Using natural conversations to classify autism with limited data: Age matters

Michael Hauser[1] **Evangelos Sariyanidi**[1] **Birkan Tunc**[1,2] **Casey J. Zampella**[1]

Edward S. Brodkin[2] **Robert T. Schultz**[1,2,3] **Julia Parish-Morris**[1,2]

[1] Center for Autism Research, Children's Hospital of Philadelphia
[2] Department of Psychiatry, University of Pennsylvania
[3] Department of Pediatrics, University of Pennsylvania

Abstract

Spoken language ability is highly heterogeneous in Autism Spectrum Disorder (ASD), which complicates efforts to identify linguistic markers for use in diagnostic classification, clinical characterization, and for research and clinical outcome measurement. Machine learning techniques that harness the power of multivariate statistics and non-linear data analysis hold promise for modeling this heterogeneity, but many models require enormous datasets, which are unavailable for most psychiatric conditions (including ASD). In lieu of such datasets, good models can still be built by leveraging domain knowledge.

In this study, we compare two machine learning approaches: the first approach incorporates prior knowledge about language variation across middle childhood, adolescence, and adulthood to classify 6-minute naturalistic conversation samples from 140 age- and IQ-matched participants (81 with ASD), while the other approach treats all ages the same. We found that individual age-informed models were significantly more accurate than a single model tasked with building a common algorithm across age groups. Furthermore, predictive linguistic features differed significantly by age group, confirming the importance of considering age-related changes in language use when classifying ASD. Our results suggest that limitations imposed by heterogeneity inherent to ASD and from developmental change with age can be (at least partially) overcome using domain knowledge, such as understanding spoken language development from childhood through adulthood.

1 Introduction

Autism Spectrum Disorder (ASD) is a neurobiologically-based condition characterized by social communication impairments and restricted, repetitive patterns of behaviors and interests [1]. Although ASD is a neurodevelopmental disorder, it is currently diagnosed using behavior alone, including spoken language. For the roughly 70 percent of individuals with ASD that have average to above-average verbal abilities [2], language is an important pathway to social connections. For clinicians and care providers, spoken language can provide a window into internal cognitive and social processing. Given that primary diagnostic tools for ASD often rely on language-mediated semi-structured interviews and play activities to elicit behaviors found in the condition [3], measuring and quantifying subtle differences in spoken language between individuals with ASD and matched typically developing (TD) controls is important for improving diagnostic speed and reliability. Furthermore, since the emergence of spoken language before age 5 is a critical predictor of later functional outcomes in ASD [4, 5, 6], characterizing spoken language development is crucial for understanding long-term developmental outcomes.

Behavioral heterogeneity in ASD is a persistent challenge for researchers and clinicians. In fact, generalizability from one individual to the next is so low that it is often said, "If you have met one person with autism, you have met one person with autism". Wide phenotypic variability has made it difficult to draw reliable statistical conclusions about ASD, and indeed, has made it challenging to study the disorder at all [7]. Significant variability is similarly present in the verbal domain, with the spoken language skills of individuals with ASD ranging from severely impaired to verbally gifted [8]. As an illustration, a recent narrative study found that intra-group variability (ASD alone) was greater than inter-group variability (between ASD and TD) [9].

Recent attempts to leverage machine learning for understanding and classifying individuals with

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 45–54
Minneapolis, Minnesota, June 6, 2019.

ASD have grappled with this phenotypic variability [10, 11]. Unfortunately, many of the most exciting machine learning models (e.g., models that are able to capture nonlinear dependencies across many dimensions), require large, well-characterized training datasets to function correctly, which are rare in ASD (and are particularly scarce for children). These two constraints in ASD research (wide variability in high dimensional spaces, and lack of large datasets), suggest that it may be useful to proactively incorporate information that psychiatrists and linguists deem important, thus guiding machine learning models to learn relevant dependencies while ignoring irrelevant ones.

2 Language in ASD

Prior research suggests that language is a valuable metric that can be used to distinguish individuals with ASD from TD controls. For example, the NEPSY narrative retelling test, in which a child listens to and retells a story while being evaluated on how many key story elements were remembered, has been explored for its utility in supporting ASD identification [12]. In an analysis of 97 children aged 4-8 years, Prud'hommeaux and colleagues found that children with ASD were more likely than TD controls to veer off topic and incorporate their own specialized interests into the narrative. Similarly, another study showed that TD children are more likely to use similar words and semantic concepts to those given in the narrative, while children with ASD will retell the narrative with different words and concepts related to their own specialized interests [9]. Although promising, these and other studies that focus on one-sided language samples, rather than more ecologically valid conversations, miss a potential source of informative variance in language in ASD: the conversational partner.

Typically, natural conversations involve dynamic adjustments on a variety of levels that facilitate rapport and communication; this is called "linguistic accommodation" or "alignment" [13]. Increased accommodation is associated with perceptions of better conversation [14], but most prior research on language in ASD has used samples from structured or semi-structured elicitation tasks - or conversations conducted with an autism specialist - rather than natural conversations [15]. Thus, it is unknown whether and how typical (non-expert) speakers adjust their conversational behaviors to accommodate social communication differences in ASD, and whether the extent of accommodation changes over the course of development. To explore this new area, the machine learning models employed in this study include include dyadic features derived from a natural conversation (such as turn-taking rates) and interlocutor (conversation partner) features, as well as features from individuals with ASD.

3 Developmental Changes in Conversation

Individuals with and without ASD continue to develop socially and cognitively throughout childhood, adolescence, and into early adulthood. For example, although Theory of Mind (or the ability to take another person's perspective) emerges in early childhood [16], it becomes increasingly sophisticated throughout typical adolescence and early adulthood [17]. Thus, age-related differences in conversation (which is inherently social) are likely to be found.

Physical and emotional changes between childhood and adolescence (e.g., puberty [18]) increase the likelihood that people's preferred topic of conversation might change over time as well. Whereas young children may be more likely to talk about family and school, older children may be more focused on peer relationships [19], and adults might naturally gravitate toward talking about occupations or romantic partners. Unfortunately, few studies have explored natural conversation across development, and normative expectations for brief conversations are poorly understood across developmental phases and ages.

4 Current Study

The purpose of the current study is to test whether separating a large sample of individuals with and without ASD into different age groups, namely middle childhood (8 to 11), adolescence (12 to 17) and adulthood (18 and up), increases the accuracy and reliability of a simple machine learning classification model for classifying ASD vs. TD, despite inevitable trade-offs in sample size.

Given the likelihood that natural conversation differs between children and adolescents in a variety of measurable ways (e.g., preferred topics), and that adolescents also converse differently than adults, we hypothesized that diagnostic classifica-

tion accuracy would improve significantly when conducted within each age group separately, as compared to the combined sample. This is in contrast to generally accepted doctrine in machine learning (i.e., that more data is better), since in our study we divide our larger dataset into three smaller datasets.

We further tested whether the specific features that best distinguished diagnostic groups differed significantly by age. Based on prior research and clinical observation, we hypothesized that the relative predictive value of specific features would differ across development.

5 Methods

5.1 Participants

One hundred forty individuals participated in the present study (ASD: N=81, TD: N=59). Participants were categorized by age into three subgroups (see Table 1): middle childhood (8-11 years), adolescence (12-17 years) and adulthood (18-50 years). Diagnoses were confirmed (ASD group) or ruled out (TD group) using the Clinical Best Estimate process [20] informed by the Autism Diagnostic Observation Schedule - Second Edition (ADOS-2) [3] and adhering to DSM-V criteria for ASD [21]. To control for non-age related phenotypic heterogeneity, age subgroups were matched on Full Scale IQ estimates (WASI-II) [22], verbal and nonverbal IQ estimates, and sex ratio (Table 1). Participants with ASD were also matched across age subgroups on autism symptom severity, based on ADOS-2 Calibrated Severity Scores [23] and scores on the Social Communication Questionnaire (SCQ) [24]. All participants were native English speakers.

5.2 Procedure

All aspects of this study were approved by the Institutional Review Boards of the Children's Hospital of Philadelphia and the University of Pennsylvania. All adult participants and parents of minor children provided written informed consent for participation. The primary experimental task for this study was a slightly modified version of the Contextual Assessment of Social Skills (CASS) [25]. The CASS is a semi-structured assessment of conversational ability designed to mimic real-life first-time encounters. Participants engaged in two three-minute face-to-face conversations with two different confederates (research staff, blind to participant diagnostic status and unaware of the dependent variables of interest). In the first conversation (Interested condition), the confederate demonstrated social interest by engaging both verbally and non-verbally in the conversation. In the second conversation (Bored condition), the confederate demonstrated boredom and disengagement both verbally (e.g., one-word answers, limited follow-up questions) and non-verbally (e.g., neutral affect, limited eye-contact and gestures). Prior to each conversation, study staff provided the following prompt to the participants and confederates before leaving the room: "Thank you both so much for coming in today. Right now, you will have three minutes to talk and get to know each other, and then I will come back into the room."

CASS confederates included 42 undergraduate students or BA-level research staff (12 males, 30 females, all native English speakers). Fourteen confederates interacted with the ASD group, 7 with the TD group, and 21 with both groups. Confederates were semi-randomly selected, based on availability and clinical judgment. Confederate sex ratios did not differ by diagnostic group (p=n.s.). In order to provide opportunities for participants to initiate and develop the conversation, and in accordance with CASS confederate instructions [25], confederates in both conditions were trained to wait 10 seconds before initiating the conversation and to speak for no more than 50% of the time. If conversational lapses occurred, confederates were trained to wait 5 seconds before re-initiating the conversation. No specific conversational topic prompts were provided to either speaker.

Audio/video recordings of CASS conversations were obtained using a specialized "TreeCam", built in-house (Figure 1), placed between the participant and confederate (seated facing one another) on a floor stand. The TreeCam has two HD video cameras pointing in opposite directions to allow simultaneous recording of participant and confederate, as well as directional microphones to record audio. For these analyses, the language sample began when the first word of the CASS was uttered, after study staff left the room, and ended when study staff re-entered.

Table 1: Sex ratio, mean age (in years) and mean IQ scores for ASD and TD children (8-11 years), adolescents (12-17 years), and adults (18-50 years), and measures of autism symptoms for ASD participants.

Dx	N	Age group	N	Sex (f/m)	Age	Full-scale IQ	Verbal IQ	Non-verbal IQ	ADOS CSS	SCQ
ASD	81	Children	22	8, 14	9.98	105	103	105	7.32	19.81
		Adolescents	24	7, 17	14.62	102	103	101	6.58	17.38
		Adults	35	5, 30	26.73	104	108	99	7.06	17.23
TD	59	Children	19	8, 11	9.58	103	104	102	·	·
		Adolescents	12	6, 6	14.17	103	101	103	·	·
		Adults	28	5, 23	28.42	109	110	106	·	·

Note: Diagnostic groups did not significantly differ on sex ratio, age, or IQ within age bins, and age bins did not differ from one another on these variables (all p=ns). In the ASD group, age bins did not differ significantly from one another on ADOS-2 calibrated severity scores (CSS) or on SCQ scores (all p=ns). Five participants with ASD had missing scores on the SCQ (1 child, 4 adults).

(a) The TreeCam audio/video capture device.

(b) Illustration of the task environment. Participants and confederates sat face-to-face while engaging in a "get to know each other" dialogue, with the TreeCam placed in between.

Figure 1: Experimental setup of the TreeCam device, as well as participants and confederates.

5.3 Audio Data Processing

Audio streams were extracted from audio/video recordings, and saved in lossless .flac format. A team of reliable annotators produced time-aligned, verbatim, orthographic transcripts of audio recordings in the transcription software XTrans [26]. Each recording was processed by two junior annotators and one senior annotator, all of whom were undergraduate students and native English speakers. Before becoming junior annotators for this cohort, each team member received at least 10 hours of training in Quick Transcription [27] modified for use with clinical interviews of participants with ASD [10, 11, 28]. In addition, annotators achieved reliability (defined as >90% in common with a Gold Standard transcript) on segmenting (marking speech start and stop times) and transcribing (writing down words and sounds produced, using the modified Quick Transcription specification) before beginning independent annotation. Training files included audio recordings of conversations between individuals with and without autism that were not used in this study.

For CASS recordings, one reliable junior annotator segmented utterances into pause groups, while the second transcribed words produced by each speaker. A senior annotator then thoroughly reviewed and corrected each file. All senior annotators had at least 6 months of prior transcription experience. Final language data were exported from XTrans as tab-delimited files that were batch imported into R. Annotations marking non-speech sounds like laughter, indicators of language errors like stutters, and punctuation were removed, while other disfluencies (including filled pauses and whole-word repetitions) remained.

5.4 Speech/Language Features

One hundred twenty-three features were calculated for each speaker (participant, confederate) in the Bored condition and the Interested condition separately, using base R [29], qdap [30], and Linguistic Inquiry and Word Count (LIWC) software [31]. There were six main feature groups: pause/overlap metrics (12), segment/turn metrics (6), speaking rate/word complexity metrics (9), LIWC categories (80), lexical entropy/diversity measures (5), and parts of speech (9). Formality and polarity (2) were also computed at the conversation level for each speaker, using all words produced by a given speaker in each condition, leading to a total of 123 linguistic features. Differences between speakers were calculated within each condition (Participant Interested - Confederate Interested, Participant Bored - Confederate Bored) and within each speaker across conditions (Participant Interested - Participant Bored, Confederate Interested - Confederate Bored), yielding $8 \times 123 = 984$ features.

LIWC [31] is a commonly used software for an-

alyzing text-based natural language data. LIWC relies on a dictionary of words that are grouped by semantic similarity into lexical categories. These word-language lexica are designated by a majority vote by human judges, as are which words that fall into each, or multiple, of these lexica. This type of text analysis has been used successfully to analyze various mental disorders [32], as well as to characterize personality traits from transcribed language or written text [33].

Lexical features are included in the current study as they have proven informative in prior ASD research. For example, the words produced by interviewing psychologists correlate significantly with ASD symptom severity [34]. Bone and colleagues conducted their analysis across a wide age range (3.58 to 13.17 years), and interlocutors were autism experts, but their research nonetheless suggests that word choice by conversational partners could be a potentially sensitive marker of ASD phenotype. In the current study, confederate word choice is captured.

Difference metrics were included in our feature set for two primary reasons. First, the original intent of the CASS task was to probe how individuals with ASD handle variations in conversational context, as compared to TD peers. Thus, within-speaker differences across two contexts (Bored interlocutor, Interested interlocutor) are pertinent relative to the original design. Second, interlocutor differences within a given condition were included as a general measure of linguistic accommodation; to study how closely the speaking rates, pause rates, and preferred conversational topics of the two speakers align. Research shows that greater linguistic accommodation is associated with social success [35] and also suggests that reduced accomodation in ASD in childhood [36] may improve by adulthood [37].

We recognize that for linear models, introducing new features as linear combinations of old features (such as the difference between the Interested and Bored conditions) is algebraically equivalent to not introducing these features at all. However, by introducing these additional features, we are guiding the model to learn dependencies that clinicians deem important and have functional value in real-world social contexts. This is especially true when using an automated feature selection technique, such as the f-value employed here, as these techniques limit the number of dimensions that can be used by a model. In the current study, rather than requiring our model to learn to take the difference across two dimensions, we are giving the model this knowledge *a priori*, and thus allowing the model to learn to use this difference with only one dimension. This type of reasoning forms the motivation for sparse coding (see below).

6 Results and Discussion

6.1 Model Design

Linear logistic regression, also known as the Maximum Entropy classifier or the softmax classifier, was used to classify ASD vs. TD. Features were down-selected before being input into the model by identifying dimensions with the highest f-value (largest mean separation between groups). The model was trained and tested according to leave one out, with an internal 5-fold cross validation to determine what percentage of the total features are kept from the f-value, selected from 0.5%, 1%, 2%, 5%, 10% or 20%. The top scoring f-test values can be seen in Figure 3 for the different age ranges. We used an ℓ_2-regularization penalty in the cost function in order to smooth out model coefficients. Our models were implemented in the Python library SciKit-Learn [38].

We use logistic regression so as to have an interpretable linear model. With more complex nonparametric and/or non-linear models, it is more difficult to understand the contribution of different variables on the model performance. We did not use a sparsity constraint in the model, such as an ℓ_1 penalty, since we are already imposing sparsity on the feature space by downsampling the feature dimension to those features with large f-values.

When designing the model, one may consider using age or gender as a covariate that automatically adjusts the model parameters, within for example a hierarchical Bayes network [39]. There are at least two difficulties with doing this in a purely data driven way. First, it introduces many additional parameters into the model one would need to learn, which on limited data is suboptimal in a statistical sense. Second, such hierarchical models are nonlinear, and thus difficult to interpret, which was an important design criteria for our model. Instead, we chose to use domain knowledge from developmental psychology to strictly define different models for different developmental age groups.

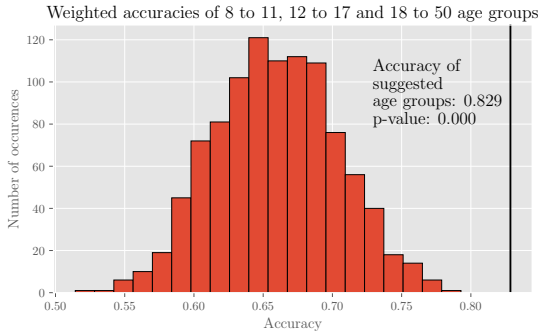


Figure 2: Comparison of the classification accuracy (weighted average of the three age groups) of the actual age-based split against 1000 randomized splits (not based on age) where sample sizes and proportions of classes in each sample were kept same as the actual split. The black vertical line shows the actual accuracy, and the red histogram shows the distribution of accuracy for random splits. The proportion of the distribution to the right of the vertical line defines the p-value.

Table 2: Classification accuracy for the three individual age groups and the entire sample. The weighted average (based on sample size) accuracy of the three age-specific models is 0.829 ($p < 0.001$, see Figure 2).

Age Range of Model	Accuracy
8 to 11	0.756
12 to 17	0.806
18 to 50	0.889
Weighted average	0.829
8 to 50	0.686

6.2 Classification Accuracy

Classification accuracy for three age-specific models, as well as the accuracy of a model for all ages together (8 and older), are shown in Table 2. Age-specific models outperformed the single model. The weighted average of the three age-specific models, weighted according to number of samples in each age group, was 0.829. In contrast, the single model for all ages achieved an accuracy of 0.686. Thus, our age-informed approach resulted in a 20.8% relative increase in accuracy, $p < 0.001$ (Figure 2). Again, this is notable as it contrasts with the standard doctrine in machine learning that training a model on more data is better; in our case we trained three models on roughly a third of the data each, yielding improved results.

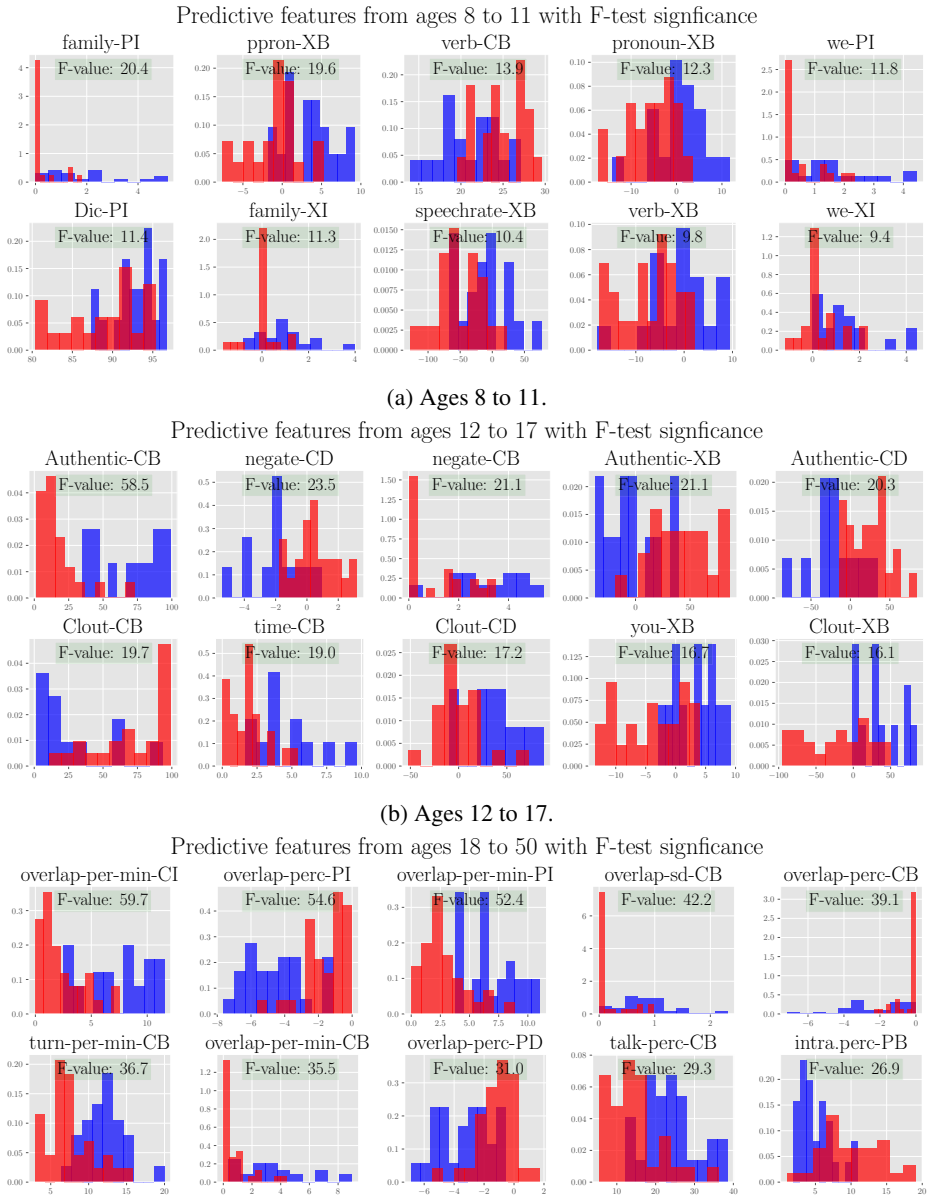
6.3 Distinguishing Features by Age Group

Different linguistic features emerged as important for distinguishing between TD and ASD participants in each age group, as seen in Figure 3.

In the 8 to 11 age group, overall pronouns and personal pronouns predicted diagnostic status, such that children with ASD produced smaller proportions of pronouns than matched TD peers. In particular, the first person plural pronoun "we" was used relatively less frequently by the ASD group, suggesting that children with ASD were less likely to describe themselves as associating with others during conversation. Children in the ASD group also tended to use more out-of-dictionary words than TD children (i.e., they produced a smaller percentage of words that were in the LIWC dictionary, relative to their total word production), which could be due to children with ASD talking about specialized, idiosyncratic interests or simply using low-frequency words or phrases. Finally, children with ASD spoke more slowly, measured in words per minute with breath pauses removed, than matched TD children, and used comparatively fewer verbs (Figure 3a).

Top linguistic features that predicted diagnosis in the 12 to 17 age group are shown in Figure 3b. The Bored condition emerged as particularly important for distinguishing between TD and ASD adolescents, as did confederate word choice. Pronouns were predictive in this age group as well. Specifically, the second person personal pronoun "you" was produced relatively more often by TD teens in relation to confederates in the Bored condition. This could indicate more attempts by the TD group to engage with an obviously bored conversational partner, and relatively diminished effort put forth by teens with ASD. Confederates speaking with autistic teens used words associated with less authenticity, but greater clout, than when speaking with TD peers, and responded more often to TD participants with negations (perhaps in response to increased questions/comments about themselves, as indicated by greater use of "you" by TD teens).

Finally, linguistic features that differentiated between conversation samples from adults with and without ASD are shown in Figure 3c. Interestingly, these features were primarily temporal; for example, top features included the number of overlapping pauses (interruptions) in the conversation, as well as the rate of pauses per minute. This suggests that whereas topics of conversation might be comparable in ASD and TD adults (i.e., similar tendencies to discuss occupations or romantic



(a) Ages 8 to 11.

(b) Ages 12 to 17.

(c) Ages 18 and older.

Figure 3: Histograms of the top 10 most discriminant features (ranked by f-test value) for the different age ranges considered, namely middle childhood, adolescence, and adulthood. In all figures, red is the ASD sample, and blue is the TD sample. Acronyms: PI = participant:interested, PB = participant:bored, PD = participant:difference (interested-bored), CI = confederate:interested, CB = confederate:bored, CD = confederate:difference (interested-bored), XI = cross:interested (participant-confederate), XB = cross:bored (participant-confederate).

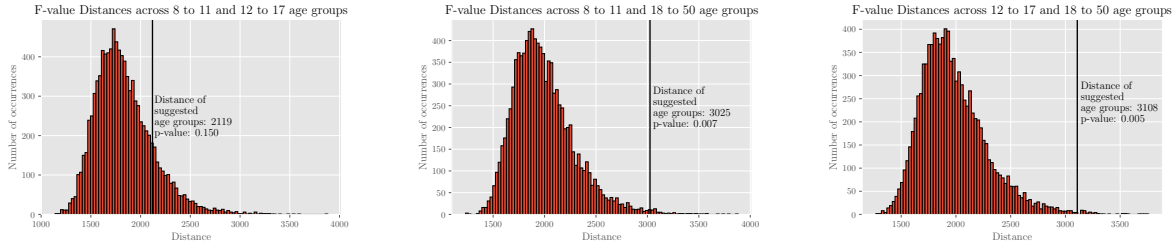
partners), the way in which conversations occur may include awkward pauses, interruptions, and other temporal atypicalities that could negatively impact conversation quality.

The linguistic features identified in our machine learning analysis are consistent with prior research, as well as with observations about ASD made by clinicians and linguists. Importantly, our analysis goes a step further by quantifying the *extent* to which each of these features is important for distinguishing diagnostic groups at each age.

6.4 Feature Consistency Across Age Groups

The purpose of this subsection is to quantify which predictive speech/language features change by age group (i.e., how many predictive features remain predictive regardless of age). To do this, we measured change in the f-value.

Suppose we have age groups $(8, 11)$ and $(12, 17)$, and would like to compare changes in



(a) Ages (8,11) and (12,17) (b) Ages (8,11) and (18,50) (c) Ages (12,17) and (18,50)

Figure 4: f-value distances ($\|f_i - f_j\|_1$) of the actual age-based split against 1000 randomized splits (not based on age) where sample sizes and proportions of classes were kept same as the actual split. The black vertical lines show the actual distance, and the red histograms show the distributions of distances for random splits.

f-values between $f_{(8,11)}$ and $f_{(12,17)}$. Since each participant is associated with $8 \times 123 = 984$ features, as mentioned in Section 6.1, then $f_{(8,11)}$ and $f_{(12,17)}$ are both 984-dimensional vectors, with each dimension containing the f-value of its corresponding feature. Measuring distances across dimensions does not make sense in this case, as each of the individual f-values are calculated in one dimension independently of each other. Thus, we use the ℓ_1-norm, sometimes referred to as the Manhattan distance, when measuring these distances, i.e. $\|f_{(8,11)} - f_{(12,17)}\|_1$.

Given that the magnitudes $\|f_{(8,11)}\|_1 = 1505$, $\|f_{(12,17)}\|_1 = 1848$ and $\|f_{(18,50)}\|_1 = 3035$, we see that the changes in magnitude of the feature importance from one age group to another are proportionally very large, and in fact often exceed, the magnitude of the features themselves. This tells us that the specific linguistic features that are important for distinguishing between ASD and TD, as defined by the f-test, vary enormously across age groups, especially when considered against the scale of the linguistic features themselves (Figures 4a- 4c, and Table 3).

Table 3: Measuring the extent to which the feature importance changes with the ℓ_1-norm, according to each feature's f-value, depending on which age group is under consideration. The p-value corresponds to distances developed from the null hypothesis where no age groups are considered, while ensuring correct proportions of ages and classes are kept.

Measurement	Value	p-value
$\|f_{(8,11)} - f_{(12,17)}\|_1$	2119	0.150
$\|f_{(8,11)} - f_{(18,50)}\|_1$	3025	0.007
$\|f_{(12,17)} - f_{(18,50)}\|_1$	3108	0.005

7 Discussion

In this study, we demonstrated that machine learning models for classifying and characterizing ASD improve significantly after incorporating domain knowledge. Specifically, we showed that models accounting for developmental changes in spoken language and conversation are more accurate for distinguishing ASD from typical development, relative to models resting on the assumption that language patterns during natural conversation remain consistent across ages. We further showed that linguistic features most strongly predicting ASD vary significantly across age groups, suggesting that specific atypicalities in the ways that individuals with ASD use language (versus TD controls) are not static across development.

These findings highlight the value of machine learning models that are clinically informed, particularly for understanding highly heterogeneous conditions like ASD. Developing separate models for different age groups (i.e., middle childhood, adolescence, and adulthood), we were able to significantly improve the models' classification performance and reliability, despite reductions in sample size. This bodes well for future applications of machine learning for studying psychiatric conditions. Future research will incorporate pitch-related features, extend classification to non-ASD psychiatric conditions, and explore the use of more complex nonlinear models for classification and prediction in larger sample sizes.

8 Conclusions and Future Work

This study has implications for our clinical understanding of ASD across the lifespan. We have identified sets of precise, objective linguistic features that are highly predictive of ASD at three different developmental stages. These features pro-

vide specific, developmentally-informed intervention targets that could be used to improve language and conversation skills in individuals with ASD. We anticipate that additional features identified through machine learning in other domains could similarly inform future efforts to develop targeted clinical interventions.

For future work, we would like to use these techniques in a longitudinal study for measuring treatment progress. This can be done by tracking feature values of an individual as they change through time. Additionally, we would like to use these techniques to see if they can be used to differentiate between other mental health disorders, such as anxiety, depression and obsessive compulsive disorder.

Acknowledgements

This work was supported by NIMH grant R34MH104407, Services to Enhance Social Functioning in Adults with Autism Spectrum Disorder (E.S. Brodkin, PI); by the National Center for Research Resources, Grant UL1RR024134, now the National Center for Advancing Translational Sciences, Grant UL1TR000003 (MPIs: E.S. Brodkin and R.T. Schultz); by the Intellectual and Developmental and Disabilities Research Center at the Childrens Hospital of Philadelphia and the University of Pennsylvania, NICHD U54HD86984 (MPIs: M. Robinson and R.T. Schultz); and by the Institute for Translational Medicine and Therapeutics (ITMAT) Transdisciplinary Program in Translational Medicine and Therapeutics (MPIs: E.S. Brodkin and R.T. Schultz), and by generous gifts from the Eagles Charitable Foundation and the Allerton Foundation to R.T. Schultz.

References

[1] Fifth Edition, American Psychiatric Association, et al. Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing*, 2013.

[2] Helen Tager-Flusberg and Connie Kasari. Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism Research*, 6(6):468–478, 2013.

[3] C Lord, M Rutter, P DiLavore, S Risi, K Gotham, and S Bishop. Autism diagnostic observation schedule–2nd edition (ados-2). *Los Angeles, CA: Western Psychological Corporation*, 2012.

[4] Christopher Gillberg and Suzanne Steffenburg. Outcome and prognostic factors in infantile autism and similar conditions: A population-based study of 46 cases followed through puberty. *Journal of autism and developmental disorders*, 17(2):273–287, 1987.

[5] Patricia Howlin, Susan Goode, Jane Hutton, and Michael Rutter. Adult outcome for children with autism. *Journal of child psychology and psychiatry*, 45(2):212–229, 2004.

[6] André Venter, Catherine Lord, and Eric Schopler. A follow-up study of high-functioning autistic children. *Journal of child psychology and psychiatry*, 33(3):489–597, 1992.

[7] Meng-Chuan Lai, Michael V Lombardo, Bhismadev Chakrabarti, and Simon Baron-Cohen. Subgrouping the autism spectrum": Reflections on dsm-5. *PLoS biology*, 11(4):e1001544, 2013.

[8] Deborah K Anderson, Catherine Lord, Susan Risi, Pamela S DiLavore, Cory Shulman, Audrey Thurm, Kathleen Welch, and Andrew Pickles. Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of consulting and clinical psychology*, 75(4):594, 2007.

[9] Masoud Rouhizadeh, Emily PrudHommeaux, Jan Van Santen, and Richard Sproat. Measuring idiosyncratic interests in children with autism. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2015, page 212. NIH Public Access, 2015.

[10] Julia Parish-Morris, Christopher Cieri, Mark Liberman, Leila Bateman, Emily Ferguson, and Robert T Schultz. Building language resources for exploring autism spectrum disorders. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2016, page 2100. NIH Public Access, 2016.

[11] Julia Parish-Morris, Mark Liberman, Neville Ryant, Christopher Cieri, Leila Bateman, Emily Ferguson, and Robert Schultz. Exploring autism spectrum disorders using hlt. In *Proceedings of the third workshop on computational lingusitics and clinical psychology*, pages 74–84, 2016.

[12] Emily Prudhommeaux and Masoud Rouhizadeh. Automatic detection of pragmatic deficits in children with autism. In *The... Workshop on Child, Computer and Interaction*, volume 2012, page 1. NIH Public Access, 2012.

[13] Cindy Gallois and Howard Giles. Communication accommodation theory. *The international encyclopedia of language and social interaction*, pages 1–18, 2015.

[14] Stanford W Gregory, Kelly Dagan, and Stephen Webster. Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43, 1997.

[15] Daniel Bone, Chi-Chun Lee, Matthew P Black, Marian E Williams, Sungbok Lee, Pat Levitt, and Shrikanth Narayanan. 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):1162–1177, 2014.

[16] Henry M Wellman. *The child's theory of mind.* The MIT Press, 1992.

[17] Iroise Dumontheil, Ian A Apperly, and Sarah-Jayne Blakemore. Online usage of theory of mind continues to develop in late adolescence. *Developmental science*, 13(2):331–338, 2010.

[18] Deborah Yurgelun-Todd. Emotional and cognitive changes during adolescence. *Current opinion in neurobiology*, 17(2):251–257, 2007.

[19] B Bradford Brown. Adolescents' relationships with peers. *Handbook of adolescent psychology*, pages 363–394, 2004.

[20] Catherine Lord, Eva Petkova, Vanessa Hus, Weijin Gan, Feihan Lu, Donna M Martin, Opal Ousley, Lisa Guy, Raphael Bernier, Jennifer Gerdts, et al. A multisite study of the clinical diagnosis of different autism spectrum disorders. *Archives of general psychiatry*, 69(3):306–313, 2012.

[21] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[22] David Wechsler. *WASI-II: Wechsler abbreviated scale of intelligence*. PsychCorp, 2011.

[23] Vanessa Hus, Katherine Gotham, and Catherine Lord. Standardizing ados domain scores: Separating severity of social affect and restricted and repetitive behaviors. *Journal of autism and developmental disorders*, 44(10):2400–2412, 2014.

[24] Michael Rutter, Anthony Bailey, and Cathrine Lord. *The social communication questionnaire: Manual*. Western Psychological Services, 2003.

[25] Allison B Ratto, Lauren Turner-Brown, Betty M Rupp, Gary B Mesibov, and David L Penn. Development of the contextual assessment of social skills (cass): A role play measure of social skill for individuals with high-functioning autism. *Journal of Autism and Developmental Disorders*, 41(9):1277–1286, 2011.

[26] Meghan Lammie Glenn, Stephanie M Strassel, and Haejoong Lee. Xtrans: A speech annotation and transcription tool. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[27] Owen Kimball, Chai-Lin Kao, Teodoro Arvizo, John Makhoul, and Rukmini Iyer. Quick transcription and automatic segmentation of the fisher conversational telephone speech corpus. In *RT04 Workshop*, 2004.

[28] Julia Parish-Morris, Mark Y Liberman, Christopher Cieri, John D Herrington, Benjamin E Yerys, Leila Bateman, Joseph Donaher, Emily Ferguson, Juhi Pandey, and Robert T Schultz. Linguistic camouflage in girls with autism spectrum disorder. *Molecular autism*, 8(1):48, 2017.

[29] RDC Team et al. R: A language and environment for statistical computing. *R foundation for statistical computing, Vienna, Austria*, 2008.

[30] Tyler W Rinker. qdap: Quantitative discourse analysis package. *University at Buffalo/SUNY*, 2013.

[31] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[32] Stanley D Rosenberg and Gary J Tucker. Verbal behavior and schizophrenia: The semantic dimension. *Archives of General Psychiatry*, 36(12):1331–1337, 1979.

[33] James W Pennebaker and Laura A King. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[34] Manoj Kumar, Rahul Gupta, Daniel Bone, Nikolaos Malandrakis, Somer Bishop, and Shrikanth S Narayanan. Objective language feature analysis in children with neurodevelopmental disorders during autism assessment. In *INTERSPEECH*, pages 2721–2725, 2016.

[35] Kate Muir, Adam Joinson, Rachel Cotterill, and Nigel Dewdney. Characterizing the linguistic chameleon: Personal and social correlates of linguistic style accommodation. *Human Communication Research*, 42(3):462–484, 2016.

[36] Zoë Louise Hopkins. *Language alignment in children with an autism spectrum disorder*. PhD thesis, University of Sussex, 2016.

[37] Katie E Slocombe, Ivan Alvarez, Holly P Branigan, Tjeerd Jellema, Hollie G Burnett, Anja Fischer, Yan Hei Li, Simon Garrod, and Liat Levita. Linguistic alignment in adults with and without aspergers syndrome. *Journal of Autism and Developmental Disorders*, 43(6):1423–1436, 2013.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[39] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.