# Vocal Interaction Dynamics of Children With and Without Autism

**Anne S. Warlaumont (awarlmnt@memphis.edu)**
**D. Kimbrough Oller (koller@memphis.edu)**
Speech-Language Pathology, University of Memphis, 807 Jefferson Ave.
Memphis, TN 38105 USA

**Rick Dale (radale@memphis.edu)**
Department of Psychology, University of Memphis, 202 Psychology Building
Memphis, TN 38152 USA

**Jeffrey A. Richards (JeffRichards@lenafoundation.org)**
**Jill Gilkerson (JillGilkerson@lenafoundation.org)**
**Dongxin Xu (DongxinXu@lenafoundation.org)**
LENA Foundation, 5525 Central Ave., Suite 100
Boulder, CO 80301 USA

## Abstract

This study examines the temporal and directional characteristics of child-adult vocal exchanges in day-long naturalistic recordings of autism and typical control groups. In both populations, adults responded frequently (on average about 40% of responses) within 1s or less, a time thought to be conducive for contingency learning by the child. However, the time to adult response tended to be longer for the autism population. In the autism group, children also tended to follow more and lead less relative to the control group, as measured by differences in diagonal recurrence profiles computed based on cross recurrence plots. The results inform on the dynamics of naturalistic communicative interaction in normal development and therefore on the social context in which language develops. They also illustrate how large datasets and modern interaction analyses can expand our understanding of differences in children with autism, a population with both social and language deficits.

**Keywords:** Social interaction; autism; temporal dynamics; cross recurrence; language development; naturalistic recording; response time; social contingency

## Introduction

In this paper, we examine fundamental issues related to the fine-grained temporal organization of vocal interaction between children and their social environment, primarily, caregivers. Recent years have seen an abundance of interest in joint action and coordinative processes in both children and adults (Galantucci & Sebanz, 2009). In the current study, we make use of latency response measures as well as the technique of cross recurrence analysis to identify leading and following patterns in the vocal exchanges between children and adults. We find a distinct signature of leading in normal children and find that a distinct breakdown of this signature is identifiable in children with autism. These results show that analysis of naturalistic recordings may reveal socio-dynamic indicators of at-risk children. We conclude with a brief discussion of the relevance of our findings to models of language acquisition in normal and disordered individuals.

### Interaction and Contingency in Language Development

The fact that language learning occurs in a dynamic and interactive social context is becoming increasingly appreciated. Children are not passive information processors nor do they learn language purely on the basis of contingent reinforcement. They are, rather, actively engaged in perceptual learning and responding to communicative acts produced by others as well as being engaged in behavioral and motor exploration for which they at least sometimes receive feedback in the form of communicative response by adults and other children in their environment.

For example, in a video-recording study performed in the participants' homes, Keller et al. (1999) found that mothers often respond within one second to their three-month-old infants' communicative acts. Relating this to the fact that one second had been previously shown to be about the amount of time within which a contingent response must occur in order for the infant to detect that contingency, the authors concluded that mothers' communicative responses to their infants' communicative attempts tend to occur within the necessary window of time for the infant to perceive them as contingent. In other words, their results support the notion that caregivers' responses to their children support infant communicative development by serving as contingent reinforcers for the infant's own communicative acts.

In a more recent study, Gros-Louis et al. (2006) observed naturalistic interactions in a laboratory setting and found that mothers responded contingently to their infant's vocalizations over 70% of the time and that the type of response they gave depended on the phonological characteristics of the infant's vocalizations. Furthermore, Goldstein, King, and West (2003) and Goldstein and Schwade (2008) have found experimentally that mothers' contingent responses do appear to shape the infant's speech-related vocal development as measured through follow-up tests.

Recently, cross recurrence analysis of time series has allowed for additional quantitative measures of interactive contingency to be measured in naturalistic child-caregiver interaction. For example, patterns of leading and following by interlocuters can be examined at multiple timescales concurrently. Dale and Spivey (2006) examined diagonal cross recurrence profiles calculated on syntactic patterns (specifically, part of speech bigrams) for three well-known child-caregiver conversation corpora. They found individual differences among the three children in their tendency to lead versus follow their caregiver. Abe (Kucjaz, 1976), who had the most advanced language out of the three children also had the greatest tendency to lead rather than follow the caregiver. This work lays foundations for application of cross recurrence analysis to other vocal interaction phenomena, to larger naturalistic datasets, and, as carried out here, to the study of populations with autism.

### Autism Spectrum Disorders (ASD)

Impaired social interaction and language learning are two components of the DSM autism diagnostic criteria. With regard to social interaction, children with ASD have exhibited differences in initiation, turn-taking, imitation, and joint attention behaviors.

In recent years, technology has become available to permit day-long naturalistic recording of infant's acoustic environments, including their own vocalizations and the speech and other environmental sounds in the infant's vicinity. Warren et al. (2009) evaluated social interaction in all-day recordings (5,256 hours over 438 sessions) in ASD and control groups. The authors discovered differences between typically developing and autistic children in the frequencies of both conversational turns and child vocalizations. These results, based on summary measures, encourage analysis at a more fine-grained level of temporal detail in order to address such issues as the directionality of the conversational exchanges and temporal characteristics of adult-child interactions. Both latency to response and diagonal cross recurrence profiles can be automatically calculated, making them suitable for application to large-scale naturalistic recordings.

### This Study

In the present study, we first looked at response latencies in a way that was similar to Keller et al. (1999). However, we evaluated much more data and used more naturalistic recordings (collected at home, daycare, and therapy as opposed to only at home in a single post-sleep, post-feeding context with experimenters present and videotaping). Other differences are that we looked at the vocal modality only, and that we evaluated age, autism, gender, and maternal education as predictive factors. We also applied cross recurrence analysis to the data and investigate leading and following tendencies in the recordings. The application of this method with large-scale recordings of adult-child speech is unique as is its application to the autism population.

## Method

### Participants

The participant recruitment, recordings and the automated labeling of them were conducted as part of previous studies. Warren et al. (2009) provide more detailed information on the procedures. The present study includes data from 26 children between 16-48 months who have been diagnosed under the classic autism subtype except for two who received Pervasive Developmental Disorder-Not Otherwise Specified subtype diagnoses; documentation of ASD diagnoses by trained professionals was provided by the children's parents. No child was reported to have a diagnosis that included echolalia (pathological repetition of previously heard speech). The study also includes data from 78 typically developing (TD) children who were selected form a larger normative database such that for each child with ASD there were three TD controls of the same gender and socioeconomic status (SES), as measured by the mother's education level, and collectively across the three controls spanning the same range of ages as the ASD child.

### Recording

Recordings were made using LENA digital language processor devices. These recorders fit into a pocket sewn into the front of custom-designed clothing and record a single channel of audio for up to 16 hours at a time. The device records the child's voice as well as other sounds within approximately a 6-10' radius of the child. Parents were mailed the devices and were instructed to begin recording when the child awoke in the morning and left the recorder on throughout the day. Recordings contexts included the home, preschool, and speech-language therapy. There were 438 recordings in total, each lasting at least 12 hours. The present study is thus based on over 5,256 hours of naturalistic recording.

### Automated Labeling

Each recording was processed using the professional version of the LENA analysis software. The software analyzes and time segments the entire recording according to the likely source of the signal, e.g., the child wearing the recorder, another child, an adult, a television or radio, silence; every part of the recording is given a label. Within child segments it also labels some sub-segments (termed *vocalizations* by the system) as speech-like or as cry/vegetative/fixed. Reliability for the automated labeling compared to human raters on TD child recordings is approximately 82% correct for adult speaker, 76% correct for key child, 75% correct for child speech-like, and 84% for child cry/vegetative/fixed (Xu et al., 2009). The software allows for exporting these sound source and child vocalization type segmentations along with other information in XML format.

We developed a set of Perl scripts to extract the specific information of interest for this study from the XML files (exported as *.its* files by the LENA software). Specifically, we extracted the start and end times of each segment labeled with relatively high confidence as coming from the child wearing the recorder (*child near*, *CHN*, segments, labeled as such because they fell near the maximum of the Gaussian mixture model that gave the likelihood that a segment was produced by the child) and of each segment labeled as coming from an adult with relatively high confidence (*female adult near*, *FAN*, or *male adult near*, *MAN*). Note that loudness and nearness to the child increase the confidence of segment coding and therefore increase the likelihood of a sound being included in the present study. Also, there were minimum duration thresholds for each segment label type; thus, a long string of vocalization by the same speaker could only be split if there was an intervening silence, TV, or other-speaker vocalization meeting the minimum duration requirement. We also identified child segments that contained only speech-like sub-segments as well as those that contained only cry/vegetative sub-segments. All the subsequent response time and cross recurrence calculations were made using only those child segments that contained no cry/vegetative sub-segments and at least one speech-like sub-segment.

## Response Time Analysis

We developed a set of programs written in Perl and R that automatically extracted and calculated response time information from the speaker labels. First, adult response times were calculated according to the following procedure. For each child segment, we determined whether an adult segment followed without any child segment intervening. Other sound source labels were permitted to intervene between the child segment and the subsequent adult segment. Then the time between the offset of the child segment and the onset of the adult segment was calculated. Child response times were calculated in exactly the same manner, except with the speaker labels reversed. Based on these response times, the median adult response time and the median child response time were calculated as well as the proportion of adult responses occurring within 1s and the proportion of child responses occurring within 1s.

## Cross Recurrence Analysis

**Cross Recurrence Plots** Cross recurrence plots (Marwan et al., 2007; Richardson et al., 2007) are matrices that indicate correspondence or lack of correspondence for every possible combination of events or times in one event time series and the events or times in another series. In our case, the vertical dimension of the matrix corresponds to the presence and absence of child segments and the horizontal dimension of the matrix corresponds to the presence and absence of adult segments (Fig. 1). Each element in the plot matrix is assigned a value of 1 (marked in black in the figure) if there is a child segment at the row corresponding to the element's row number as well as an adult segment at the segment corresponding to the element's column number.

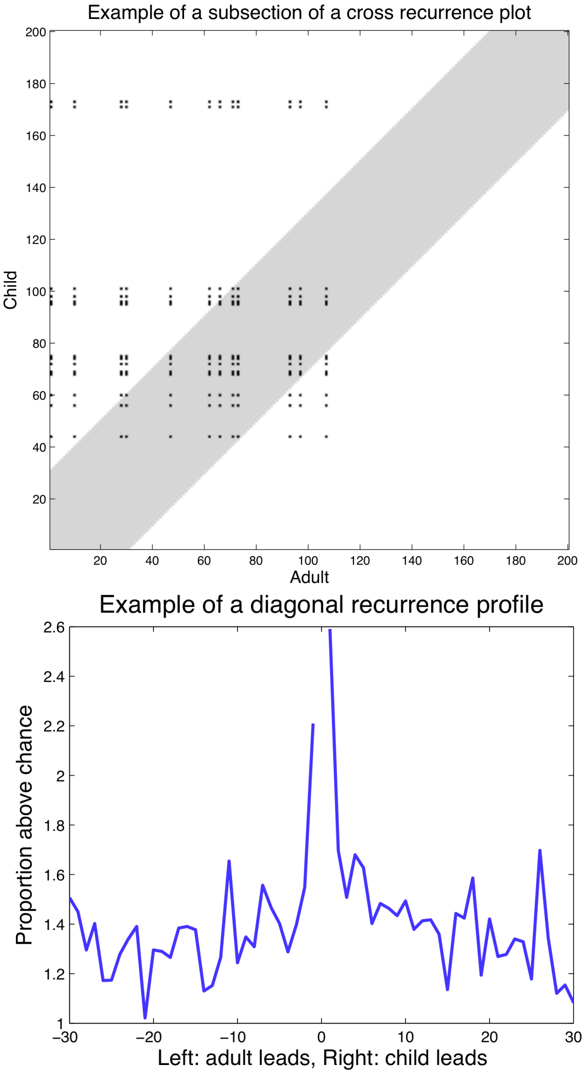

Figure 1: On the top is the cross recurrence plot for the first 200s of one of the recordings. The gray region indicates the portion from which the diagonal recurrence profile is calculated. On the bottom is the diagonal recurrence profile for the entire recording.

The time series that were used for making these charts were broken into 1s chunks. When either the child or an adult was speaking, a 1s chunk was coded as 1 in the speaker's series and as null value in the non-speaker's series. Regardless of the actual length of the segment, it was coded as lasting 1s so that long speaker segments would not be treated as having long lags to segments by the other speaker. When neither child nor adult were speaking, both time series were coded as null values for the duration of the no-speaker time, rounded down to the nearest second. The recurrence plot is square since both the vertical and the horizontal dimensions have length equal to the total number of 1s chunks in the recording.

**Diagonal Cross Recurrence Profiles** A number of measures, such as recurrence rate, determinism, etc. can be extracted directly from the cross recurrence plot (Marwan et al., 2007). However, in this study we focused on measures that were extracted from the plot's diagonal recurrence profile (explained below) after it had been derived from the recurrence plot. In the physical sciences, this is sometimes referred to as the recurrence probability or the recurrence spectrum (Marwan et al., 2007). Richardson and Dale (2005) and Richardson et al. (2007) have used this measure in analyses of linguistic coordination. It can be interpreted as a lag profile that reflects co-occurrence patterns between utterances at varying relative lags. We provide some further description here.

The diagonal on the recurrence plot running from the origin to the final event on both axes reflects when the child and caregiver are speaking at the same time. Sometimes this main line is referred to as the "line of synchronization," since any points on this line reflect matching on/off states for the child and the adult(s). However, since the automatic labeling procedure does not allow overlapping speaker labels, there will never be a match along this diagonal.

The next diagonal line just below-right of the primary diagonal contains the matches between the child's on/off states and those of the adult series one step into the future. In other words, the elements of this adjacent below-right diagonal are given a point on the plot when a given child segment was immediately followed by an adult segment (i.e., the adult spoke one time step later during the interaction). Conversely, the elements of the adjacent above-left diagonal line have a point when a given adult segment was immediately followed by a child segment Moving to diagonals further below-right or above-left give indication of when the adult followed the child at larger lags and when the child followed the adult at larger lags, respectively.

For each diagonal line parallel to the primary diagonal, the number of 1's can be added and divided by the total number of elements in that diagonal to give the proportion cross recurrence for the speaker order and lag amount corresponding to that diagonal. These proportions can then be plotted to create a diagonal recurrence profile (Fig. 1). By randomly shuffling the speaker labels and recalculating the diagonal recurrence profile, and by repeatedly doing this and averaging across the shuffled label profiles, one can obtain a bootstrapped estimate of the baseline diagonal recurrence profile that would be expected if were no systematic leading-following relationship between the speakers. Dividing the actual diagonal recurrence profile by the baseline estimate gives a normalized diagonal recurrence profile that represents proportion above chance leading/following tendencies.

In this study, we measured three characteristics of the normalized diagonal recurrence profile for a given recording. The first was the height of the profile at the point immediately right from center. This gives an indication of how often the adult vocalized immediately after a child vocalization. The second was the height at the point immediately left of center; it tells how often the child's vocalizations immediately followed the adult's. The third measure is the ratio of the sum of values on the right side of the profile (which is higher when the adult tended to follow the child) to the sum of values on the left side of the profile (higher when the child tended to follow the adult). This gives a measure of the general balance between leading and following across the two speakers.

# Results

## Response Time Results

The adult response times and child response times for the ASD and control groups are plotted as averaged histograms in Figure 2.

For each of the four response time independent measures (adult median response time, adult proportion within 1s, child median response time, and child proportion within 1 s) we ran a mixed model regression with participant ID as a random effect and ASD status, age in weeks, gender, and mother's education level (a measure of the family's socioeconomic status) as fixed effects.

Adult median response time was significantly longer for children with ASD (*M* = 2.32s, *SD* = 1.22) than for the controls (*M* = 1.65s, *SD* = 0.78), $p < 0.001$, $\beta = 0.331$, and was significantly shorter as maternal education increased, $p < 0.001$, $\beta = -0.234$. Adult proportion of responses within 1s was significantly smaller for children with ASD (*M* = 0.37, *SD* = 0.10) than for the controls (*M* = 0.43, *SD* = 0.08), $p < 0.001$, $\beta = -0.348$, and was significantly larger as maternal education increased ($p < 0.001$, $\beta = 0.284$).

Child median response time was longer for children with ASD (*M* = 2.70s, *SD* = 1.38) than for the controls (*M* = 2.37s, *SD* = 0.92) though this did not reach statistical significance, $p = 0.063$, $\beta = 0.161$, and was significantly shorter as maternal education increased, $p = 0.016$, $\beta = -0.153$. Child proportion of responses within 1s was significantly larger as maternal education increased, $p = 0.010$, $\beta = 0.174$.

Age and gender did not significantly predict any of the four independent variables.

## Cross Recurrence Results

The averaged diagonal recurrence profiles for the ASD and control groups are plotted in Figure 3. As with the response time measures, each of the three dependent variables (height immediately right of center, height immediately left of center, and ratio of right side to left side) was regressed on participant ID as a random effect and ASD status, age in weeks, gender, and mother's education level as fixed effects.

The height at the point immediately right of center was only significantly predicted by age, decreasing as age increased, $p < 0.001$, $\beta = -0.228$.

The height at the point immediately left of center, which represents the frequency with which the child immediately

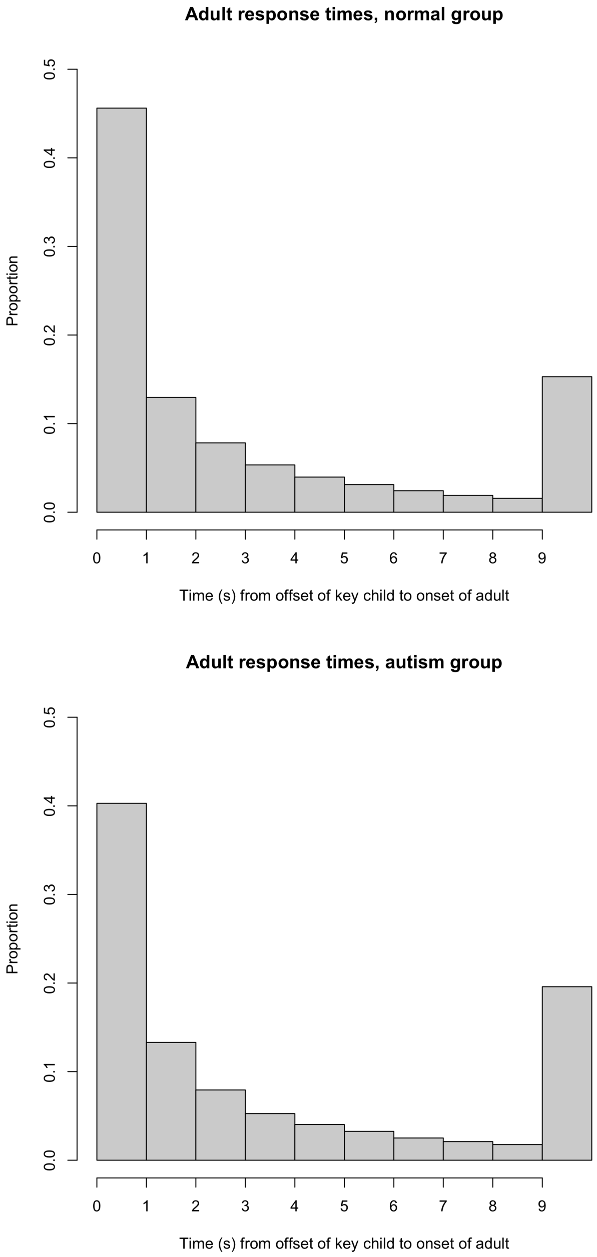

Figure 2: Histograms of adult response latencies for children with and without ASD.

followed the adult, was significantly higher for the ASD group (*M* = 1.73, *SD* = 0.91) than for the control group (*M* = 1.53, *SD* = 1.09), $p < 0.001$, $\beta = 0.207$. The height at this point was also significantly lower as maternal education increased, $p = 0.017$, $\beta = -0.183$, and was lower as age increased, $p = 0.002$, $\beta = -0.214$.

The ratio of the right side (from lag 1 through lag 10) to the left side (from lag 1 through lag 10) of the diagonal cross recurrence profile was smaller for the ASD group (*M* = 1.06, *SD* = 0.18) than for the TD group (*M* = 1.25, *SD* = 0.30), $p < 0.001$, $\beta = -0.398$, indicating that the general tendency for the child to lead and for the adult to follow was lessened in the autism group. A small but significant increase was accounted for by age, $p = 0.01$, $\beta = 0.136$.

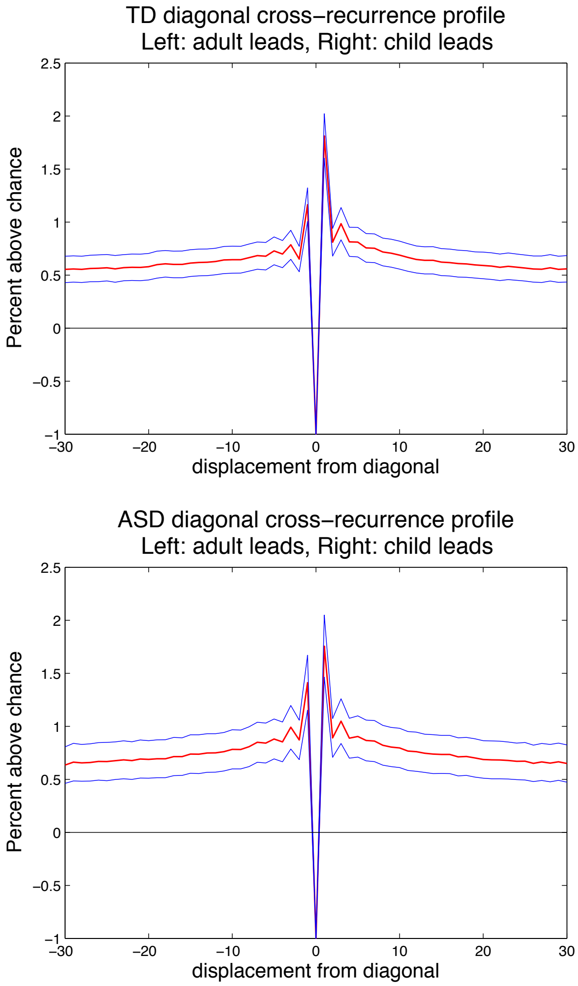

Figure 3. Diagonal recurrence profiles averaged across all recordings in the TD group (top) and all recordings in the ASD group (bottom). In each profile, the red line indicates the mean values across recordings. Blue lines indicate 95% confidence intervals. Displacement from diagonal is in seconds.

## Discussion

This study provides new information from automated analysis over large naturalistic recordings in support of the idea that social interaction is impaired in ASD. Interestingly, the strongest trend concerned the adult's responses to the child rather than the child's responses to the adult. There were differences in both the dynamics and the directionality of adult-child interaction in ASD. The length of time before an adult responded to an ASD child's speech or speech-like vocalization was larger in ASD than for TD children with a smaller percentage of responses occurring within the 1s window considered ideal for contingency detection. In addition, ASD children's speech and speech-like vocalizations were more of a tendency to follow the adult vocalizations than TD children's.

The shift of the balance toward child following (and adults leading) and increased latency of adult responses to the child when they did occur could be due to less initiation of communication on the part of the ASD children and/or to reduced communicative content or other deficiencies in the vocalizations of children with ASD. It could also be due to adults' reduced attentiveness to the vocalizations of children with the disorder. This pattern of following vs. being followed may have feedback effects on the child's language development, reducing the quality of the contingency-based input available to the child with ASD as they acquire speech, language, and other communication skills. At present, there are very few computational models that attempt to capture the interplay among cognitive agents in a realistic way (one exception may be language evolutionary models; see Cangelosi & Parisi, 2002, for examples). The dynamic interplay between cognitive agents during development, such as speech-contingency patterns, may produce feedback loops that substantially impact learning within an individual system.

The present work is relevant to theoretical, including computational, modeling of speech-language development. Language learning occurs in the context of social interactions during which the child hears what other speakers say but also receives contingent reinforcement for their own vocalizations. Understanding the typical dynamics of these interactions may help guide the development of models that take into account the dynamic interactive social context of language learning. They may also help inform models of autism. Some of the deficits present in autism may be the result of a negative feedback loop in which children with autism produce fewer or lower-quality conversation initiations, leading to adults' responding with lower frequency and more latency, which in turn leads to poorer learning of language and communication-related skills by the child.

From a practical standpoint, measures of conversational dynamics, both at short and long timescales could potentially be applied for early identification of autism or other communicative disorders. Being a disorder that involves profound social and cognitive impairments, differences in patterns of communicative interaction, such as in leading-following and elicitation of quick responses, might indicate risk for autism. For example, automatically computed interaction-based measures (such as the ones used in the present study) could supplement the acoustic measures used in an existing autism screening tool (Xu et al., 2009)

## Acknowledgments

Thanks go to the LENA Foundation, particularly to Terry and Judith Paul, for funding and executing the development of the LENA recording and automated labeling system as well as the autism recordings database used in this study. Author ASW was supported by a Department of Energy Computational Science Graduate Fellowship. DKO was supported by the Plough Foundation.

## References

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. London: Springer.

Dale, R. & Spivey, M. J. (2006). Unraveling the dyad: using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*, 391-430.

Galantucci, B., & Sebanz, N. (2009). Joint action: current perspectives. *Topics in Cognitive Science*, *1*, 255-259.

Goldstein, M. H., Kin, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 8030-8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515-523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*, 509-516.

Keller, H., Lohaus, A., Völker, S., Cappenberg, M., Athanasios, C. (1999). Temporal contingency as an independent component of parenting behavior. *Child Development*, *70*, 474-485.

Kuczaj, S. (1976). –ing, -s, and –ed: a study of the acquisition of certain verb inflections. Unpublished doctoral dissertatio, University of Minnesota.

Marwan, N., Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*, 237-329.

Richardson, D. C. & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*, 39-54.

Richardson, D. C., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18*, 407-413.

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2009). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*. Advance online publication. doi: 10.1007/s10803-009-0902-5

Xu, D., Richards, J. A., Gilkerson, J., Yapanel, U., Gray, S., Hansen, J. (2009). Child vocalization composition as discriminant information for automatic autism detection. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2518-2522). Minneapolis, MN: IEEE.

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA$^{TM}$ Language Environment Analysis System in young children's natural home environment* (LENA Foundation Technical Report LTR-05-02). Retrieved from *http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2*