Exploring Autism Spectrum Disorders Using HLT

Julia Parish-Morris[★], Mark Liberman[◦], Neville Ryant[◦], Christopher Cieri[◦], Leila Bateman[★], Emily Ferguson[★], Robert T. Schultz[★]

[◦]Linguistic Data Consortium. University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA, 19104 USA

[★]Center for Autism Research. Children's Hospital of Philadelphia
3535 Market Street, Suite 860, Philadelphia, PA, 19104 USA

Abstract

The phenotypic complexity of Autism Spectrum Disorder motivates the application of modern computational methods to large collections of observational data, both for improved clinical diagnosis and for better scientific understanding. We have begun to create a corpus of annotated language samples relevant to this research, and we plan to join with other researchers in pooling and publishing such resources on a large scale. The goal of this paper is to present some initial explorations to illustrate the opportunities that such datasets will afford.

1 Introduction

Autism Spectrum Disorder (ASD) is a highly heterogeneous, brain-based developmental disorder affecting approximately 1.5% of the population (Christensen, 2016). Primary diagnostic indicators include impairments in social communication and reciprocity, as well as the presence of repetitive behaviors and restricted patterns of interests (American Psychiatric Association, 2013). Despite significant symptom overlap in the core domains of social communication and repetitive behaviors, individuals diagnosed with ASD can look very different from one person to the next. Clinical presentation varies substantially depending on age, context, IQ, intervention history, and presence or absence of common comorbidities such as ADHD and anxiety disorder. The heterogeneous presentation with respect to overall severity and pattern of co-occurring conditions makes research aimed at improving treatments and isolating biological mechanisms much more complicated.

The phenotypic heterogeneity of ASD contributes to conflicting research findings that paint a confusing picture in the literature. For example, depending upon the characteristics of a particular sample, groups of children with ASD can look as if they have face processing impairments or not (Weigelt, Koldewyn, & Kanwisher, 2012), perceptual processing biases or not (D'Souza, Booth, Connolly, Happé, & Karmiloff-Smith, 2015), and persistent social language differences or not (Fein et al., 2013). In response to reproducibility issues, one strategy has been to shift away from research based solely on a categorical conceptualization, such as schizophrenia, ADHD, and ASD, to a domain-based dimensional approach that cuts across traditional diagnostic categories. This approach to understanding mental disorder is explicitly encouraged by the National Institute of Mental Health through the Research Domain Criteria effort (RDoC; (Insel, 2014)). The RDoC approach is trans-diagnostic and grounded in the study of process (where there is clear or emerging support on underlying biological processes), such as specific neural systems that relate to dimensions of behavior in model systems and in humans.

Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 74–84,
San Diego, California, June 16, 2016.

A complementary approach to improve reproducibility is to focus on large sample sizes so as to be able to more easily generalize results to all individuals with autism. Most research groups lack the resources to obtain large samples, and thus pooled efforts and data sharing become key. Large samples also provide the statistical power necessary to control for a larger array of possible confounding variables. In an effort to increase data sharing and power to parse the heterogeneity of ASD, the National Institutes of Health established the National Database for Autism Research (NDAR; ("National Database for Autism Research - Home," n.d.)). This database provides de-identified data for large N secondary data analyses. However, aside from common characterization variables (see Bone, Goodwin, et al., 2014), NDAR will not have sufficient data for more specialized needs such as human language technology research.

In this paper, we describe a new opportunity for data sharing in a format designed to facilitate research on speech and language in ASD, and explore the possibilities associated with this sort of database. The Children's Hospital of Philadelphia Center for Autism Research (CAR) collected the samples analyzed here, and established a collaborative project with the University of Pennsylvania Linguistic Data Consortium (LDC). We focus on recorded conversations that took place from 2008-2015 during the course of clinical evaluations for autism.

1.1 The Autism Diagnostic Observation Schedule

The ADOS is a semi-structured, conversation- and play-based evaluation tool used by expert clinicians to help inform diagnostic decision-making. There are 4 versions of the ADOS, one of which is selected for administration based on an individual's language ability at the time of evaluation. Module 3 requires phrase speech, and includes a large section devoted to conversation. During this part of the evaluation, clinicians ask questions about social-emotional concerns. These questions are designed to elicit language or behavior that differentiates individuals with social communication difficulties from those without (e.g., "What does being a friend mean to you? Do you ever feel lonely?"). Importantly, the samples arising from this section are similar in form and content to samples used in past clinically-oriented HLT research.

One benefit of targeting language produced during the ADOS for HLT research is ubiquity; the ADOS is widely included in research-grade Gold Standard diagnostic evaluations, both inside and outside the United States, and is routinely recorded for clinical reliability purposes. Many of these audio-video recordings are associated with clinical metadata such as age, sex, clinical judgment of ASD status, autism severity metrics, IQ estimates, and social/language questionnaires, as well as genetic panels, brain scans, behavioral experiments, and infrared eye tracking. The quality of recording is variable, with a multitude of recording methods employed. A substantial number of these recordings have yet to be assembled into a large, shareable resource. We view this as a largely untapped opportunity for data sharing that could facilitate advancements in clinically oriented HLT research and autism research more broadly.

1.2 The present study

In 2013, CAR and the Linguistic Data Consortium (LDC) began a project aimed at analyzing ADOS recordings from more than 1200 toddlers, children, teens, and adults, most of which were ultimately diagnosed with ASD. These recordings are associated with rich characterization data in the form of interviews and questionnaires, cognitive and behavioral assessments, eye tracking, brain scans, and genetic tests. Our initial goal was to determine whether automated analysis of language recorded during the ADOS could predict diagnostic status, although our aims have since expanded to include identifying correlates of phenotypic variability within ASD. This second aim is particularly meaningful in the clinical domain and in our search for causes of autism; if we can accurately and objectively quantify the linguistic signal, we have a much better chance of reliably mapping it to real-world effects and to connecting it with biological mechanisms.

The current paper reports on our work-in-progress, and provides preliminary results from a cohort of 100 children. We analyze a small subset of possible lexical and acoustic features in combination with clinical measures. Our goal is to spur

interest in growing and sharing valuable resources like this one.

2 Dataset

To date, our corpus includes natural language samples from 100 participants engaged in the conversation and reporting section of ADOS Module 3 (mean length of recording ~20 minutes).

2.1 Subjects

Three diagnostic groups were included: ASD (N=65, mean age: 10 years), non-ASD mixed clinical (N=18, mean age: 10.39 years), and typically developing (TD; N=17; mean age: 11.29 years). ASD is more common in males than females (Wing, 1981), and our clinical groups have more boys than girls (ASD: 75% male; non-ASD mixed clinical: 94% male; TD: 47% male). Mothers and fathers had a median post-high school educational level of 4 years (Bachelor's degree) for the ASD and non-ASD mixed clinical groups, and 2 years (Associate's degree) for the TD group. Median household income was $60,000-$99,000.

The ASD group was determined to have an autism spectrum disorder according to DSM-IV criteria after a Gold Standard evaluation that included the ADOS, cognitive testing, parent interviews, and questionnaires. After undergoing the same rigorous evaluation as their peers with ASD, the non-ASD mixed clinical group was determined not to meet diagnostic criteria. This group is highly heterogeneous, with some participants exhibiting subthreshold ASD symptoms and others diagnosed with anxiety or ADHD. Due to the small sample size of this group and the TD group, analyses should be interpreted with caution. The TD group had no reported history of ASD, no significant neurological history, no first-degree family members with ASD, and did not meet clinical cutoffs on a common ASD screener (Social Communication Questionnaire; SCQ; (Rutter, Bailey, & Lord, 2003)).

2.2 Clinical measures

Participants were administered a variety of behavioral and cognitive tests during in-person visits at the Center for Autism Research. Parents completed questionnaires about their child's social and behavioral functioning either directly before or during the visit. Means, standard deviations, and ranges are provided in Table 1.

Autism Diagnostic Observation Schedule (ADOS; (Lord et al., 2012)). In addition to providing natural language samples, the ADOS is a scored instrument. Highly trained clinicians rate various aspects of children's behavior on a scale of 0-3 (higher = more autism-like). A subset of these ratings are combined using an algorithm that results in a total score for each of two domains: social affect (SA) and repetitive behaviors/restricted interests (RRB). Three comparison scores can also be calculated, which roughly index the severity of autism symptoms for a given child overall, in the social affect domain, and in the repetitive behaviors/restricted interests domain (see Table 1).

Table 1. Means and standard deviations for cognitive test scores, clinical observation ratings, and parent questionnaires.

	ASD	Non-ASD	TD
Full-scale IQ	105.31 (14.88)	97.77 (11.01)	104.06 (14.68)
Verbal IQ	106.91 (14.41)	100.78 (12.64)	108.24 (14.07)
Nonverbal IQ	105.94 (13.95)	95.06 (10.29)	100.94 (14.24)
ADOS severity score	6.49 (2.47)	2.72 (1.56)	1.47 (0.94)
ADOS SA severity score	6.29 (2.42)	3.06 (1.92)	2.06 (1.3)
ADOS RRB severity score	7.08 (2.54)	4.72 (2.91)	2.53 (2.18)
SRS t-score	80.6 (16.46)	81.22 (17.91)	39.82 (5.05)
CCC-2 GCC	81.44 (14.13)	77.24 (14.84)	115 (8.24)
CCC-2 SIDI	-9.75 (8.07)	-5.7 (12.68)	5.4 (6.01)

Differential Abilities Scales – 2nd Edition (DAS-II; (Elliott, 2007)). Overall (full-scale) IQ, non-verbal IQ, and verbal IQ were assessed via the DAS-II. DAS-II IQ measures have a mean of 100.

Children's Communication Checklist – 2nd Edition (CCC; (Norbury, Nash, Baird, & Bishop, 2004)). The CCC-2 is a norm-referenced parent re-

port questionnaire focused on aspects of structural and pragmatic language. The Global Communication Composite (GCC) is an overall measure of parent impressions of child communication competency, and the Social Interaction Difference Index (SIDI) score is designed to flag children in need of further evaluation for ASD or other disorders (negative scores indicate risk).

2.3 Interviews

Research reliable PhD-level clinical psychologists and/or psychology trainees administered the ADOS module 3 to all participants in quiet neutral rooms. Evaluations were videotaped using a single feed or PiP from 3 corner-mounted cameras, and audio was recorded through a ceiling microphone. After we obtained consent from participants to use their sessions for research purposes, entire video recordings were copied from their original media onto a shared file system accessible only to project members with current certifications for research on human subjects. Audio was extracted from the video stream and saved in lossless FLAC format. Except for extraction and format conversion, the data was identical to the original recording.

The ADOS is a semi-structured interview, so questions from the conversation and reporting section were occasionally spread throughout the entire interview (which lasts approximately 45-60 minutes). More often, they were clustered together in a section that lasts ~20 minutes. A knowledgeable member of study staff selected the largest chunk of continuous conversation and reporting questions for transcription and annotation.

2.4 Transcription and annotation

As described in a prior methods paper (Parish-Morris et al., 2016), transcription teams at LDC and CAR created time aligned, verbatim, orthographic transcripts of the conversation and reporting section for each participant. The LDC transcription team consisted of two junior and two senior transcribers, all college educated native speakers of American English. The junior transcribers performed segmentation of the audio files into pause groups and transcription. The senior transcribers corrected the initial transcripts and occasionally did transcription from scratch.

For this effort, LDC created a new transcription specification that resembles those used for conversational speech. The principal differences are that the current specification requires that participants be labeled only by their role (Interviewer and Participant) and that the boundaries between speech and non-speech be placed rather accurately because (inter-)turn duration is a factor of interest.

After LDC established the transcription process and pilot results were found to be promising, CAR developed a team to extend the corpus and begin evaluating inter-annotator agreement. The CAR team consists of multiple pairs of college educated native speakers of American English that transcribe the conversation and reporting section of the ADOS independently, a third more senior transcriber responsible for comparing and adjudicating the work of the first two, and a fourth transcriber who compares CAR and LDC transcripts when the latter are available, and adjudicates remaining disagreements. In this way, 4 transcribers and 2 adjudicators with complementary goals produce a "gold standard" transcript for analysis and for evaluation/training of future transcriptionists.

2.5 Quality control

LDC transcribed 52 files, and CAR transcribed 100 including independent transcriptions of the 52 that LDC transcribed. A simple comparison of word level identity between CAR's adjudicated transcripts and LDC's transcripts revealed 93.22% overlap on average, before a third adjudication resolved differences between the two. In the case of files that were transcribed by CAR only (N=48), pre-adjudication overlap in word-level comparisons between transcribers averaged 92.18%. We are confident that two or three complete transcriptions plus one or two complete adjudications has resulted in a reliable data set.

2.6 Forced alignment

Segmentations for the transcribed turns of each ADOS evaluation were produced by forced alignment using an aligner trained on all turns in the corpus. The aligner was trained with the Kaldi ASR toolkit (Povey et al., 2011) using the CMUdict lexicon with stress markings removed; pronunciations for out-of-vocabulary (OOV) words were generated with the Sequitur G2P toolkit

(Besacier, Barnard, Karpov, & Schultz, 2014) using a model trained on CMUdict. The acoustic frontend consisted of 13 mel frequency cepstral coefficient (MFCC) features extracted every 10 ms using a 25 ms Hamming window plus first and second differences; all features were normalized to zero mean and unit variance on a per-speaker basis. A standard 3-state Bakis model was used for all speech phones and a 5-state models allowing forward skips used to model non-speech phones (silence, breaths, coughs, laughter, lipsmacks, and other non-speech vocalizations), untranscribable regions, and out-of-vocabulary words (words which were not in CMUdict and for which grapheme-to-phoneme transduction failed). To improve segmentation accuracy, special 1-state boundary models were inserted at each phone transition as in Yuan et al. (2013). Acoustic modeling was performed using a deep neural network consisting of 4 layers of 512 rectified linear units with input consisting of an 11 frame context (5-1-5).

Feature extraction. In this first analysis, we focused on child features (lexical and acoustic). Planned future analyses will assess interviewer features, and integrate across both speakers to assess variables such as synchrony and accommodation.

Word choice. Prior research suggests that individuals with ASD produce idiosyncratic or unusual words more often than their typically developing peers (Ghaziuddin & Gerstein, 1996; Prud'hommeaux, Roark, Black, & Van Santen, 2011; Rouhizadeh, Prud'Hommeaux, Santen, & Sproat, 2015; Rouhizadeh, Prud'hommeaux, Roark, & van Santen, 2013; Volden & Lord, 1991); and may repeat words or phrases (van Santen, Sproat, & Hill, 2013). Using a lexical feature selection approach (Monroe, Colaresi, & Quinn, 2008), we calculated the frequency of each word in a child's transcript. We used this feature to classify samples as ASD or TD.

Disfluency. Differential use of the filler words "um" and "uh" has been found across men and women, older and younger people, and in ASD (Irvine, Eigsti, & Fein, 2016; Lunsford, Heeman, & Van Santen, 2012; Wieling et al., 2016). Here, we compared the percentage of UM relative to UM+UH across groups.

Speaking rate. In our pilot analysis, we found slower speaking rates in children with ASD vs. TD (Parish-Morris et al., 2016). We attempted to replicate this finding in a larger sample by calculating the mean duration of each word produced by participants in a speech segment (a stretch of speaking between silent pauses).

Latency to respond. Children with ASD have been reported to wait longer before responding in the course of conversation (Heeman, Lunsford, Selfridge, Black, & Van Santen, 2010). To explore this feature in our own sample, we calculated the elapsed time between clinician and child turns.

Fundamental frequency. Prior research has found that pitch variables distinguish language produced by children with ASD from language produced by typically developing children (Asgari, Bayestehtashk, & Shafran, 2013; Kiss, van Santen, Prud'hommeaux, & Black, 2012; Schuller et al., 2013). Here we compared the prosody of participants by calculating mean absolute deviation from the median (MAD) as an outlier-robust measure of dispersion in F0 distribution.

3 Preliminary analysis and results

The analyses and figures below are meant to spur interest and give a hint as to potential avenues to explore using a larger data set. A subset (N=46) of the current sample was described in a forthcoming paper (Parish-Morris et al., 2016).

3.1 Diagnostic classification

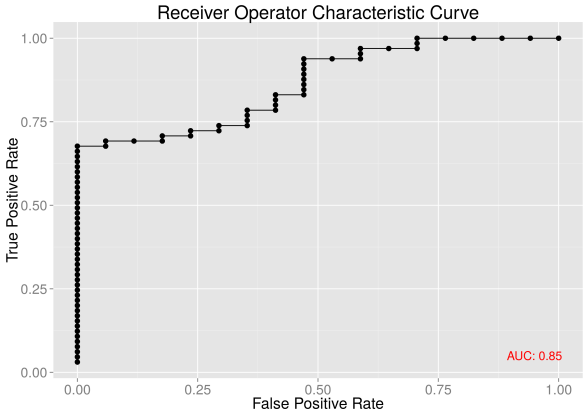
We found that word choice alone served surprisingly well to separate the ASD and TD groups. Naïve Bayes classification, using leave-one-out cross validation and weighted log-odds-ratios calculated using the "informative Dirichlet prior" algorithm of Monroe et al. (2008), correctly classified 68% of ASD patients and 100% of typical participants. Receiver Operating Characteristic (ROC) analysis revealed good sensitivity and specificity using this classification metric, with AUC=85% (Figure 1).

The 20 most "ASD-like" words in this analysis were: *{nsv}, know, he, a, now ,no , uh, well, is, actually, mhm, w-, years, eh, right, first, year, once, saw, was* (where {nsv} stands for "non-speech vocalization", meaning sounds that with no lexical counterpart, such as imitative or expressive noises). Of note, "uh" appears in this list, as does "w-", a stuttering-like disfluency.

At the other end of the scale, we found that the 20 least "ASD-like" words in this analysis were:

like, um, and, hundred, so, basketball, something, dishes, go, york, or, if, them, {laugh}, wrong, be, pay, when, friends. Here, the word "um" appears, as does the word "friends, and laughter.

Figure 1. Receiver operating characteristic on word choice separates ASD from TD.



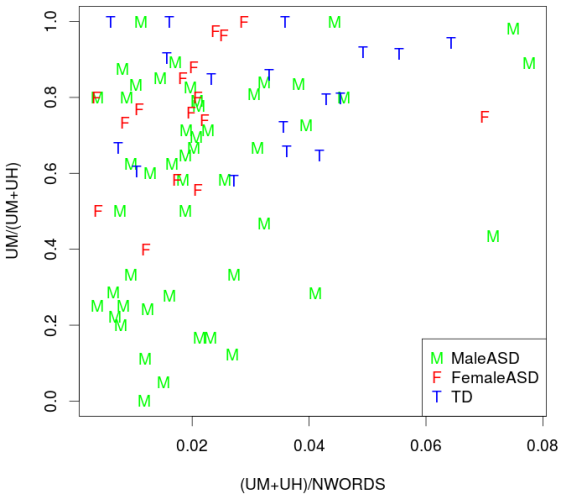
As we discuss below, many linguistic and phonetic features showed systematic differences among the diagnostic groups, and feeding combinations of these features into modern machine-learning algorithms will certainly do an even better job of classifying the participants in our dataset than a simple "bag of words" model. However, we feel that focus on classification at this stage is premature, because of the previously-referenced phenotypic diversity and uneven diagnostic group sizes in our sample. Rather, we believe that similar analysis of much larger datasets will enable us to place individuals in a space with several significant dimensions of relevant variation, rather than trying to force them into discrete categories.

3.2 Other feature differences

Disfluency. We compared rates of um production across the ASD and TD groups (um/(um+uh)). The ASD group produced UM as 61% of their filled pauses (CI: 54%-68%), while the TD group produced UM as 82% of their filled pauses (CI: 75%-88%). The minimum value for the TD group was 58.1%, and 23 of 65 participants in the ASD group fell below that value.

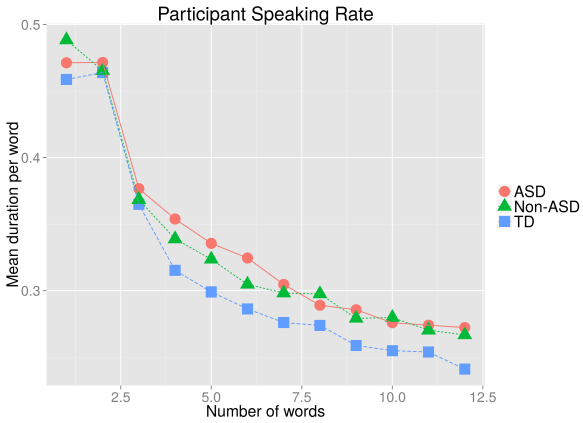
Given prior research showing sex differences on this variable (Wieling et al., 2016), we marked data points as originating from males or females for the purposes of visualization. Figure 2, plotting overall rate of filled pauses against the proportion of filled pauses that are UM, illustrates this interaction of sex and diagnostic category. This naturally raises the question of what other characteristics might also be correlated with these differences; and it underlines the opportunity to use data of this type to discover and explore new dimensions of relevant variation.

Figure 2. Disfluencies in the ASD and TD groups.



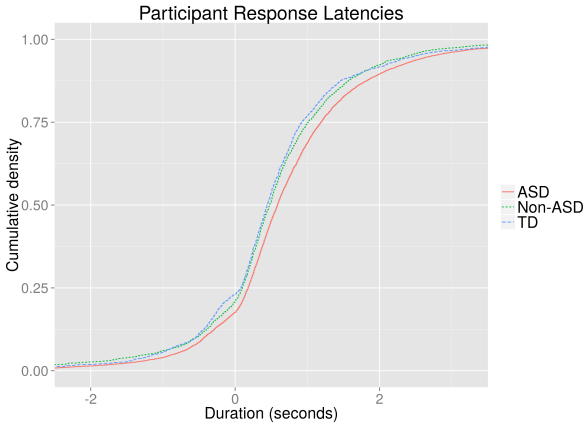
Speaking rate. A comparison of mean word duration as a function of phrase length revealed that TD participants spoke the fastest (overall mean word duration of 376 ms, CI 369-382, calculated from 6891 phrases), followed by the non-ASD mixed clinical group (mean=395 ms; CI 388-401, calculated from 6640 phrases), followed by the ASD group with the slowest speaking rate (mean=402 ms; CI: 398-405, calculated from 24276 phrases).

Figure 3. Mean word duration as a function of phrase length differed among all three groups.



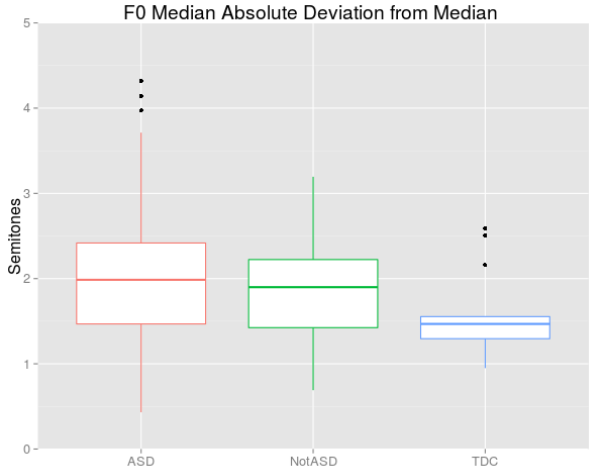
Child latency to respond. Our analyses revealed that children with ASD were slower to respond to interviewer bids for conversation than TD participants, with children in the non-ASD mixed clinical group falling in the between.

Figure 4: Cumulative distribution of response latencies for three diagnostic categories.



Fundamental Frequency. To compare the prosody of participants we examined an outlier-robust measure of dispersion in their F0 distribution: mean absolute deviation from the median (MAD). F0 contours were extracted for every ADOS session using an implementation of the Kaldi pitch tracking algorithm (Gharemani et al. 2014) using a 10 ms step, 10 ms analysis window width, and search range of 50 to 600 Hz, with all frames identified as belonging to a voiced phone in the forced alignment retained. After then dropping frames from speech segments (as defined in Section 2.7) of duration less than 500 ms and and longer than 5 seconds, F0 values were transformed from Hz to semitones using the 5th percentile of each speaker as the base, which served as input for computation of MAD. As depicted in the box-and-whisker plot in Figure 5, MAD values for F0 are both higher and more variable within the ASD and non-ASD mixed clinical group than the TD group (ASD: median: 1.99, IQR: 0.95; non-ASD: median: 1.95, IQR: 0.80; TD: median: 1.47, IQR: 0.26).

Figure 5: Median absolute deviation from median F0 in semitones relative to speaker's 5th percentile.



3.3 Correlations with clinical and demographic measures

Our relatively large group of 65 children with ASD offered an opportunity to examine within-group correlations. Due to space constraints, we focus on disfluency and response latency. Future analyses with a larger sample will explore these relationships in TD and non-ASD mixed clinical participants.

Disfluency. We explored relationships between the percent of um/uh disfluencies that were "um", and age/sex/IQ. No significant relationships were found with age or IQ (full-scale, verbal, or non-verbal). As suggested by Figure 2, we found significant sex differences in "um" fillers. Males with ASD filled pauses with "um" instead of "uh" at significantly lower rates (M=56%) than females with ASD (M=75%; Welch's t=-3.20, p=.003). This finding mirrors sex differences found in larger samples of typically developing adults (Wieling et

al., 2016). More "um" use was also associated with lower ADOS severity scores (Spearman's Rho=-.25, p=.045; males and females did not differ on autism severity), but parent ratings of social and communication competence as measured by the CCC-2 were unrelated to "um" use. This discrepancy could be due to the nature of the prolonged observation on the part of parents (judgments are based on years of observation, during which time parents may become used to their child's disfluencies) versus the short, time-constrained observations of clinicians.

Child latency to respond. The mean length of transitions from interviewer to participant did not correlate with age or any measure of IQ, nor did it differ by participant sex. It did, however, correlate positively with overall ADOS calibrated severity scores (Pearson's r =.28, p=.02). An examination of subscale severity scores suggests some measure of specificity; the social communication severity score of the ADOS correlated with response latency (Pearson's r=.31, p=.01), while the repetitive behaviors/restricted interests severity score did not (Pearson's r =.04, p=.73). As in the case of disfluencies, response latency did not correlate with parent reports of social communication competence.

3.4 Discussion

Our preliminary exploration of this new data set indicates that word choice produced during ADOS evaluations can be used to differentiate children with ASD from typically developing children with good sensitivity and specificity. Using a variety of features, including word choice, inter-turn pause length, and fundamental frequency, we were able to characterize the linguistic signal at a highly granular level. Importantly, we not only found that these features discriminate groups, but also showed that certain features also correlate with clinical presentation. This relationship suggests language-clinical connections that inform personalized approaches to social communication intervention.

Classification sensitivity and specificity using word choice went down relative to prior work with a smaller pilot sample (AUC: 92%; (Parish-Morris et al., 2016)). This may be due to at least two factors that underline the need for a larger corpus than the one we have at present. First, we increased the variability of our ASD sample by adding more heterogeneous participants. Our first pilot sample consisted of carefully selected "clean" groups of children with classic ASD and typically developing controls, whereas the extension reported here was much more realistic and clinically unclear (e.g., we included ASD participants with a milder phenotype or clinical comorbidities). Second, participants in our first pilot groups were individually matched on a variety of characteristics (age, IQ, sex, parent education, income). Our extended sample tripled our ASD sample, did not increase our TD sample significantly, and did not match individually. It is unusual for TD participants to be administered the ADOS evaluation in most labs, partly due to the expensive and time-consuming nature of the assessment. Large, comparable samples from TD participants are essential to research in this area, and will require efforts to improve accessibility and reduce cost in future studies.

4 Future directions

4.1 New sources, more participants

The ASD sample reported here is large relative to much prior work, but our analyses were nonetheless constrained by smaller comparison groups. We are actively annotating additional ASD samples from past studies conducted at the Center for Autism Research, collecting new data from an expanded age range in our lab (including preschoolers and adults), and seeking out collaborators who wish to contribute language samples to this collection. (Due to privacy laws, final transcripts and audio recordings from this and all other collections must be carefully wiped of personally identifiable information prior to sharing, a process that is currently underway for the present sample.) In particular, we are searching for diverse, well-phenotyped samples enriched for typically developing participants, participants with non-ASD clinical diagnoses, and females with ASD.

Fewer girls than boys are diagnosed with ASD (Christensen, 2016), and they have been historically understudied. Significant linguistic features interact with sex, however (e.g., differences in the use of disfluencies such as um/uh), making this variable especially important to study. We aim to build a cohort of samples from females with and

without ASD, to explore the effects of sex-diagnosis interactions on language features.

One immediate goal for our team is to begin telephone collection of ADOS-like samples. Research-grade ADOS recordings, while retrospectively ubiquitous, are not inexpensive or easy to obtain. At present, participants and families must meet with a highly trained clinician, often traveling long distances to do so. We are in the process of developing a protocol that can be administered over the telephone, with relatively untrained conversational partners. We aim to explore the relative classification and characterization value of this method versus rigorous lab-based ADOS recordings.

4.2 Interviewer Analysis

Our current analysis is far from comprehensive. Most notably, we constrained our analyses to child features. Given that the ADOS evaluation is a conversation, it is essential to analyze interviewer speech and language characteristics as well (Bone et al., 2012; Bone, Lee, Black, et al., 2014; Bone, Lee, Potamianos, & Narayanan, 2014). Future analysis plans include assessing dynamic relationships between interviewer and child features over the course of the evaluation.

4.3 Additional types of annotation

We saw interesting patterns in the use of UM and UH emerge in the present analysis, and in the list of the most ASD-associated words, we saw one example of a fluent self-correction, namely the partial word *w-*. This suggests that a more comprehensive annotation of disfluencies, including their semantic, morpho-syntactic, phonetic, and prosodic affinities, would be informative.

Word frequencies were surprisingly diagnostic – perhaps the frequency of syntactic and semantic word categories will also be interesting, including things like parts of speech, negations, and contractions. It is likely to be worthwhile to distinguish the semantic categories of referents, e.g. to individuals, groups, places, and so on. Various other semantic categories may also be interesting – concreteness of reference, span of co-reference relations, definiteness and indefiniteness, and so on.

We saw some signal in simple counts of turn length and speech-segment length – it is plausible that we would learn more from an analysis of syntactic features such as clause length, depth of embedding, frequency of various sorts of modification, etc. Modern analysis techniques can make it relatively cheap to get high-quality analyses of this type.

We could multiply examples almost indefinitely. Our main point in starting the list is that when we have a large body of sharable data of this type, then researchers with new ideas can add their own layers of annotation and explore the resulting patterns. Modern techniques for tagging, parsing, and other sorts of analysis will make such explorations increasingly efficient – as long as a large body of appropriate data is available.

Acknowledgments

This project was supported by an Autism Science Foundation Postdoctoral Fellowship awarded to J. Parish-Morris (Mentors: R.T. Schultz, M. Liberman, C. Cieri), by the NICHD funded Intellectual and Developmental Research Center grant (# U54 HD86984, M Yudkoff, PI), a grant from the Pennsylvania Department of Health (SAP # 4100047863) to R. Schultz and by NIMH RC1MH08879 (R. Schultz, PI). We gratefully acknowledge all the children and families that participated in our research, as well as the clinicians, staff, and students that collected/compiled data at the Center for Autism Research.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5* (5 edition). Washington, D.C: American Psychiatric Publishing.

Asgari, M., Bayestehtashk, A., & Shafran, I. (2013). Robust and accurate features for detecting and diagnosing autism spectrum disorders. In *INTERSPEECH* (pp. 191–194). Retrieved from http://www.cslu.ogi.edu/~zak/is13-autism.pdf

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. http://doi.org/10.1016/j.specom.2013.07.008

Bone, D., Black, M. P., Lee, C.-C., Williams, M. E., Levitt, P., Lee, S., & Narayanan, S. (2012). Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist. In *INTERSPEECH*. Retrieved from http://sail.usc.edu/~dbone/Bone_spontaneousProsody_ADOS_IS2012.pdf

Bone, D., Goodwin, M. S., Black, M. P., Lee, C.-C., Audhkhasi, K., & Narayanan, S. (2014). Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises. *Journal of Autism and Developmental Disorders*. http://doi.org/10.1007/s10803-014-2268-6

Bone, D., Lee, C.-C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: Insights From a Study of Spontaneous Prosody. *Journal of Speech Language and Hearing Research*, *57*(4), 1162. http://doi.org/10.1044/2014_JSLHR-S-13-0062

Bone, D., Lee, C.-C., Potamianos, A., & Narayanan, S. (2014). An Investigation of Vocal Arousal Dynamics in Child-Psychologist Interactions using Synchrony Measures and a Conversation-based Model. In *Fifteenth Annual Conference of the International Speech Communication Association*. Retrieved from https://mazsola.iit.unimiskolc.hu/~czap/letoltes/IS14/IS2014/PDF/AUTHOR/IS140377.PDF

Christensen, D. L. (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR. Surveillance Summaries*, *65*. Retrieved from http://www.cdc.gov/mmwr/volumes/65/ss/ss6503a1.htm

D'Souza, D., Booth, R., Connolly, M., Happé, F., & Karmiloff-Smith, A. (2015). Rethinking the concepts of "local or global processors": evidence from Williams syndrome, Down syndrome, and Autism Spectrum Disorders. *Developmental Science*. http://doi.org/10.1111/desc.12312

Elliott, C. D. (2007). Differential Ability Scales®-II - DAS-II. San Antonio, TX: Harcourt Assessment. Retrieved from http://www.pearsonclinical.com/education/products/100000468/differential-ability-scales-ii-das-ii.html

Fein, D., Barton, M., Eigsti, I.-M., Kelley, E., Naigles, L., Schultz, R. T., … Tyson, K. (2013). Optimal outcome in individuals with a history of autism: Optimal outcome in individuals with a history of autism. *Journal of Child Psychology and Psychiatry*, *54*(2), 195–205. http://doi.org/10.1111/jcpp.12037

Ghahremani, P., BabaAli, B., Povey, D., Ried hammer, K., Trmal, J., & Khudanpur, S. (2014, May). A pitch extraction algorithm tuned for automatic speech recognition. In *IEEE ICASSP 2014*.

Ghaziuddin, M., & Gerstein, L. (1996). Pedantic speaking style differentiates Asperger Syndrome from High-Functioning Autism.

Heeman, P. A., Lunsford, R., Selfridge, E., Black, L., & Van Santen, J. (2010). Autism and interactional aspects of dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 249–252). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1944551

Insel, T. R. (2014). The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. *American Journal of Psychiatry*, *171*(4), 395–397.

Irvine, C. A., Eigsti, I.-M., & Fein, D. A. (2016). Uh, Um, and Autism: Filler Disfluencies as Pragmatic Markers in Adolescents with Optimal Outcomes from Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *46*(3), 1061–1070. http://doi.org/10.1007/s10803-015-2651-y

Kiss, G., van Santen, J. P., Prud'hommeaux, E. T., & Black, L. M. (2012). Quantitative Analysis of Pitch in Speech of Children with Neurodevelopmental Disorders. In *INTERSPEECH*. Retrieved from http://20.210-193-52.unknown.qala.com.sg/archive/archive_papers/interspeech_2012/i12_1343.pdf

Lord, C., Rutter, M., DiLavore, P. S., Risi, S., Gotham, K., & Bishop, S. L. (2012). Autism diagnostic observation schedule, second edition (ADOS-2). Torrance, CA: Western Psychological Services.

Lunsford, R., Heeman, P. A., & Van Santen, J. P. (2012). Interactions Between Turn-taking Gaps, Disfluencies and Social Obligation. In *INTERSPEECH* 2012.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, *16*(4), 372–403.

National Database for Autism Research - Home. (n.d.). Retrieved April 7, 2016, from https://ndar.nih.gov/

Norbury, C. F., Nash, M., Baird, G., & Bishop, D. (2004). Using a parental checklist to identify diagnostic groups in children with communication impairment: a validation of the Children's Communication Checklist--2. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *39*(3), 345–364. http://doi.org/10.1080/13682820410001654883

Parish-Morris, J., Cieri, C., Liberman, M., Bateman, L., Ferguson, E., & Schultz, R. T. (2016). Building Language Resources for Exploring Autism Spectrum Disorders. *Proceedings of the Language Resources and Evaluation Conference 2016*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … others. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. Retrieved from http://infoscience.epfl.ch/record/192584

Prud'hommeaux, E. T., Roark, B., Black, L. M., & Van Santen, J. (2011). Classification of atypical language in autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 88–96), ACL.

Rouhizadeh, M., Prud'Hommeaux, E., Santen, J. V., & Sproat, R. (2015). Measuring idiosyncratic interests in children with autism. Presented at the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL-IJCNLP 2015.

Rouhizadeh, M., Prud'hommeaux, E. T., Roark, B., & van Santen, J. P. (2013). Distributional semantic models for the evaluation of disordered language. In *HLT-NAACL* (pp. 709–714). Citeseer.

Rutter, M., Bailey, A., & Lord, C. (2003). SCQ: The Social Communication Questionnaire. Los Angeles, CA: Western Psychological Services. Retrieved from https://www.wpspublish.com/store/Images/Downloads/Product/SCQ_Manual_Chapter_1.pdf

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., … others. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.

van Santen, J. P. H., Sproat, R. W., & Hill, A. P. (2013). Quantifying Repetitive Speech in Autism Spectrum Disorders and Language Impairment: Repetitive speech in ASD and SLI. *Autism Research*, *6*(5), 372–383. http://doi.org/10.1002/aur.1301

Volden, J., & Lord, C. (1991). Neologisms and idiosyncratic language in autistic speakers. *Journal of Autism and Developmental Disorders*, *21*(2), 109–130. http://doi.org/10.1007/BF02284755

Weigelt, S., Koldewyn, K., & Kanwisher, N. (2012). Face identity recognition in autism spectrum disorders: A review of behavioral studies. *Neuroscience & Biobehavioral Reviews*, *36*(3), 1060–1084. http://doi.org/10.1016/j.neubiorev.2011.12.008

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., & Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, *Forthcoming*.

Wing, L. (1981). Sex ratios in early childhood autism and related conditions. *Psychiatry Research*, *5*(2), 129–137.

Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., & Wang, W. (2013). Automatic phonetic segmentation using boundary models. In *INTERSPEECH* 2013.