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.

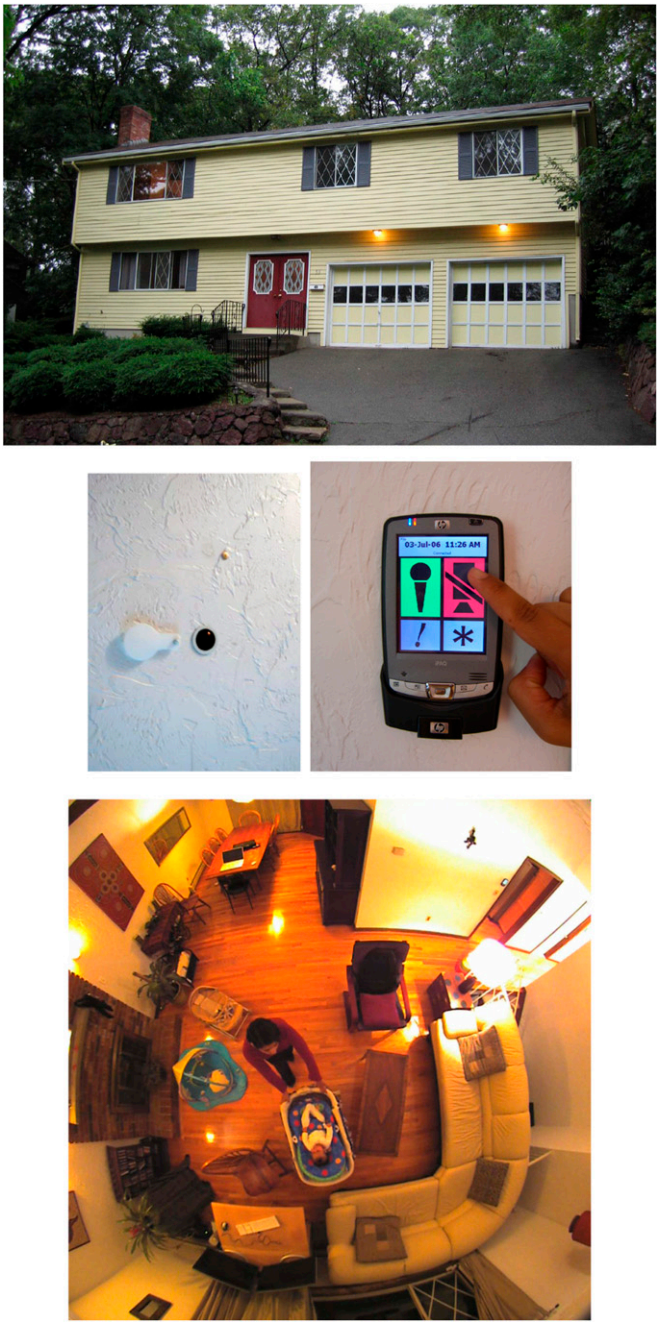


Fig. S1. Site of the Human Speechome Project, where all recording took place. Also shown is the ceiling-mounted camera with an open privacy shutter, the microphone, the recording controller, and a view into the living room.

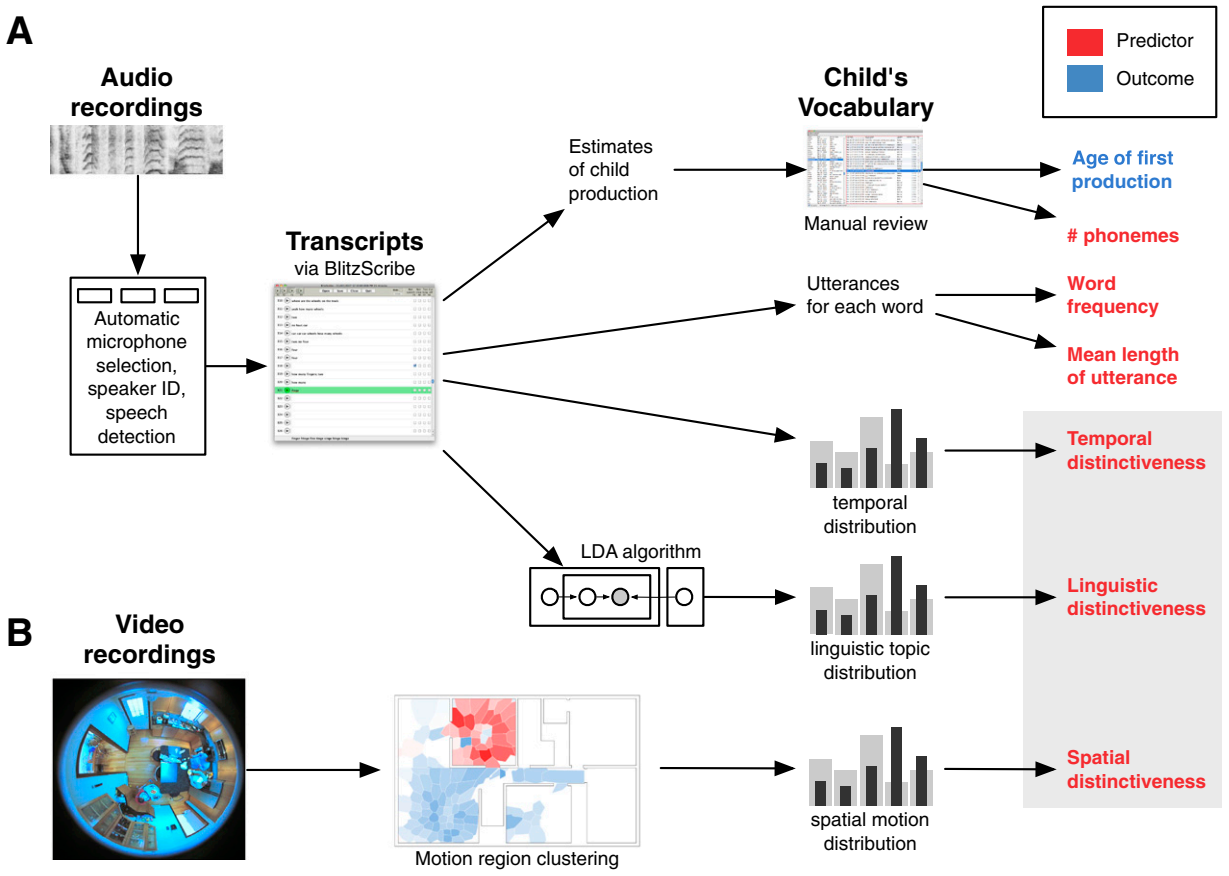


Fig. S2. Schematic of data collection and processing for our dataset, leading to our outcome (blue) and predictor (red) variables. (*A*) Audio recordings are filtered automatically for speech and speaker identity and then transcribed. Transcripts are used for the identification of the child's productions, 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.

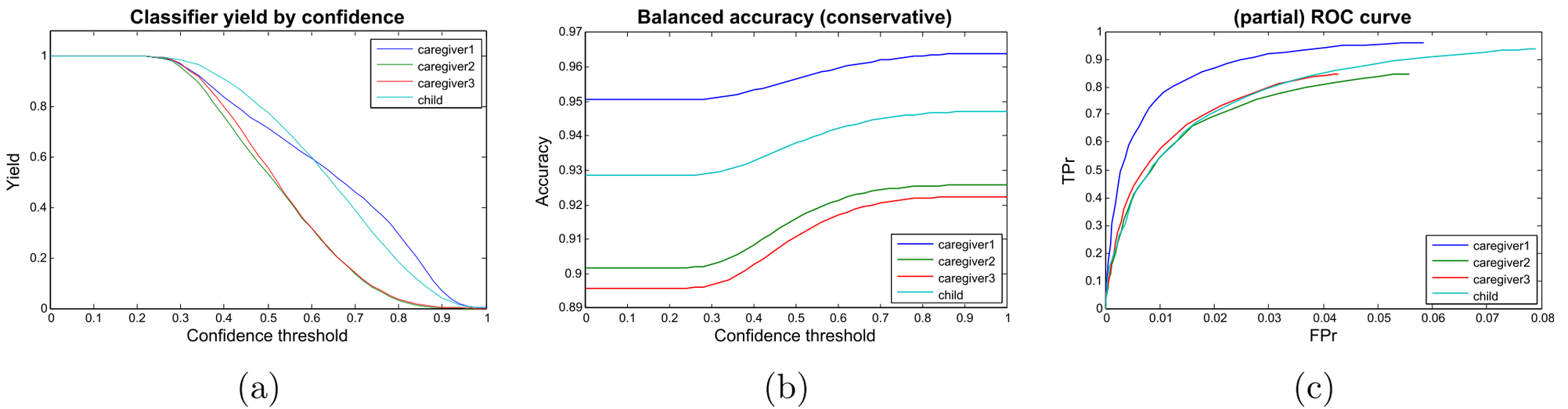


Fig. S3. Speaker identification performance curves. Classifier yield is the fraction of the speaker classifications above a confidence threshold (*A*), and accuracy is the fraction of above-threshold classifications correct for each speaker (*B*). (*C*) Receiver operating characteristic (ROC) curve displays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

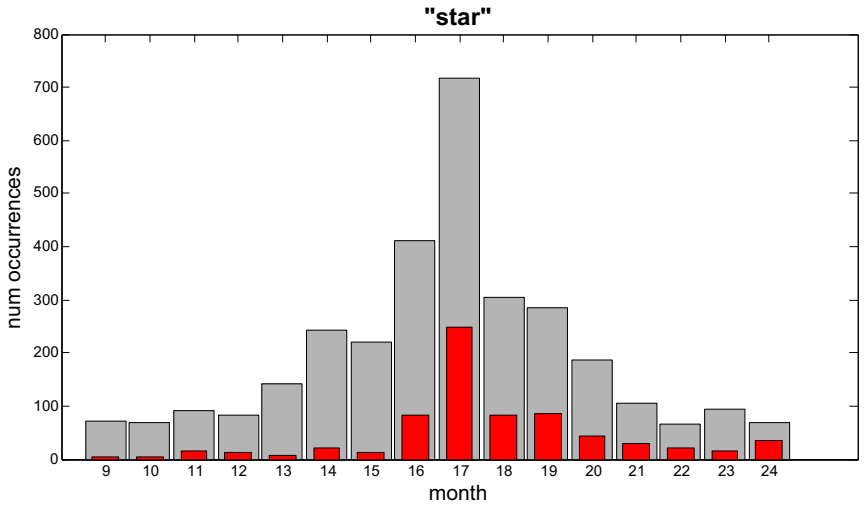


Fig. S4. Counts for the word "star" by month. Child-labeled counts are shown in red, whereas total counts across all speakers are shown in gray.

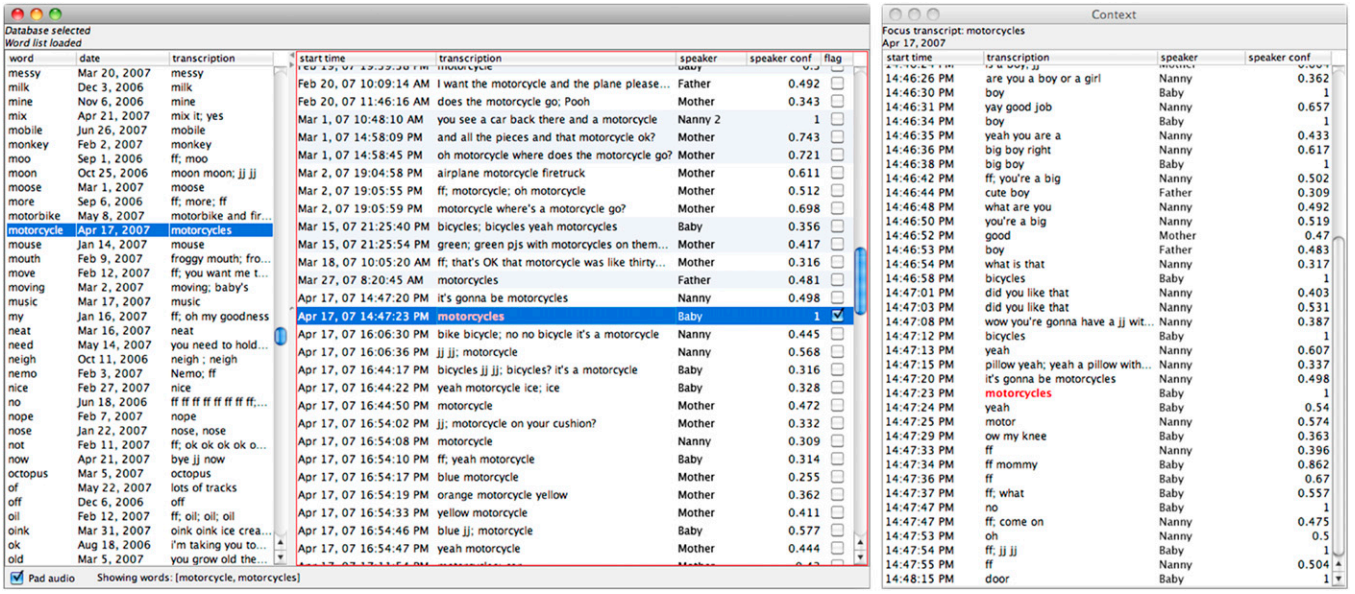


Fig. S5. Screen shot of the Word Birth Browser tool showing the main window (*Left*) and context window (*Right*). In the main window, the left pane is used to select a word to review and the right pane presents all utterances containing the target word, which can be sorted by different attributes. The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

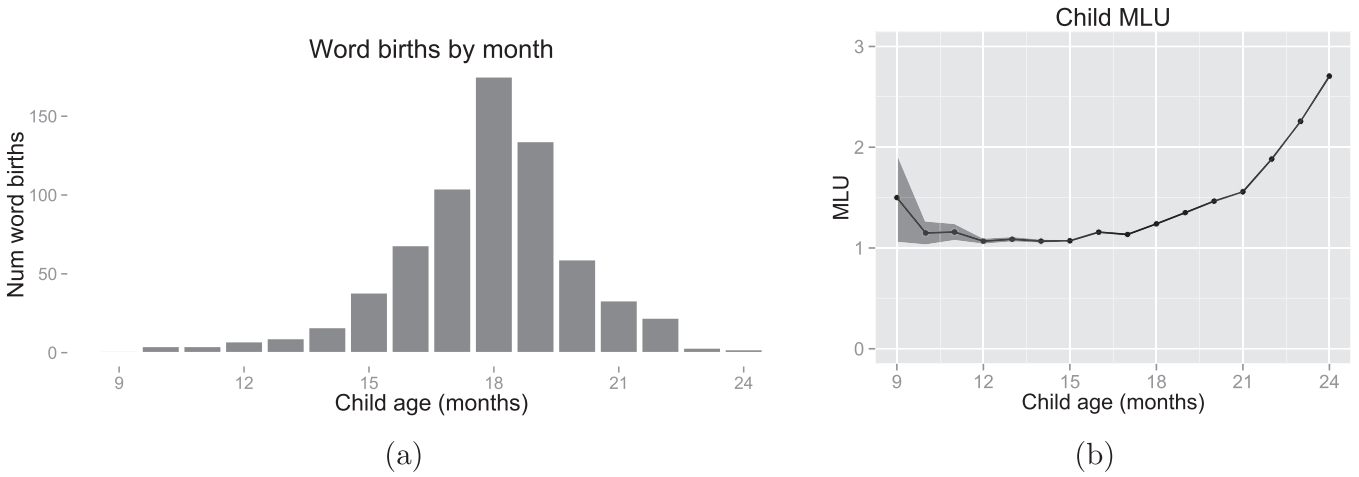


Fig. S6. Child's word birth count (*A*) and MLU (*B*) by month (95% confidence interval shaded). The child's total vocabulary is increasing across the full 9- to 24-mo age range, but the growth rate exhibits an increase up to 18 mo of age, followed by a decline. However, MLU remains relatively flat (at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

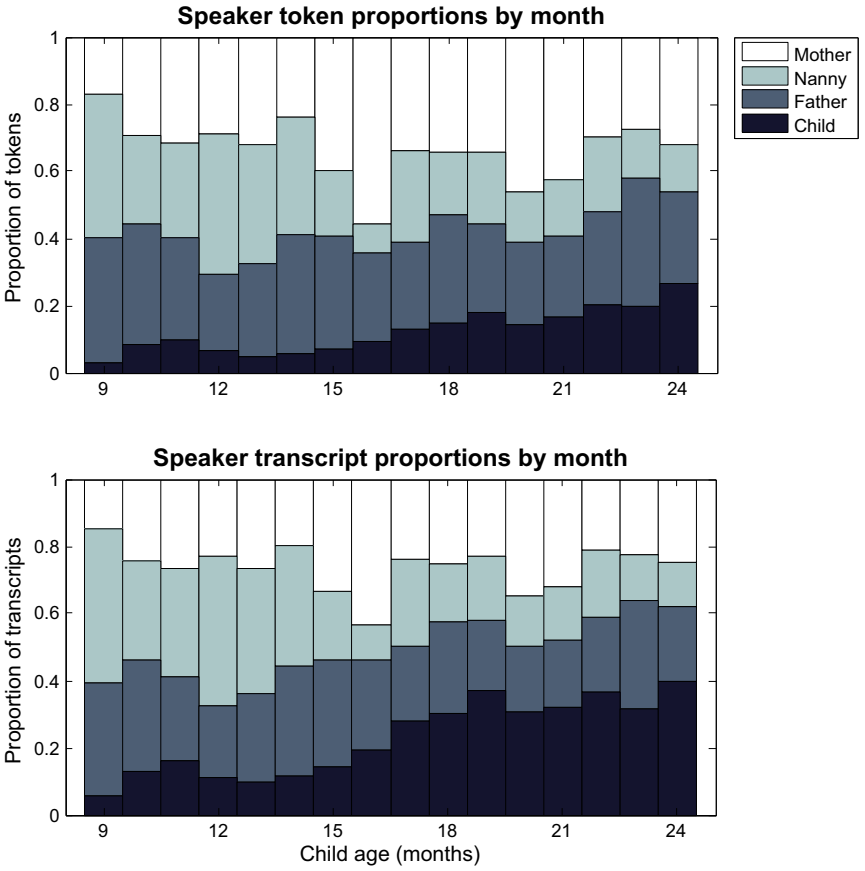
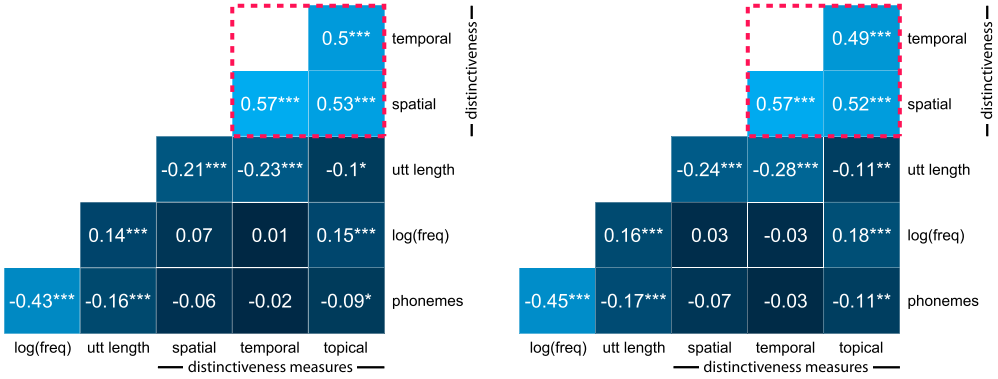


Fig. S7. Overall breakdown of spoken language over time for each speaker. The proportion of word tokens produced (*Top*) and the proportion of transcripts produced (*Bottom*) are shown.



(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