UNDERSTANDING PATTERNS IN INFANT-DIRECTED SPEECH IN CONTEXT: AN INVESTIGATION OF STATISTICAL CUES TO WORD BOUNDARIES

by

ROSE M. HARTMAN

A DISSERTATION

Presented to the Department of Psychology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2016

DISSERTATION APPROVAL PAGE

Student: Rose M. Hartman

Title: Understanding Patterns in Infant-Directed Speech in Context: An Investigation of Statistical Cues to Word Boundaries

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Psychology by:

Dare Baldwin	Chair
Caitlin M. Fausey	Core Member
Ulrich Mayr	Core Member
Vsevolod M. Kapatsinski	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2016

© 2016 Rose M. Hartman
This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivs (United States) License.

DISSERTATION ABSTRACT

Rose M. Hartman

Doctor of Philosophy

Department of Psychology

December 2016

Title: Understanding Patterns in Infant-Directed Speech in Context: An Investigation of Statistical Cues to Word Boundaries

People talk about coherent episodes of their experience, leading to strong dependencies between words and the contexts in which they appear. Consequently, language within a context is more repetitive and more coherent than language sampled from across contexts. In this dissertation, I investigated how patterns in infant-directed speech differ under context-sensitive compared to context-independent analysis. In particular, I tested the hypothesis that cues to word boundaries may be clearer within contexts.

Analyzing a large corpus of transcribed infant-directed speech, I implemented three different approaches to defining context: a top-down approach using the occurrence of key words from pre-determined context lists, a bottom-up approach using topic modeling, and a subjective coding approach where contexts were determined by open-ended, subjective judgments of coders reading sections of the transcripts. I found substantial agreement among the context codes from the three different approaches, but also important differences in the proportion of the corpus that was identified by context, the distribution of the contexts identified, and some characteristics of the utterances selected by each approach. I discuss

implications for the use and interpretation of contexts defined in each of these three ways, and the value of a multiple-method approach in the exploration of context.

To test the strength of statistical cues to word boundaries in context-specific sub-corpora relative to a context-independent analysis of cues to word boundaries, I used a resampling procedure to compare the segmentability of context sub-corpora defined by each of the three approaches to a distribution of random sub-corpora, matched for size for each context sub-corpus. Although my analyses confirmed that context-specific sub-corpora are indeed more repetitive, the data did not support the hypothesis that speech within contexts provides richer information about the statistical dependencies among phonemes than is available when analyzing the same statistical dependencies without respect to context. Alternative hypotheses and future directions to further elucidate this phenomenon are discussed.

CURRICULUM VITAE

NAME OF AUTHOR: Rose M. Hartman

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR
University of Wisconsin, Madison, WI

DEGREES AWARDED:

Doctor of Philosophy, Psychology, 2016, University of Oregon
Master of Science, Psychology, 2013, University of Oregon
Bachelor of Science, Linguistics, 2008, University of Wisconsin - Madison

AREAS OF SPECIAL INTEREST:

Developmental Psychology
Quantitative Methods

GRANTS, AWARDS AND HONORS:

Graduate Teaching Fellowship, Psychology, 2016
Graduate Education Committee research award, 2016
Graduate Teaching Fellowship, Center for Assessment, Statistics, and Evaluation, 2015 to 2016
Centurion Award, 2015
ICPSR Scholarship for Developmental, Child, and Family Psychology, 2014
Gregores Research Award, 2013 and 2014
Clarence and Lucille Dunbar Scholarship, 2013
Gary E. Smith Summer Professional Development Award, 2013
Graduate Education Committee travel award, 2012
Carolyn M. Stokes Memorial Scholarship, 2012 and 2014
Graduate Teaching Fellowship, Psychology, 2010 to 2015
Academic Excellence Scholarship in Letters and Science, 2006

PUBLICATIONS:

Maier, R. & Baldwin, D. (2016). Exploring Some Edges: Chunk-and-Pass Processing at the very Beginning, across Representations, and on to Action. *Behavioral and Brain Sciences*.

Baldwin, D. & Maier, R. (2014). Natural Pedagogy in Communicative Development. In Brooks, P. J., & Kempe, V. (Eds.). *Encyclopedia of language development*. Thousand Oaks, CA: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781483346441

Vendlinski M. K., Javaras K. N., Van Hulle C. A., Lemery-Chalfant K., Maier R., Davidson, R. J., & Goldsmith, H. H. (2014) Relative Influence of Genetics and Shared Environment on Child Mental Health Symptoms Depends on Comorbidity. *PLoS ONE*, 9(7): e103080. doi:10.1371/journal.pone.0103080

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The speech infants hear moment to moment is not composed of utterances randomly sampled from the language — their linguistic environment is shaped by, among other things, the contexts in which it occurs. For example, utterances containing words like "breakfast" or "yum" may be more likely to be heard by infants when in the kitchen than in other parts of the house; "cow", "jump", and "moon" may be particularly likely in the early afternoon, when a caregiver reads a favorite book of nursery rhymes just before nap time. Infants' daily activities are often highly routinized, and this includes caregivers' speech (Bruner, 1975). As Bruner emphasized, regularities in speech occur within a larger landscape of regularities in infants' experience, and this concert of related probabilistic cues may help in important and non-obvious ways to make language, as complex as it is, readily learnable.

Recent work underscores the importance of considering infants' linguistic input in this inclusive way. It is clear that infants can — and do — process the speech they hear together with a host of co-occurring cues (e.g., Baldwin, 1991; Baldwin & Meyer, 2007; Gogate & Maganti, 2016; Gogate, Prince, & Matatyaho, 2009; Goldstein et al., 2010; Horst, 2013; Smith, Suanda, & Yu, 2014; Smith & Yu, 2008). For example, a variety of context cues have been shown to guide infants' and toddlers' word learning, including social cues and routines (Baldwin, 1991; Campbell & Namy, 2003), location (Benitez & Smith, 2012; Roy, Frank, DeCamp, Miller, & Roy, 2015), visual background (Vlach & Sandhofer, 2011), time of day (Roy et al., 2015), and linguistic context (Goldberg, Casenhiser, & Sethuraman, 2004; Horst, Parsons, & Bryan, 2011; Roy et al., 2015). In many cases, capitalizing

on regularities across a variety of dimensions can make an otherwise daunting learning task tractable (e.g., Bahrick & Lickliter, 2000; Gogate et al., 2009). Computational modeling results support the findings from behavioral studies, demonstrating that simultaneously processing several streams of probabilistic cues can make the patterns in each of them clearer (e.g., Andrews, Vigliocco, & Vinson, 2009; Christiansen, Allen, & Seidenberg, 1998; S. Frank, Keller, & Goldwater, 2013; Räsänen & Rasilo, 2015). Examining the speech stream in isolation may misrepresent infants' input and potentially exaggerate the challenges posed to learners in coming to understand it.

The purpose of the present investigation is to apply this framework to one particular language acquisition problem — word segmentation — to explore how context shapes patterns relevant to discovering important linguistic structure in the speech infants hear.

Word Segmentation

The successful use of language is a multifaceted skill, and language acquisition poses not one problem to the learner, but many. For the purposes of this study, however, I will focus on learning to identify words within fluent speech as a useful linguistic microcosm within which to explore relevant acquisition issues. Fluent speech lacks obvious, overt acoustic cues (such as pauses) that reliably mark word boundaries, in contrast to the way that spaces demarcate words in writing (Cole & Jakimik, 1979). As a result, the identification of words within fluent speech is an impressive feat. And yet, infants can and do recognize individual words from fluent speech, and this ability is apparent in behavioral studies by six or seven months of age (Bergmann & Cristia, 2015; Jusczyk & Aslin, 1995). Investigation

into the cues infants use to achieve identification of word boundaries in speech continues to be an active area of research.

A large literature demonstrates that infants can use a variety of cues to identify word boundaries in speech without needing to rely on pauses (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Christiansen, Onnis, & Hockema, 2009; Jusczyk, 1999; Jusczyk, Houston, & Newsome, 1999; Soderstrom, Nelson, & Jusczyk, 2005). Many of the cues that predict word boundaries are language-specific, however, and must therefore be learned from the input. This creates a chicken-and-egg problem, where learners cannot know which cues are predictive of word boundaries — or in what way — until they have already successfully segmented a sufficiently large inventory of words (Jusczyk, 1999; Thiessen & Saffran, 2003). For example, in English, stress patterns are an excellent cue to word boundaries: Most multi-syllabic words are stressed on the first syllable (e.g. PA-per, TALK-ing, BOUND-ar-y), so positing a word boundary before stressed syllables generally yields correct segmentation. Not all languages show this stress pattern, though, and some (such as French) show the opposite, where first syllables are generally unstressed. So lexical stress could only be a useful cue to word boundaries after sufficient units have been identified to notice the predominant stress pattern in that language, if one exists (Swingley, 2005; Thiessen & Saffran, 2003). There are many such language-specific cues to word boundaries, including prosodic features like lexical stress (E. K. Johnson & Seidl, 2009; Shukla, Nespor, & Mehler, 2007), rules for phonotactics such as which consonant clusters can or cannot begin a word (Christiansen et al., 2009; Onnis, Monaghan, Richmond, & Chater, 2005), and the use of learned words to identify further units when they occur together in the speech stream (Bortfeld et al., 2005).

Although more experienced infants may certainly take advantage of such learned, language-specific cues, other explanations are required to account for initial word segmentation. Sensitivity to distributional properties of speech — patterns in the sequence of speech sounds — provides an attractive mechanism for early word segmentation because it does not rely on the learner possessing an existing inventory of segmented units (Saffran, Aslin, & Newport, 1996; Swingley, 2005; Thiessen & Saffran, 2003, 2007). Under this hypothesis, infants track statistical regularities among speech sounds (often operationalized as syllables or phonemes), treating strings of sounds with high statistical coherence as word-like units. Importantly, tracking co-occurrence patterns among speech sounds need not rely on any prior inventory of segmented units. This is underscored by the fact that after even a brief exposure to a novel language with no cues to word boundaries other than the statistical coherence of strings of syllables presented therein, both adults (Perruchet & Desaulty, 2008; Saffran, Newport, & Aslin, 1996) and infants (Pelucchi, Hay, & Saffran, 2009a, 2009b; Saffran, Aslin, & Newport, 1996) show sensitivity to the "words" presented. In this way, infants could segment fluent speech with no language-specific knowledge in place other than the ability to identify and track the most basic units in the speech stream.

In many studies, the basic units infants track are assumed to be syllables, but demonstrations with phonemes, either as individuals or processed in short sequences (diphones or triphones), exist as well (Baayen, Shaoul, Willits, & Ramscar, 2015; Christiansen et al., 2009; Daland & Pierrehumbert, 2011). While there is substantial evidence to suggest that even very young infants are able to track phonemes and syllables as units (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Eimas, 1999;

Jusczyk & Derrah, 1987; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995; Phillips & Pearl, 2015), it is important to note that a statistical learning account of word segmentation can be applied to co-occurrence patterns among units at lower levels as well. For example, in an analysis of carefully-controlled recordings of fluent speech, Räsänen (2011) found that a computational model using statistical learning in the form of transitional probabilities calculated over atomic acoustic events in the raw signal yielded reasonably accurate word segmentation without ever needing to refer to phoneme or syllable units.

Statistical Learning and Word Segmentation. Statistical learning is an attractive learning mechanism because of its simplicity and power: Considerable evidence suggests that it is a general feature of our cognition that is available across domains (Baldwin, Andersson, Saffran, & Meyer, 2008; Bulf, Johnson, & Valenza, 2011; Kirkham, Slemmer, & Johnson, 2002; Kirkham, Slemmer, Richardson, & Johnson, 2007; Räsänen, 2014; Romberg & Saffran, 2013), across the lifespan (Janacsek, Fiser, & Nemeth, 2012; Weinert, 2009) and even across species (Hauser, Newport, & Aslin, 2001). The level of scientific investment in the statistical learning proposal for word segmentation is underscored by the fact that Google Scholar currently lists approximately 3,000 papers citing the seminal Saffran, Aslin, and Newport (1996) paper originally demonstrating statistical learning in 8-month-old infants [1]. The original effect has been widely replicated in a variety of applications (for a review, see Romberg & Saffran, 2010). Because of the robustness of the evidence for statistical learning ability and the fact that it does not need to appeal to acquired language-specific knowledge, it is a popular

[1]With a little over 50 of those citations occurring in the first two years after publication, this article is an outlier even among *Science* papers, which currently lists a 2-year impact factor of 34.

and compelling explanation for how infants begin to segment fluent speech into word-like units.

Although it has been established that infants can use statistical regularities among syllables to identify word boundaries in the lab, it is as yet unclear the extent to which these findings accurately characterize language acquisition as it occurs naturally. The tightly controlled experimental designs typical of the statistical learning literature (e.g., Pelucchi et al., 2009b; Saffran, Aslin, & Newport, 1996) differ from infants' actual linguistic experience in a number of important ways. First, the language used in laboratory stimuli is often designed so the statistical regularities among syllables are the only cues to word boundaries; this design enables researchers to isolate that mechanism (Estes & Lew-Williams, 2015; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996), but it has been shown in more complex designs that infants' use of those regularities depends in part on the availability of other cues. Pauses (Lew-Williams, Pelucchi, & Saffran, 2011), stress patterns (E. K. Johnson & Seidl, 2009; Thiessen & Saffran, 2003), other prosodic markers (Gout, Christophe, & Morgan, 2004; Shukla et al., 2007), the presence of familiar words (Bortfeld et al., 2005), and cues from other modalities (Seidl, Tincoff, Baker, & Cristia, 2014; Thiessen, 2010) can all interact with infants' use of statistical learning to identify words within fluent speech. Moreover, natural linguistic input occurs in a complex multimodal environment with many demands on attention, multiple speakers, a mix of infant-directed speech (IDS) and overheard speech, Zipfian (rather than uniform) distribution of words and utterances presented in a non-random order, and the like.

The success of statistical learning as a mechanism for word segmentation hinges on the availability of statistical cues to word boundaries in the speech

infants actually hear. A primary goal of research on statistical learning accounts of word segmentation, then, is an accurate description of the patterns available to infants in their natural language exposure. The present research aims to add to the existing body of work attempting to accurately characterize the nature of the linguistic patterns infants encounter.

Modelling Word Segmentation from Statistical Cues. Many studies have applied computational models of word segmentation to corpora of natural speech, articulating a range of assumptions about how language is processed, learned, and stored. The comparison of models making different assumptions or relying on different cues in the input has proven to be a valuable way to test the plausibility of different theories of word segmentation on natural linguistic input (Brent, 1999b; Monaghan & Christiansen, 2010). Computational models also have the ability to reveal something about the structure in the input itself, however — a model that successfully segments natural speech using a particular cue provides evidence of the availability and potential potency of that cue in the input. Models focused on cues such as repeated words (Brent, 1999a), phonotactics at utterance boundaries (Monaghan & Christiansen, 2010), and statistical co-occurrence of speech sounds (Goldwater, Griffiths, & Johnson, 2009), for example, have all proven successful to various degrees at segmenting corpora of infant-directed speech. From this perspective, Bayesian word segmentation models provide a particularly attractive option because they are "ideal" learners; they optimally represent the patterns in the input according to whatever structure they use. While this may be problematic in attempts to model actual human performance in segmentation tasks (see discussion in M. C. Frank, Goldwater, Griffiths, & Tenenbaum, 2010), it is well suited to summarizing the strength of

statistical cues in the input. Bayesian models that use the statistical co-occurrence of speech sounds to identify word-like units in speech, therefore, can provide an estimate of the strength of the statistical signal to word boundaries in a given corpus.

Two Bayesian word segmentation models have received considerable attention in the literature: the hierarchical Dirichlet process (HDP) model (Goldwater et al., 2009) and the collocation-syllable adaptor grammar (M. Johnson, 2008). Both rely on the statistical co-occurrence of speech sounds (phonemes) to posit word-like units in fluent speech. The two models are similar in many respects, but there are a few interesting differences as well. For example, the HDP model makes no assumptions about the structure of words, whereas the adaptor grammar has built-in constraints such that words must be composed of syllables, which are in turn composed of an optional onset, a vocalic nucleus, and an optional coda — constraints that are helpful in languages like English where words are always composed of syllables, but which may not apply to languages where that is not the case. The differences between the HDP model and the adaptor grammar affect segmentation performance and have potential implications for cross-linguistic generalizability. Given the differing advantages of each type of model, both models were incorporated in the present research toward the goal of achieving an understanding of patterns in the input that are potentially available to infant language learners.

What is the Input?. One limitation of many of the existing studies regarding statistical learning and language acquisition is that they focus on the speech stream as the only source of information to learners. Infants, on the other hand, experience language embedded in a set of rich, multimodal cues, as

mentioned earlier. Moreover, infants are known to be able to take account of such cues for many language-learning purposes (Baldwin, 1991; Benitez & Smith, 2012; Bruner, 1975; Gogate, Bahrick, & Watson, 2000; Gogate et al., 2009; Horst et al., 2011; Samuelson, Smith, Perry, & Spencer, 2011; Seidl et al., 2014; Vlach & Sandhofer, 2011). There are meaningful regularities in experience beyond language, and patterns in the experienced environment correlate with patterns in language (Andrews et al., 2009). Moreover, while cues in the distributional properties of speech and experience are often redundant, they also have the potential to be informative in different ways making the joint processing of linguistic and environmental patterns especially effective (Riordan & Jones, 2011). In other words, cues in the environment may make cues in the linguistic input more informative, and vice versa.

In the case of word segmentation, there is already some evidence to suggest that environmental cues may facilitate infants' use of statistical cues within the speech stream. In a recent behavioral study, Seidl et al. (2014) demonstrated that the co-occurrence of touch with statistical units can boost 4-month-old infants' ability to segment fluent speech in an artificial language statistical learning task. Exploring the role of environmental cues in word segmentation more generally, Synnaeve, Dautriche, Börschinger, Johnson, and Dupoux (2014) found that Bayesian computational models (implementations of Johnson's adaptor grammar) yielded more accurate segmentation when the models could segment speech differently for different activity contexts. They operationalized activity context using topics from topic modeling (Latent Dirichlet Analysis, LDA; Blei, Ng, & Jordan, 2003), resulting in a much more diffuse sense of environmental cues than the very concrete, local cues used by Seidl and colleagues in their behavioral task.

Synnaeve and colleagues' modeling results suggest that statistical cues relevant to word segmentation vary with environmental cues embodied in activity context (otherwise there could be no change in the models' performance as a result of processing contexts separately).

Examining the speech stream while ignoring broader experience in which it is embedded — which I will collectively call "context" — may lead to underestimation of the available signal for language learners. Therefore, recharacterizing infants' language exposure as including context may help bring theoretical models of word segmentation closer to the reality of infants' experience with language.

Language and Context

Speech varies by context. The words we use depend on, for example, the topic of conversation, such that in any given sample of speech, the words that do occur in that sample are likely to be over-represented relative to their global frequencies and all other words are necessarily under-represented (Altmann, Pierrehumbert, & Motter, 2009; Church & Gale, 1995; Ramscar & Port, 2016). To illustrate, in a language sample that contains a word like 'frequency' (such as this document), it is likely that that word will occur at a much higher frequency in that sample than would be expected given its overall frequency in the language. This is true not only for content words that relate transparently to the topic of a language sample, but also for less obvious words like 'said' or 'well' (Church & Gale, 1995). This applies to child-directed speech (CDS) as well, with some words used preferentially in particular contexts (Roy et al., 2015; Roy, Vosoughi, & Roy, 2014). Although the observation that word use varies by context may seem so

obvious as to be uninteresting, it has important implications for the structure of linguistic patterns in speech (i.e., the potential input for the developing system).

Indeed, recent investigations demonstrate that the co-occurrence of words with particular contexts may facilitate word learning. In their analysis of a very dense longitudinal corpus of one child's linguistic experience from 9 to 24 months, Roy et al. (2015) found that words occurring preferentially in specific contexts were produced earlier by the child than words that occurred across contexts. In addition, contextual distinctiveness predicted age of first production even after accounting for frequency. In fact, higher frequency was associated with earlier production only for nouns, while contextual distinctiveness predicted earlier production for all word classes. Interestingly, Roy et al's finding contradicts results of an analysis of many pairs of infants and their caregivers from the *Child Language Data Exchange System* (CHILDES, MacWhinney, 2000), which demonstrated that greater contextual diversity (rather than less) predicts earlier normative age of acquisition of early nouns, and contextual diversity is a better predictor than frequency (Hills, Maouene, Riordan, & Smith, 2010). [2] There were several differences between the Roy et al. (2015) analyses and Hills et al. (2010) that could have contributed to this discrepancy, including differences in the age examined (9 to 24 months in Roy et al.'s analyses, and 12 to 60 months in Hill et al.'s), the outcome used (age of first production for Roy et al. and normative age of acquisition for Hills et al.), and the operationalization of linguistic contextual diversity (LDA topics in Roy et al. and co-occurrence within a moving window for Hills et al.). One additional interesting possibility is suggested by the fact that the Roy et al. corpus comprised all-day

[2]Jones, Johns, and Recchia (2012) also found contextual distinctiveness to be positively related to earlier word learning in a corpus analysis of non-CDS, an empirical word learning study with an artificial language, and results from several modeling demonstrations.

recordings, capturing the natural range of activities making up that child's days, whereas the CHILDES corpora used by Hills et al. were mostly shorter recordings done at a time of day that was convenient for both caregivers and researchers, and when the child was likely to be awake and cooperative. Many of the CHILDES recordings also focus on play time in particular, under-sampling other activities like bathing, dressing, etc. These differences in the corpora being analyzed mean that the contexts identified by LDA in Roy and colleagues' Speechome corpus versus linguistic context in the pooled CHILDES corpora may not be comparable.

In addition to studies on context specificity and children's production of new words, there is also evidence that reliable contextual cues can facilitate word learning assessed through comprehension measures. In a series of behavioral experiments, Benitez and Smith (2012) found that 16- to 18-month-old infants in a word learning task showed better memory for novel word-object pairs at test when the novel objects were always named in consistent locations compared to infants who received the same training but with varying locations for each object. There are, of course, multiple possible explanations for why context-specific words may be easier to learn. One possibility is that infants capitalize on contextual overlap across different occurrences of a word to determine the word's referent in what might otherwise be ambiguous labeling events (Smith & Yu, 2008). For example, a toddler who usually hears 'ball' while kicking a ball in the hallway may have an easier time associating that word with its referent than he would for a word like 'with' which occurs across a wider range of situations. There is also evidence to suggest that spatial regularities help infants to efficiently direct their attention during naming events, and therefore allow them more opportunity to encode referents along with their labels (Benitez & Smith, 2012; Samuelson et al., 2011).

Although this converging evidence highlights the importance of contextual cues in word learning, less is known about the role of context in earlier stages of language acquisition such as word segmentation. To date, behavioral studies of the role of complex input factors in word segmentation have focused on how patterns *within the speech stream* interact in infants' identification of word boundaries (Altvater-Mackensen & Mani, 2013; Bortfeld et al., 2005; E. K. Johnson & Jusczyk, 2001; E. K. Johnson & Seidl, 2009; E. K. Johnson & Tyler, 2010; Lew-Williams et al., 2011; Lew-Williams & Saffran, 2012; Thiessen, Hill, & Saffran, 2005; Thiessen & Saffran, 2003), rather than how contextual cues may shape infants' word segmentation. Nevertheless, the substantial evidence for context effects in word learning underscores the plausibility of the hypothesis that context may shape patterns relevant to other aspects of language acquisition as well. In particular, context-specific processing of language may reveal clearer cues to word boundaries — the more homogeneous, coherent speech within contexts may provide richer information about the statistical dependencies among speech sounds than is available when analyzing the same statistical dependencies without respect to context. If so, this would suggest that the signal available to infants to identify units in fluent speech may be stronger than has been suggested (Yang, 2004).

What is Context?. This thesis is specifically directed toward examining the role of context in the patterns available in the speech infants hear. What I mean in general terms by *context* is a pattern of environmental cues that repeats over time. For example, 'bath time' may be a context characterized by cues including location (in the tub), time of day (evening), physical sensation (being wet), sounds (running water, splashes), the presence of a particular caregiver (e.g., mother), etc., and these cues are likely to occur together each time bath time

occurs. Relevant cues might be on any number of dimensions, and likely depend not only on the probability of those cues occurring together but on their collective conditional probabilities; cues that are likely to occur together and unlikely to occur anywhere else would be especially informative. For example, in the case of 'bath time' as described above, location (being in the tub) may be a particularly strong cue to that context because most infants are unlikely to be in the tub except during bath time. The presence of a particular caregiver, on the other hand, would be much less informative since infants interact with their caregivers in a variety of contexts that do not share other cues with bath time.

Language itself can also serve as a contextual cue. Even before knowing the meanings of words infants may associate sequences of speech sounds with the contexts in which they occur (e.g., Smith & Yu, 2008). Since many words occur preferentially in some contexts over others (Altmann et al., 2009; Church & Gale, 1995; Roy et al., 2015), these associations may become reliable cues to context. For example, an utterance like 'splish splash splosh!' may be part of a playful routine that often occurs during bath time and is unlikely to occur during any other activity; hearing that sequence of speech sounds (even without understanding its meaning) can become a cue that helps to define bath time.

A pattern of contextual cues may repeat in infants' experience for a variety of reasons. Many of the cues that make up a particular context may occur with that context simply because of recurring situational or logistic factors. For example, being in the tub will reliably be part of bath time because it is physically necessary to complete that goal. Similarly, time of day may be a good predictor of family dinner because of the regular constraints of family members' schedules. Beyond such practical concerns, contexts may be further standardized by caregivers as part

of their attempts to coordinate with their infants. In a pilot study of six infants with their mothers followed longitudinally over approximately six months, Bruner (1975) found that mothers standardized many of their joint actions with their infants during typical daily contexts (meal time, bath time, and play time), making recurrences of a given activity more similar to each other and more predictable. He posited that standardizing joint action in this way can make interpretation of intentions easier for both mothers and infants, provide easier ways for infants to successfully (non-linguistically) express intention, and make it easier for infants to calibrate their attention with their mothers'. More particularly relevant for the current study, highly standardized contexts may make it easier for infants to identify the contexts themselves as they repeat across time (Qian, Jaeger, & Aslin, 2012).

Of course, context can be defined on multiple timescales, from a matter of seconds (e.g. linguistic frames; Arnon & Clark, 2011; Mintz, 2003) to contexts that persist over months or years (e.g. socioeconomic status; Hart & Risley, 1995; Weisleder & Fernald, 2013). Depending on the learning phenomenon under investigation, coarser or finer temporal patterns of context may be more relevant. For the purposes of the current study, I will focus on contexts that persist over several minutes to as long as half an hour or so, corresponding to the rough timescale for typical infant daily activities examined in the language acquisition literature (e.g. Bruner, 1975; Fausey, Jayaraman, & Smith, 2015; Hoff-Ginsberg, 1991; Roy et al., 2015; Soderstrom & Wittebolle, 2013). Contexts at this scale are of particular interest because they have been shown to be associated with variation in important characteristics of infant-directed speech. For example, Hoff-Ginsberg (1991) recorded interactions of mothers with their toddlers in four different context

settings (meal time, toy play, dressing, and book reading), and found significant differences in all measures of maternal speech calculated (rate, lexical diversity, grammatical complexity, etc.). Soderstrom and Wittebolle (2013) also found differences in the amount of caregivers' and toddlers' speech depending on activity context (10 possible activities, coded in 5-minute bins) in natural recordings taken either at home or in daycares.

Context in the Language Acquisition Literature. Context (at the scale of typical daily activities) has been defined using a variety of different methods. Many of these methods rely on researchers' knowledge and intuition about what activities commonly make up infants' days in order to develop top-down context categories, which are then used for parent report (e.g., Fausey et al., 2015) or researcher coding of a corpus (e.g., Soderstrom & Wittebolle, 2013). Another approach is to use an automatic, data-driven procedure for defining context based on the words that occur, such as topic modeling (Latent Dirichlet Allocation, LDA; Blei et al., 2003). This approach capitalizes on the fact that many words occur preferentially in specific contexts, and uses that to infer the 'topic' (in this case, activity context) from the distribution of word frequencies. Importantly, the goal in this approach is generally not to model the linguistic context *per se*, but rather it is often (explicitly or implicitly) assumed that LDA topics indirectly measure activity (e.g., "We use topics from a Latent Dirichlet Allocation model as a proxy for 'activities' contexts... We do not posit that the infants learn the topic models on linguistic cues while bootstrapping speech and segmentation, but rather that they get activity context from non-linguistic cues." Synnaeve et al., 2014, p.2326-2327). A third approach to defining context is to rely on naive subjective

judgments, allowing parents (Place & Hoff, 2011) or coders (Roy, Frank, & Roy, 2012) to describe activity context in an open-ended way.

Unfortunately, there have been very few attempts to compare multiple approaches to defining context in the same corpus, so little is known about the extent to which different approaches to defining context capture the same latent construct that is of interest in all of these studies: activity context. Two recent analyses of the Speechome corpus are notable exceptions — Roy et al. (2012) found reasonable agreement between topic modeling contexts and activity contexts as judged by human coders, and Roy et al. (2015) reported that topic modeling contexts were also correlated with both time of day and location of the child within the house.

The Present Study

This dissertation addresses the main question: Are statistical cues to word boundaries in speech available to infants clearer within activity contexts? In other words, is infant-directed speech more easily segmentable when processed context by context rather than all together?

Although statistical learning provides a promising explanation for how infants may begin to segment the speech they hear into word-like units, it is only plausible to the extent that statistical cues to word boundaries are actually available in infant-directed speech during the period when infants learn to segment speech. One possibility is that, because the speech infants hear is shaped by context (Hoff-Ginsberg, 1991; Roy et al., 2015; Soderstrom & Wittebolle, 2013; Synnaeve et al., 2014), the relevant patterns in the statistical co-occurrence of syllables also vary by context. If so, collapsing across contexts and analyzing statistical cues to word boundaries without respect to context may underestimate

the information available to infants. This would suggest that descriptions of statistical cues to word boundaries *within context* may reveal a stronger signal.

Of course, the interpretation of any test for how speech patterns vary by context depends on how 'context' is operationalized. As described earlier, context has been operationalized a number of different ways in the existing literature, and it is as yet largely unclear how different methods of defining context relate to each other. To address this, I began my analyses with a methodological comparison of three different approaches to defining context in the same corpus: top-down (using key words from pre-defined context categories), bottom-up (using topic modeling), and subjective coding (using open-ended coder judgments). In the top-down approach, contexts were determined by the occurrence of key words from pre-defined context categories, such that when a key word occurred, that utterance and those immediately around it were included in that context category. In the bottom-up approach, I applied a topic model to the corpus to automatically categorize utterances by context using the topics inferred by the topic model. For the subjective coding, research assistants coded the entire corpus for activity context, providing short, open-ended descriptions of the context for short portions of the corpus presented in random order. In each case, I used the resulting context codes for each utterance to extract context-specific sub-corpora for each approach to defining context.

I then used the context sub-corpora to test the hypothesis that statistical cues to word boundaries are clearer within context, comparing the strength of statistical cues to word boundaries in the context sub-corpora to the same statistical cues in the corpus without respect to context. To measure the strength of statistical cues to word boundaries, I used the two Bayesian computational

models alluded to earlier — Johnson's adaptor grammar (M. Johnson, 2008) and Goldwater's Hierarchical Dirichlet Process model (Goldwater et al., 2009) — to segment the speech in each sub-corpus, using the relative success of the segmentation as an index of the segmentability of the sub-corpus. Context-specific subsets of the corpus may be more easily segmentable than the corpus as a whole because of the relative increase in the frequency of context-specific words, leading to less lexical diversity within contexts. The more homogeneous, coherent speech within contexts may provide richer information about the statistical dependencies among phonemes than would be available when analyzing the same statistical dependencies without access to context.

CHAPTER II

DEFINING CONTEXT

The Corpus

The Korman (1984) corpus comprises dense longitudinal recordings from 6 infants and their middle class mothers at home in the United Kingdom. Mothers were instructed to keep the recording apparatus near the child and on for as much of the day as possible. Experimenters dropped off the recording equipment around noon and picked it up around noon the next day. There are six recordings for each infant, spanning the age range from 6 to 16 weeks (for more details on the participants and recording methods, see Korman, 1984). This corpus is notable for being one of very few publicly available corpora providing recordings of the linguistic environment for infants this young. This is especially important for the present project because behavioral studies demonstrate that typically-developing infants as young as six or seven months are able to successfully recognize individual words presented in fluent speech (Bergmann & Cristia, 2015; Jusczyk & Aslin, 1995). To understand how this skill develops, it is important to understand infants' linguistic input before seven months with respect to cues to word boundaries.

The original corpus (available via CHILDES, MacWhinney, 2000) is transcribed orthographically, but the present study treats the statistical cues to word boundaries in fluent speech. Phonetic transcription, which represents words directly by the speech sounds that make them up, was therefore necessary. This is the English corpus used by Swingley (2005); I used the same phonetic approximations from that study as well (for details about the phonetic approximation, see Swingley, 2005). Using the phonetic approximations of words in the corpus instead of the orthographic transcription allowed me to track

statistical patterns among speech sounds in a way that more plausibly maps onto infants' processing of speech. For example, in the phonetic approximation, the first syllable of 'mummy' and 'money' is transcribed the same way even though the spellings are different, and homographs (e.g. 'read' in present tense and 'read' in past tense) are transcribed differently even though the spellings are the same.

Because parts of my analyses included the family as part of the model (e.g. the estimation of the structural topic model), unlike Swingley, I needed a sufficiently large corpus within each family. One of the six families (child "hi") generated substantially shorter transcripts than the other five, with each of the transcripts fewer than 200 utterances and the shortest just 30 utterances long. The sample of speech for that family was small enough that any attempt to model family-specific variation in topics for those transcripts would be fruitless. Therefore, I have excluded that family from all of the analyses.

Three Approaches to Defining Context

I used three different methods for defining contexts, each designed to reflect a different approach to coding context used in the language acquisition literature. I refer to these approaches as top-down, bottom-up, and subjective coding.

Top-Down: Defining Context by Key Words. The top-down approach began with eight pre-defined contexts based on infant activities commonly coded in the literature (e.g., Bruner, 1975; Fausey et al., 2015; Hoff-Ginsberg, 1991; Place & Hoff, 2011; Roy et al., 2015, 2012; Soderstrom & Wittebolle, 2013): bath time, bed time, body touch (cuddling, tickling, etc.), diaper/dressing, fussing, meal time, media (TV, radio, etc.), and play. For each context, I generated a list of key words from the words on the Oxford CDI (a UK adaptation of the MacArthur CDI, Fenson et al., 1994; Hamilton, Plunkett, & Schafer, 2000). Words that were clearly

semantically related to one of the context categories were assigned to that category — for example, *wash* is on the list for bath time words, and *nappy* is on the list for diaper/dressing words. There are many words on the Oxford CDI that did not relate clearly to any of the context categories (e.g., *door, give, hello, don't*); those words did not appear on any key word list.

The choice of the eight contexts and the inclusion of words on each of those lists was inherently very subjective and reflected my intuition as someone with general expertise in child development and language acquisition. This method was intended to represent a plausible approach for defining context based on researchers' knowledge and the existing literature about infants' activities. The complete key word lists are available in Appendix A.

Utterances in the corpus were then tagged for each context (not mutually exclusively) when those key words occurred, including the two utterances immediately before and after a tagged utterance as well. This procedure yielded eight context-specific sub-corpora.

Bottom-Up: Defining Context by Topic Modeling. The bottom-up method relied on topic modeling (Blei & Lafferty, 2007; Blei et al., 2003) to determine 'topics' in the transcripts.

Topic modeling is a general approach with several different specific implementations (for a review of several popular versions, see Blei, 2012). The simplest and most common version is latent Dirichlet allocation (LDA, Blei et al., 2003). LDA models words in each document as arising from a mixture of underlying latent topics. It outputs a list of the topics (defined by which words are most closely associated with each topic) and loadings for each document that quantify how closely it matches each topic. This procedure is often used

to automatically identify linguistic topics or contexts in child-directed speech (e.g., S. Frank, Feldman, & Goldwater, 2014; Roy et al., 2015; Synnaeve et al., 2014). LDA is limited in that the topics themselves are modeled using a Dirichlet distribution, and are therefore are assumed to be independent of one another. In most applications, this assumption is unlikely to be met; to illustrate this point, Blei and Lafferty (2007) analyze all of the abstracts from *Science* from 1990-1999 and showed that topics of abstracts were naturally correlated, so a paper about genetics was much more likely to also be about disease than X-ray astronomy. As applied to my project, a document involving meal time would be probably likely to also include fussing, and less likely to include bath time. Contexts themselves are correlated in experience, so forcing an independence assumption on the topics discovered by LDA results in a limited ability to model the activity structure in the data. A solution to this issue is to use correlated topic modeling (CTM, Blei & Lafferty, 2007), which uses the same logic to associate documents with a mixture of latent topics, but allows those topics to covary. Because I expected that activity contexts in the corpus would covary naturally, I used CTM rather than LDA to estimate topics.

Using CTM instead of LDA also permits extending the analysis to a structural topic model (STM), which allows covariates to predict the content and prevalence of the topics in the documents (Roberts, Stewart, Tingley, & Airoldi, 2013). With no covariates in the model, STM reduces to CTM. When there are known metadata about the documents, however (such as the identity of the infant-mother dyad in each document), covariates can be added to the model [1]. Content

[1]Note that the STM model uses priors that draw the influence of covariates to zero unless the data strongly suggests otherwise. This reduces the risk of over-fitting the data by including covariates; if there is naturally little variation in the prevalence or content of topics by dyad, the STM model will return results very similar to a plain CTM.

covariates predict which words are most closely associated with each topic. It is important to have flexibility in topic content dyad to dyad because there could be systematic differences in the words each family used during each activity. One obvious example would be the infants' names (and nicknames, such as 'lulu' or 'treasure') — these are dyad specific, so allowing the model to estimate dyad differences in topic content may improve model results for topics where substantial name use occurs. There could easily be other systematic differences in the words different families used during the same activity context, and including dyad identity as a content covariate allowed the model to estimate those differences while still modeling what was the same for each topic across families. Prevalence covariates predict how heavily particular topics will load on each document. For example, if the model discovered a topic that corresponded to 'meal time', it could be the case that some dyads had more documents in the meal context than other dyads (perhaps they ate more often, or had longer meals so each meal spanned multiple documents). Estimating differences dyad to dyad in the prevalence of each topic was especially important for any contexts that were much more likely in some dyads compared to others; for example, 'taking photos' could be a distinct, coherent, frequent activity in some families (they may pose their infants, take lots of pictures in a row, etc.) but not in others. Using dyad identity as a covariate for both content and prevalence can improve the model fit and the quality of the resulting topics by allowing the model to be sensitive to differences family to family in the rhythm of daily activities and what words they use during them.

Topic modeling analysis requires a set of 'documents', and the words in each document are assumed to arise from a mixture of underlying topics. The model infers latent topics based on the sets of words that tend to occur together

in documents. The ideal length of the documents varies based on the desired granularity of topics — if the topics of interest (in this case activity contexts) are expected to change every 10 minutes or so, then the documents analyzed should be approximately that length. For this project, I divided the corpus into documents with a moving window of 30 utterances beginning at the start of each transcript and shifting it forward by 30 utterances until the end of each transcript. This divided the corpus into roughly 450 documents, respecting the natural boundary at the end of each recording. The decision to use 30 utterances was based on exploration of the corpus, which revealed that 30 utterances was typically enough to capture about one activity, although of course this varied widely. Roy et al. (2015) used a 10-minute sliding window, which would be roughly the same size (with substantial variation due to fluctuations in the rate of speech over the day). Given the high percentage of short utterances, 30 utterances also approximates the optimal size of attentional frame for words' context in CHILDES (5-50 words) as revealed by analyzing predictors of the normative age of acquisition for early nouns (Hills et al., 2010).

Another important consideration is the number of topics to search for. Traditionally, the researcher sets a specific number of topics to find and inputs that number as a parameter in the topic modeling algorithm (Blei et al., 2003). Again, the ideal choice depends on the desired qualities of the resulting topics. Searching for fewer topics would discover more global differences between topics (e.g. searching for just two topics might discover something like 'fussing' and 'not fussing' as the two activity contexts), whereas searching for a larger number of topics would discover more subtle differences (potentially distinguishing several different types of 'bath time', several different types of 'meal time', etc.). Of course,

the effect of the number of topics interacts with the content of the corpus being analyzed — searching for 10 topics in a corpus composed completely of free play interactions will capture subtle differences between 10 different kinds of free play, whereas searching for 10 topics in a corpus that includes day-long recordings of infants' natural experience will likely discover only coarse differences in types of play and will instead discover a range of more distinct activities. Similarly, the number of topics appropriate for a given granularity may vary by the child's age. The range of a newborn's daily activities may be captured with just a handful of topics (e.g. eating, fussing, sleeping, diaper/dressing, and alert interaction with a caregiver), whereas a toddler's regular activities may include a large number of routines, favorite games, etc. which could not be captured at the same granularity by only a handful of topics.

In their analysis of the Providence corpus, Synnaeve et al. (2014) used seven topics to capture activity context. Providence is a collection of hour-long recordings of children (1- to 3-year-olds) with their mothers at home, generally during the middle of the day and often capturing free play behavior. Roy et al. (2015) used 25 topics to capture activity context in their dense longitudinal corpus of most of one child's waking experience from 9 to 24 months.[2] For the current project, I expected that roughly 10 topics would adequately capture contexts at the desired granularity. This estimate represented a trade-off between the fact that the Korman corpus is more representative of the range of daily activities (because it is composed of day-long recordings), and the fact that the infants in the Korman corpus are much younger than in other existing work (only 6 to 16 weeks old). I

[2]S. Frank et al. (2014) searched for 50 topics in their analysis of the C1 section of the Brent corpus (recordings of 9- to 15-month-old infants at home with their caregivers), but their goal was to capture more fine-grained differences than the studies reviewed here.

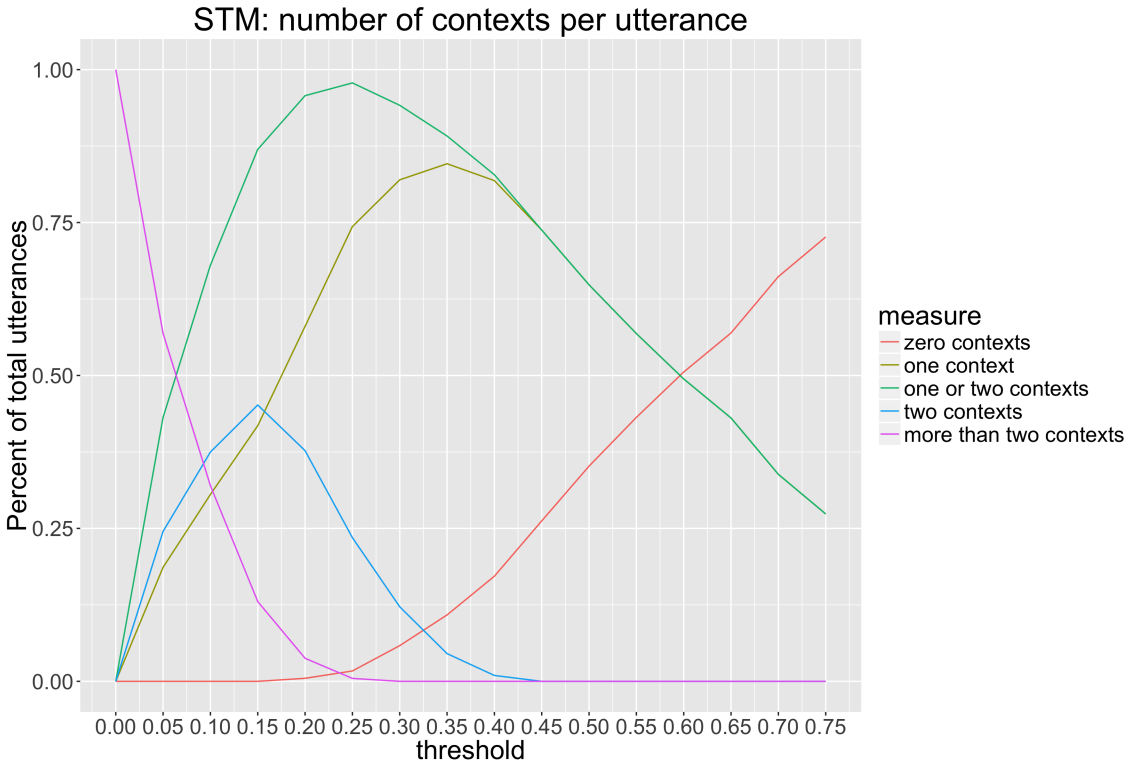
ran the topic model with 8 to 20 topics, examining the results from each for fit to the corpus as well as my judgment of the quality of the resulting topics based on interpretability, exclusivity, and semantic coherence, as is typical in topic modeling analysis (Blei, 2012; Blei & Lafferty, 2007; Blei et al., 2003; Roy et al., 2015). Exclusivity refers to how distinct the topics are from each other, and semantic coherence refers to how similar each topic is across its different occurrences. On these metrics, a solution with 12 topics appeared, based on informal qualitative analysis, to provide an optimal description of the corpus.

The topic modeling results included a description of each topic (which words are most closely associated with each topic) and estimates of how much each topic was represented in each document (the 30-utterance windows used for the analysis). The topic loadings for a document represent the percentage of words attributable to each latent topic. For example, in a model with five topics, a given document might have loadings of .05, .10, .05, .45 and .35, indicating that most of the words in that document could be attributed to the last two topics. Ideally, each document would have near-zero loadings for most topics and higher loadings for only one or two topics, giving a clear sense of what the document is about.

I used the topic loadings for each document to include the utterances in that document in one or more context sub-corpora. In the original paper introducing STM, Roberts et al. (2013) used STM to analyze themes in a set of open-ended survey responses and compared the STM themes to human coder judgments on the same data. Since the coder judgments were categorical decisions rather than continuous loadings, Roberts et al. set a threshold for the topic loadings and considered any document with a loading above that threshold to be judged as "about" that topic by the STM model (note that as long as the threshold was set

below .5, it was possible for a document to be above threshold on more than one topic, in which case it would be included as both). They used a threshold of .2, but the most appropriate threshold depends on the distribution of topic loadings for a particular model and corpus — an ideal threshold should be high enough that most documents are categorized with no more than one or two topics, but not so high that many documents are categorized with no topics at all.

Figure 1. The resulting number of utterances with zero, one, two, or more context codes for a range of thresholds for structural topic modeling topic loadings. A threshold of .25 maximizes the number of utterances with one or two context codes.



For the current study, a threshold of .25 yielded the best balance of coverage of the corpus (few documents with no topic) and exclusivity (few documents with multiple topics), as depicted in Figure 1. I used this threshold to assign each

document (30-utterance window) to one or more topics, generating a context sub-corpus of utterances for each topic.

Subjective Coding: Defining Context by Coder Judgments. The subjective coding method differed from the other two approaches in relying on manual coding of the corpus rather than on an automatic coding procedure. This procedure used an R script that separated the corpus into 30-utterance windows (as with the topic modeling documents), and then printed a randomly selected window to the screen with the instructions for the coder to read the transcript and type in one or more contexts describing what was happening. Coders could enter anything at all — they were not provided with a list of contexts to work from, so that they could provide naive judgments about what was happening in the parent-child interaction apparent within the portion of the transcript they were reading (similar to the video coding procedure used by Roy et al., 2012). Coders were free to enter one context, multiple contexts, or 'none' for no context (used when they were unable to make a guess about what was happening) for each window.

The 30-utterance segments were determined by a sliding window shifting every 10 utterances (e.g., one window would include utterances 1-29, and the next would include utterances 10-39). An individual coder was never presented with the same window twice, but there was no such restriction across coders. Each utterance was coded many times, by different coders and by the same coder in different windows (e.g. utterance 15 might appear to the same coder once in a window including utterances 1-29 and again later in a window from 10-39). Responses from exactly five coders were used for each utterance. For each utterance that was coded by more than five coders, five coders were randomly selected.

The raw codes underwent two stages of cleaning. First raw codes were standardized, correcting spelling differences, standardizing redundant forms (e.g., 'getting dressed', 'dressing' and 'putting on clothes' all became 'dressing'), and collapsing clear synonyms (e.g., 'soothing', 'consoling', and 'comforting' all became 'soothing'). Then, standardized codes were combined into activity categories, grouping together highly similar codes into one activity category (for example, 'dressing', 'undressing', and 'brushing hair' all become 'dressing'). This resulted in 42 unique activity categories, but only 11 of them accounted for 96% of the codes assigned. Each utterance was finally assigned to each context (or not) depending on whether or not at least three of the five coders on that utterance independently generated codes for that context.

To illustrate this process, consider the following example utterance, which occurred about halfway through the first transcript: "let's dry you". Coders read this utterance presented in a window of 30 utterances and provided codes describing the whole window. All together, this utterance was coded 14 times by 7 different coders; while each coder never saw the same window more than once, they did see this utterance repeated in different windows. Because this was more than the necessary 5 unique coders, 5 codes were randomly selected from the 14 provided, with the constraint that each of the 5 selected codes were provided by a different coder. In this case, that left the raw codes for this utterance shown in Table 1. These raw codes then underwent the two-step cleaning process, first standardizing forms (e.g. 'after bath' and 'post bath' both became 'post-bath'), and then grouping together similar codes into context categories.

At this point, the utterance could be assigned to the 'bath-time' sub-corpus because a majority of the coders (all 5, in this case) coded it as 'bath-time'. It did

Table 1. Illustrative example of the processing of codes from the coder judgment approach to defining context. The utterance in this example is "let's dry you" and the final context code ends up being "bath-time" since that is the only context category that at least three of the five coders provide.

	Raw Codes	Cleaned Codes	Context Categories
Coder 1	bath time	bathtime	bathtime
Coder 2	washing; diapering; preparing for bed	washing; diaper change; pre-bedtime	bathtime; diaper change; sleep
Coder 3	post-bath	post-bath	bathtime
Coder 4	drying	drying	bathtime
Coder 5	bath time	bathtime	bathtime

not reach criterion on 'diaper change' or 'sleep', since each of those was given by only one coder. So this utterance was included in the 'bath-time' sub-corpus and no other.

Assessing Agreement Between Methods of Defining Context.

For both the topic modeling results and the subjective coding results, I had originally intended to have two versions of the output: continuous loadings for each utterance on each context, and a binary decision (yes/no) for each utterance as to whether or not it was included in each context sub-corpus. The binary measures corresponded to the information that was actually used in the subsequent hypothesis testing, but the continuous measures may have provided richer information, potentially increasing statistical power. However, for the continuous topic modeling loadings, it was not possible to use standard statistical methods with all topic modeling contexts entered in the same model because each utterance was constrained to have a total probability of 1 across all topics, making the matrix of topic modeling loadings rank deficient; in effect, there is perfect multicollinearity among the topic modeling contexts. Any model with all of the loadings in it will be impossible to estimate analytically. Rather than attempting work-arounds to be

able to use the continuous topic loadings (such as dropping one or more topics from the models), I focused my attention on the agreement between the binary versions of contexts for each method. Although it would be ideal to be able to analyse the continuous loadings as well as the binary context tagging, the binary data were derived directly from the continuous loadings, so it was unlikely that the results would diverge too dramatically. All analyses reported henceforth are on the binary context coding only.

I measured agreement between each pair of methods using contingency tables. For example, to examine the agreement between the top-down word list approach to defining context and the topic modeling approach, I constructed a contingency table tallying the number of utterances that had each combination of word list code and topic modeling code (e.g. How many utterances were tagged as word list "bath" and topic modeling topic 1? How many as word list "bath" and topic modeling topic 2?). Because context codes were not mutually exclusive within each approach to defining context, the standard Pearson chi-squared coefficient could not be applied to the full table of co-occurrence counts. That would have counted the same observation multiple times when it co-occurred with multiple other contexts. In order to retain accurate margins in the contingency tables, each observation (i.e. utterance) must sum to one within each approach to defining context. The simplest way to achieve this was to drop from the analysis any utterance with more than one context code within an approach. In the word list approach, that excluded 712 utterances (5.7% of the corpus), in topic modeling approach, 2998 utterances (24%), and in the coder judgments approach, 123 utterances (1%). Then I analyzed the remaining utterances using standard measures of association for contingency tables (e.g. chi-squared test

of independence, Cramer's V). I complemented these analyses with a second approach, which was to include all of the utterances and use statistical methods appropriate to multiple-response variables, for example, testing simultaneous pairwise marginal independence (Agresti, 2007). This had the advantage of retaining all of the information in the data, but meant standard methods for exploring and interpreting the results could not be applied. I used the test of independence for multiple-response categorical variables by Bilder and Loughin (2004), available in the MRCV package in R (Koziol & Bilder, 2007). It is an intuitive extension of the Pearson chi-squared statistic, testing the null hypothesis that all of the options from one set (e.g. all context codes from the word list approach) are independent of all options from the second set (e.g. all context codes from the topic modeling approach).

For each of the three possible pairs of approaches to defining context (word list and topic modeling, word list and coder judgments, and topic modeling and coder judgments), I created contingency tables excluding utterances with multiple codes and used a chi-squared test of independence to test the hypothesis that the context coding from the two methods were unrelated. I followed that up with calculation of Cramer's V to assess the strength of the relationship between approaches. Cohen (1992) suggested values of .1, .3, and .5 as small, medium, and large effect sizes for chi-squared based measures of effect size like Cramer's V. I then computed the adjusted chi-squared statistic for multiple-response categorical variables testing for simultaneous pairwise marginal independence.

Finally, I used finite mixture modeling (latent class analysis, LCA) to determine areas of agreement among all three methods. This analysis took the context codes from all three approaches at once and analyzed them all together

to infer latent classes explaining patterns of agreement across variables. I used the poLCA package in R (Linzer & Lewis, 2013), entering each of the three coding approaches as a categorical variable with as many possible outcomes as it had context codes. As with the contingency tables, the analysis excluded utterances with multiple context codes within the same approach. Because the LCA was estimated with an iterative, probabilistic EM algorithm, however, utterances with missing data (such as from having multiple context codes) on one or more approaches could still be included in the analysis; the algorithm just uses as many variables as are available for each case (Linzer & Lewis, 2011).

Results. There were some contexts with only a very small number of utterances, making calculations relying on the occurrences of those contexts (such as contingency tables) unreliable. For example, a standard chi-squared test of independence is only appropriate with expected cell counts of at least 5 per cell — otherwise the resulting statistic is not chi-squared distributed, rendering the p value meaningless. For contexts with only a handful of utterances, there simply was not enough information available to track their co-occurrence with other context codes.

The analyses reported here were therefore restricted to contexts with at least 60 utterances (enough to have at least 5 per cell for all expected counts). This excludes the contexts 'TV', 'touching', 'hiccups', 'taking pictures', and 'outside' from the subjective coder judgment approach and 'media' from the word list approach, composed of between 2 and 23 utterances each.

Before assessing agreement between the different approaches to defining context in this corpus, I examined the three approaches on their own. The most striking difference was the proportion of the corpus covered by each method.

Context codes from topic modeling loadings provided nearly complete coverage of the corpus (98.3% of utterances in the corpus are included in one or more contexts). This was to be expected since the threshold to convert the continuous topic loadings into binary context code decisions for each utterance was decided by maximizing the number of utterances with one or two context codes. The word list method covered only 36.1%, and coder judgments covered 50.1%. For the word list method, utterances with no context code were simply utterances that did not occur within 2 utterances of a key word from the context lists. Using the coder judgment approach, utterances with no context code were ones on which coders did not agree; in order for an utterance to be coded for a given context, at least three of the five coders for that utterance needed to generate a description that would fall into that context category (see Methods for more details). Since coders' responses were completely unconstrained (e.g. there was no list of categories from which they selected codes), the context needed to be fairly obvious and unambiguous in the transcript for the independent coders to spontaneously generate sufficiently similar descriptions. Utterances with no context code from the coder judgments were not clear enough from the transcripts for independent coders to agree[3].

In the cases where there was no context code, in either the word list or coder judgment approaches, it may have been because of a lack of sensitivity in the method (i.e. an activity context really was occurring, but it was not reflected in the transcripts in such a way that the coding approach could pick up on it) or it may reflect real gaps in the infants' activities. This may be similar to the

[3] Coders did have the option to enter a context of 'none' if they could not tell what was happening in the transcript, which could potentially provide a direct way to capture utterances coders agree are ambiguous. Only 26 utterances (0.2%) had at least 3 out of 5 coders generate a code of 'none', however. This suggests that coders felt like they could offer guesses about the context for most utterances, but they did not reach sufficient agreement for many of them.

"transition time" coded in Soderstrom and Wittebolle (2013), used to mark time between clear activity contexts. There is nothing distinctive about transition time *per se* that would make it similar each time it occurs; rather, it is defined by a lack of any other activity occurring. Soderstrom and Wittebolle (2013) note that this was particularly obvious in their daycare recordings, where contexts were often very clearly delineated with the possibility of a delay between the end of one clear context and the beginning of the next (e.g. snack-time ends and there is transition time while the teacher clears up snack materials before playtime begins). Note, however, that transition time made up less than 5% of the total time in Soderstrom and Wittebolle's daycare recordings and even less in the at-home recordings, a much smaller portion than the context-less sections of the corpus in the current study. This discrepancy could be the result of a difference in the granularity of coding. Soderstrom and Wittebolle coded for activities in five minute bins, with the code for each bin corresponding to the activity that made up the majority of that five minutes. If 'transition time' is generally relatively fleeting, it could have happened often throughout the day but not surface as the dominant activity for many five minute bins.

The fact that the topic modeling approach yielded much more complete coverage of the corpus than either of the other two approaches limited the amount of agreement between the topic modeling context codes and the others, because there was necessarily a large number of utterances that were tagged with a context from the topic modeling but did not have a corresponding tag in the other approaches. It may also inform the interpretation of the activity contexts across methods, since utterances that were ambiguous or uninformative in the word list and coder judgment approaches were nevertheless coded by the topic modeling,

implying that one or more topics may have captured linguistic patterns typical of these hard-to-code utterances.

Looking a little more closely at the portion of the corpus that did have context codes for each approach to defining context, there appeared to be differences in the distribution of contexts as well. Contexts as identified by the coder judgments showed a dramatically skewed distribution, with one very common context and many others that were much less common (see Figure 2). The contexts from the word list method also appeared to be skewed, although less so than those from the coder judgments. In contrast, the contexts resulting from the topic modeling analysis were closer uniform in distribution, with several of the most common contexts all relatively close to each other in prevalence.

Because this corpus included five families and there were five recordings for each family, it was possible to examine patterns of context within and between families as well. The types of contexts that were the focus of this work should be common enough and general enough to apply across families, to be able to connect with existing work on the topic (e.g., Fausey et al., 2015; Hoff-Ginsberg, 1991; Roy et al., 2015; Soderstrom & Wittebolle, 2013). Indeed, for all three approaches, contexts did not appear to be idiosyncratic to any particular families or transcripts. Across all contexts in all three approaches, there was no context that occurred only in one transcript or only in one family; all contexts occurred in multiple transcripts and in multiple families. Except for the smallest context sub-corpus (the 'housework' context from the coder judgments approach, which is just 103 utterances total), each context occurred in the transcripts of at least four of the five families. Figure 3 shows the breakdown by context for each transcript within each family when contexts are defined using the word list approach, and Figure 4 and

Figure 5 show similar data for the topic modeling and coder judgments approaches, respectively.

To assess agreement between the approaches to defining context, I examined the context codes from two approaches at time (word list and topic modeling, word list and coder judgments, topic modeling and coder judgments). I constructed contingency tables on the unambiguously tagged utterances, i.e. those utterances with at most one context tagged for each approach to defining context. This excluded a substantial portion of the corpus for the topic modeling approach in particular, so to test the analogous hypothesis on the full corpus (not just the non-ambiguously tagged utterances), I also tested for simultaneous pairwise marginal independence.

When including only utterances with at most one context code per approach, a chi-squared test of independence rejected the null hypothesis that the contexts from the word list approach were unrelated to the context codes resulting from the topic modeling analysis ($\chi^2(84, N = 10155) = 3201.48$, $p < .001$), with a small to medium effect size ($V = 0.21$). In other words, there is evidence of a relationship between context codes from the word list and topic modeling coding approaches. The bootstrap analysis of simultaneous pairwise marginal independence conducted on the entire corpus echoed these results, also indicating a significant deviation from independence between the word list context codes and the topic modeling context codes, $p < .001$. There was also a significant relationship between context codes from the word list approach and the coder judgment approach on non-ambiguous utterances ($\chi^2(63, N = 7809) = 5509.3$, $p < .001$), with a medium effect size ($V = 0.32$). This was also reflected in the bootstrapped test of simultaneous pairwise marginal independence, $p < .001$.

Finally, the context codes from the coder judgments and from the topic modeling were also related, both by the chi-squared test of independence on the non-ambiguous codes ($\chi^2(108, N = 10879) = 6060.25$, $p < .001$), with a small to medium effect ($V = 0.25$), and by the bootstrapped test of simultaneous pairwise marginal independence on the whole corpus, $p < .001$. Taken together, these findings suggest that the context codes resulting from these three different approaches to defining context were clearly related, but not redundant; there were discrepancies as well as points of agreement.

To explore the points of agreement and disagreement among the three approaches to defining context more thoroughly, I used a latent class analysis (LCA). LCA assumes the overlap among context codes is caused by an underlying latent construct, "context", with R different levels, each corresponding to a different class. Each observation (utterance, in this case) is assumed to belong to one latent class. When there is no strong theoretical basis for setting the number of classes in advance, it is standard to choose the number of classes based on model fit (Linzer & Lewis, 2011). One popular index of model fit is the Bayesian Information Criterion (BIC), which balances model fit with the number of estimated parameters, preferring models that are more parsimonious (Schwartz, 1978). A model with six latent classes yielded the best fit (see Fig. 6). I will report on this model more fully to explore the relationships between the latent "context" classes and the context codes from each of the three approaches to defining context in this corpus.

First, it is important to note that not all of the latent classes were the same size. The selected 6-class model identified one relatively large class (Class 2), which covered approximately 43% of the corpus, and to a lesser extent Classes 1, 3, and 6

which cover an additional 17%, 15% and 16% of the corpus, respectively. Together, these top 4 classes cover most of the corpus (91%). The two smaller classes, Class 4 and Class 5 cover 7% and 2% of the corpus, respectively.

In order to interpret the nature of each class, it is helpful to examine the class-conditional probabilities. The class-conditional probability for a context code is the probability of an utterance being tagged with that code, given that it is in a particular latent class. Utterances that belong to latent Class 1, for instance, had a 56.5% probability of being tagged as "bath" and a 1.7% probability of being tagged "bed" in the word list approach. The class-conditional probabilities for all of the the context codes are depicted in Fig. 7.

The latent classes captured similar patterns of responses across utterances. To illustrate, consider context codes that were related to bath time. Class 1 was characterized by utterances that were tagged with "bath-time" according to the coder judgments, had key words from the "bath" list in the word list approach, and were identified as topic 7 or 8 by the structural topic model. In other words, there is a set of utterances in the corpus that were likely to be tagged as "bath" by word lists and coders, and to show up as topic 7 or 8 in the topic modeling contexts. To aid in interpretability, each context from the topic modeling approach can be defined by the words that were most probable in that topic. Topic 7 and topic 8 were both defined by words that are consistent with bath activity (the top five words from topic 7: hair, water, bad, fun, minute; topic 8: bum, bath, splash, swim, shake). Similarly, the 'meal' context from the word list method and the 'mealtime' tag from coder judgments were both high probability in Class 3, as was topic 4 (top five words for topic 4: door, lunch, minute, sit, lucky). Class 2 (the largest class) was high probability for play contexts, Class 5 for sleep, and Class 6 for

fussing. Class 4 (one of the smaller classes, covering 7% of the corpus) was less clearly defined, with no particularly high probability context codes, but reasonable probability for play and diapering/dressing contexts.

The chi-squared tests of independence and simultaneous pairwise marginal independence demonstrated that there was significant alignment between each of the approaches to defining contexts; examination of the latent classes confirms that that alignment conforms to expected patterns.

Discussion. These analyses compared the contexts from three different coding approaches on the same corpus: tagging utterances by the occurrence of key words from the Oxford CDI, human coders providing subjective judgments on utterances, and contexts derived from topic modeling loadings. There was substantial agreement across the approaches (Cramer's V's of $.2 - .3$), confirming that these three very different methods were still homing in on similar phenomena.

Importantly, it is also clear that these three approaches to defining context were not interchangeable; there were points of divergence as well as similarities. The context codes resulting from topic modeling in particular did not perfectly match the context codes from the other approaches. Although previous work has noted that topic modeling and coder judgments of context from video yield similar results (Roy et al., 2012), that association has not been closely examined. A major contribution of this work is the increased clarity around the degree and nature of the agreement (and disagreement) among approaches to defining context. Topics are typically defined by the most probable words within each topic (Blei et al., 2003), but this analysis reveals potential limitations of that strategy for the current application. For example, the most probable words in topic 5 (tickle, toe, feet, din, tick, tum) suggested an activity context like tickling and playing. Class 2

had high probabilities for the 'playtime' tag from coder judgments and the 'body touch' category from the word list method, which contained key words including body parts and touching verbs like 'tickle', 'hug', 'cuddle', and 'kiss', with the 'play' category from the word list approach also relatively high probability. Topic 5 would seem to be a natural fit, but actually topics 2 and 10 had higher probabilities for Class 2. From their most probable words, topic 2 appeared to be about play (top words: bop, hello, monkey, give, hi, thing), while topic 10 seemed to capture scolding (top words: hand, see, naughty, fed, can, bite). In other words, Class 2 identified a set of utterances that were consistently tagged with play and touch contexts in two approaches (coder judgments and word list) and play and scolding topics from the topic model. This may indicate a systematic disagreement between approaches, or simply misleading interpretation of the topics based on their top words; the 'scolding' words may be used in jest, such that someone reading the transcripts themselves (as the coders did) would identify those interactions as playful rather than scolding.

One source of disagreement may have been the different constraints on the various methods. In particular, topic modeling prefers models with relatively uniform prevalence (how often each topic occurs). This preference makes sense in many topic modeling applications, where the topics might reasonably be expected to be equally likely, but may not be ideal for capturing activity contexts. The word list approach and coder judgments had no such preference, and in fact, both resulted in skewed context prevalence, with a few very highly prevalent contexts and many more less prevalent ones. This echoes recent work on parent reports of infants' daily activities showing that, especially for the youngest infants studied, the distribution of activities was naturally skewed, with roughly half of the day

spent sleeping, substantial proportions eating and playing, and much less time in a variety of other activities (Fausey et al., 2015).

It is interesting to note that some particular contexts may be easier targets for agreement across these three different approaches than others. For example, bath utterances appeared to be clearly identified by both the word list and coder judgment approaches, as were play, meal, sleep, and fussing utterances. Diaper change was much more varied, though, with a less clear pattern of responses across methods. This has important implications for researchers interested in differences in caregiver speech context to context (e.g., Hoff-Ginsberg, 1991; Soderstrom & Wittebolle, 2013) because it suggests that the reliability of context identification itself may vary by context. If context definitions themselves are noisier for some contexts, then that reduces the reliability of measures taken in those contexts as well. Any comparison of, for example, lexical diversity in caregiver speech in one context to another will have to take into account the difference in reliability of those two measurements.

Figure 2. The number of utterances in each context for each approach to defining context.

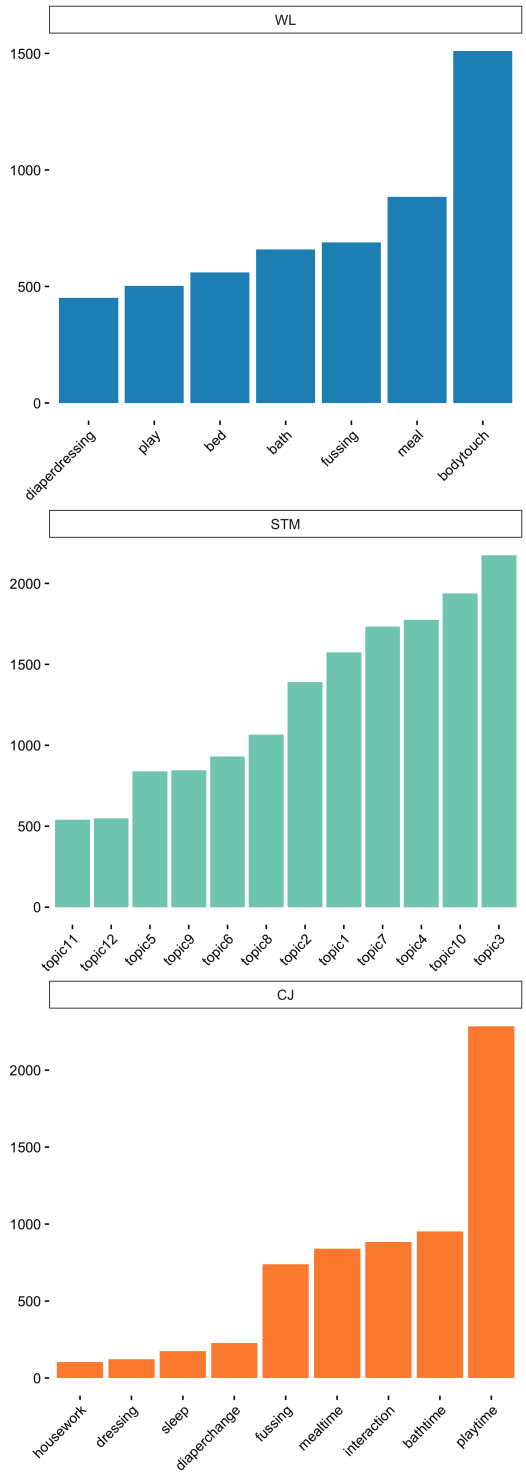


Figure 3. The proportion of utterances in each context by transcript, for each family. Contexts are defined by the occurrence of key words selected from the Oxford CDI.

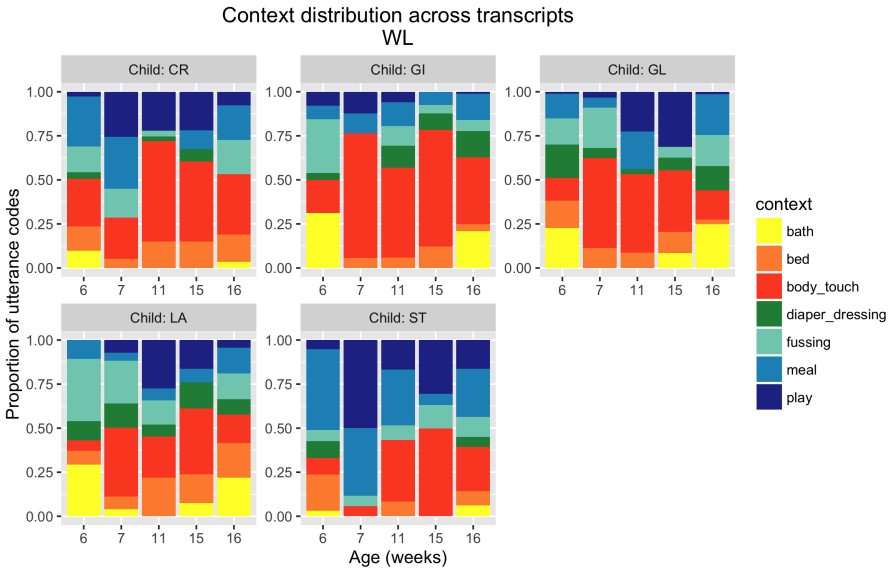


Figure 4. The proportion of utterances in each context by transcript, for each family. Contexts are defined by loadings from topic modeling analysis.

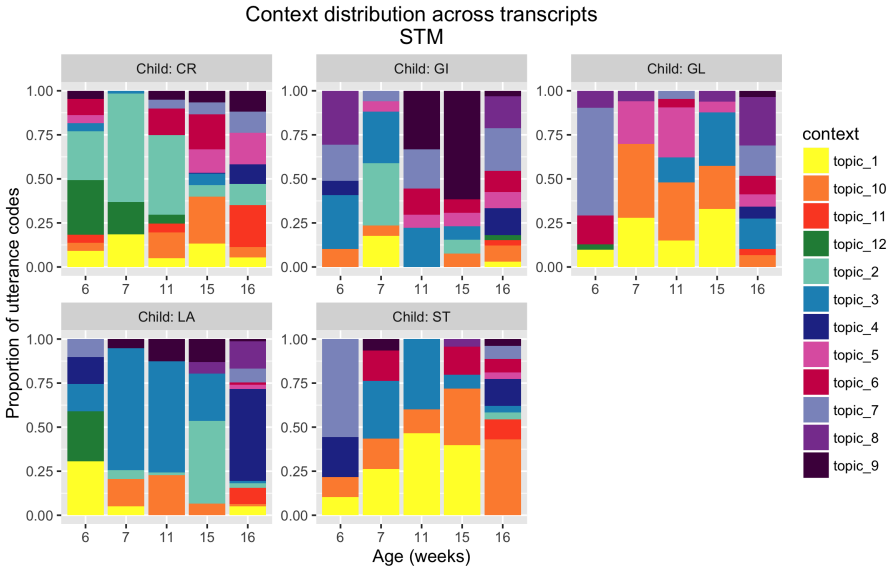


Figure 5. The proportion of utterances in each context by transcript, for each family. Contexts are defined by human coders providing open-ended judgments, which are then categorized into context codes.

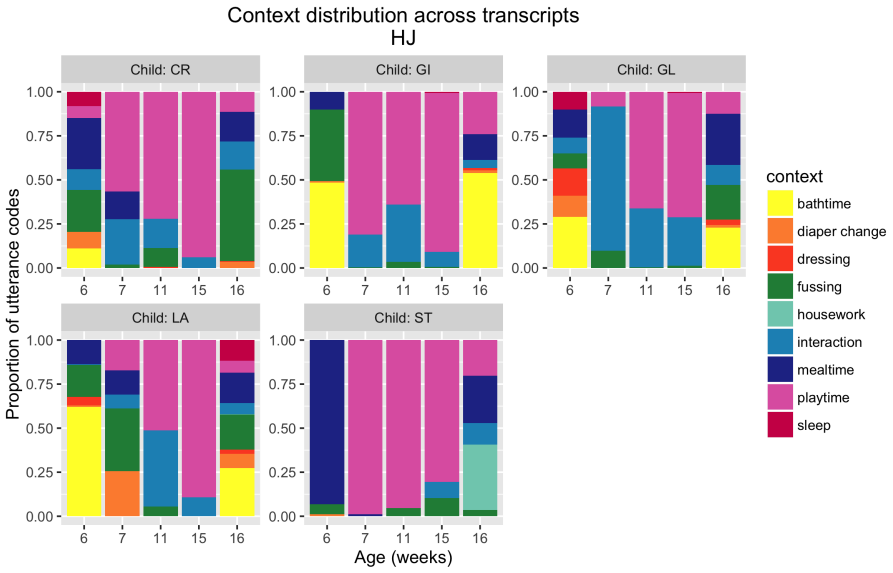


Figure 6. Model fit (Bayesian Information Criterion) for latent class analysis models for context fit with a range of classes. Lower BIC indicates better model fit.

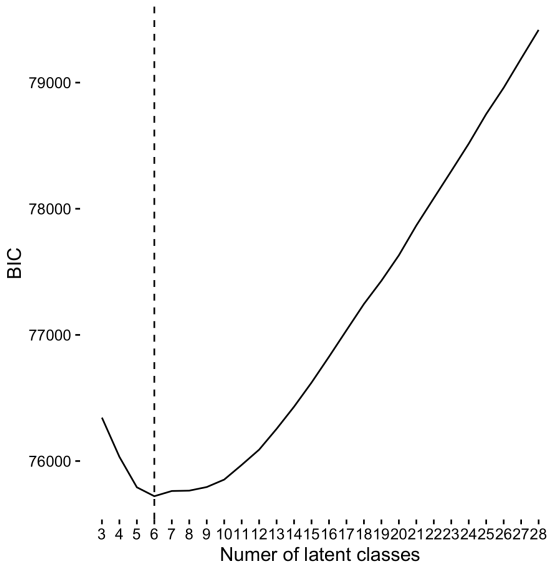
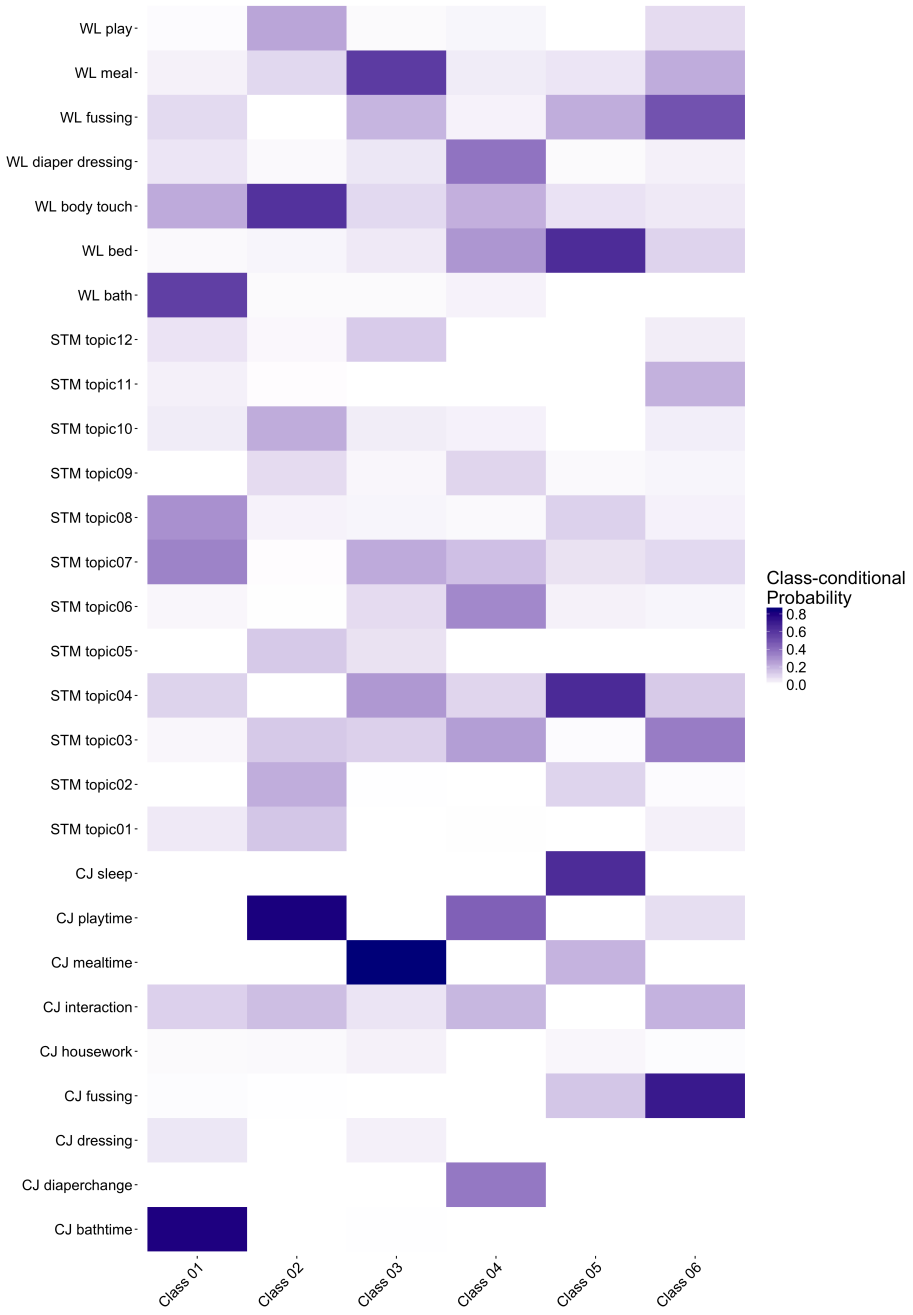


Figure 7. Class-conditional probabilities for each context code. WL = word list approach, STM = structural topic modeling approach, CJ = coder judgment approach.



CHAPTER III

STATISTICAL CUES TO WORD BOUNDARIES WITHIN CONTEXT

The previous section introduced three approaches to defining context in transcripts of infant-directed speech: by the occurrence of key words, with topic modeling, and with subjective coder judgments. I showed that while there is substantial agreement across methods, they are not redundant; differences in the proportion of the corpus covered, the relative prevalence of the contexts, and other methodological characteristics all contribute to divergence among the three approaches. While the previous section focused on which utterances were included in each sub-corpus, the current section characterizes the context sub-corpora themselves with a set of metrics relevant to a major task of early language acquisition — word segmentation. I analyzed the context sub-corpora resulting from each of the three approaches to test the hypothesis that context-specific patterns in word use may yield clearer statistical cues to word boundaries within context sub-corpora compared to the corpus as a whole. To do this, I measured several relevant descriptive statistics on the sub-corpora and made use of computational models of word segmentation to directly assess segmentability itself.

Assessing Segmentability. Bayesian word segmentation models provide an attractive option for segmenting corpora based on the statistical patterns in speech. Because Bayesian models are "ideal" learners, they optimally represent the patterns in the input, according to whatever structure they use. While this may be problematic in attempts to model actual human performance in segmentation tasks (see discussion in Frank et al. 2010), it suited my needs well since my goal was simply to characterize patterns in the input. I used two Bayesian

word segmentation models — the hierarchical Dirichlet process (HDP) model (Goldwater et al., 2009) and a collocation-syllable adaptor grammar[1] (M. Johnson, 2008) — to ensure that any effects I observed were not particular to one specific model.

The HDP model and the adaptor grammar are similar in many respects. They both use the co-occurrence of phonemes in the speech stream as the primary cue to identify word-like units, they both incorporate hierarchical structure to allow for the fact that there are statistical dependencies between words as well as within words, and they are both estimated using Bayesian methods. The adaptor grammar differs from the HDP model in how it represents the structure of words, however, by including a constraint that words must be composed of syllables, and syllables are in turn composed of an optional onset, a vocalic nucleus, and an optional coda. Because the models operate over phonemes as the unit of analysis, the HDP model can — and does (Goldwater et al., 2009) — identify sub-syllabic units as candidate words, such as the morpheme /z/ frequently used for pluralization or to mark possession. The adaptor grammar, on the other hand, is constrained such that candidate words must be composed of syllables, so it cannot make that kind of over-segmentation error. This constraint improves performance of the adaptor grammar for segmenting a language like English where all words are composed of one or more syllables, but may not generalize to other languages as readily. Moreover, the adaptor grammar tracks which phonemes occur in word-initial onsets and word-final codas and can incorporate this information into its decisions to posit word boundaries, simultaneously learning patterns in the

[1] Note that there are several different versions of Johnson's adaptor grammar available. I used the model that exhibited the best performance in Börschinger, Demuth, and Johnson (2012), the collocation-syllable adaptor grammar with 3 levels of collocations (called 'colloc3Syll').

language's phonological system and processing speech into units, and allowing the two processes to bootstrap each other. Again, this improves performance, at least on English language corpora (note the difference in performance between the "colloc" and "collocSyll" models in Börschinger et al., 2012). Nine-month-old infants (English and Dutch) also track these kinds of phonotactic patterns (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993), although it is unclear whether they would have sufficient expertise to make use of those cues earlier in infancy to support early word segmentation.

Another important difference is in the way the HDP model and adaptor grammars measure relationships between words. The HDP model relies on sequence learning, where one word is allowed to predict the next, whereas the adaptor grammar uses collocations (sets of words that tend to occur next to each other) to chunk speech hierarchically. There is no theoretical limit on the number of layers of collocations — a model could potentially group candidate words into initial collocations, and then also allow those collocations to be further grouped into larger collocations, and so on (the model as implemented in this study allows three levels of collocations). This reflects the natural hierarchical structure in language noted by linguists, and also nicely mirrors "chunking" models of processing that may characterize both language processing and acquisition (e.g. Christiansen & Chater, 2016). The fact that the collocation adaptor grammar can include several levels of hierarchical dependencies connects with evidence that infants and children are sensitive to multiword segments of speech (Arnon & Clark, 2011; Bannard & Matthews, 2008; Soderstrom, Seidl, Nelson, & Jusczyk, 2003). Several recent demonstrations have suggested that a chunking model of word segmentation may be the best match to infants' (Monaghan & Christiansen, 2010) and adult's

performance (M. C. Frank et al., 2010). Either strategy — sequential prediction (Goldwater's HDP model) or hierarchical collocations (as in the adaptor grammars) — capture important non-independence in the ordering of words in speech. For example, "that" is much more likely to come after "what's" than after "my". With both models, the statistical dependency in "what's-that" can still be modeled as a word boundary, either as "what's" predicting "that" (HDP) or as a collocation of words (adaptor grammar). This ameliorates the issue of under-segmenting noted in models that allow only for dependencies within words and do not also capture dependencies between words (see discussion in Goldwater et al., 2009).

Although they make different assumptions about the structure of words and the relationships between words, both the HDP model (Goldwater et al., 2009) and the collocation-syllable Adapator Grammar (M. Johnson, 2008) were reasonable strategies to assess the segmentability of sub-corpora in the present study. Importantly, unlike many other computational models, both have been shown to work reasonably well at relatively small corpus sizes: Börschinger et al. (2012) tested both of these models (and several others) on subsections of the Providence corpus ranging from just under 1,000 utterances up to about 25,000. The HDP showed some improvement in performance over that range, beginning with a token F-score of about 60% and reaching about 70% at the largest corpus sizes, but the adaptor grammar implemented here was relatively stable at about 85% from the smallest size. While some of the sub-corpora used in the present study were smaller than 1,000 utterances (especially for the word list approach to defining context, see Fig. 2), the fact that the adaptor grammar in particular has been shown to be effective at 1,000 utterances suggests applying it at even smaller sizes may not be unreasonable.

Testing Contexts Against Nontexts. For the purposes of the present study, the object of applying the computational models was not just to get an estimate of segmentability for each context sub-corpus, but to test the hypothesis that statistical cues to word boundaries are clearer within context. The necessary comparison is context-based segmentation to non-context-based segmentation. The whole corpus would seem to be the obvious way to conceptualize the appropriate corpus for examining non-context-based segmentation. However, a comparison between context-based and non-context-based segmentation conducted in that manner would be problematic because of large differences in sample size, especially given that many important corpus metrics are sensitive to corpus size, including segmentation performance of the computational models (Börschinger et al., 2012). Any difference discovered between the context subsets and the whole corpus could be attributed simply to the difference in size instead of anything about the statistical patterns in the context sub-corpora *per se*.

Instead, I used a bootstrapping procedure to simulate an empirical null distribution for each context sub-corpus, providing a null comparison that retained the structure of the global corpus but was size-matched to the context sub-corpora. For each context sub-corpus, I took a random sample of the same number of utterances from the corpus, generating a matched "nontext" for each context sub-corpus. Then, in each nontext sub-corpus, I applied computational models to measure the segmentability of the sample, saving the resulting segmentation estimates. This procedure was repeated many times for each sub-corpus, generating a size-matched empirical null distribution for each metric of interest. Because this procedure was repeated with randomly selected utterances, the actual content of the random sub-corpora varied from sample to sample; in the limit, the random

"nontexts" reflected the structure available in the whole corpus (since they were randomly sampled from it). The "nontexts" were always the same size as the context sub-corpora they matched, however, making the segmentation results calculated on them a suitable null comparison for the context segmentation results.

I assessed the performance of the computational models by comparing their segmentation "solution" to the actual English words (tokens) in the sub-corpus analyzed. This comparison can be quantified in two complementary ways: the proportion of tokens correctly segmented out of all tokens segmented by the model ("precision", or "accuracy"), and the proportion of tokens correctly segmented out of all tokens that were available in that sub-corpus ("recall", or "completeness"). For example, if the English phrase *big fat tummy* is segmented by a model as *bigfat tummy*, that would yield a precision of 1/2 and a recall of 1/3. Prioritizing both precision and recall balances evaluation of a model between under-segmenting (which increases precision at the cost of discovering fewer words) and over-segmenting (which, because of the prevalence of monosyllabic words in English, correctly recovers more tokens but penalizes precision by generating multiple incorrect tokens for each over-segmented multisyllabic word). The F-score (the harmonic mean of precision and recall) is a useful compromise and is often reported in assessments of word segmentation models.[2] It incorporates both precision and recall and has the attractive property of naturally penalizing models with a large difference between precision and recall, so models that dramatically under- or over-segment have lower F-scores than models that strike an appropriate balance.

[2]Börschinger et al. (2012) used the harmonic mean of precision and recall as their key metric, while Goldwater et al. (2009) used the closely-related geometric mean instead. The interpretation for the two measures is very similar, but the calculation is different. For the present study, token F-score is calculated as the harmonic mean of precision and recall for both models.

For the purposes of this study, token F-score from each of the computational models is used as an indicator of 'segmentability' for each sub-corpus. This is a novel application of computational models for word segmentation. While there have been several investigations comparing different models (or versions of models) on the same language samples to assess the models (Börschinger et al., 2012; Goldwater et al., 2009, *inter alia*), in this case the focus was on assessing the language samples themselves, in relation to whether taking context into account enhanced segmentability. Because of this, I made no attempt to modify model parameter settings to maximize performance on this corpus,[3] as has been done in previous reports on these models (Goldwater et al., 2009; M. Johnson, 2008). I used the parameter settings reported by Goldwater et al. (2009) and Börschinger et al. (2012). For Goldwater's HDP model, the parameters were $\alpha_0 = 3000, \alpha_1 = 100, p_{boundary} = .2$. The parameters of the adaptor grammar (a and b) were determined by the model during fitting, using weak priors (a uniform beta prior for a and a $Gamma(100, 0.01)$ prior for b), so they could be automatically adjusted for the corpus at hand. Both models were originally optimized with respect to the Bernstein-Ratner-Brent corpus (Brent, 1999b) and so may be expected to perform worse on other corpora, although the additional flexibility of the adaptor grammar may make it more robust to such changes. The F-scores obtained in the present investigation were lower than those reported by M. Johnson (2008) and Goldwater et al. (2009); this was expected both because of the application of the models to a different corpus than the one they had been optimized for, and because of the reduced corpus size (although the Korman corpus at 12511

[3]This is similar to Börschinger et al. (2012), who were interested in the effect of input corpus size rather than maximizing model performance. They applied a set of models — including both of the models used in the present study — to new corpora without tuning any model parameters.

utterances is larger than the Bernstein-Ratner-Brent corpus at 9790 utterances, the context sub-corpora analyzed for this study range in size from 103 to 2286 utterances). The goal in this case was not to achieve the best segmentation possible, but rather to examine how differences in the sub-corpora analyzed related to differences in segmentation performance in order to understand the relationship between structure in the input — in particular, context-based structure — and segmentability. Because of this, these results do not speak directly to the quality or validity of one model over another and should not be interpreted in that way. Instead, the segmentation performance for each context should be compared to its bootstrapped null distribution, within each model.

The comparison of segmentability estimates (token F-scores) from context-specific sub-corpora to estimates from exactly the same models run on randomly generated sub-corpora provided a rigorous, tightly-controlled test to answer the following question: Was the speech from utterances within a given context more segmentable than the same number of utterances randomly sampled from the same corpus without respect to context? The parameter settings for each model were held constant. Because the null distributions were built with samples from the *same* corpus as the context estimates, this method also controlled for all corpus factors including infants' age, gender, SES, and a host of potentially influential factors that may be harder to estimate or for which no metadata may exist. If token F-scores were significantly higher in context-specific sub-corpora compared to random sub-corpora of the same size, it would have provided evidence that the act of subsetting the corpus by context increased segmentability.

In addition to measuring segmentability itself, I included the analysis of several other descriptive statistics that may be related to the segmentability of a

corpus. There is evidence that several aspects of a corpus's structure may impact how readily it is segmented, including repetition (Brent, 1999a; M. C. Frank et al., 2010; Onnis, Waterfall, & Edelman, 2008), words in isolation (Brent & Siskind, 2001; M. C. Frank et al., 2010; Lew-Williams et al., 2011; Monaghan & Christiansen, 2010), utterance length (M. C. Frank et al., 2010), and skew (Kurumada, Meylan, & Frank, 2013). For the present study, I operationalized these with type-token ratio, proportion of one-word utterances, mean number of words per utterance, and the proportion most-frequent-word, respectively. For each, I used the same bootstrapping procedure described above to measure differences between context subsets and size-matched random samples from the same corpus. I hypothesized that more repetition, more isolation, longer utterances, and more skew would be associated with higher segmentability. If context sub-corpora differed from their size-matched random "nontext" sub-corpora on these features, that may explain any difference in segmentability.

Bootstrapping. For each of the context sub-corpora and for each metric of interest, I resampled random utterances from the corpus over and over to build each bootstrapped null distribution. For both the HDP model and the adaptor grammar, note that the model estimation process itself was also iterative (Gibbs sampling): For a given sub-corpus, a model would adjust its representation (the segmented units it was considering), assess fit between that representation and the observed input corpus (log-likelihood), adjust again to improve fit, and so on for the desired number of iterations. In Gibbs sampling, once a model has reached the best representation it can achieve for the given input, there will be very little change with subsequent iterations (the model is said to have reached convergence). The number of iterations needed for a model to reach convergence depends on

a number of factors including the input data and the complexity of the model; for the HDP model I used 5,000 iterations with simulated annealing (see details in Goldwater et al., 2009) and for the adaptor grammar I used 500 iterations. For the random "nontext" distributions for segmentability, each bootstrapped sample was run for the specified number of iterations with the Gibbs sampler and the final token F-score was recorded as one observation in the bootstrapped null distribution.

Recommendations for an acceptable number of bootstrapped samples for a test vary widely, from as few as 19 samples (Dufour & Kiviet, 1998) to 100,000 (Chernick & LaBudde, 2014). Because of the random error introduced during resampling, a p value estimated from any finite number of bootstrapped samples will have some error around it, with p values estimated from fewer samples having more error (Davidson & MacKinnon, 2000). When computation time is of no concern, therefore, it makes sense to run as many bootstrapped samples as possible as this will result in more precise p values. In hypothesis testing, however, precise estimation of the p value itself is of less importance than certainty about whether it falls above or below a critical threshold α (e.g. .05). For p values well below α or well above it, more error is acceptable, but for p values close to the cutoff, high precision is important: A p value of .04 with +-.03 error is insufficiently precise to interpret the hypothesis test, whereas a p value of .84 with the same error is fine. When computation time is expensive, it makes sense to run sufficient samples to confidently place the bootstrapped p value either above or below the critical cutoff α. Davidson and MacKinnon (2000) propose a method for determining whether a test has sufficient bootstrapped samples to interpret the hypothesis test by modeling the counts of bootstrapped estimates above and below the observed

estimate as coming from a binomial distribution with probability of success set at α (.05 for a one-tailed test, .025 for a two-tailed test). Effectively, this tests whether bootstrapped p values greater than α are significantly greater than α (in which case the null is retained), and for p values less than α whether they are significantly less than α (in which case the null is rejected). If the binomial test itself is not significant, then there are insufficient samples to determine whether p is above or below α; Davidson and MacKinnon (2000) recommend increasing the number of bootstrapped samples to increase the precision of p, and then re-testing it against α. This process continues until p is sufficiently precise or until the maximum feasible number of samples has been reached. When the true p value is very close to α, it may not be feasible to determine whether it is above or below α.

For the present study, I compared each metric estimated for each context sub-corpus to a bootstrapped null distribution generated by estimating the same metrics on sub-corpora made by randomly sampling the same number of utterances from the whole corpus. The null hypothesis in each case was that, for the metric in question, selecting utterances by activity context was no different from selecting them randomly. For each test, I used the procedure outlined in Davidson and MacKinnon (2000) to ensure that I had a sufficiently precise estimate of the p value to interpret the hypothesis test. Unless otherwise noted, all bootstrapped p values were either significantly above or below α. Across all context sub-corpora defined by all three approaches, there were two tests from the descriptive statistics where p was too close to α to feasibly determine whether it was above or below, and one test of segmentability using the computational models where that was the case. For all such tests, I opted to take the more conservative approach and retain the

null. Throughout, I provided estimates of effect size (distance from the null mean in standard deviations) to facilitate interpretation of significant effects.[4]

Results. There are some contexts that are made up of only a very small number of utterances, potentially rendering calculations on those context subsets (such as type-token ratio and mean length of utterance) unreliable (Malvern & Richards, 1997). The analyses reported here are therefore restricted to contexts with at least 100 utterances in them. This excludes the contexts 'TV', 'touching', 'hiccups', 'taking pictures', and 'outside' from the subjective coder judgment approach and 'media' from the word list approach, composed of between 2 and 23 utterances each.

Descriptive Statistics of Contexts versus Nontexts. I assessed the context subsets on several measures that may be related to how easily a language sample can be correctly segmented into words. These measures included repetition (type-token ratio), proportion of isolated words, mean utterance length in number of words, and the proportion of tokens accounted for by the most frequent type (an index of how skewed the frequency distribution is).

Across methods and contexts, there was a substantial difference in repetition (type-token ratio, TTR) in context subsets compared to subsets with the same number of utterances taken randomly from the corpus. Type-token ratio is strongly related to corpus size, and so naturally varies substantially across context sub-corpora (0.09 to 0.37). In order to aggregate results across context subsets, context estimates were converted into Z-scores using the mean and standard deviation measured in that context's size-matched bootstrapped null distribution. This

[4]Note that this is effectively a Z-score, but because the distributions do not match a theoretical normal distribution, significance tests for Z-scores do not apply (i.e. it is not necessarily the case that 5% of cases fall outside of 1.96 standard deviations from the mean).

means that each context was represented according to where it fell within its own bootstrapped sampling distribution; it represents how extreme that context estimate was relative to estimates calculated on random samples of the same number of utterances. Under the null hypothesis that sampling utterances by context would be no different from sampling utterances randomly, the estimates from the context sub-corpora should generally have fallen within the bulk of the distribution of estimates from size-matched random sub-corpora. In the case of repetition (TTR), it is clear that the context sub-corpora were not at all typical in the distribution of random sub-corpora; they were several standard deviations more repetitive than typically occurred in random sub-corpora of the same size. In all three approaches to defining context, context subsets had dramatically lower type-token ratio (more repetition) than size-matched random samples, with Z-scores ranging from -4.17 to -10.73, as depicted in Figure 8. This underscores the fact that — across all three approaches to defining contexts — the resulting context subsets were composed of utterances that were more similar to each other, reusing a smaller number of unique words, than would be expected by chance.

In contrast, there was no systematic difference between context subsets and their size-matched random null distributions on what proportion of the tokens were the most frequent word, an index of skew in the frequency distribution of word types. As shown in Figure 9, there was variability across the samples, of course, but the context estimate Z-scores were spread across the bootstrapped null distributions, with 54% falling within 2 standard deviations of their null distribution means. Nine contexts had estimates significantly above their null distributions from the topic modeling approach approach, but five were significantly below. The more extreme estimates did not fall above 4.38 or below -4.32

standard deviations from their means. Overall, some contexts showed significantly higher concentration of the most frequent word than chance (more skew) and others had significantly less than chance, but fell well within their null distributions. This may have been due to the fact that the most common words in any sample of infant-directed speech are unlikely to be context-specific. In this corpus, the word "you" emerged as the most frequent type in almost every sub-corpus, regardless of size. If there were differences in the shape of the frequency distributions by context, it seems unlikely that a measure relying on the most frequent type would be sensitive to those differences. Instead, it may be necessary to characterize the skewness of the distribution more completely (which would require larger samples than many of the smaller context sub-corpora examined here).

Instead of a broad, general difference between contexts and "nontexts" (or lack thereof), the pattern of results for the proportion of one-word utterances (isolation) varied by approach to defining context. Context estimate Z-scores from the subjective coder judgment contexts (shown in Fig. 10) were mostly small, with 67% falling within 2 standard deviations of their null distribution mean. The contexts that did show a significant difference from their null distributions trended toward a comparatively lower proportion of one-word utterances, although the effects were modest (the most extreme context, mealtime, was 3.37 standard deviations below its null distribution mean). Fewer of the context estimate Z-scores from the topic modeling contexts were small, with 25% falling within 2 standard deviations of their null distribution mean. The contexts that differed significantly from their null distributions did not show a clear overall trend, however, with 6 contexts showing a proportion of one-word utterances significantly lower than would be expected by chance, and 3 contexts significantly higher than chance,

ranging from at most 8.31 standard deviations below the null mean (topic 7) to as much as 8.38 standard deviations above it (topic 3). Contexts from the word list approach, on the other hand, showed a more robust pattern. Only 1 context (fussing) was within 2 standard deviations of its null mean. The rest of the contexts all showed significantly less isolation (lower proportion of one-word utterances) than would be expected by chance, with estimates ranging from 2.28 to 6.21 standard deviations below their null means. In general, it appears that contexts generated using the word list approach and (to a lesser extent) the subjective coder judgments approach tended to have fewer one-word utterances compared to random sub-corpora of the same size. Contexts from the topic modeling topics varied widely, with some showing significantly higher proportions of one-word utterances than would be expected by chance and others showing significantly lower proportions of one-word utterances.

Differences in utterance length (mean number of words per utterance) also varied by approach to defining context (see Fig. 11). As with the proportion of one-word utterances, context estimate Z-scores for utterance length from the subjective coder judgment contexts were mostly small, with 78% falling within 2 standard deviations of their null distribution mean. Two contexts did display a significant difference from their null distributions, housework and mealtime, trending toward comparatively longer utterances, with estimates 6.59 and 3.36 standard deviations above their null means, respectively. Context estimates for utterance length from the topic modeling topics were mixed, as was the case for proportion of one-word utterances. 25% of the topic modeling context estimates fell within 2 standard deviations of their null distribution means. The contexts that did differ significantly from their null distributions were more or less evenly split

in terms of the direction of the effect, with 4 contexts showing a mean utterance lengths significantly lower than would be expected by chance, and 5 contexts significantly higher than chance, ranging from at most 7.51 standard deviations below the null mean (topic 2) to as much as 6.87 standard deviations above it (topic 10). Echoing results on the proportion of one-word utterances, contexts from the word list approach, showed a clear trend towards longer utterances. Only one context (fussing) was within 2 standard deviations of its null mean. The rest of the word list contexts all had significantly longer utterances (mean number of words per utterance) than would be expected by chance, with estimates ranging from 11.03 to 5.57 standard deviations above their null means.

Taken together, these results suggest that contexts generated using the word list approach and (to a lesser extent) the subjective coder judgments approach tend toward longer utterances, with a lower proportion of words in isolation and higher mean number of words per utterance compared to random sub-corpora of the same size. Contexts from the topic modeling topics vary widely, with some contexts characterized by longer utterances on average and a lower proportion of isolated words, while other contexts show the opposite pattern. These two measures are closely related ($r(26) = -0.77$, $p < .001$), as could be predicted by the fact that increasing the proportion of one-word utterances would naturally lower the mean utterance length for a given corpus. While context sub-corpora do not appear to differ systematically from randomly sampled sub-corpora on skew (the proportion of all tokens belonging to the single most frequent type), across the board they show dramatically more repetition (lower type-token ratio).

Segmentability of Contexts versus Nontexts. Results from the adaptor grammar (Börschinger et al., 2012; M. Johnson, 2008) showed very little

evidence of an advantage for context-specific sub-corpora compared to random samples of the same number of utterances. Just as with the corpus descriptive estimates (type-token ratio, etc.), estimates from context-specific sub-corpora were expressed as Z-scores, computed using the mean and standard deviations from their bootstrapped sampling distributions. As shown in Figure 12, the F-scores from the adaptor grammar for context-specific sub-corpora mostly fell within their bootstrapped null distributions, with nearly all of them falling within 2 standard deviations of their null distribution mean. There were a few contexts with token F-score estimates significantly above their null distributions, including one from the word list approach (play) and three from the topic modeling approach (topic 2, topic 3, topic 7), of which one falls at least 2 standard deviations above its null mean (topic 2, top words: bop, hello, monkey, give, hi), with an F-score 2.81 standard deviations above its null distribution mean. As a few isolated cases in a set of many tests of the same hypothesis, these few significant departures would need to be replicated before they can be interpreted as reliable.

Results from the HDP bigram model (Goldwater et al., 2009) are mixed. For contexts from the topic modeling topics or subjective coder judgments, there is no strong evidence of an advantage for context-specific sub-corpora compared to random samples of the same number of utterances. As shown in Figure 13, the F-scores from the HDP bigram model for context-specific sub-corpora fall well within their bootstrapped null distributions, with 90% of them falling within 2 standard deviations of their null distribution mean (all of which are also non significant by the bootstrapped p value). The contexts resulting from tagging utterances by the occurrence of key words showed a different pattern, however, with all context F-scores falling above their null distribution means, by 0.83 to 2.96

standard deviations, 4 of which are significantly above their null distributions by the bootstrapped p value, with 3 of them falling more than 2 standard deviations above their null distribution means.

On the one hand, the findings regarding word-list-related context-driven segmentability as measured by the HDP model may indicate that the context sub-corpora were indeed more segmentable, but the HDP model was more sensitive to that difference than the adaptor grammar and the effect was small and fragile enough not to be detected with the other approaches to defining context. On the other hand, a clear and plausible alternative account is available that would render these findings considerably less interesting. The selective effect for the word list contexts may be driven by the fact that the word list approach to defining contexts had a clear tendency to systematically select for longer utterances, unlike the topic modeling and coder judgment approach. Indeed, there was a strong correlation between Z-scored model performance (token F-score) and Z-scored mean utterance length for the HDP bigram model, $r(26) = 0.68$, $p < .001$, but not for the adaptor grammar, $r(26) = 0.04$, $p = 0.836$, as displayed in Figure 14. Recall that the adaptor grammar can learn information about syllable structure, taking advantage of rich phonotactic information at the edges of utterances; this may counteract what would otherwise simply be a loss of information due to shorter utterances. The HDP bigram does not use phonotactics, possibly resulting in better performance on corpora with relatively longer utterances; this is as yet only an intriguing hypothesis, which is not conclusively explored here and could easily be a study in its own right. It is consistent with the results observed here, however. The only advantage for context-specific sub-corpora arose for one approach to defining context (by key word occurrence) and for one model (HDP bigram), suggesting it

may have been an interaction between an artifact of the sampling procedure in that approach to defining context and the particulars of the HDP bigram model.

Discussion. Taken together, these results suggest that context sub-corpora were no more easily segmentable than would be expected by chance. The lack of evidence for improved segmentability in context-specific subsets using the adaptor grammar in particular may appear surprising in light of other recent findings showing improved token F-scores from the very same model when activity context information is provided in the form of context labels for utterances derived from topic modeling topics (Synnaeve et al., 2014). Importantly, Synnaeve and colleagues found that while the versions of the model that could build separate context-specific vocabularies did out perform implementations that could not (including the model used in this study), this advantage only appeared after the model had had access to a sufficiently large input corpus, at least about 10,000 utterances from the Naima section of the Providence corpus. They interpreted this as a natural consequence of the fact that the context-sensitive models had more complex structure to learn (several vocabularies rather than just one), requiring more data. In the current study, the models were the same whether applied to context subsets or random subsets; there is no reason that the models should require more input when applied to context subsets relative to random subsets. However, it may still be the case that larger corpus samples would reveal a context advantage for the adaptor grammar. Any difference in segmentability in context-specific sub-corpora compared to random sub-corpora could be reinforced and compounded with larger and larger sample sizes, leading to a clearer difference between segmentability of context-specific sub-corpora and size-matched random sub-corpora.

Figure 8. Repetition (type-token ratio, TTR) in context sub-corpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

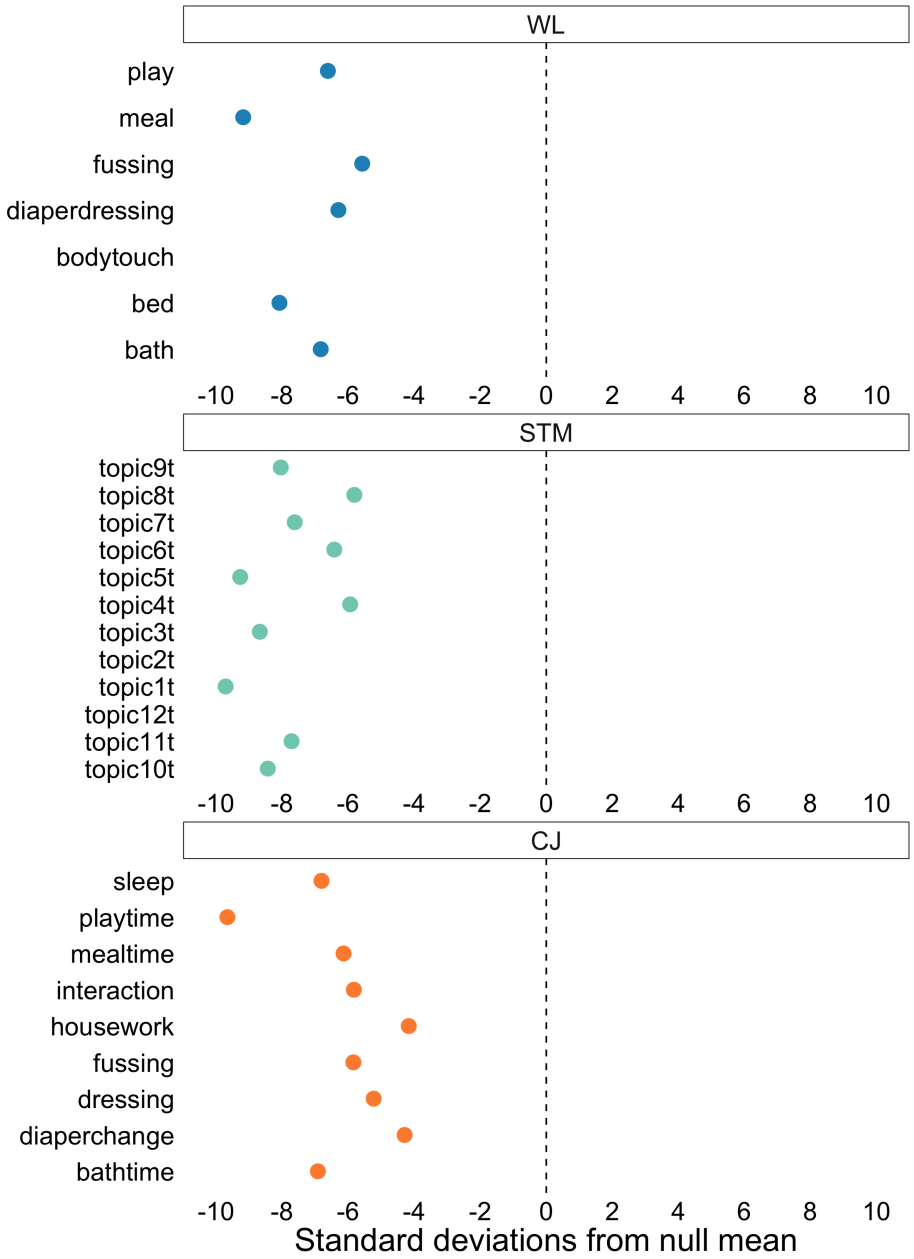


Figure 9. Skew (proportion of tokens accounted for by the single most frequent type) in context sub-corpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

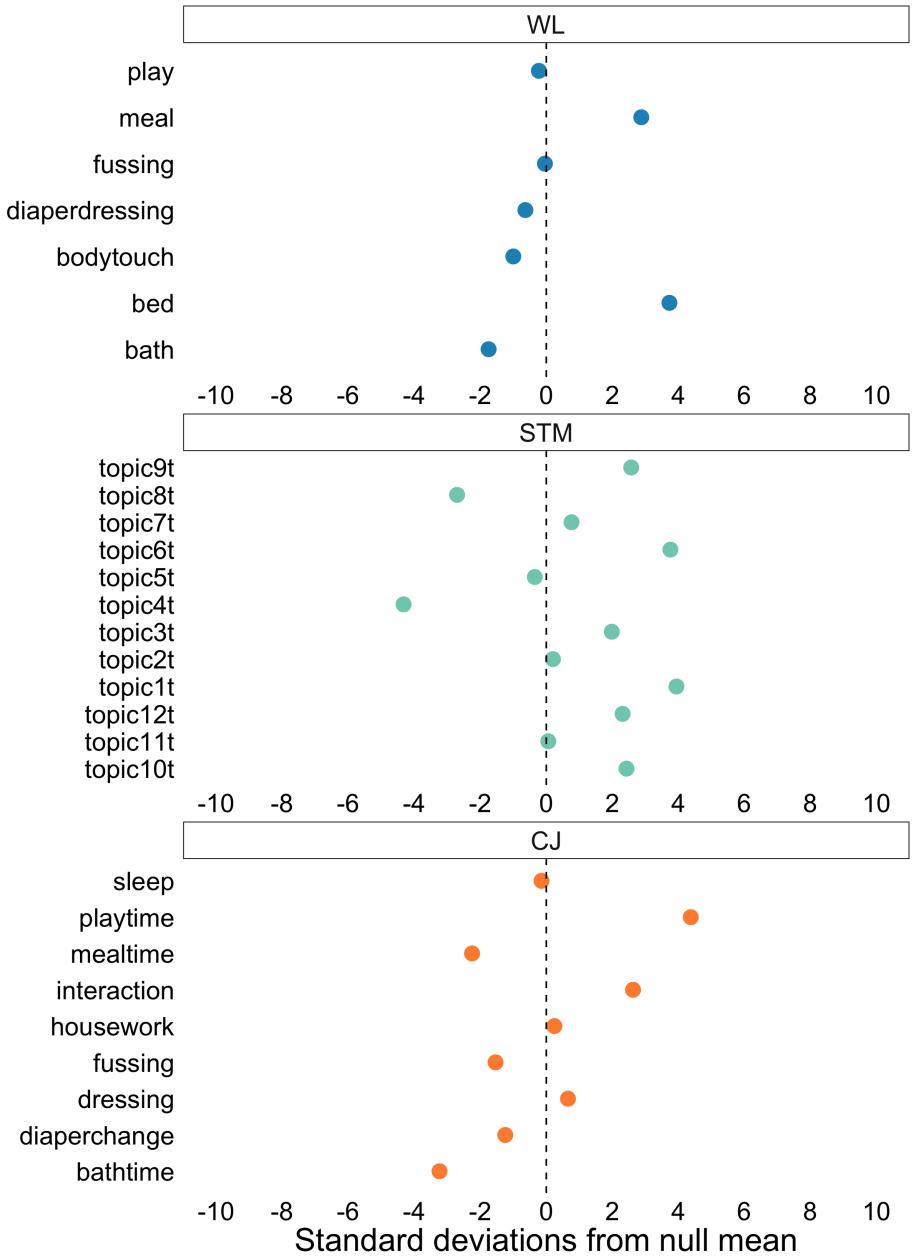


Figure 10. Isolated words (the proportion of one-word utterances) in context sub-corpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

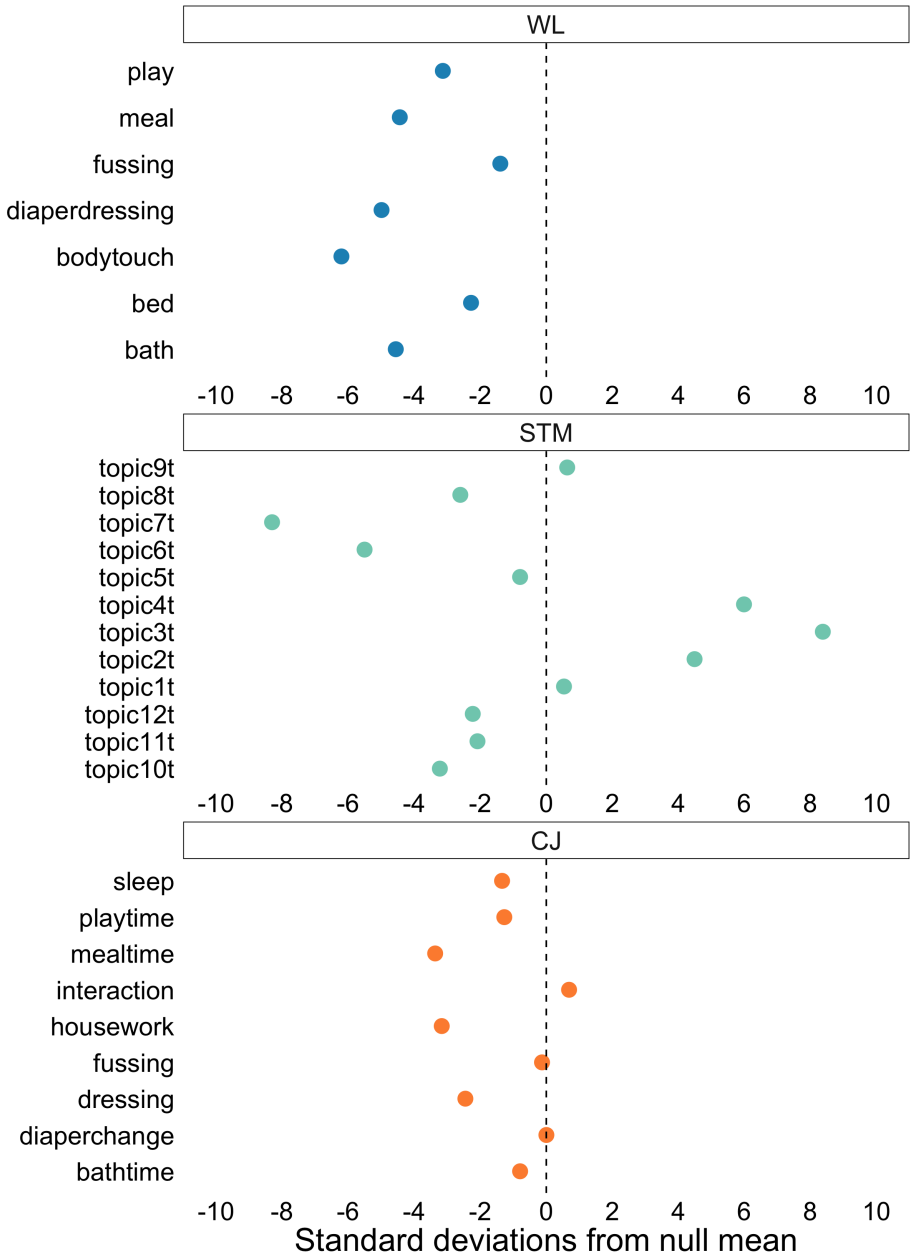


Figure 11. Utterance length (mean number of words per utterance) in context sub-corpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

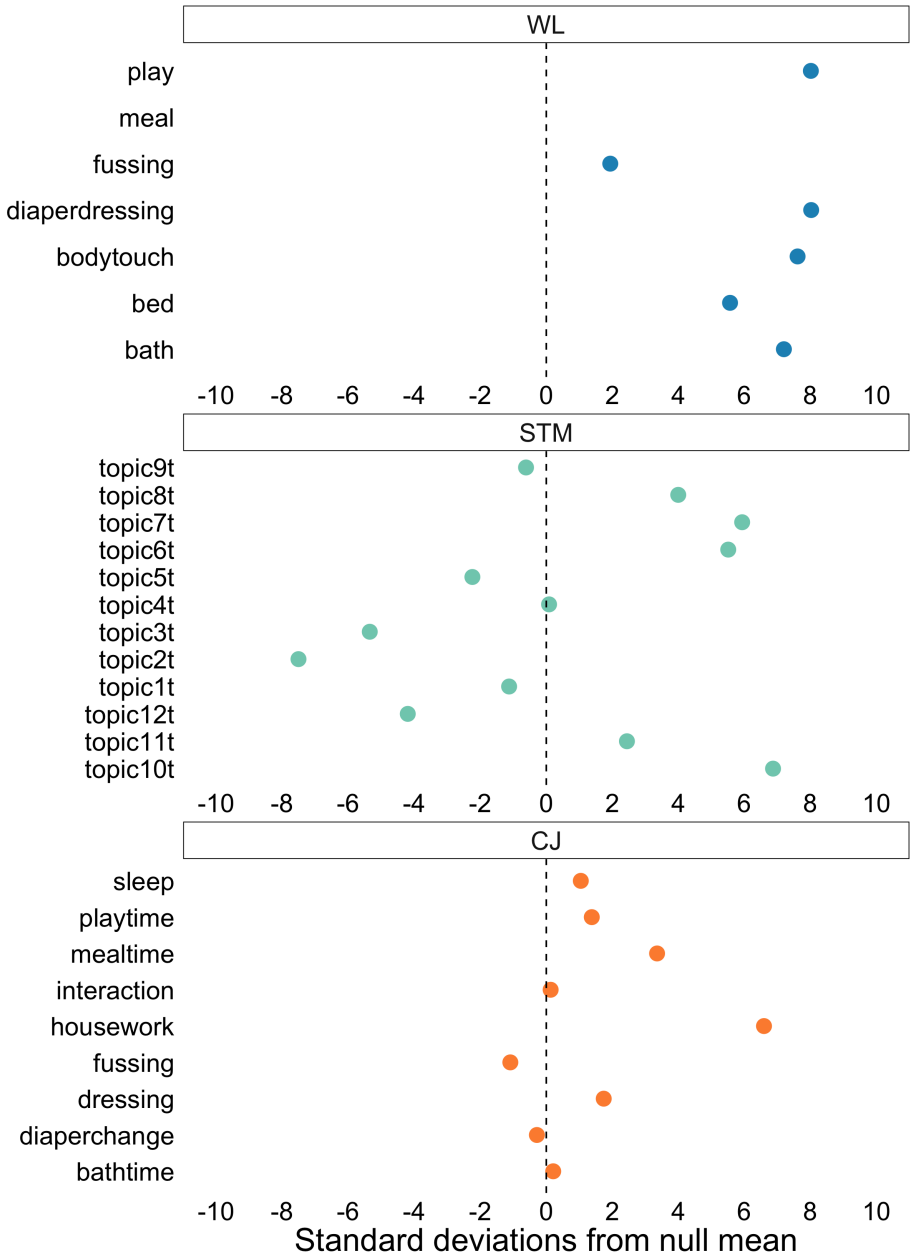


Figure 12. Segmentability (the adaptor grammar token F-scores) of context sub-corpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

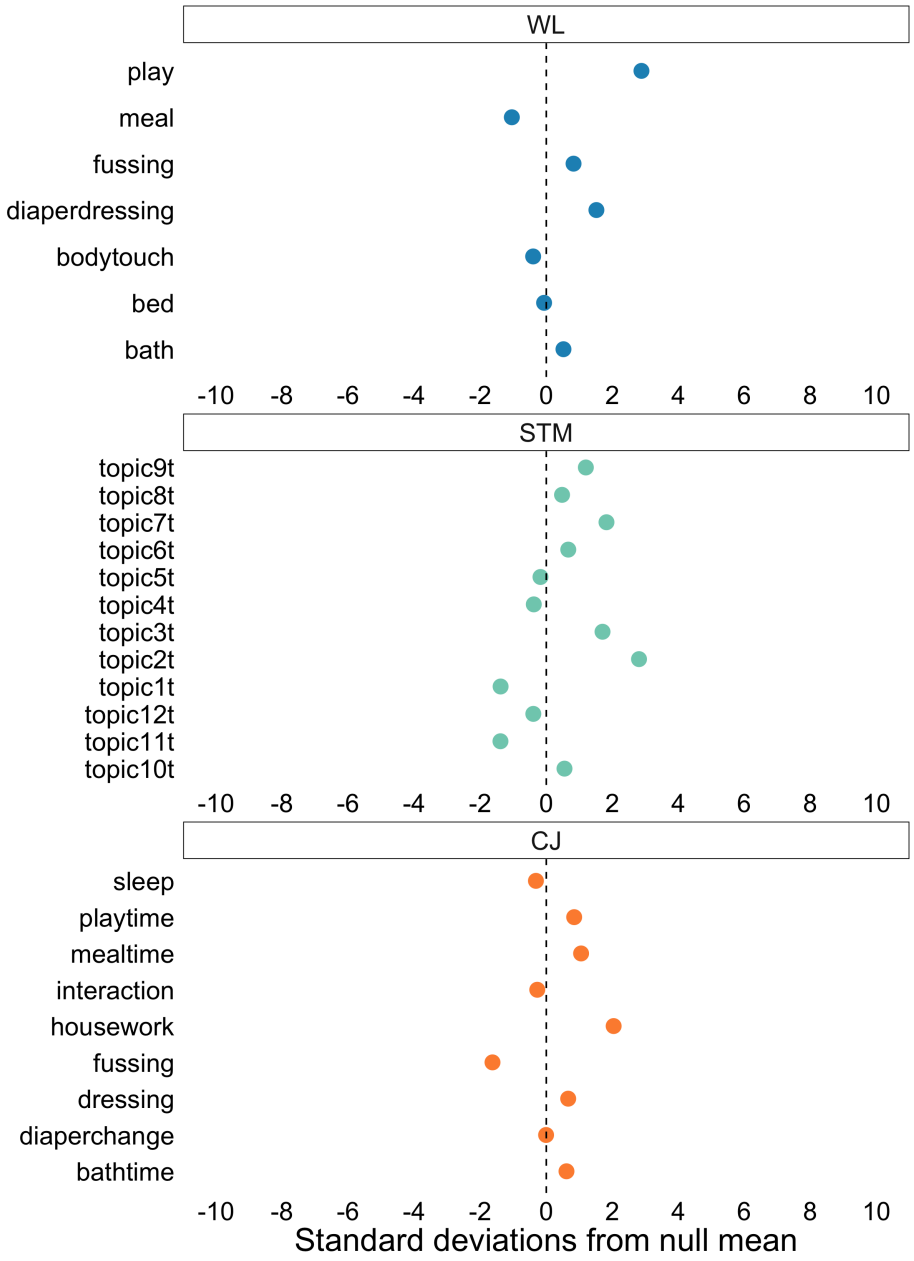


Figure 13. Segmentability (the HDP bigram model token F-scores) of context subcorpora, as compared to size-matched bootstrapped null distributions. Plotted by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments. To facilitate comparison across contexts, each context estimate is standardized (Z-scored) using the mean and standard deviation of its null distribution.

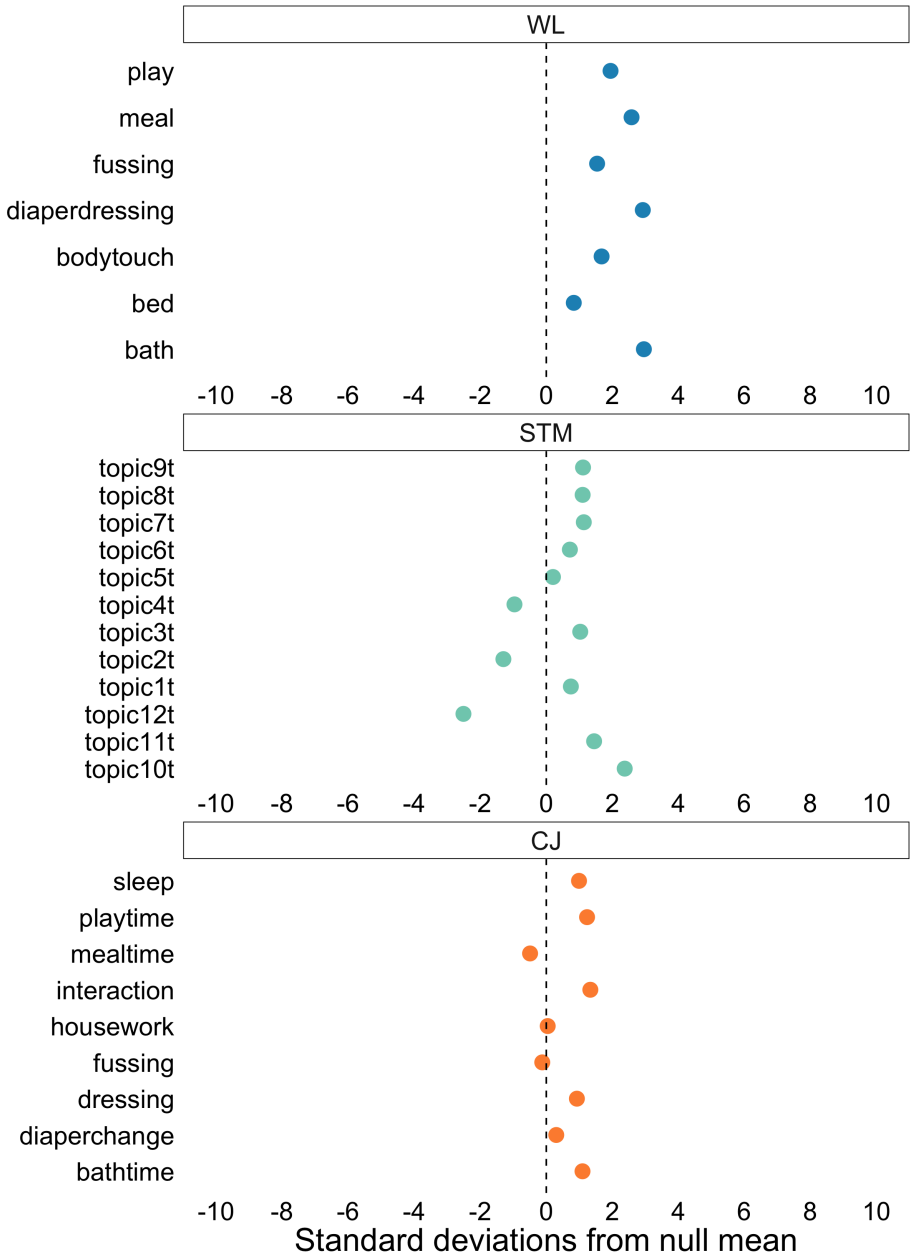
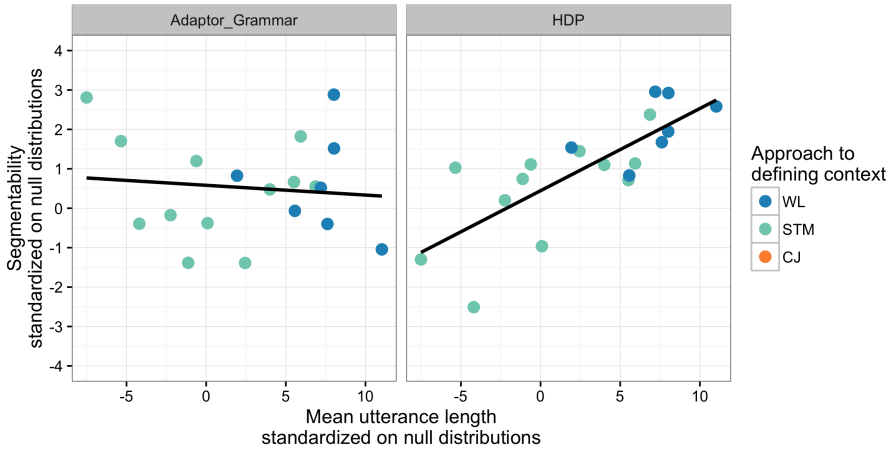


Figure 14. Segmentability and mean utterance length, both standardized (Z-scored) using the mean and standard deviation of their null distributions. Colored by approach to defining contexts: WL = Word List, STM = Structural Topic Modeling, CJ = Coder Judgments.



CHAPTER IV

GENERAL DISCUSSION

The goal of this project was to explore the role of activity contexts in the structure of infant-directed speech. To investigate this, I began by defining contexts three different ways (by the occurrence of key words, with topic modeling, and with subjective judgments by coders) and measuring the extent to which the three approaches identified the same underlying contexts. The fact that there was substantial agreement among the three different approaches to defining context supports the conclusion that each method measured (imperfectly) the same underlying construct, interpreted as activity contexts. The contexts from the three approaches also diverged, however, highlighting possible method artifacts introduced by each method. I then used the context-specific sub-corpora from each approach and analyzed the speech within each context to test the hypothesis that statistical cues to word boundaries may be clearer within contexts, a hypothesis that was not supported by the data. Together, this series of analyses fills a gap in the literature by providing a detailed comparison of multiple approaches to defining context and contributes to a growing body of evidence on how infants may apply statistical learning to parse the speech they hear in their natural environments.

Defining Contexts

Activity contexts have been defined many ways in the language development literature, including time of day (Roy et al., 2015; Soderstrom & Wittebolle, 2013), conversational topic (S. Frank et al., 2014; Roy et al., 2015; Synnaeve et al., 2014), conversational partner (Brown-Schmidt, Yoon, & Ryskin, 2015; Hoff, 2010), coder judgments (Soderstrom & Wittebolle, 2013), researchers' decisions about when to record (Bruner, 1975; Hoff-Ginsberg, 1991; Weizman & Snow, 2001),

and parental report (Fausey et al., 2015). Since it is not clear from the existing literature how closely different methods for diagnosing context would map onto infants' own subjective experience of context, it was not possible to assess my three approaches to defining context against an obvious and accepted criterion. Instead, I examined patterns in how the three methods I opted to undertake (i.e., defining context by the occurrence of key words, with topic modeling, and with subjective judgments by coders) related to each other in order to build evidence about how each relates to the unmeasurable latent construct 'context', i.e. their construct validity (Cronbach & Meehl, 1955). The comparison of contexts from all three approaches on the same corpus of infant-directed speech is a valuable methodological contribution for those wishing to add activity context data to corpus analyses of child-directed speech, as has become increasingly common. I found that the choice of approach to defining context has consequences for the following: what proportion of the corpus was coded for context, systematic biases in the features of utterances selected for, and interpretability of the resulting contexts. Implications of each of these differences are explored in more depth below.

The three approaches to defining context yielded very different levels of coverage of the corpus, with topic modeling resulting in context codes for nearly all of the utterances while coder judgment contexts and word list contexts covered roughly half and one third of the corpus, respectively. This raises the question, "What proportion of a corpus *should* we expect to be coded for context?" The answer no doubt depends on a number of factors including the nature of the corpus itself (day-long recordings at home will differ in the rhythm of contexts relative to 30-minute sessions in the lab) and the granularity of coding (coding large blocks

of time with the "main" context occurring during that block may overestimate duration of each context if, for example, a 20 minutes of play during a 30-minute block of time means the whole 30 minutes are coded as 'play'). Crucially, it also depends on what "context" is taken to mean. In the language learning literature, context often refers to clear, (relatively) discrete activities, perhaps with periods of less well-defined time — such as Soderstrom and Wittebolle's 'transition time' — in between (Bruner, 1975; Fausey et al., 2015; Hoff-Ginsberg, 1991; Soderstrom & Wittebolle, 2013; Weizman & Snow, 2001). The use of topic modeling to infer contexts is an extension of tools designed for text analysis (Blei et al., 2003), where there is no expectation that some documents will simply lack a true underlying topic. Contexts resulting from the application of topic modeling to corpora of speech may diverge from contexts from other approaches because the topic model attempts to explain *the entire* corpus with topics, including sections that may be judged by other methods to be context-less 'transition time'. In the best case scenario, all of the 'transition time' is similar enough in word use patterns that the topic model captures it with a limited number of identifiable 'garbage' topics [1]. A more problematic situation could occur if word use patterns during 'transition' times between contexts are not similar enough to each other; this could potentially result in a large number of 'garbage' topics targeting different kinds of transition times, or a distortion of other topics that might otherwise cleanly capture more meaningful contexts. The data presented here do not speak directly to *why* the contexts from one approach differ from those from another, but it is clear that contexts from topic modeling include many more utterances than contexts

[1]For example, see the 'garbage' topic reported in Synnaeve et al. (2014). Similarly, Roy et al. (2012) provide the top words for only 16 of the 25 topics they estimated, suggesting that the remaining topics were less interpretable.

identified by the occurrence of key words or as judged by coders. This difference in coverage certainly contributes to divergence between the topic modeling contexts and those from the other two approaches (and, indeed, the level of agreement, as measured by Cramer's V, is lower between topic modeling contexts and either of the other two approaches than between word list contexts and coder contexts). Future investigations could more thoroughly address the question of how to identify sections of a corpus that exist between contexts (or, more to the point, between contexts that are of interest to the researcher), and how to use that information to guide the inference of contexts using topic modeling.

Another difference between the three approaches to defining context was the introduction of systematic biases in the kinds of utterances included in context subsets. Unlike the topic modeling approach or the coder judgment approach, the word list approach selected for longer utterances (or possibly just fewer one-word utterances). This was likely because the key words themselves (originally taken from the Oxford CDI) tended to occur in longer utterances. This highlights the subtle effects of the choice of key words. I selected words from the Oxford CDI that could be transparently related to one of the context categories; words with no obvious activity context association (e.g. 'all', 'big', 'look') were excluded. The content-heavy words that made it onto the word lists were mostly nouns, verbs, and adjectives (e.g. 'bath', 'clean', 'soap', 'wash' were on the list for bath time words). This had the unintended side-effect of biasing utterance selection in the word list approach towards utterances with those kinds of words, which tended to be longer. Of course, a different set of key words (or this set of key words applied to a new corpus) may or may not result in the same bias toward longer utterances. What this demonstrates is the value of a multi-method approach to defining a latent

construct like context. A simpler study, restricted to a single approach to defining context, would have been unable to diagnose the increase in average utterance length (and the associated increase in segmentability using the HDP model) as an effect of the word list method in particular rather than an effect of context *per se*.

An additional advantage of using multiple approaches to defining context on the same corpus is that it provides a check on what would otherwise be the default interpretation of each context. The word list approach began with a set of contexts and words associated with them and built sub-corpora of utterances containing those words (and the utterances immediately around them); the assumption was that the sub-corpora would reflect the original context categories. The topic modeling approach was almost the reverse process — it began with the words that occurred together in the corpus and inferred topics based on that, which are then traditionally defined by the key words that are most closely associated with them (Blei et al., 2003). In each approach, interpreting contexts for each method in isolation relied on the assumption that the context categories reflected the activity contexts happening in the utterances that made up those sub-corpora. By comparing contexts across different approaches, I was able to check the extent to which similar contexts from different approaches identified the same utterances in the corpus. This was especially apparent in the latent class analysis (LCA), which compared contexts from all three methods simultaneously. For the most part, contexts across approaches lined up in predictable ways and the classes identified by the LCA grouped together sensible contexts. For example, class 1 had high probabilities for the 'bath-time' context from coder judgments, the 'bath' context from the word list approach, and topics 7 and 8 from the topic modeling approach,

both of which included top words associated with bathing and water. Comparison was not always so straightforward, though — class 2 grouped together the 'play' and 'body touch' (cuddling, tickling, etc.) contexts from the word list approach with the 'playtime' context from the coder judgments, but the tickling context from the topic modeling (topic 5) was much lower probability. Instead, topic 10 was the highest probability for that class, defined by words that suggest a context like 'scolding' (naughty, bite, look). In this case, interpretation based on the top words for these topics clashed with how utterances in that topic were interpreted by coders[2]. This underscores the limitations of interpreting topic modeling topics based solely on the top words for each topic. Some recent topic modeling software (the stm R package Roberts, Stewart, & Tingley, 2016) has additional functionality designed to ameliorate this issue, facilitating the retrieval of documents or parts of documents that are representative of a given topic. This analysis provides insight into similarities and differences in approaches to defining contexts from transcripts, enabling future researchers to make more informed decisions about which approach makes the most sense for their needs and with their resources.

Cues to Word Boundaries within Contexts

Examining agreement among the three different approaches to defining context provided evidence for the validity of each as a measurement of activity contexts in transcripts of infant-directed speech. The second series of analyses built on that groundwork by illustrating an application of context information in a corpus analysis of statistical cues to word boundaries. This analysis was motivated

[2] A post-hoc justification is inviting: Caregivers' use of words like 'naughty' with their infants may often be playful, especially for infants as young as those in this corpus (1-4mos), in which case there is no conflict between the topic modeling context and those from the other two methods. This may very well be the case. The point stands that the interpretation of topic modeling topics from the top words *in the absence of alternative context codes for those utterances* may be misleading.

by the hypothesis that context-specific subsets of the corpus may be more easily segmentable than the corpus as a whole because of the increased repetition / decreased lexical diversity in context-specific subsets compared to the corpus as a whole — the more homogeneous, coherent speech within contexts may provide richer information about the statistical dependencies among phonemes than is available when analyzing the same statistical dependencies without respect to context.

In order to compare structure within contexts to general structure in the corpus, I used resampling to build null distributions for each measure in each context. This allowed me to test for differences between the context sub-corpora and the whole corpus while controlling for the number of utterances contributing to each measure.

For the most part, context sub-corpora were no more easily segmentable than random sub-corpora of the same size. The clear exception was contexts from the word list method, which were significantly more easily segmentable than random sub-corpora of the same size, but only when using the HDP model, not the adaptor grammar. While this may suggest that a context-based segmentation may indeed be easier than non context-based segmentation under some circumstances, a plausible alternative hypothesis is that the longer utterances selected for by the word list contexts provided an advantage for the HDP model.

There are a few explanations for why I may have found no compelling evidence in support of the hypothesis that speech within contexts is more easily segmentable. One possibility is that there was a real difference in the segmentability of context sub-corpora compared to size-matched random sub-corpora, but that the difference was too subtle for the computational models to

capitalize on it with so little data. Synnaeve et al. (2014) found that providing context information (topics from a topic model) to the adaptor grammar improved its segmentation, but only after the model had had access to a sufficiently large input corpus, at least about 10,000 utterances in their study. The entire Korman corpus is 13,000 utterances, so none of the context sub-corpora approached the size at which Synnaeve and colleagues noted a reliable context advantage. Unfortunately, the Korman corpus is the largest publicly available English corpus with infants in the first half of their first year — the time during which infants' experience processing speech would be crucial for laying the groundwork for word segmentation. Composed of two day-log recordings and three shorter recordings each for five infants, this corpus only includes the equivalent of up to[3] about two weeks of infants' natural experience, but exciting new efforts to improve organization and sharing of extensive, natural recordings (HomeBank: Warlaumont, VanDam, & MacWhinney, 2015) will make larger corpora available in the coming years. Run on a much larger corpus, it is possible the computational models would begin to show a segmentation advantage for context sub-corpora.

Another possibility is that any advantage of context-specific segmentation depends on also being able to integrate cues across contexts, using both context-specific and global patterns. While Synnaeve et al. (2014) found that splitting the adaptor grammar processing by context improved segmentation, the best segmentation was achieved by a model that maintained both context-specific vocabularies and a global vocabulary. In the current study, contexts were completely separated from each other during analysis, so the computational models could not combine local and global information to make segmentation decisions.

[3]The total time is potentially much less, since mothers had the recorders for 24 hours but sometimes turned them off for unknown lengths of time.

A third possibility is that the process for defining context was too noisy, so that the resulting sub-corpora were, in fact, essentially random samples from the corpus. That would result in more or less the pattern of results observed: estimates of segmentability within context sub-corpora fall within the bulk of the distribution of estimates of segmentability from random sub-corpora of the same size. Two facts make this explanation unlikely, however. The first is that the context sub-corpora *did* differ significantly from their null distributions on other measures, in particular type-token ratio. This suggests that context sub-corpora are not effectively random, but rather highly homogeneous and repetitive (lower type-token ratio) relative to random samples of the same size. The second relevant fact is that I observed substantial agreement across methods in context codes. The convergence of all three methods suggests that they were each (to some degree) tapping into the same underlying construct. While it is unlikely that the context sub-corpora were effectively random samples, there definitely was some amount of noise. It is possible that the degree of noise in the context defining process was enough to dramatically reduce power to detect any difference between context sub-corpora and random sub-corpora. Depending on the level of error in context assignment and the true size of the difference (which could be close to zero) in segmentability for contexts compared to random sub-corpora, noise in context assignment could be driving the null results. As with concerns around the sensitivity of the computational models on small corpora, the solution for this issue would be to conduct these analyses on a much larger corpus.

Limitations and Future Directions

Unlike other recent work on the effects of context in language acquisition (e.g., Roy et al., 2015), my three approaches to defining context all relied solely on

the transcribed speech. This is a limitation in that there is undoubtedly relevant information that was not available (e.g. having video available would certainly improve the human coding judgments of context), but it is also a strength in that the methods presented here are all readily applicable on any existing transcribed corpus. Most of the corpora available on CHILDES do not include video, either because it was never recorded or because the researchers do not have participants' permission to share videos, which pose a much greater risk to privacy violations than transcribed speech. It is also important to remember that there is no data source (speech, video, location, time of day, parent report, etc.) that would provide the 'ground truth' of context from the infant's perspective; any attempt to measure context will be a proxy. The research presented here speaks to how well we can infer activity context from transcribed speech alone. Many researchers only have access to transcribed corpora, and even in cases with ready access to rich, multimodal data, the resources required to recode it for context information may be prohibitive, making the analysis of the affordances of transcribed speech on its own an important area of investigation. An important future direction will be replicating these analyses on a corpus that also provides additional data sources to define contexts (such as video, location, or time of day), allowing the analysis of context-specific patterns in the speech stream when the contexts themselves are not also defined by the speech stream.

The corpus used in this study is represented using phonetic approximations (i.e. dictionary pronunciations) of the transcribed speech. This phonetic approximation is a simplified and idealized version of what the actual phonetic material would have been — a given word is represented with the identical pronunciation each time it occurs, ignoring probable irregularities due to co-

articulation, prosody, etc. Crucially, this makes the (obviously questionable) assumptions both that infants can use adult-like phonemic categories to identify speech sounds (e.g. distinguishing the difference between *ba* and *pa*) and that they recognize repetitions of phonemes over time (e.g. recognizing *cat* in both *your cat* and *that cat*). Neither of these assumptions are likely to be unambiguously met. There is a tremendous amount of variability in how a given phoneme is realized in natural speech, depending on the sounds that come before and after it, the speaker, prosody, affect, and so on. Moreover, infants' phonemic categories are not adult-like until much later (Rivera-Gaxiola, Silva-Pereyra, & Kuhl, 2005; Werker, Yeung, & Yoshida, 2012), so it is unlikely that they would ignore within-phoneme variability the way adult listeners do. Because the corpus analysed here is represented in idealized phonetic approximations, it is unclear the extent to which these results will bear on the structure of infants' real language exposure. An important future direction will be analyzing the statistical cues to word boundaries in the raw acoustic signal, rather than in transcribed speech. There have been several exciting developments on this front in recent years (e.g., McInnes & Goldwater, 2011; Räsänen, 2011, 2014; Räsänen, Doyle, & Frank, 2015), demonstrating that the principles of statistical learning can successfully be applied to a more realistic acoustic stimulus and still identify word boundaries. As tools for analyzing raw acoustic information improve, the analyses presented here could be replicated using more realistic models of infants' speech process. Recordings of caregivers' speech could be divided into context-specific sub-corpora and the raw acoustic material processed for cues to word boundaries as a measure of segmentability, instead of applying models that use phonetic transcriptions, as was done here.

Another limitation of this work is that it attempts to isolate one facet of language acquisition — word segmentation — and study it as a process independent of all of the other learning (linguistic and otherwise) that 1- to 4-month-old infants are engaged in. An important direction for this and any other such research is to bring models of infant learning closer to the rich, complex problems infants actually face, and to flesh out descriptions of the structure in infants experience to better approximate the data infants actually have to work with. The incorporation of activity context information is one step, but the data and analyses reported here are still far from capturing infants' processing of speech as it would occur naturally.

Given the current discrepancy between the results presented here and the advantage of context-specific word segmentation found by Synnaeve et al. (2014), the most immediate next step should be to investigate that mystery. This could be undertaken in several ways. One approach would be to extend Synnaeve and colleagues' analyses to other definitions of context. They used contexts defined by an unusual two-step topic modeling approach, but the models they used could be applied to any corpora with context-annotated utterances. Contexts could be defined using more traditional LDA topic modeling (as used by Roy et al., 2015), structural topic modeling (used in the present study), coder judgments, or context word lists. It is possible that the effects they observed are specific to the method they used to identify contexts; extending their analysis to other definitions of context could rule that out. If the key difference between Synnaeve and colleagues' implementation and my own is the greater sophistication of their implementation of the model (which learned segmentation in all contexts simultaneously and, in one case, built both context-specific vocabularies and global shared vocabularies,

allowing information from each context to support the others), then applying their model to additional approaches to defining context should yield the same context-specific advantage. A related line of research could extend the current study to larger corpora, such as the Providence corpus used by Synnaeve and colleagues, or the Thomas corpus (the largest English language corpus currently available on CHILDES). While neither Providence nor Thomas is ideal for investigating input to early word segmentation (they begin at 11 months and 24 months, respectively), their size would make it possible to more thoroughly examine the role of corpus size in the context advantage for word segmentation. It is possible that the key difference between the current study and Synnaeve and colleagues' analysis is not the flexibility of their model but simply the amount of input available. If so, application of the current methods to larger corpora should reveal an advantage for context-specific word segmentation. These and related studies would build understanding of the circumstances under which context-specific processing facilitates word segmentation, and provide additional insight into the mechanism of the effect.

Conclusion

This work makes several important contributions. It is one of the first studies to expand the scope of studies of infant word segmentation beyond patterns in the speech stream to include contextual cues, along with existing behavioral (Seidl et al., 2014) and computational modeling (Synnaeve et al., 2014) work. It is the only existing description of statistical cues to word boundaries by context in natural recordings of infant-directed speech, adding a new dimension to ongoing discussions of whether or not statistical regularities among syllables provide sufficiently strong cues to word boundaries for infants to begin to segment speech

(Swingley, 2005; Yang, 2004). Moreover, the results reported here provide unique insight into how different approaches to defining context relate to each other; the results from the analysis of agreement across context methods have the potential to be a useful resource for researchers studying context in infant-directed speech for a variety of applications, not just the study of statistical cues to word boundaries. Context likely plays a pervasive role throughout language acquisition (Bruner, 1975) — this work both moves forward our understanding of the role of context in one particular aspect of the structure in infant-directed speech, and also facilitates further work on this and other important questions in the study of language acquisition.

APPENDIX

CONTEXT KEY WORDS LISTS

context	key words
bath	bath, bath, bathroom, bathrooms, baths, bathtub, bathtubs, clean, clean, cleaned, cleaner, cleanest, cleaning, cleans, dried, drier, dry, drying, soap, soaped, soaping, soaps, soapy, splash, splashed, splashes, splashing, splashy, swam, swim, swimmer, swimming, swims, towel, toweling, towels, towled, wash, washed, washes, washing, wet
bed	asleep, bed, bedroom, bedrooms, beds, blanket, blankets, nap, napped, napping, naps, night, night_night, nights, pillow, pillows, sleep, sleepier, sleepiest, sleeping, sleepy, slept, tired
body_touch	arm, arms, belly_button, belly_buttons, cheek, cheeks, cuddle, cuddled, cuddles, cuddling, face, faces, feet, finger, fingers, foot, hand, hands, head, heads, hug, hugged, hugging, hugs, kiss, kissed, kisses, kissing, knee, knees, leg, legs, nose, noses, tickle, tickled, tickles, tickling, toe, toes, tummies, tummy, tummy_button, tummy_buttons
diaper_dressing	brush, brushed, brushes, brushing, button, buttoned, buttoning, buttons, comb, combed, combing, combs, dress, dressed, dresses, dressing, hair, hat, hats, jacket, jackets, jeans, jumper, jumpers, nappie, nappies, nappy, pjs, potties, potty, pyjamas, shirt, shirts, shoe, shoes, shorts, sock, socks, sweater, sweaters, trousers, wipe, wiped, wipes, wiping, zip, zipped, zipping, zips
fussing	bad, cries, cried, cry, crying, hurt, hush, nasty, naughty, sad, scared, shh, shush, sick, ssh
meal	all_gone, apple, apples, ate, banana, bananas, bib, bibs, biscuit, biscuits, bottle, bottles, bowl, bowls, bread, breakfast, butter, cake, cakes, carrot, carrots, cereal, cheese, chicken, chips, cup, cups, dinner, dish, dishes, drank, drink, drink, drinkies, drinking, drinks, drunk, eat, eating, eats, egg, eggs, fed, feed, feeding, feeds, fish, food, fork, forks, fridge, fridges, high_chair, high_chairs, hungry, ice_cream, jam, juice, lunch, meat, milk, milky, orange, oranges, pasta, peas, plate, plates, refrigerators, refrigerator, spaghetti, spoon, spoons, sweet, sweets, tea, tea, thirsty, toast, yum, yummy
media	radio, television, TV
play	ball, balloon, balloons, balls, block, blocks, brick, bricks, buggies, buggy, dance, danced, dances, dancing, doll, dolls, fire_engine, fire_engines, jump, jumped, jumping, jumps, pat-a-cake, peekaboo, play, play_pen, play_pens, played, playing, plays, pushchair, pushchairs, ride, rides, riding, rode, sang, sing, singing, sings, sung, swing, swinging, teddies, teddy, teddy_bear, teddy_bears, threw, throw, throwing, throws, toy, toys, train, trains

REFERENCES CITED

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley.

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). doi: 10.1371/journal.pone.0007678

Altvater-Mackensen, N., & Mani, N. (2013). Word-form familiarity bootstraps infant speech segmentation. *Developmental Science*, *16*(6), 980–990. doi: 10.1111/desc.12071

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, *116*(3), 463–498. doi: 10.1037/a0016261

Arnon, I., & Clark, E. V. (2011). Why Brush Your Teeth Is Better Than Teeth Children's Word Production Is Facilitated in Familiar Sentence-Frames. *Language Learning and Development*, *7*(2), 107–129. doi: 10.1080/15475441.2010.505489

Baayen, R. H., Shaoul, C., Willits, J. A., & Ramscar, M. (2015). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, Cognition, and Neuroscience*.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory Redundancy Guides Attentional Selectivity and Perceptual Learning in Infancy. *Developmental Psychology*, *36*(2), 190–201. doi: 10.1037//0012-1649.36.2.190

Baldwin, D. A. (1991). Infants' Contribution to the Achievement of Joint Reference. *Child Development*, *62*(5), 875–890.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382–407. doi: 10.1016/j.cognition.2007.07.005

Baldwin, D. A., & Meyer, M. (2007). How Inherently Social is Language? In E. Hoff & M. Shatz (Eds.), *Handbook of language development* (pp. 87–106). Cambridge, UK: Blackwell Publishers.

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning. *Psychological Science*, *19*(3), 241–248. doi: 10.1111/j.1467-9280.2008.02075.x

Benitez, V. L., & Smith, L. B. (2012). Predictable locations aid early object name learning. *Cognition*, *125*(3), 339–352. doi: 10.1016/j.cognition.2012.08.006

Bergmann, C., & Cristia, A. (2015). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, 1–17. Retrieved from `http://inworddb.acristia.org`

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of experimental psychology: General*, *117*(1), 21–33. doi: 10.1037/0096-3445.117.1.21

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, *29*(4), 711–721. doi: 10.1037/0012-1649.29.4.711

Bilder, C. R., & Loughin, T. M. (2004). Testing for Marginal Independence between Two Categorical Variables with Multiple Responses. *Biometrics*, *60*(1), 241–248.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. doi: 10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, *1*(1), 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993–1022. doi: 10.1162/jmlr.2003.3.4-5.993

Börschinger, B., Demuth, K., & Johnson, M. (2012). Studying the effect of input size for Bayesian Word Segmentation on the Providence Corpus. *Proceedings of the 24th International Conference on Computational Linguistics (COLING2012)*, 325–340.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005, apr). Mommy and Me: Familiar Names Help Launch Babies Into Speech-Stream Segmentation. *Psychological Science*, *16*(4), 298–304. doi: 10.1111/j.0956-7976.2005.01531.x.Mommy

Brent, M. R. (1999a). An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning*, *34*, 71–105.

Brent, M. R. (1999b). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, *3*(8), 294–301. doi: 10.1016/S1364-6613(99)01350-9

Brent, M. R., & Siskind, J. M. (2001, sep). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*(2), B33–B44. doi: 10.1016/S0010-0277(01)00122-6

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In *Psychology of learning and motivation, advances in research and theory* (Vol. 62, pp. 60–92).

Bruner, J. S. (1975). The Ontogenesis of Speech Acts. *Journal of Child Language*, *2*, 1–19.

Bulf, H., Johnson, S. P., & Valenza, E. (2011, oct). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–32. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/21745660` doi: 10.1016/j.cognition.2011.06.010

Campbell, A. L., & Namy, L. L. (2003). The role of social-referential context in verbal and nonverbal symbol learning. *Child development*, *74*(2), 549–63.

Chernick, M. R., & LaBudde, R. A. (2014). *An introduction to bootstrap methods with applications to r*. John Wiley and Sons.

Christiansen, M. H., Allen, J. P., & Seidenberg, M. S. (1998). Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, *13*(2-3), 221–268. doi: 10.1080/016909698386528

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never Bottleneck: A Fundamental Constraint on Language. *The Behavioral and brain sciences*, 1–52. doi: 10.1017/S0140525X1500031X

Christiansen, M. H., Onnis, L., & Hockema, S. A. (2009). The secret is in the sound: from unsegmented speech to lexical categories. *Developmental science*, *12*(3), 388–95. doi: 10.1111/j.1467-7687.2009.00824.x

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*, 1–24. doi: 10.1017/S1351324900000139

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159.

Cole, R. A., & Jakimik, J. (1979). *A model of speech perception.*

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302.

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. doi: 10.1111/j.1551-6709.2010.01160.x

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews*, *19*(1), 55–68. doi: 10.1080/07474930008800459

Dufour, J.-M., & Kiviet, J. F. (1998). Exact inference methods for first-order autoregressive distributed lag models. *Econometrica*, 79–104.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, *105*(3), 1901–1911.

Estes, K. G., & Lew-Williams, C. (2015). Listening Through Voices: Infant Statistical Word Segmentation Across Multiple Speakers. *Developmental Psychology*, *51*(11), 1–12.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2015). The changing rhythms of life: Activity cycles in the first two years of everyday experience. In *Society for research in child development.* Philadelphia, PA.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5).

Frank, M. C., Goldwater, S. J., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–25. doi: 10.1016/j.cognition.2010.07.005

Frank, S., Feldman, N. H., & Goldwater, S. J. (2014). Weak semantic context helps phonetic learning in a model of infant language acquisition. In *Proceedings of the 52nd annual meeting of the association of computational linguistics.*

Frank, S., Keller, F., & Goldwater, S. J. (2013). Exploring the utility of joint morphological and syntactic learning from child-directed speech. In *Proceedings of the conference on empirical methods in natural language processing.*

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures. *Child development*, *71*(4), 878–894.

Gogate, L. J., & Maganti, M. (2016). The dynamics of infant attention: Implications for crossmodal perception and word-mapping. *Child development*. doi: 10.1111/cdev.12509

Gogate, L. J., Prince, C. G., & Matatyaho, D. J. (2009). Two-month-old infants' sensitivity to changes in arbitrary syllable-object pairings: The role of temporal synchrony. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 508–519. doi: 10.1037/a0013623

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3), 289–316. doi: 10.1515/cogl.2004.011

Goldstein, M. H., Waterfall, H. R., Lotem, A., Halpern, J. Y., Schwade, J. A., Onnis, L., & Edelman, S. (2010). General cognitive principles for learning structure in time and space. *Trends in Cognitive Sciences*, *14*(6), 249–258. doi: 10.1016/j.tics.2010.02.004

Goldwater, S. J., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. doi: 10.1016/j.cognition.2009.03.008

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access: II Infant data. *Journal of Memory and Language*, *51*(4), 548–567. doi: 10.1016/j.jml.2004.07.001

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory: Lower scores in the uk than the usa. *Journal of Child Language*, *27*, 689–705.

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

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*(3), B53–B64. doi: 10.1016/S0010-0277(00)00132-3

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The Associative Structure of Language: Contextual Diversity in Early Word Learning. *Journal of memory and language*, *63*(3), 259–273. doi: 10.1016/j.jml.2010.06.002

Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Language*, *30*(3-4), 461–472. doi: 10.1177/0142723710370525

Hoff-Ginsberg, E. (1991). Mother-Child Conversation in Different Social Classes and Communicative Settings. *Child Development*, *62*(4), 782–796.

Horst, J. S. (2013, jan). Context and repetition in word learning. *Frontiers in psychology*, *4*(April), 149. doi: 10.3389/fpsyg.2013.00149

Horst, J. S., Parsons, K. L., & Bryan, N. M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*(FEB), 1–11. doi: 10.3389/fpsyg.2011.00017

Janacsek, K., Fiser, J., & Nemeth, D. (2012). The Best Time to Acquire New Skills: Age-related Differences in Implicit Sequence Learning across Human Life Span. *Developmental science*, *15*(4), 496–505. doi: 10.1111/j.1467-7687.2012.01150.x.The

Johnson, E. K., & Jusczyk, P. W. (2001). Word Segmentation by 8-Month-Olds: When Speech Cues Count More Than Statistics. *Journal of Memory and Language*, *44*, 548–567. doi: 10.1006/jmla.2000.2755

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. *Developmental science*, *12*(1), 131–41. doi: 10.1111/j.1467-7687.2008.00740.x

Johnson, E. K., & Tyler, M. D. (2010, mar). Testing the limits of statistical learning for word segmentation. *Developmental science*, *13*(2), 339–345. doi: 10.1111/j.1467-7687.2009.00886.x

Johnson, M. (2008). Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. *Proceedings of the Association for Computational Linguistics*(June), 398–406.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *66*(2), 115–124. doi: 10.1037/a0026727

Jusczyk, P. W. (1999). How infants begin to extact words from speech. *Trends in Cognitive Sciences*, *3*(9), 323–328.

Jusczyk, P. W., & Aslin, R. N. (1995). *Infants' detection of the sound patterns of words in fluent speech* (Vol. 29) (No. 1). doi: 10.1006/cogp.1995.1010

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*(5), 648–654. doi: 10.1037/0012-1649.23.5.648

Jusczyk, P. W., Friederici, A. D., Wessels, J. M. I., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sounds patterns of native language words. *Journal of Memory and Language*, *32*, 402–420.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in english-learning infants. *Cognitive psychology*, *39*(3-4), 159–207. doi: 10.1006/cogp.1999.0716

Jusczyk, P. W., Jusczyk, A. M., Kennedy, L. J., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 822–36. Retrieved from URL|

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35–42.

Kirkham, N. Z., Slemmer, J. A., Richardson, D. C., & Johnson, S. P. (2007). Location, Location, Location: Development of Spatiotemporal Sequence Learning in Infancy. *Child development*, *78*(5), 1559–1571.

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First language*, *5*, 44–45.

Koziol, N., & Bilder, C. (2007). MRCV: A Package for Analyzing Categorical Variables with Multiple Response Options. *The R Journal*, *6*(June), 144–150.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–53. doi: 10.1016/j.cognition.2013.02.002

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental science*, *14*(6), 1323–9. doi: 10.1111/j.1467-7687.2011.01079.x

Lew-Williams, C., & Saffran, J. R. (2012, feb). All words are not created equal: expectations about word length guide infant statistical learning. *Cognition*, *122*(2), 241–6. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3246061{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1016/j.cognition.2011.10.007

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*, *42*(10), 1–29. doi: 10.18637/jss.v042.i10

Linzer, D. A., & Lewis, J. B. (2013). poLCA: Polytomous variable latent class analysis [Computer software manual]. Retrieved from http://dlinzer.github.com/poLCA (R package version 1.4)

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk (third edition)*. Lawrence Erlbaum Associates.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58–71.

McInnes, F. R., & Goldwater, S. J. (2011). Unsupervised Extraction of Recurring Words from Infant-Directed Speech. In *Proceedings of the 33rd annual conference of the cognitive science society.*

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117. doi: 10.1016/S0010-0277(03)00140-9

Monaghan, P., & Christiansen, M. H. (2010, jun). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of child language*, *37*(3), 545–64. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/20307344` doi: 10.1017/S0305000909990511

Onnis, L., Monaghan, P., Richmond, K., & Chater, N. (2005). Phonology impacts segmentation in online speech processing. *Journal of Memory and Language*, *53*(2), 225–237. doi: 10.1017/CBO9781107415324.004

Onnis, L., Waterfall, H. R., & Edelman, S. (2008). Learn locally, act globally: Learning language from variation set cues. *Cognition*, *109*(3), 423–430. doi: 10.1016/j.cognition.2008.10.004

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009a). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244–7. doi: 10.1016/j.cognition.2009.07.011

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009b). Statistical learning in a natural language by 8-month-old infants. *Child development*, *80*(3), 674–85. doi: 10.1111/j.1467-8624.2009.01290.x

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & cognition*, *36*(7), 1299–1305. doi: 10.3758/MC.36.7.1299

Phillips, L., & Pearl, L. (2015). The Utility of Cognitive Plausibility in Language Acquisition Modeling: Evidence From Word Segmentation. *Cognitive Science*, *39*, 1824–1854. doi: 10.1111/cogs.12217

Place, S., & Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child development*, *82*(6), 1834–49. doi: 10.1111/j.1467-8624.2011.01660.x

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012, jan). Learning to represent a multi-context environment: More than detecting changes. *Frontiers in psychology*, *3*(July). doi: 10.3389/fpsyg.2012.00228

Ramscar, M., & Port, R. F. (2016). How Spoken Languages Work in the Absence of an Inventory of Discrete Units. *Language Sciences*, *53*, 58–74. doi: 10.1016/j.langsci.2015.08.002

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, *120*(2). doi: 10.1016/j.cognition.2011.04.001

Räsänen, O. (2014). Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. In *Proc. 36th annual conference of the cognitive science society* (pp. 2817–2822). Quebec, Canada.

Räsänen, O., Doyle, G., & Frank, M. C. (2015). Unsupervised word discovery from speech using automatic segmentation into syllable-like units. In *Proceedings of interspeech.*

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*(September).

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. doi: 10.1111/j.1756-8765.2010.01111.x

Rivera-Gaxiola, M., Silva-Pereyra, J., & Kuhl, P. K. (2005). Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. *Developmental science*, *8*, 162–172. doi: 10.1111/j.1467-7687.2005.00403.x

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). stm: R package for structural topic models [Computer software manual]. Retrieved from `http://www.structuraltopicmodel.com` (R package version 1.1.3)

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIREs Cogn Sci*, 906–914. doi: 10.1002/wcs.78

Romberg, A. R., & Saffran, J. R. (2013). All Together Now: Concurrent Learning of Multiple Structures in an Artificial Language. *Cognitive science*, 1–31. doi: 10.1111/cogs.12050

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1419773112

Roy, B. C., Frank, M. C., & Roy, D. (2012). Relating Activity Contexts to Early Word Learning in Dense Longitudinal Data. In *Proceedings of the 34th annual cognitive science conference.*

Roy, B. C., Vosoughi, S., & Roy, D. (2014). Grounding language models in spatiotemporal context. In *Fifteenth annual conference of the international speech communication association.*

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, *35*(4), 606–621. doi: 10.1006/jmla.1996.0032

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PLoS ONE*, *6*(12). doi: 10.1371/journal.pone.0028095

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Seidl, A. H., Tincoff, R., Baker, C., & Cristia, A. (2014, apr). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental science*, 1–10. doi: 10.1111/desc.12182

Shukla, M., Nespor, M., & Mehler, J. (2007, feb). An interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive psychology*, *54*(1), 1–32. doi: 10.1016/j.cogpsych.2006.04.002

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, *18*(5), 251–258. doi: 10.1016/j.tics.2014.02.007

Smith, L. B., & Yu, C. (2008, mar). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–68. doi: 10.1016/j.cognition.2007.06.010

Soderstrom, M., Nelson, D. G. K., & Jusczyk, P. W. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development*, *28*(1), 87–94. doi: 10.1016/j.infbeh.2004.07.001

Soderstrom, M., Seidl, A. H., Nelson, D. G. K., & Jusczyk, P. W. (2003, aug). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*(2), 249–267. doi: 10.1016/S0749-596X(03)00024-X

Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE*, *8*(11). doi: 10.1371/journal.pone.0080646

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, *50*(1), 86–132. doi: 10.1016/j.cogpsych.2004.06.001

Synnaeve, G., Dautriche, I., Börschinger, B., Johnson, M., & Dupoux, E. (2014). Unsupervised Word Segmentation in Context. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 2326–2334).

Thiessen, E. D. (2010). Effects of Visual Information on Adults' and Infants' Auditory Statistical Learning. *Cognitive Science*, *34*(6), 1093–1106. doi: 10.1111/j.1551-6709.2010.01118.x

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental psychology*, *39*(4), 706–716. doi: 10.1037/0012-1649.39.4.706

Thiessen, E. D., & Saffran, J. R. (2007). Learning to Learn: Infants' Acquisition of Stress-Based Strategies for Word Segmentation. *Language Learning and Development*, *3*(1), 73–100.

Vlach, H. A., & Sandhofer, C. M. (2011). Developmental differences in children's context-dependent word learning. *Journal of Experimental Child Psychology*, *108*(2), 394–401. doi: 10.1016/j.jecp.2010.09.011

Warlaumont, A., VanDam, M., & MacWhinney, B. (2015). *The homebank system.* Retrieved 2016-10-22, from `homebank.talkbank.org`

Weinert, S. (2009). Implicit and explicit modes of learning: Similarities and differences from a developmental perspective. *Linguistics*, *47*(2), 241–271. doi: 10.1515/LING.2009.010

Weisleder, A., & Fernald, A. (2013, nov). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143–52. doi: 10.1177/0956797613488145

Weizman, Z. O., & Snow, C. E. (2001). Lexical Input as Related to Children's Vocabulary Acquisition: Effects of Sophisticated Exposure and Support for Meaning. *Developmental Psychology*, *37*(2), 265–279. doi: 10.1037/0012-1649.37.2.265

Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How Do Infants Become Experts at Native-Speech Perception? *Current Directions in Psychological Science*, *21*(4), 221–226. doi: 10.1177/0963721412449459

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–6. doi: 10.1016/j.tics.2004.08.006