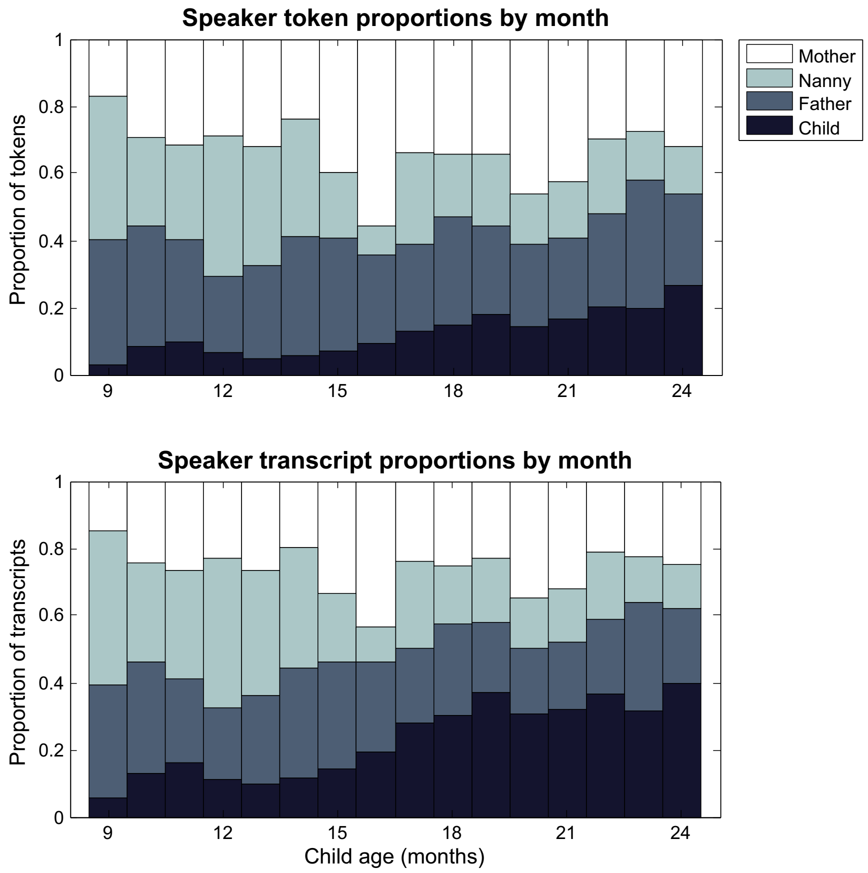
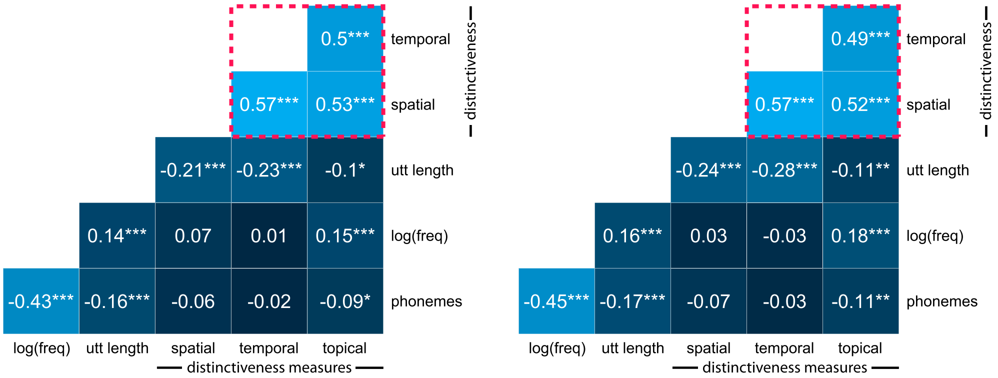


Fig. S7. Overall breakdown of spoken language over time for each speaker. The proportion of word tokens produced (*Top*) and the proportion of transcripts produced (*Bottom*) are shown.



(a) Pearson corr. coefficients (b) Spearman corr. coefficients

Fig. S8. Pearson (*A*) and Spearman (*B*) correlation (corr.) coefficients between all pairs of predictors. Frequency and number of phonemes are most strongly correlated, an indication that longer words tend to be used less frequently [first noted by Zipf (43)]. The red box shows correlations for distinctiveness predictors. freq, frequency; utt, utterance. ***$P \leq 0.001$; **$P \leq 0.01$; *$P \leq 0.05$.

Table S1. Top 10 words on which the three distinctiveness predictors differ in their predictions

Rank	Word	Deviation	Spatial	Linguistic	Temporal
1	Diaper	14.63	4.03	0.94	−1.36
2	Chase	8.07	−1.00	2.01	−1.80
3	Change	7.10	2.96	0.16	−0.62
4	Light	7.08	3.49	0.28	0.19
5	Breakfast	6.41	−1.06	−1.30	1.92
6	Living	4.90	−1.00	1.74	−0.94
7	Door	4.85	2.08	−0.59	−0.63
8	Poop	4.84	2.46	0.17	−0.51
9	Medicine	4.64	−0.63	−0.91	1.86
10	Downstairs	4.63	1.89	−0.84	−0.65

Table S2. Baseline regression model

Variable	Estimate	SE	*t* value	Pr(>|*t*|)
(Intercept)	555.150	2.353	235.927	<0.001
s.cmu.phon	15.710	2.629	5.977	<0.001
sln.freq.pre	−6.657	2.647	−2.515	0.012
s.uttlen.pre	16.341	2.430	6.725	<0.001

Pr, probability.

Table S3. Regression model, including spatial distinctiveness predictor

Variable	Estimate	SE	*t* value	Pr(>|*t*|)
(Intercept)	554.131	2.210	250.757	<0.001
s.cmu.phon	14.936	2.504	5.964	<0.001
sln.freq.pre	−3.079	2.691	−1.144	0.253
s.uttlen.pre	16.313	2.670	6.111	<0.001
srl.sp.KL	−22.053	2.258	−9.767	<0.001

Table S4. Regression model, including temporal distinctiveness predictor

Variable	Estimate	SE	*t* value	Pr(>|*t*|)
(Intercept)	555.015	2.239	247.842	<0.001
s.cmu.phon	14.942	2.503	5.969	<0.001
sln.freq.pre	−5.874	2.531	−2.321	0.021
s.uttlen.pre	11.802	2.406	4.905	<0.001
srl.temp.KL	−19.330	2.297	−8.414	<0.001

Table S5. Regression model, including linguistic distinctiveness predictor

Variable	Estimate	SE	*t* value	Pr(>|*t*|)
(Intercept)	553.592	2.313	239.343	<0.001
s.cmu.phon	15.009	2.608	5.754	<0.001
sln.freq.pre	−5.405	2.788	−1.939	0.053
s.uttlen.pre	18.483	2.629	7.031	<0.001
srl.topic.KL	−14.267	2.350	−6.072	<0.001

Table S6. Regression model, including all distinctiveness predictors

Variable	Estimate	SE	*t* value	Pr(>|*t*|)
(Intercept)	553.134	2.218	249.394	<0.001
s.cmu.phon	14.281	2.517	5.673	<0.001
sln.freq.pre	−4.519	2.772	−1.630	0.104
s.uttlen.pre	14.814	2.731	5.424	<0.001
srl.topic.KL	−1.252	2.746	−0.456	0.649
srl.temp.KL	−9.033	2.833	−3.189	0.002
srl.sp.KL	−15.997	2.883	−5.549	<0.001

Table S7. Baseline (number of phonemes, MLU, and frequency) + speaker distinctiveness models for each word class

Word class	Speaker distinctiveness	SE	*t* value	Pr(>|*t*|)
All ($N = 678$)	−4.573	2.396	−1.909	0.057
Nouns ($N = 379$)	−7.818	2.930	−2.668	0.008
Predicates ($N = 201$)	5.884	3.896	1.510	0.133
Closed class ($N = 64$)	2.542	8.383	0.303	0.763