Children gesture when speech is slow to come

Daniel Yurovksy, Madeline Meyers, Nicole Burke, and Susan Goldin-Meadow
{yurovsky, mcmeyers, nicoleburke, sgm}@uchicago.edu
Department of Psychology
University of Chicago

Abstract

We test this prediction in a corpus of videos of parent-child interaction in the home recorded longitudinally from 14- to 34-months. naturalistic parent-child o

Keywords: communication; language acquisition; gesture; corpus analysis

Introduction

Children learn a striking amount of language in their first few years of life–thousands of sounds, words, grammatical categories, and the combinatoric properties that allow them to be combined to produce meaningful utterances (E. V. Clark, 2009). They also come to understand what all of this language *is for*: communicating with other people (Zipf, 1949). There is good reason to think that these two problems are deeply intertwined. The language that children hear is rarely a running commentary on the world around them–when a child's parents return home from work, they are much more likely to say "whatcha been doing all day?" than "I am opening the door" (Gleitman, 1990). Understanding that their parent is not trying to tell them about the door may go a long way to learning what the words they are hearing mean.

The understanding that speakers productions are intended to communicate information is at the core of the kinds of inferences that adults routinely make when processing language. These pragmatic inferences, for instance, are the reason that hearing a speaker say that they ate "some of the cookies," causes us to think that some cookies still remain on the plate (Grice, 1969). Children's ability to perform these kinds of complex inferences appears relatively late in language development (Noveck, 2001). However, a growing body of empirical evidence shows that a basic understanding of the communicative purpose of language is already present in the first year of life (Tomasello, 2000). For instance, children appear to understand that speakers communicate information to other adults, even if they themselves do not understand the words being said (Vouloumanos, Martin, & Onishi, 2014; Vouloumanos, Onishi, & Pogue, 2012). But is this understanding of communicative goals present in children's *language production* as well?

The core of extended communicative interactions is taking turns: participants each contribute to the discourse, but only one at a time (Sacks, Schegloff, & Jefferson, 1974). Turn taking is not only universal among both modern and indigenous cultures, the length of time between turns is highly stereotyped, and predicted by the same factors across cultures (Stivers et al., 2009). Evidence from both early observational studies and more recent experiments suggests that tracking of turn boundaries emerges early in infancy–perhaps in the course of scripted interactions like patty cake (Bruner, 1983; Casillas & Frank, 2017).

The regularity of these turns makes communication inherently time constrained: If you stop talking for too long, you lose your turn. Adults are sensitive to this time pressure, for instance producing filled pauses like "um" when they are having difficulty retrieving the words they want to produce in order to signal their desire to hold onto their turn (H. H. Clark & Fox Tree, 2002). If retrieval is still unsuccessful, linguistically-proficient adults can opt for an alternative word or even a description that gives their interlocutor enough information to help retrieve the word for them (H. H. Clark & Schaefer, 1989). Children still learning their native language, for whom such strategies are unavailable, might resort to an alternative mode of communication: pointing.

Children produce deictic gestures early in infancy, and appear to understand that these gestures both direct attention and communicate intentions by the time they are 12-months-old (Liszkowski, Carpenter, & Tomasello, 2007; Tomasello, Carpenter, & Liszkowski, 2007). Around the same time, infants begin producing their first spoken words (Bloom, 2000). Over the next few years, infants will produce many more words, and need to rely less on deictic gesture to communicate. However, while children master some words early, others which are less frequent may remain difficult to retrieve and produce. If children, like adults, are sensitive to the time pressures of communication, then then they may use gesture even for *known* words if these words are slow to come.

Communication as a race between modalities

When children wish to share their interest in an object with a caregiver, they have two modalities available to them. One possibility is to use spoken language, producing the canonical label for it (e.g. "ball"). Alternatively, they can use a deictic gesture, e.g. a point, to draw the caregiver's attention to it. When should children use each of these modalities?

If the child does not know that the object is called "ball," they have no choice but to point. However, if they do know its label, time pressure on communication produces a race between modalities. If the child can recall the word quickly, they should prefer to use language–as speech is less effortful than pointing (Zipf, 1949). However, if the process of recalling and producing the word is progressing too slowly, the child risks losing their conversational turn and should instead point.

This kind of race model can be formalized nicely as two competing accumulators (see e.g. S. D. Brown & Heathcote,

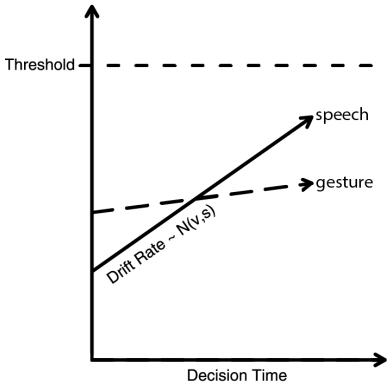

Figure 1: Reference as as a race between modalities. The drift rate of pointing should be independent of referent, but speech should vary with aproperties of words, e.g. frequency

2008). Each modality accumulates activation at its own independent rate, and whichever is the first to reach threshold wins the race and is chosen as the referential modality (Figure 1). Although the difficulty of pointing may vary due to issues of proximity of the speakers to each-other, the location of the target referent, etc., the difficulty of pointing should in general be independent of the thing being pointed to. On the other hand, the difficulty of recalling and producing a word varies from word to word. In adults, this difficulty is influenced by many features of the word, including the phonology and orthography of both the word and its neighbors in the lexicon (see e.g. Vitevitch, 2008). Here we focus on just one–contributor: Input frequency (Wingfield, 1968). The more frequently we hear a word, the easier it is for us to retrieve and produce it. Children's *language processing* shows similar effects of frequency–children's speed and accuracy of known words increases as they become more frequent (Swingley, Pinto, & Fernald, 1999). If their *language production* is similarly affected by frequency, than the rate of the speech accumulator should increase as frequency increases, resulting in it winning the race for reference more often.

The utility of this framework is that it makes detailed predictions about the relationship between modality and production time as features of the target referent change. We test three specific predictions of this model in children's spontaneous productions from 14- to 34-months:

1. As the frequency of a referent in children's input increases, they should be relatively more likely to use speech, and less likely to use gesture to communicate about it.

2. As children develop and learn more language, words should be known better and thus be easier to retrieve. Thus, speech should win the race more often–especially for low frequency words.

3. Recent use of a word should make it easier to retrieve, thus children should be relatively more likely to use speech for low frequency referents in if they have occurred previously in the same discourse than at baseline.

person	utterance	gesture	spoken	gestured
parent	do you want to read a book quick with mom		book;mom	
child	no			
child	mommy		mom	
parent	no			
parent	oh you want to wear your necklaces		necklace	
parent	uhoh			
parent	I think it's stuck			
parent	you need some help			
child		hold		necklace
parent	why don't you just say help instead of yelling			
parent	can you say help mommy		mom	

Table 1: An example of the output of data processing

Method

The data analyzed here are transcriptions of recordings parent-child interactions in the homes of 10 children from the Chicagoland area. Each recording was ˜90min long, and participants were given no instructions about how to interact–the goal was to observe the natural ecology of language learning. Each child was recorded 6 times at 4-month intervals starting at 14-mo. and ending at 34-mo.

Participants

These children's data was drawn from the larger Language Development Project dataset pseudo-randomly to preserve the socio-economic, racial, and gender diversity representative of the broader Chicago community (Goldin-Meadow et al., 2014). Of the 10 children, 5 were girls, 3 were Black and 2 were Mixed-Race. Families spanned a broad range of incomes, with 2 families earning $15,000 to $34,999 and 1 family earning greater than $100,000. The median family income was $50,000 to $74,999.

Data Processing

The original Language Development Project transcripts consist of utterance-by-utterance transcriptions of the 90 minute recordings, as well as a transcription of all communicative gestures produced by children and their conversational partners, including conventional gestures (e.g. waving "bye"), representational gestures (e.g. tracing the shape of a square), and deictic gestures (e.g. pointing to a ball).

For each of these communicative acts, we coded all concrete noun referents indicated in either the spoken or gestural modality (see Table 1). As it is difficult both to gesture about, and to code, gestures for abstract entities like "weekend," we focused only on nouns that could be referred to in either gesture or speech. Spoken referents were coded only if a noun label was used (e.g. no pronouns were included), and only deictic gestures were counted as referential to minimize ambiguity in coding. Synonyms, nicknames, and proper nouns were all coded according to a manual that can be found in the linked github repository.

Reliability

In order to ensure the integrity of the coded data for further analyses, we first assessed inter-rater reliability, and then assessed whether the coded referents were present in the scene.

Inter-Rater Reliability To assess the reliability of referent coding, 25% of the transcripts were coded by a second independent coder. Reliability between coders was good (Cohen's $\kappa = 0.76$). Issues and discrepancies in coding decisions were discussed and resolved during the formation of a coding manual.

Presence of referents Although we coded for concrete referents that had the potential to be produced either in speech or in gesture, we found that these referents were not always physically present in the environment. If a referent was not present, it could not be referred to in the gestural modality–potentially biasing our analyses. After coding all referents from the transcripts, the primary coder judged whether each was likely to be present in the scene according to a list of criteria described in the coding manual. Across the 60 transcripts, 90% of referents were judged to be present. There was a nonsignificant effect of child's age ($p < 0.115$) and whether the subject was the parent or child ($p < 0.58$), but there was a significant interaction effect of age and whether the subject was the child or parent. With an increase in age, Absent referents were included in estimates of input frequency, but excluded from analysis of production modality.

Reliability for referent presentness was calculated for 5% of the data by comparison of judgments created by the primary coder and observations of video data. The reliability was acceptable for child-produced referents (Cohen's κ= 0.79), as well as for all referents in the dataset (Cohen's κ=0.72).

Results

The key predictions of our race model of reference all connect the ease of recall of a referent's spoken label. Although ease of recall is likely related to a number of factors (e.g. phonotactic probability, neighborhood density, etc), we focus here on one easily quantifiable and well-attested predictor: input frequency (Wingfield, 1968).

Estimating Frequency

To estimate the input frequency of each referent in the corpus, we summed its frequency of use across all children and parents and across both the speech and gestural modalities. This estimator is of course imperfect– It assumes, for instance, that every child receives the same input, and that input frequency is stationary across development. Nonetheless, because of the difficulty of estimating these frequencies well, especially from a corpus of this size, we felt that a more complex estimator would introduce more error than it was worth.

Figure 2 shows the frequency distribution of the 1533 individual referents in this corpus across all recordings. As with many other frequency distributions in language, referential

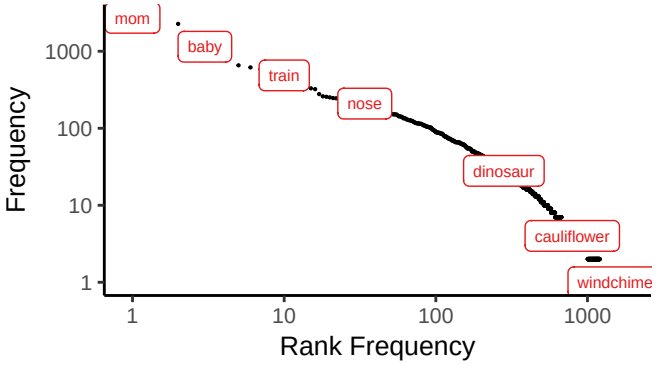

Figure 2: Referents varied widely in their frequency of use, appearing approximately Zipfian. We predict that referents frequent in the input–like baby–should be be more likely to emerge in speech than infrequent referents like cauliflower

frequencies were approximately Zipfian, appearing approximately linear on a log-log scale (Piantadosi, 2014). These frequency estimates were used to test the first key prediction of the race model of communication: Labels for referents that are easier to retrieve from memory are more likely to be produced in speech.

Predictions 1 and 2: The effects of frequency

If the modality that children use for referential communication is the result of a race between speech and gesture, factors that facilitate lexical retrieval and word production should make speech win the race more often. As more frequent words lead to faster retrieval in adults, we hypothesized that frequency should have the same effect for young children. Consequently, when children want to refer to things that are talked about more often, they should be more likely to use speech (Prediction 1). Further, since exposure to language increases over development, older children should be relatively more likely than younger children to use speech for the same referent. (Prediction 2).

Figure 3 shows how the probability of speech and gesture changed with referents' frequency and over development. We performed all statistical analyses on continuous frequency data, but to facilitate visualization here divide referents into four quartiles from most frequent (1) to least frequent (4). Children were relatively more likely to use speech for more frequent referents, and more likely to use speech over development. These data appear consistent with both of the first two predictions of our race model for communication.

To test these predictions statistically, we used as our dependent variable modality of production for each individual referential event by every child at all six ages. This binary outcome–speech or gesture–was predicted with a mixed effects logistic regression with fixed effects of frequency, age, and their interaction, and a random slope of frequency for each child. As the effect of frequency on memory and processing tends to be linear in log scale, frequency was log-

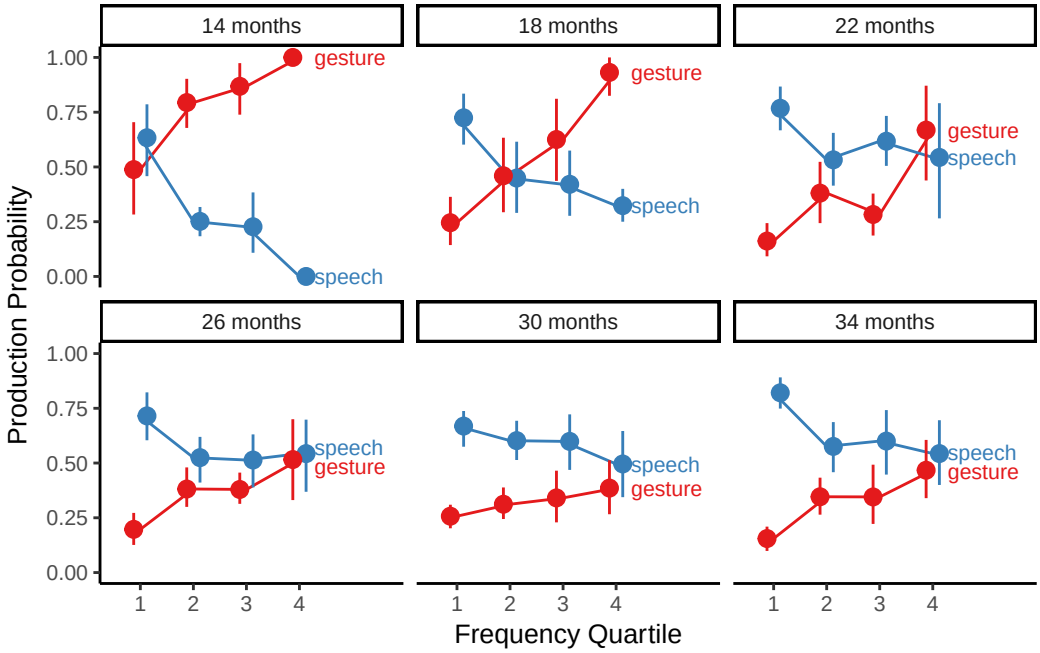

Figure 3: Probability of referential events being expressed in speech (blue) vs. gesture (red) as a function of frequency and children's age. For ease of visualization, referents were divided into four quartiles (1-most frequent). Points show group averaed proportions, error bars show 95% confidence intervals computed by non-parametric bootrstrap

term	estimate	Z-value	p-value
Intercept	-.59 (0.16)	-3.67	<.001
log frequency	.27 (0.01)	22.26	<.001
age	.72 (0.09)	8.00	<.001
log frequency * age	-.09 (0.01)	-7.60	<.001

Table 2: Coefficient estimates for a mixed-effects logistic regression predicting probability of prodution in speech for a referential event. The model was specified as `speech ~ log(freq) * scale(age) + (scale(age)|subj)`

transformed. In addition, age was scaled to improve model estimation. Both main effects were highly reliable predictors, as was the interaction between them (Table 2). Children were significantly more likely to use speech to refer to more frequent referents, more likely to use speech as they got older, and the effect of frequency decreased over development–presumably because the easiest to retrieve referents already win the race even for younger children. Because these analyses were performed on all references for all children, some referents were produced only one or a few times, and thus only in a single modality. When this modality was gesture, we cannot know whether children knew the spoken labels for these referents, and thus whether there was a race at all. To ensure that our results were not driven by words that children did not know, we subset the data down to only referents that children produced in *both* speech and gesture in a single session. All predictors remained significant ($p < .001$) in the same direction, and numerically similar except for age, which increased. Even by this more conservative analysis, both predictions of the race model were confirmed: Children's are more likely to use speech for more frequent referents, and more likely to do so as they get older. Even for known words, the speed of lexical retrieval and production predict whether speech will win the race against gesture.

Prediction 3: Recent referents get a boost

If children's referential communications are produced by a system that is senstiive to the time-pressures on communication, speech should emerge more often as labels become easier to retrieve and produce. The previous analyses confirm this relationship for one predictor of ease of retrieval: Lexical frequency. However, these references do not occur in a vaccum: they are embedded in broader communicative discourses. A key feature of these discourses is that referential events come in bursts: If a something is referred to in one utterance, it is likely to be referred to again in the next utterance (Altmann, Pierrehumbert, & Motter, 2009).

These topical bursts likely occur for functional reasons–once something interesting has entered the discourse, there is no reason to drop it right away. But they also have an important consequence for production. Although low-frequency words are harder to retrieve the first time, subsequent retrievals in the same discourse become easier–these words get a recency boost (Pickering & Ferreira, 2008). If the drift rate for speech is a function of ease of retrieval, then it should be affected by these burts as well. Consequently, we predict that

children should be relatively more likely to speech to refer to low-frequency referents within a discourse burst than if their reference is the first introduction of the low-frequency referent into the discourse.

In order to test this prediction, we needed to operationalize the boundaries between these discourse bursts. When a referent appears for the first time in a transcript, it is easy to tag as new to the discourse. When it is immediately referred to again, it is also easy to determine that it is part of the same discourse burst. However, whenever the referent appears again after 5 minutes, it is less obvious whether this is a part of the previous discourse burst or a new one. To resolve this coding issue, we defined a simple bag of referents model for discourse.

Under this model, when a new referential event occurs, its target is drawn randomly from a all possible referents the target of each referential event is an independent draw from the set of all referents with probability proportional to its frequency (Altmann et al., 2009). The recurrence time (τ) between two successive occurrences of the same referent this Poisson sampling process is described by an Exponential distribution: $\lambda e^{-\lambda\tau}$, where (λ) is the proportion of all referential events for which this referent is the target. The expected recurrence time for a referent is just the reciprocal of its proportional frequency: If DOG occurrs 50 times in a discourse in which there are a total of 1000 referential events, it should on average occur every $\tau = 20$ events.

The bag of referents model then serves as a null model: Discourse bursts are very low probability events, as they consist of a run of short reccurrence times. We thus define the probability of an event being part of previous discourse as the probability of drawing a recurrence time at least that extreme from its bag words distribution. Figure 4 shows a Gleitman plot of a segment of one parent-child interaction (Frank, Tenenbaum, & Fernald, 2013). Each tile represents the occurrence of a reference in each utterance over time. The colors of the tiles show the probabilities assigned by this bag of referents that these occurrences are new discourse bursts.

Because referents that have occurred recently should be easier to retrieve, the race model predicts that referents within a discourse burst should have faster drift rates and consequently come out in speech than if they were retrieved for the first time to begin a discourse burst. We test this prediction by adding the new discourse burst probability from the bag of words model to our previous mixed-effects model predicting the probability that a child's reference will use the speech modality 3. Both frequency and age remained highly significant predictors, as did their interaction. In addition, referents starting a new discourse burst were reliably less likely to be produced in speech, and this effect interacted with both frequency and age: Discourse novelty lead to gesture particularly for infrequent referents and for younger children. The three-way interaction was mot significant and did not improve model fit.

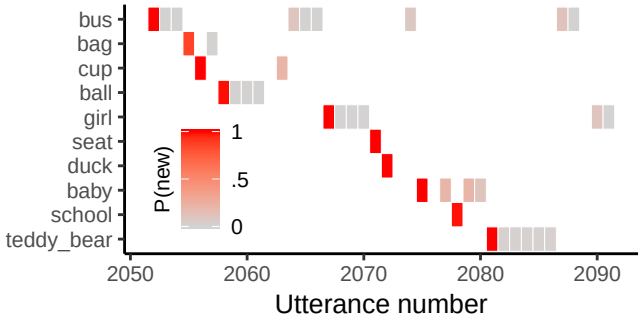

Figure 4: A Gleitman plot of a slice of parent-child interaction. Tiles show which objects are referents of each utterance. Tile color shows predicted probability of being a new discourse topic under the bag of referents model

term	estimate (SE)	Z-value	p-value
Intercept	-.25 (0.17)	-1.47	0.14
log frequency	.20 (0.02)	11.83	$<.001$
age	.67 (0.09)	7.25	$<.001$
new discourse	-.85 (0.14)	-6.01	$<.001$
log freq * age	-.10 (0.01)	-8.15	$<.001$
log freq * new disc	.16 (0.03)	5.49	$<.001$
age * new disc	.23 (0.05)	4.55	$<.001$

Table 3: Coefficient estimates for a mixed-effects logistic regression predicting probability of prodution in speech for a referential event. The model was specified as `speech ~ log(freq) * scale(age) + new_discourse * log(freq) + new_discourse * scale(age) + (scale(age)|subj)`

Discussion

Young children are inundated with language, hearing on the order of 30 million words by the time they are four years old (Hart & Risley, 1995). From the statistical relationships within and among these words, children must discover the latent structures that allow them to become fluent speakers of their native language. Some of these words will be overheard, addressed by one parent to another or a sibling to a friend. However, some will be directed to the child, and these child-directed words maybe be particularly supportive of learning (Weisleder & Fernald, 2013). Child-directed speech is not merely a corpus of well-formed language; It is contingent on the child's own attention, interests, and prior knowledge, and thus can be directed *by the child themself*.

Young children are notorious for asking questions "why?" In her analysis of 5 children from the Brown and Kuczaj corpora in CHILDES, Chouinard (2007) reports children asking over 100 questions per hour they interaction with adults over the 2-5 year range. These questions are powerful because they allow children to learn about two important things simultaneously: The causal relationships in the world around them, and also about the structure of language itself. By driving the discourse into predictable areas of content, they can

reduce referential ambiguity in learning new language for this content.

Long before they can explicitly direct their input with wh-questions, children can sometimes achieve a similar outcome simply by referring to objects in their environment. Having observed a referential event, parents will often follow-in with expansions and additional information about the child's target of interest (H. H. Clark, 2014)

All code for these analyses are available at
`https://github.com/dyurovsky/gesture`

Acknowledgements

This research was funded by a James S. McDonnell Foundation Scholar Award to DY and National Institutes of Health P01HD40605 to SGM.

References

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. *PLoS ONE*, *4*(11), e7678–7.

Bloom, P. (2000). *How children learn the meanings of words*. MIT press: Cambridge, MA.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Bruner, J. (1983). *Child's talk: Learning to use language* (pp. 111–114). Norton.

Casillas, Marisa, & Frank, M. C. (2017). The development of children's ability to track and predict turn structure in conversation. *Journal of Memory and Language*, *92*(C), 234–253.

Chouinard, M. M. (2007). Children's questions: A mechanism for cognitive development. *Monographs of the Society for Research in Child Development*, *72*(1), 1–112.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Clark, H. H. (2014). How to talk with children. In I. Arnon, M. Casillas, C. Kurumada, & B. Estigarribia (Eds.), *Language in interaction: Studies in honor of eve v. clark* (pp. 333–352). John Benjamins.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to Discourse. *Cognitive Science*, *13*(2), 259–294.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1–24.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588–599.

Grice, H. P. (1969). Utterer's meaning and intention. *The Philosophical Review*, *78*, 147–177.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Reference and attitude in infant pointing. *Journal of Child Language*, *34*, 1–20.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

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

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn taking for conversation. *Language*, *50*, 696–735.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., ... Levinson, S. C. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, *106*(26), 10587–10592.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, *71*(2), 73–108.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, *10*, 401–413.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, *78*(3), 705–722.

Vitevitch, M. S. (2008). What Can Graph Theory Tell Us About Word Learning and Lexical Retrieval? *Journal of Speech, Language, and Hearing Research*, *51*(2), 408–422.

Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate? *Developmental Science*, *17*(6), 872–879.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*, 12933–12937.

Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152.

Wingfield, A. (1968). Effects of frequency on identification and naming of objects. *The American Journal of Psychology*, *81*(2), 226–234.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.