Relating Activity Contexts to Early Word Learning in Dense Longitudinal Data

Brandon C. Roy
The Media Laboratory
Massachusetts Institute of Technology
bcroy@media.mit.edu

Michael C. Frank
Department of Psychology
Stanford University
mcfrank@stanford.edu

Deb Roy
The Media Laboratory
Massachusetts Institute of Technology
dkroy@media.mit.edu

Abstract

Early word learning is contingent on linguistic input, but a child's linguistic experience is also embedded in the larger, natural structure of everyday life at home. We investigate the activity structure of life in the home of one young child, and link this structure to the child's early word learning. Our analysis is based on the dense, naturalistic, longitudinal corpus collected for the Human Speechome Project. To study activity structure, we apply probabilistic topic modeling techniques to the corpus. The emergent topics capture not only linguistic structure, but also spatial and temporal regularities indicative of coherent activity contexts. We consider the child's word learning with respect to caregiver word usage frequency and word distributions across activity contexts. We find that frequency and consistency of use across context are predictive of age of acquisition. Words that are used more frequently and in more contextually constrained settings are learned earlier, suggesting that activity contexts may be an important aspect of the child's natural learning environment and worthy of further study.

Keywords: Language acquisition; word learning; non-linguistic context; topic modeling.

Introduction

Children's early word learning is a remarkable achievement, the result of powerful learning processes unfolding in the natural setting of a child's first years of life. Cultural and individual variability in children's early environments has led researchers to question the contributions of the child's innate faculties relative to the role of the environment. But to the extent that children are *learning* language, the environment must provide appropriate conditions for learnability: There must be some consistent underlying structure for learning mechanisms to build upon.

In lexical development in particular, the linguistic environment—what words a child hears, and how often—provides essential input for the young learner. Yet the child's natural environment consists of other dimensions in addition to language: spatial, physical and social dimensions, to name a few. Learners are exposed to their input in the rich, multimodal domain of everyday experience. In this work, we begin to investigate the activity structure of day-to-day life and its contributions to early word learning. Based on the idea that words and referents are more predictable in sufficiently constrained situations, we hypothesize that words associated with a limited range of recurrent activities will tend to be learned earlier. That is to say, consistent linguistic input across a narrower range of activities poses a simpler learning problem.

The effect of overall linguistic input on lexical development was investigated by Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991). They were the first to document positive correlation between the quantity of child-directed speech and a child's vocabulary size and growth rate. For individual words, increased frequency of use was also tied to earlier acquisition of those words; our own (Roy, Frank, & Roy, 2009) and other (Goodman, Dale, & Li, 2008) findings replicate this pattern. In addition to frequency, words presented in single word utterances (Brent & Siskind, 2001) and with prosodic stress (Echols & Newport, 1992; Vosoughi, Roy, Frank, & Roy, 2010) are also acquired earlier.

In addition to studying linguistic input, work in cross-situational word learning has investigated how words can be linked to referents through their consistent co-occurrence across a range of situations. In the face of referential uncertainty, a learner sensitive to the statistics of which words and referents co-occur can learn correct word-referent pairings (Yu & Smith, 2007). But the idea of learning by gradually accumulating word-referent co-occurrences was challenged by Medina, Snedeker, Trueswell, and Gleitman (2011), on the grounds that the sheer number of possible pairings in everyday experience, coupled with memory limitations, leads to an intractable learning problem. Their data suggest a different learning strategy based on early binding between words and referents, with errors corrected through natural processes of forgetting.

While the natural environment is complex, it does provide structure notably absent from many laboratory-based word learning experiments. Bruner (1985) emphasized the importance of naturally occurring, predictable *formats* of interaction that support communication. To study the role of formats in language acquisition, Bruner moved his research into the "clutter of life at home" via naturalistic, observational methods. One format that Bruner studied was the game of "peek-a-boo", a recurring, rule-bound activity that occurs across a wide developmental period. Language works in concert with the game to help reveal the meaning of words.

With Bruner's formats in mind, the goal of the present study is to investigate the activity structure of a child's first years of life, how the child's linguistic input links

Note: this draft corrects an error in Figure 4.

to these activities, and how such language in context relates to vocabulary growth. Bruner's formats are complex, with deep rule-governed structure and social roles, patterns that recur over time during the child's early life. They are difficult to study in detail, especially since they must be observed and deconstructed from longitudinal observations of natural behavior. To avoid this difficulty, we study a simplified representation of formats: consistent *activity contexts*.

We operationalize the idea of an activity context using data mining and machine learning techniques, applied to the multimodal, dense longitudinal recordings collected for the Human Speechome Project (Roy et al., 2006). We apply Latent Dirichlet Allocation models (Blei, Ng, & Jordan, 2003) to the transcribed speech in the Speechome Corpus, obtaining a set of "topics" that connect groups of related words. Inspection of these topics along linguistic, spatial, and temporal dimensions demonstrates that many correspond to coherent, everyday activity contexts such as `mealtime`, `diaper-change`, and so on. We then consider the child's vocabulary growth relative to both the standard input frequency and measures of a word's diversity across activity contexts.

The Human Speechome Corpus

The Human Speechome Project (HSP) (Roy et al., 2006) was launched in 2005 to study early language development through analysis of audio and video recordings of the first three years of one child's life. The house of one of the authors (DR, who had a newborn child), was outfitted with eleven omnidirectional cameras, fourteen microphones, and a custom recording system designed for large-scale audio/video recording. The cameras and microphones, embedded in the ceilings, provided near complete coverage of the house while remaining unobtrusive, and the practice of simply turning the system on in the morning and leaving it on all day facilitated adoption of the system and helped to minimize observer effects. The nature of this project required extreme sensitivity to the family's privacy: They had full control over recordings and the ability to "back-delete" recordings if an embarrassing moment was captured. Audio was recorded using boundary-layer microphones which yield high quality audio, even for whispered speech. Video was recorded at approximately 1 megapixel, 15 frames per second, using high dynamic-range cameras for the wide range of lighting conditions. On average, the family recorded 10 hours per day, from the child's birth to age three. Altogether, the recordings span roughly 120,000 hours of audio and 90,000 hours of video, capturing an estimated 70% of the child's waking hours.

Data Annotation

To date, the focus of our annotation and analysis has been on the subset of data spanning the child's 9-24 month age range. For this subset, our long-term goal has been to transcribe *all* speech both heard and produced by the child, but this is a significant challenge using traditional transcription methods. To address this, we have developed BlitzScribe, a new tool for fast, semi-automatic speech transcription (Roy & Roy, 2009). The BlitzScribe system processes raw, unstructured audio and automatically finds speech, segments it into manageable segments, and presents those candidate speech segments to a human transcriber in a simplified user interface. We then use a fully automatic speaker ID system to identify the speaker in an utterance.

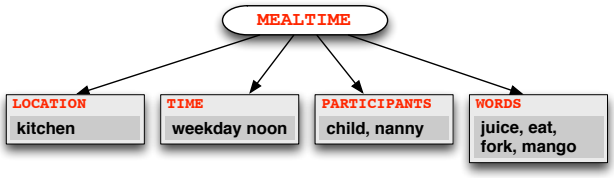

Figure 1: A schematic illustrating how four different dimensions of observable data can depend on a common latent activity context. Viewed as a generative model, the activity context `mealtime` gives rise to the four kinds of observed data.

Human annotators label which room the child is in (and whether he is awake) over the course of the day. This step ensures that what is transcribed is effectively "child available speech" (CAS), or speech that could be considered linguistic input. Although many studies focus on child-directed speech (CDS) for input-uptake analysis, CDS is much more difficult to obtain at a large scale than CAS. Using BlitzScribe, we have transcribed more than 80% of the CAS audio collected in the 9-24 month age range, which we refer to as the Speechome Corpus. Currently we have transcribed approximately 8 million words, and when fully transcribed we expect the corpus to consist of about 10 million words. However, since some post-processing is required for the latest transcripts, the work described here uses an earlier version of the corpus consisting of approximately 5 million words.

The Child's Lexicon

The density and coverage of the Speechome Corpus enables a detailed look at lexical development, including both caregiver speech and the child's vocabulary over time. In earlier work (Roy et al., 2009), using a smaller version of the corpus, we identified a *word birth* as the first productive use of a word by the child in our transcripts. For our purposes, this served as the age of acquisition (AoA) of each word in the child's lexicon. We repeated this procedure using the current, larger corpus and identified a large set of candidate word births. We then manually reviewed each of these, removing morphological variations like plurals, dropping invalid word

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ic	ok	ball	i	ba	duck	fish	spider	he	bye	butterfli	book	chew	water	la	cow
cream	put	it	thei	diaper	quack	e	twinkl	like	gaga	crunch	cat	eat	okai	baa	moo
you	come	push	just	poo	bear	farm	piggi	wa	ya	two	sam	mango	no	wool	sun
want	pant	go	know	beep	doggi	turtl	woof	he'	hi	four	fox	it	abar	sheep	moon
mango	your	catch	to	chang	giraff	jellyfish	meow	him	ee	three	read	chees	milk	[noise]	done
what	on	bounc	but	bath	dog	jiggli	star	yeah	NANNY	five	hat	juic	merrili	sir	diddl
shake	go	throw	that	grape	camel	sea	oink	i	rock	ladybug	chip	yum	row	[crying]	all
babi	let'	kick	year	pee	ruff	jelli	neigh	it	wibbl	tigger	knox	pea	tea	[babbling]	jump
do	it	get	like	tama	bird	jenni	bitsi	but	bye-by	bee	ham	bite	cup	[laugh]	mulberri
drum	shower	basketbal	dollar	crayon	eleph	fishi	itsi	just	cradl	pooh	beetl	more	cooki	dame	frederick

Figure 2: The top 10 words (orthographic substitutions are due to the stemming process) for 16 of the top 25 topics.

births, and adjusting birth dates. We identified 670 unique forms in the child's lexicon at 24 months.[1]

Activity Contexts

The detailed record of development contained in the Speechome Corpus includes the child's first words up to multiword utterances. But in addition, the basic routines of daily life are also captured, providing a backdrop for early development. What activities does a child participate in during his first years, and how can they be found in a large, unstructured collection of recordings?

Our approach is to view an *activity context* as a hidden or *latent* variable that explains a set of observable data. An activity such as `mealtime` typically takes place in the kitchen, around noon or in the early evening and involves the whole family, with the speakers often uttering food- and eating-related words. A particular combination of observed *time*, *location*, *words* and *participants* may be best explained by the `mealtime` activity context, illustrated by Figure 1. Thus, an activity context is a latent variable identified by observations across modalities. We wish to identify a set of latent activity contexts from these observables across the entire Speechome Corpus.

Automatic methods for inferring latent variables have been successfully used in data mining applications like document modeling. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is one such technique, which finds a set of latent "topics" that best capture the thematic content of a collection of documents. In LDA, each document is represented as an unordered collection ("bag") of words; the inferred topics are modelled as distributions over words. Topics group related words together and documents are represented as sparse mixtures of topics. Often, a human can interpret and label the topics simply by inspecting the topic words. As a first exploration of activity structure in the Speechome corpus, we apply LDA directly to transcripts. We then assess the relationship between LDA topics and activity contexts using data from time and location.

Applying LDA to the Speechome Corpus

To apply LDA to the Speechome Corpus, we partitioned the transcripts into "documents" using a sliding window procedure. Beginning at the 9-month mark we advanced a 10 minute window over the corpus, shifting the window forward by 10 minutes up to the 24 month mark. All transcribed speech in a window was output as a document for processing by LDA, skipping empty time windows that didn't contain speech, resulting in 13,672 documents. After some experimentation, we found stemming to be a useful preprocessing step, normalizing word forms to a common root using the Porter stemmer (Porter et al., 1980), and only accepting those words occurring in more than five documents (and occurring more than five times in the corpus.) This yielded a vocabulary of 6,583 unique word types.

In the case of standard LDA, the number of topics to produce is a parameter of the algorithm, and we found 25 topics to be a manageable number while still producing coherent topics. Extensions to LDA such as Hierarchical Dirichlet Processes (Teh, Jordan, Beal, & Blei, 2006) can automatically select the number of topics, and informal experiments with this method also resulted in 20 – 30 topics. To interpret the resultant topics, a common starting point is to review the top words in each topic. We ranked words using the method in (Blei & Lafferty, 2009), which roughly measures the informativeness of the word for the topic relative to the other topics (Figure 2).

From Topics to Activities

Do topics capture activities? We investigate two methods to make the link: via correlations in time and space, and via human-annotated activities.

Activities in time and space LDA outputs topic mixture weights for each document; since documents also have spatial and temporal attributes, we can exploit this to measure how topics are distributed in time and space. Each topic's time distribution was calculated by weighting the time of day of each document by the topic's con-

[1] We did not annotate or study *receptive* AoA, which is often documented in diary studies but is much more difficult with a large corpus. Identifying word births is challenging in its own right, since the child's word form may differ from the adult form. In describing the diary study of her daughter's early lexical development, Dromi (1987) reviews these and other challenges. In our case, the original audio, video, and access to caregivers were all helpful resources.

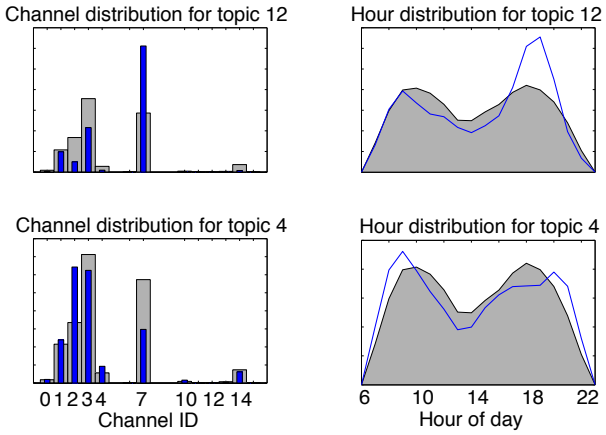

Figure 3: Spatial and temporal distributions for topics 12 and 4. The gray background graphs show overall averages, the blue foreground graphs the "conditional" distributions for the topic (distributions reweighted by the selected topic activity.) The channels of interest here are 7 (kitchen) and 2 (baby bedroom.)

tribution to that document. The spatial distribution of topics across rooms of the house was calculated similarly.

Figure 3 shows the temporal and spatial distributions for topics 12 and 4, relative to the average temporal and spatial distributions. When topic 12 is active, we see that recording channel 7 (the kitchen) is well above average, and the temporal distribution peaks at about 7pm. Inspecting the words in topic 12 shows that it captures food and eating related terms. So, topic 12 appears to be a `mealtime` activity context, or perhaps is more specific to `dinnertime`. Topic 4 is most active, relative to the average, in the early morning and late evening, and in channel 2, the baby's bedroom. This topic appears to capture the `diaper-change` activity. Thus, at least a subset of topics appear to follow coherent spatial and temporal distributions.

Human annotated activities In concert with our efforts to automatically identify activity contexts, we are also manually annotating activities. Using BlitzScribe, annotators now transcribe assignments spanning 15 minutes of "house time," then list the activities that took place. When we began this annotation project, we gave little instruction to transcribers, asking them to make up their own activity tags as necessary. Nevertheless, we found consistency in the activities that emerged. After conventionalizing tag names, we obtained roughly 30 activities for around 300 annotated assignments. These annotations can provide another means for validating LDA topics as proxies for activity contexts.

To test for relationships between LDA topics and activity contexts, we examined the correlations between individual topics and the human-annotated activities. While these correlations remain speculative due to the sparsity of the human-annotated activities, several significant correlations emerged. In the case of the `diaper-change` activity, for example, only topic 4 was significantly correlated (with words like "diaper," "poo," and "change" highly active). In the case of `eating`, topics 12, 16 and 0 are significantly positively correlated, with 12 being the strongest (e.g. "chew," "eat," and "mango"). For `reading`, a number of topics were active, including 5, 6, 10, and 11, all of which contained words related to different books that were read to the child. And for `crying`, topic 17 (e.g. "daddy," "blanket," and "ssh") was most active. In summary, although at present human annotation of activities is too limited to provide full coverage, the relationships between activities and topics makes us optimistic that our topics are capturing at least some aspects of the varying activity contexts in the child's environment.

Table 1: Coefficients on a multilevel linear regression model predicting age of acquisition (months) on the basis of log frequency in child-available speech, topic entropy of the word, and their interaction.

Predictor	Coefficient	Std. Err.	t-value
Intercept	18.49	0.28	65.83
Log frequency	-0.83	0.12	-7.08
Topic entropy	0.54	0.10	5.44
Log freq × entropy	0.06	0.13	0.48

Word Learning

If LDA topics act as a proxy for activity contexts, then we should be able to use them to test a primary hypothesis of interest: that words that appear in consistent activity contexts are learned relatively earlier than those that appear across a range of contexts. Said another way, words with high *topic entropy*—that do not appear consistently in one or a small set of topics—should be produced later by the child.

We used multilevel linear regression (Gelman & Hill, 2007) to predict age of acquisition (AoA, in months) on the basis of word frequency and topic entropy. AoA measures are described above. For word frequency, we measured the total number of utterances of a target word in our sample up to the age of acquisition of the word, normalized by the number of days of transcripts up until that time to allow these measurements to be compared for words with different AoA.[2] For topic entropy,

[2]We measure only up until the acquisition of the word to avoid a confound: the child's production of a word could change the adult use of the word. Note that this change, the exclusion of words from the topic model, and several other minor changes make regression coefficients for frequency slightly lower compared with our previous work.

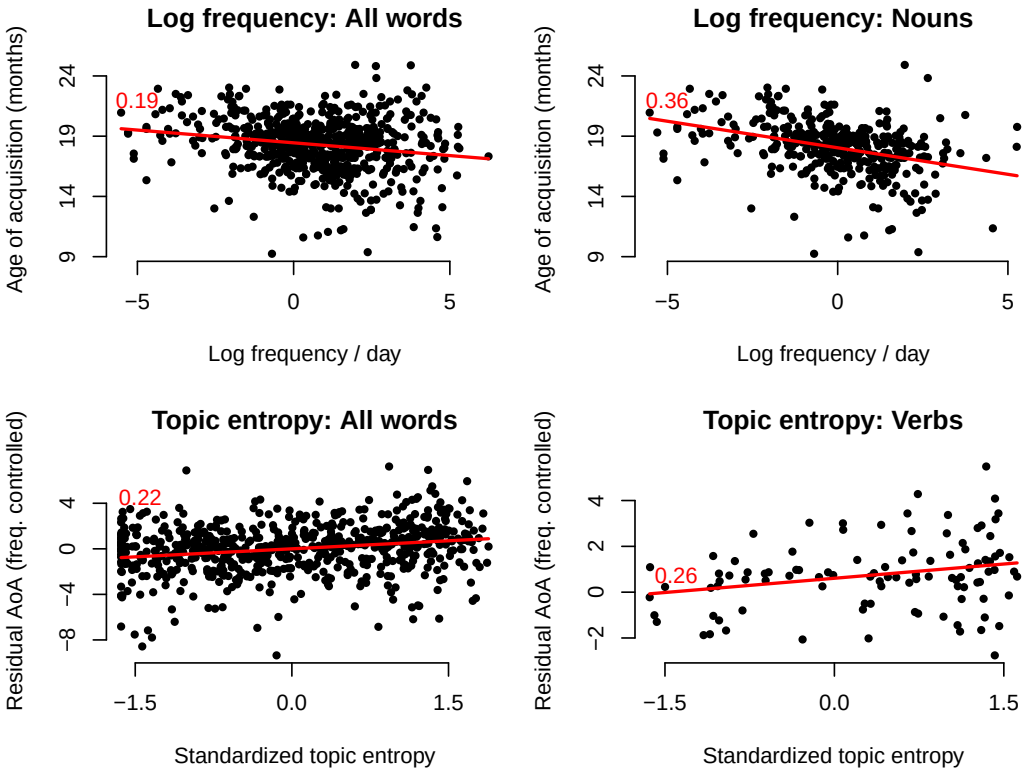

Figure 4: Each plot shows age of acquisition in months for individual words plotted by a predictor. (top) Age of acquisition for words plotted by standardized log word frequency per day (prior to the child's first production of the word). Left plot shows all words, right plot shows nouns. (bottom) Residual age of acquisition for words plotted by standardized topic entropy. Left plot shows all words, right plot shows verbs. Red lines show best fitting linear function; numbers indicate correlation coefficients.

we computed the entropy of the word's weight distribution over the 25 different topics found by the LDA model. We z-scored the units for both of these predictors to be able to compare coefficients for each.

Our model included fixed effects for topic entropy, frequency, and their interaction. The model also included random terms for each syntactic category, and its frequency and topic entropy, and their interaction. Coefficients are shown in Table 1. Coefficient weights can be interpreted as months in AoA per standard deviation in log frequency or topic entropy. We assessed reliability for individual coefficients by testing whether they increased model likelihood. Log frequency had a large negative effect on AoA: more frequent words were learned earlier ($p < .005$). Topic entropy had almost as large an effect: words in more constrained activity contexts (lower entropy) were learned earlier as well ($p < .005$). Their interaction did not significantly increase model fit ($p = .69$).

Figure 4 shows the relationships described by this model. Without regressing out frequency, topic entropy is relatively uncorrelated with age of acquisition. When both terms are entered in a model, however, the effect is much larger. Figure 4 displays this conditional relationship by plotting residual AoA (controlling frequency) by topic entropy.

Topic entropy and part of speech are likely correlated: closed class words like "if" are likely to occur in every topic (topic entropy of 1.3 SDs above mean), while nouns like "pasta" only appear frequently in one context (1.6 SDs below). A key part of this analysis was the use of multilevel models to control for part-of-speech effects. Without including random effect terms for part-of-speech, the interaction between frequency and topic entropy was large, probably because topic entropy and frequency for closed class words is high. Adding the random effects terms eliminated this interaction, however.

To summarize this analysis: we found that the consistency of the contexts within which words appeared was almost as strong a predictor of age of acquisition as pure frequency.

Discussion and Future Work

Early word learning is a product of powerful learning mechanisms coupled with the rich experience of early childhood. Linguistic input is of critical importance to

lexical development, but it is situated in the larger structure of daily life. The importance of social activity structures was emphasized by Bruner (1985), yet large-scale, quantitative study of their effect on language acquisition has proven difficult. To address this, we used document modeling techniques to operationalize activity contexts. We found evidence that many of the resultant topics captured coherent, interpretable patterns of linguistic, temporal and spatial activity. These activity contexts then provided a useful source of information in modeling lexical acquisition: we found that more contextually focused words were learned earlier.

In future work, we plan to add location, participants, and time to models of latent activity contexts. An interesting question for these extensions is whether some contexts are of more value than others for general word learning or for learning particular words. In addition, a study of episodes of a particular activity may help build intuitions about how activities develop and change over time, and how this progression relates to the child's development.

Our study here represents a first step towards a more complete model of lexical acquisition, one that incorporates elements of social and physical context. In a similar vein, Miller (2011) and Shaw (2011) studied the spatial distribution of language in the Speechome Corpus. Miller (2011) found that more spatially localized words correlated with earlier AoA, noting that many of the most salient locations were directly interpretable in terms of the activities known to take place at those locations. Our work builds on this intuition, targeting activities via their linguistic manifestations. While both of these methods are at best proxies for as-yet-unseen structures, our hope is that by continuing to develop methods for identifying activity contexts, we can gain some insight into the crucial role these social structures play in early language learning.

Acknowledgments

Many thanks to our team of annotators, and to Cybelle Smith and anonymous reviewers for helpful comments.

References

Blei, D. M., & Lafferty, J. (2009). Topic models. In A. Srivastava & M. Sahami (Eds.), *Text mining: Classification, clustering, and applications* (pp. 71–93). Chapman & Hall.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003, January). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 33–44.

Bruner, J. (1985). The role of interaction formats in language acquisition. In J. P. Forgas (Ed.), *Language and social situations* (pp. 31–46). Springer-Verlag.

Dromi, E. (1987). *Early lexical development.* Cambridge University Press.

Echols, C., & Newport, E. (1992). The role of stress and position in determining first words. *Language acquisition*, *2*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models* (Vol. 648). Cambridge University Press New York.

Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014.

Miller, M. (2011). *Semantic Spaces: Behavior, Language and Word Learning in the Human Speechome Corpus.* Unpublished master's thesis, Massachusetts Institute of Technology.

Porter, M., et al. (1980). *An algorithm for suffix stripping.* Program.

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

Roy, B. C., & Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech.* Brighton, England.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., ... Gorniak, P. (2006). The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 2059–2064). Mahwah, NJ: Lawrence Erlbaum.

Shaw, G. (2011). *A taxonomy of situated language in natural contexts.* Unpublished master's thesis, Massachusetts Institute of Technology.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

Vosoughi, S., Roy, B. C., Frank, M. C., & Roy, D. (2010). Effects of caregiver prosody on child language acquisition. In *Fifth international conference on speech prosody.* Chicago, IL.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414.