Computationally Characterizing Communicative Content and Context in Autistic Children

Grace Olive Lawley
B. A. Mathematics, Lewis & Clark College, 2017

Presented to the
Computer Science & Engineering Graduate Program
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science & Engineering

March 2024

Computer Science & Electrical Engineering Graduate Program
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Ph. D. dissertation of

Grace Olive Lawley

has been approved.

Steven Bedrick, Ph. D.
Associate Professor, OHSU
Thesis Advisor

Peter A. Heeman, Ph. D.
Associate Professor, OHSU

Jill K. Dolata, Ph. D., CCC-SLP
Adjunct Associate Professor, OHSU

Eric Fombonne, M. D.
Professor Emeritus, OHSU

Meysam Asgari, Ph. D.
Associate Professor, OHSU

This dissertation is dedicated to my Dad,
who always encouraged me to pursue knowledge.

Acknowledgments

I've had the unique experience of having two advisors during my PhD, Steven Bedrick and Jan van Santen. To Jan van Santen, thank you for believing in me and for showing me that pursuing a PhD was not only a possibility for me, but for helping it become a reality. To Steven Bedrick, thank you for picking up where Jan left off and for always challenging me to challenge myself; your mentorship has been invaluable.

I would like to thank my committee members, Peter Heeman, Jill Dolata, Eric Fombonne, and Meysam Asgari, for their guidance and enthusiasm throughout this entire process. I am also thankful for Alison Hill, Jackie Wirz, and Katy McKinney-Bock, whose advice was instrumental in building my confidence as a woman in STEM, especially during the first few years of my PhD.

I would like to express my immense gratitude for my past teachers, who helped me become the individual that I am today: my college professors, especially Paul Allen, Yung-Pin Chen, and Iva Stavrov; my high school teachers, especially Nancy Feidelman, Thaddeus Lisowski, and Naoko Akiyama; my middle school math teacher, Nick Grener.

Gaines Hall would not have been the same if not for my colleagues: Alex Salem, Heather MacFarlane, Joel Adams, and Rosemary Ingham. Thank you for the laughs, support, and advice.

Last but not at all least, I would like to thank my fiancé, Derek Sweet, whose endless patience and compassion has been a constant among many variables. Thank you for always cheering me on and celebrating my accomplishments.

Contents

List of Tables

List of Figures

Abstract

Computationally Characterizing Communicative Content and Context in Autistic Children

Grace Olive Lawley

Doctor of Philosophy
Oregon Health & Science University
School of Medicine

March 2024
Thesis Advisor: Steven Bedrick, Ph. D.

Pragmatic language difficulties are common among autistic children, and assessment of pragmatic language skills over time is an important predictor of quality of life outcomes during adulthood. Current metrics for pragmatic language are qualitative in design and are expensive in terms of time and resources. With the use of Natural Language Processing (NLP) methods, robust measures of pragmatic language features can be obtained in an automated, reliable, and relatively inexpensive fashion. Such metrics can be used to augment traditional pragmatic language assessments. Improving our understanding of how autistic individuals use language not only helps us learn how to become better conversational partners ourselves, but also enables us to build language tools that accommodate for pragmatic language differences.

In this dissertation, we leverage traditional statistical methods to adapt and augment established NLP techniques to investigate three areas of pragmatic language that autistic children are known to have difficulty with. We use a corpus of transcribed Autism Diagnostic Observation Schedule (ADOS) sessions for 117 autistic children (98 male) and 65 Typically Developing (TD) children (37 male), aged 4 to 15 years old. We first compare how autistic children use the fillers *um* and *uh* differently than their TD peers during conversations. After controlling for age, sex, and IQ, we found that autistic children used less *um* frequently than their TD peers and that structural language scores predicted *um* usage while social affect and pragmatic language scores did not. Next, we investigate differences in topic maintenance ability. We present a novel statistical approach for investigating group difference in the document-topic distribution vectors created by Latent Dirichlet Allocation (LDA). After transforming the vectors using Aitchison geometry, we use multivariate analysis of variance (MANOVA) to compare sample means and calculate effect size using partial eta-squared. We validate our method on a subset of the *20Newsgroup* corpus and then apply our method to our clinical corpus. We found that the topic distribution vectors of autistic children significantly differed from those of TD children when responding to questions about social difficulties. Lastly, we investigate differences in backchannel usage (i.e., *right*, *okay*, *uhhuh*) between autistic and TD children. After adjusting for age, sex, and IQ, we found that autistic children used less backchannels than their TD peers and were less likely to produce a backchannel with a greater overlap length.

Chapter 1

Introduction

1.1 Motivation

Children with Autism Spectrum Disorder (ASD) frequently have difficulties with social communication and pragmatic language, such as differences in filler usage (Gorman et al., 2016; Irvine et al., 2016), topic maintenance (Baltaxe and D'Angiola, 1992; Paul et al., 2009), turn-taking (Capps et al., 1998; Paul et al., 2009), and non-verbal cues (de Marchena and Eigsti, 2010; Paul et al., 2009). Since conversational ability of autistic children is an important predictor of quality of life outcomes during adulthood, such as friendship and vocational independence (DaWalt et al., 2019; Friedman et al., 2019), targeted evaluation of pragmatic language skills and remedial intervention is essential. In addition, by improving our understanding of how autistic individuals use pragmatic language, we can become more informed and accommodating conversational partners.

Existing tools for assessing language ability in autistic children, such as the Clinical Evaluation of Language Fundamentals (CELF; Semel et al., 2004), are costly, labor intensive, and provide limited evaluation of pragmatic language skills. These structured language evaluations require a trained professional, such as a speech pathologist, to administer them. In addition, some areas of pragmatic language usage, such as pause length, are difficult for humans to reasonably measure accurately and consistently over the course of a language sample. Existing structured language assessments often fail to capture pragmatic language features that are only observable in naturalistic conversational contexts (Adams, 2002), and may therefore fail to capture the pragmatic differences associated with reduced conversational reciprocity in ASD. Alternative caregiver-reported language assessments, such as the Children's Communication Checklist, version 2 (CCC-2; Bishop, 2003), are more successful at measuring such aspects of pragmatic language ability (Dolata et al., 2022; Volden and Phillips, 2010). These assessments rely on the knowledge of the child's caregiver who can offer a better representation of the child's communication ability over time and across contexts. Although such language assessments are constrained by the caregiver's accuracy as a language observer, they are more reflective of the child's day-to-day language usage and ability than alternative clinician-derived instruments.

Another promising approach to measuring pragmatic language ability in autistic children involves leveraging Natural Language Processing (NLP) methods to create metrics that can be used to augment traditional pragmatic language assessments. These methods rely on computational analyses of transcribed language samples and allow for refined and automated measurement of aspects of language (e.g., word usage, grammatical errors, sentence complexity, sentence structure, etc.) that cannot be reasonably and accurately measured manually unless a substantial amount of time and resources are invested into the task. By using NLP methods to automate language analysis tasks, researchers are able to analyze larger quantities of text in a shorter amount of time. Additionally, while manual annotation methods often require additional processing steps after annotation in order to properly compensate for any differences between transcribers or examiners, computational methods are able to produce highly reliable and consistent results as is. As such, they can be repeated often, providing a truly blind assessment in the context of treatment studies. Moreover, NLP methods also have the ability to capture the subtle signals in language that are difficult for a human to detect and measure, such as the precise amount of time two utterances overlap. Previous studies have successfully used NLP methods to capture various aspects of pragmatic language differences in ASD, including atypical language (Prud'hommeaux et al., 2011; Salem et al., 2021), topic maintenance (Adams et al., 2021; Goodkind et al., 2018; Prud'hommeaux et al., 2017), repetitive speech (Rouhizadeh et al., 2015; van Santen et al., 2013), and disfluency usage (Gorman et al., 2016; Irvine et al., 2016; MacFarlane et al., 2017; Parish-Morris et al., 2017).

Although there are many positives to using NLP methods to assist in pragmatic language assessment metrics, there are a few potential obstacles that should be highlighted. While there has been a lot of recent progress in the field of NLP overall, evaluation methods have failed to keep pace and have continuously lagged behind. Existing evaluation metrics are typically narrow in scope and are designed to assess concepts such as model performance or classification accuracy, which are often not of clinical interest. Even when standard and commonly used evaluation metrics are reported, these metrics can still fall short of being informative and useful for future reproduction and comparison studies (Reimers and Gurevych, 2017). In addition, existing evaluation methods frequently lack hypothesis-driven inferential statistical approaches (Dror et al., 2018). As a majority of NLP methods were not designed with clinical applications in mind, such methods typically require additional development and augmentation to meet the statistical standards expected in clinical research.

1.2 Problem statement

Pragmatic language difficulties are common among autistic children. Current assessments used to measure pragmatic language ability are expensive both in terms of time and resources, and often fail to capture pragmatic language features commonly observed in naturalistic conversational contexts. NLP methods can be used to develop automated language metrics that are robust, reliable, and relatively inexpensive. Such metrics can be used to augment traditional language assessments and to assist clinicians, caregivers, and teachers. However, in order to apply traditional NLP methods to ASD research, adaptation and inclusion of statistical evaluation is required as there is a lack of statistical evaluation methods for existing NLP methods. In this dissertation, we leverage traditional statistical methods to adapt and augment established NLP techniques to investigate and computationally characterize three areas of pragmatic language that autistic children are known to have difficulty with: filler usage (*um* and *uh*), topic maintenance, and backchanneling.

1.3 Research objectives

The work presented in this dissertation spans three areas of pragmatic language that autistic children have difficulties with. Using NLP techniques paired with traditional statistical methods, we will first quantify these aspects of language and then report group differences, if any. The three areas of pragmatic language explored and the analyses performed are briefly summarized below.

Filler usage. We will investigate differences in the usage of the fillers *um* and *uh* between the ASD and TD groups. For each child, we will compute three measures of filler usage: *um-rate*, *uh-rate*, and *um-ratio*. First, we will examine general usage differences of *um* and *uh* between the ASD and TD groups using nonparametric Wilcoxon-Mann Whitney tests and rank-biserial correlations to estimate effect size. Next, we will examine usage differences after controlling for participant age, sex, and IQ by using mixed effect logistic regression models with a per-participant random intercept. Lastly, within the ASD group, we will use mixed effect logistic regression models to investigate whether filler usage is associated with the difficulties with social communication, pragmatic language, and structural language characteristic of ASD.

Topic maintenance. We will investigate differences between the distribution of topics discussed by the ASD and TD groups. Using the topic modeling technique known as Latent Dirichlet Allocation (LDA), we present a novel statistical method for quantifying group difference in topic distributions that involves Aitchison geometry and multivariate analysis of variance

(MANOVA). We report the results of validating our statistical method on a subset of the *20Newsgroup* corpus, a well-known corpus that is commonly used for topic modeling. Using our clinical sample, we first use LDA, a well-established topic modeling method, to represent each child as a topic distribution vector. Next, we apply our novel statistical method to examine group difference and report the results.

Backchanneling patterns. We will investigate backchanneling differences between the ASD and TD groups. We present our method for determining whether an utterance is a backchannel or overlapping-backchannel, and compute the usage rates of both for each child. We first compare the overall usage rates of backchannels and overlapping-backchannels using nonparametric Wilcoxon-Mann Whitney tests and rank-biserial correlations to estimate effect size. Next, we used mixed effects logistic regression models, investigate whether group differences in backchannel rates are robust to participant-level age, sex, and IQ as well as utterance-level overlap length.

1.4 Organization of the dissertation

The following two chapters contain the relevant background and sample information for the chapters to come. In Chapter 2, we discuss ASD, the associated pragmatic language differences, and two standardized assessments that are frequently used during diagnostic assessment for the possibility of ASD. In Chapter 3, we describe the clinical sample analyzed throughout this dissertation. We describe the participants and the language samples analyzed. Chapters 4, 5, and 6 describe the analyses performed on filler usage, topic maintenance, and backchanneling, respectively. In Chapter 4, we investigate filler usage differences between autistic children and their TD peers as well as the clinical measures associated with this difference. In Chapter 5, we present a novel statistical approach for quantifying group difference in the topic distributions produced by topic models and then use this method to investigate topic maintenance differences between autistic children and their TD peers. In Chapter 6, we examine whether backchannels are used at different rates by autistic children than their TD peers and whether this difference is related to overlap and overlap length.

1.4.1 Terminology

We recognize that the meaning and connotations of words change over time, and the terminology used in this dissertation may be outdated in the near future. We use identity-first language (i.e., autistic children) in this dissertation instead of person-first language (i.e., children with autism), since the former is the current preference among autistic individuals (Brown, nd). We use sex to refer to sex assigned at birth.

1.4.2 Funding

This work was supported in part by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under awards R01DC012033 and R01DC015999.

ADHD	Attention Deficit Hyperactivity Disorder
ADOS	Autism Diagnostic Observation Schedule
ADOS-2	Autism Diagnostic Observation Schedule, 2nd ed.
ANOVA	analysis of variance
ASD	Autism Spectrum Disorder
BEC	best estimate clinical
CCC-2	Children's Communication Checklist, version 2
CELF	Clinical Evaluation of Language Fundamentals
C-unit	Communication-unit
df	degrees of freedom
DSM-IV-TR	Diagnostic and Statistical Manual of Mental Disorders, 4th ed., text revision
DSM-V	Diagnostic and Statistical Manual of Mental Disorders, 5th ed.
GCC	Global Communication Composite (CCC-2)
ILR	isometric logratio transformation
IQ	intelligence quotient
LDA	Latent Dirichlet Allocation
IQR	interquartile range
MANOVA	multivariate analysis of variance
MANCOVA	multivariate analysis of covariance
NLP	natural language processing
MLUM	mean length of utterance in morphemes
RRB	Restricted and Repetitive Behavior (ADOS)
SA	Social Affect (ADOS)
SALT	Systematic Analysis of Language Transcripts
SCQ	Social Communication Questionnaire
sd	standard deviation
SLI	Specific Language Impairment
TD	typically developing; typical development
VEM	variational expectation-maximization
WISC-IV	Wechsler Intelligence Scale for Children, 4th ed.
WPPSI-III	Wechsler Preschool and Primary Scale of Intelligence, 3rd ed.
OR	odds ratio

Table 1.1: Acronyms and initialisms used in this dissertation.

Chapter 2

Background

The technical background for the analyses presented in this dissertation can be found in Chapter 4, 5, and 6, respectively. Here, we will provide additional background information on Autism Spectrum Disorder (ASD) and standardized assessments commonly used during diagnostic evaluation.

2.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a developmental disorder that is characterized by difficulties with social communication as well as by restricted and repetitive patterns of behavior (American Psychiatric Association, 2013). Recent surveys conducted by the Centers for Disease Control and Prevention (CDC) reported an ASD prevalence rate of 1 in 36 among 8 year old children (approximately 4% of boys and 1% of girls) (Maenner et al., 2023). While difficulty with social communication is a core symptom of ASD, the extent and severity of these impairments varies throughout the autism spectrum (Tager-Flusberg and Kasari, 2013). Some autistic children have language skills within the normal range (Landa, 2000), whereas others remain minimally verbal throughout life (Tager-Flusberg and Kasari, 2013). For those who fall between these two extremes, difficulties with structural language (i.e., vocabulary, phonology, syntax) are common (Tager-Flusberg and Caronna, 2007; Wittke et al., 2017). In contrast, pragmatic language (i.e., social aspects of language) is universally impaired (Landa, 2000; Tager-Flusberg et al., 2005; Volden et al., 2008; Young et al., 2005).

2.2 Standardized assessments

In the following section, we will describe the administration and scoring of two standardized assessments: the Children's Communication Checklist, version 2 (CCC-2; Bishop, 2003); the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2; Lord et al., 2000). In addition to being administered to all participants in the sample analyzed in this dissertation, these assessments are also commonly used during diagnostic evaluation for the possibility of ASD by clinicians across the United States, increasing the likelihood that others will be able to replicate our work in the future.

2.2.1 Language ability

The CCC-2 (Bishop, 2003) is a caregiver-completed checklist that is used to measure a child's communication skills. It is designed to assess the everyday communication skills of children 4:0 to 16:11 years old.[1] The caregiver is given a checklist of 70 statements describing how children communicate. For each statement, they are asked to recall how often they have observed the child do the described behavior and respond using the following scale:

- 0 = less than once a week (or never)
- 1 = at least once a week, but not every day
- 2 = once or twice a day
- 3 = several times (more than twice) a day (or always)

The 70 items are divided into 10 subscales: (A) Speech; (B) Syntax; (C) Semantics; (D) Coherence; (E) Inappropriate initiation; (F) Stereotyped language; (G) Use of context; (H) Nonverbal communication; (I) Social relations; (J) Interests.

Each raw subscale score is first converted to an age-standardize score and then three primary measures are calculated: Structural Language scale score, Pragmatic Language scale score, and Global Communication Composite (GCC). The Structural Language score is derived from subscales A to D, and evaluates articulation and phonology, language structure, vocabulary, and discourse. The Pragmatic Language score is derived from subscales E to H, which cover pragmatic areas of communication that are difficult to capture using standard language assessments. The GCC score is the sum of the scaled scores of A through H. For all three measures, lower scores indicate more communication problems. Although the CCC-2 is not intended to be used as diagnostic tool, children who have low GCC scores alongside low Pragmatic Language scores are recommended for further ASD screening (Bishop, 2013).

[1]Here, 16:11 corresponds to 16 years and 11 months old.

2.2.2 ASD symptom severity

The ADOS-2 (Lord et al., 2000) is a semi-structured standardized assessment between an examiner and child that is designed to provide opportunities for the examiner to observe speech and behavior that are characteristic of ASD as defined by the DSM-IV-TR (American Psychiatric Association, 2000). There are five modules of the ADOS-2, depending on the age and language and development level of the individual being assessed:

- Toddler module, designed for children who do not consistently use phrase speech and are between 12 to 31 months old.
- Module 1, designed for children who do not consistently use phrase speech and are 31 months or older.
- Module 2, designed for children who use phrase speech but do not yet have fluent speech.
- Module 3, designed for children and adolescents who have fluent speech.
- Module 4, designed for adolescents and adults who have fluent speech.

Since the analyses presented in this dissertation are performed on transcribed ADOS-2 Module 3 sessions, we will only discuss this module in detail. Module 3 of the ADOS-2 is comprised of 14 activities (see Table 2.1). These activities are typically administered in the order provided, however, the examiner is able to change the order if clinically indicated. Overall, administration usually takes 30 to 60 minutes.

Immediately following the administration of the test, the examiner completes 29 coding items. From these, three measures are calculated: Social Affect (SA) score (10 items; range 0-20); Restricted and Repetitive Behavior (RRB) score (4 items; range 0-8); Overall Total score (14 items; range 0-28). For all three scores, a higher value indicates more severe ASD symptoms. The Overall Total score is the combination of the SA and RRB score. The coding items used to calculate the SA and RRB scores are listed in Table 2.2.

1. *Construction Task*
2. *Make-Believe Play*
3. *Joint Interactive Play*
4. *Demonstration Task*
5. *Description of a Picture*
6. *Telling a Story From a Book*
7. *Cartoons*
8. *Conversation and Reporting*
9. *Emotions*
10. *Social Difficulties and Annoyance*
11. *Break*
12. *Friends, Relationships, and Marriage*
13. *Loneliness*
14. *Creating a Story*

Table 2.1: Activities in the ADOS-2, Module 3.

Social Affect (SA)
(A-7) Reporting of events
(A-8) Conversation
(A-9) Descriptive, Conventional, Instrumental, or Informational Gestures
(B-1) Unusual Eye Contact
(B-2) Facial Expressions Directed to Examiner
(B-4) Shared Enjoyment in Interaction
(B-7) Quality of Social Overtures
(B-9) Quality of Social Response
(B-10) Amount of Reciprocal Social Communication
(B-11) Overall Quality of Rapport
Restricted and Repetitive Behavior (RRB)
(A-4) Stereotyped/Idiosyncratic Use of Words or Phrases
(D-1) Unusual Sensory Interest in Play Material/Person
(D-2) Hand and Finger and Other Complex Mannerisms
(D-4) Excessive Interest in or References to Unusual or Highly Specific Topics or Objects or Repetitive Behaviors

Table 2.2: Coding items used to calculate Social Affect (SA) and Restricted and Repetitive Behavior (RRB) scores for the ADOS-2, Module 3.

Chapter 3

Data

3.1 Participants

The clinical sample analyzed in Chapters 4-6 consists of participants from two separate, larger studies, both conducted at Oregon Health & Science University in Portland, Oregon, USA. Various subsets of both studies have been analyzed in prior publications from our group (Adams et al., 2021; Gorman et al., 2016; Lunsford et al., 2010, 2012; MacFarlane et al., 2017, 2023; Salem et al., 2021; van Santen et al., 2013). We will first describe these two separate studies and then describe the combined sample analyzed in this dissertation.

3.1.1 Study 1

Participants aged 4 to 8 years old with ASD, TD, or Specific Language Impairment (SLI) were recruited from the Portland, Oregon metropolitan area for an expressive and receptive language study at Oregon Health & Science University. Recruitment was done through a variety of community and health care resources, including local healthcare specialists, autism clinics, and educational service districts. Data collection took place between 2005 and 2009. Participants were excluded from the study if they had any of the following: known metabolic, neurological, or genetic disorder; gross sensory or motor impairment; brain lesion; orofacial abnormalities (e.g., cleft palate); intellectual disability. Absence of speech intelligibility impairments was confirmed by a certified speech language pathologist. All participants were native English speakers, had an IQ ≥ 70, and a mean length of utterance in morphemes (MLUM) ≥ 3.0.

The study was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award R01DC007219. The study was approved by the Oregon Health & Science University Institutional Review Board and all research was performed in accordance with their relevant guidelines and regulations. Participating families were fully informed about the study procedures and provided written consent.

All participants completed a series of tests, including experimental tasks and cognitive, language, and neuropsychological assessments. Intellectual level for participants younger than 7 years old was estimated using the Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III; Wechsler, 2002), while intellectual level for participants ages 7 and older was estimated using a short form of the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler, 2003). Following Sattler and Dumont (2004), full scale IQ was estimated from the sum of the scaled scores of the three subsets administered: Information, Block Design, and Vocabulary. Expressive and receptive language ability was estimated using the Clinical Evaluation of Language Fundamentals (CELF). Participants younger than 6 years old were administered the CELF-Preschool-2 (Semel et al., 2004), while participants 6 years and older were administered the CELF-4 (Semel et al., 2003). Pragmatic and structural language skill was assessed using the Children's Communication Checklist, version 2 (CCC-2; Bishop, 2003). All participants, regardless of diagnostic group, were administered the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2; Lord et al., 2000). ADOS sessions were recorded and transcribed at a latter date by a team of trained transcribers who were blind to participant diagnosis and intellectual ability. Transcription guidelines were based on conventions used by Systematic Analysis of Language Transcripts (SALT) guidelines (Miller and Iglesias, 2012). Both examiner and child speech was transcribed.

Participants were included in the ASD group if the following criteria were met: (1) best estimate clinical (BEC) consensus judgment by an experienced panel – two clinical psychologists, one speech-language pathologist, and one occupational therapist (all of whom had specific expertise with ASD) – using DSM-IV-TR criteria (American Psychiatric Association, 2000); (2) overall total scores above the ASD classification threshold on both the ADOS-2 and Social Communication Questionnaire (SCQ; Rutter et al., 2003). Participants were included in the SLI group according to the following criteria: (1) documented history of either language delay, deficits, or both; (2) BEC consensus judgment of language impairment in the absence of ASD based on all available evidence (i.e., medical and family history, prior assessments, educational records, study assessments); (3) core language score on the CELF below 85 (one standard deviation below the mean). Additional study details can be found in Hill et al. (2015) and van Santen et al. (2013).

In total, the original sample consists of 110 children (50 ASD, 43 TD, 17 SLI). From this, a subset of 74 participants (34 ASD, 40 TD), aged 4 to 8, were included the sample analyzed in this dissertation.

3.1.2 Study 2

Participants aged 7 to 17 with ASD, TD, or Attention Deficit Hyperactivity Disorder (ADHD) were recruited for an fMRI study at Oregon Health & Science University. Recruitment was done by community outreach and referrals from Oregon Health & Science University's specialty clinics, with data collection taking place between 2012 to 2018. Participants were excluded from the study if they had any of the following: seizure disorder; cerebral palsy; pediatric stroke; history of chemotherapy; sensorimotor handicaps; closed head injury; thyroid disorder; schizophrenia; bipolar disorder; current major depressive episode; fetal alcohol syndrome; Tourette's disorder; severe vision impairments; Rett's syndrome; current use of psychoactive medications. All participants were native English speakers and had an IQ ≥ 70.

The study was supported by the National Institute of Mental Health of the National Institutes of Health under awards R01MH115357 and R01MH086654. The study was approved by the Oregon Health & Science University Institutional Review Board and all research was performed in accordance with their relevant guidelines and regulations. During an initial screening visit, informed written consent and assent was obtained from all participants and their parents or guardians.

Participants were directly assessed by experienced child psychiatrists and clinical psychologists and diagnoses of ASD and ADHD were confirmed using BEC consensus judgment using DSM-IV-TR and DSM-5 criteria (American Psychiatric Association, 2000, 2013). Intellectual level of participants was estimated with the WPPSI-III or WISC-IV. (Wechsler, 2003, 2002). Full scale IQ was estimated from the sum of the scaled scores of the three subsets administered: Information, Block Design, and Vocabulary (Sattler and Dumont, 2004). Pragmatic and structural language ability for all participants was measured using the CCC-2 (Bishop, 2003). All participants, regardless of diagnostic status, were administered the ADOS-2 (Lord et al., 2000). Assessments were scored using the revised algorithms (Gotham et al., 2009), and sessions were recorded to be transcribed at a later date. Transcription was done by a team of trained transcribers who were blind to participants' diagnosis status and intellectual ability. For additional study details, see Salem et al. (2021).

A total of 83 participants (102 ASD, 45 ADHD, 28 TD) were included in the main neuroimaging study. From the original study sample, a subset of 108 participants (83 ASD, 25 TD), aged 7 to 15, were included in the sample analyzed in this dissertation.

3.1.3 Current study

The analyses in this dissertation were performed on a combined sample of participants from the two studies described above. This combined sample consisted of a total of 182 participants (117 ASD, 65 TD) aged 4 to 15 years old. All participants were native English speakers, had an IQ $\geq$ 70, had fluent speech, and had a MLUM $\geq$ 3.0. Demographic and clinical sample characteristics are summarized in Table 3.1. The standardized measures included in this table from the ADOS are

	ASD *(n = 117, 98 males)*				TD *(n = 65, 37 males)*				
	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	p
Age (in years)	4.54	15.6	10.03	2.82	4.21	14.5	8.22	2.83	<0.001
IQ	72	138	102.19	15.77	90	147	116.94	12.37	<0.001
ADOS SA	3	19	9.18	3.48	0	8	0.95	1.47	<0.001
ADOS RRB	0	8	3.59	1.53	0	2	0.45	0.64	<0.001
ADOS Total	7	24	12.77	3.73	0	10	1.40	1.79	<0.001
CCC-2 Pragmatic	1.5	10.8	4.96	1.69	7.5	15.8	12.05	1.73	<0.001
CCC-2 Structural	1	12	7.01	2.29	8.5	15	11.73	1.57	<0.001
CCC-2 GCC	45	103	75.13	11.0	87	143	115.18	12.09	<0.001

Table 3.1: Demographic and clinical sample characteristics.

the Social Affect (SA) score, Restricted Repetitive Behavior (RRB) score, and the Overall Total score. From the CCC-2, the standardized measures included are the Pragmatic Language score, Structural Language score, and the Global Communication Composite (GCC) score. Further details regarding the administration and scoring of the ADOS and the CCC-2 can be found in Section 2.2. There were 98 male participants in the ASD group (83.8%) and 37 male participants in the TD group (56.9%). As expected, autistic participants had significantly lower IQ scores than their TD peers ($p < 0.001$), although their mean score was close to the population mean. The ASD group also had significantly higher scores for the three measures of autism symptom severity (ADOS SA; ADOS RRB; ADOS Total) compared to the TD group ($p < 0.001$). Lastly, as anticipated, all three scores for language ability (CCC-2 Pragmatic Language; CCC-2 Structural Language; CCC-2 GCC) were significantly lower for the ASD group than the TD group ($p < 0.001$).

3.2 Language samples

The language samples analyzed in this dissertation consisted of transcribed ADOS-2 Module 3 sessions (Lord et al., 2000). This module is designed for children and adolescents with fluent speech. All ADOS interviews were administered by research assistants or by a senior clinical psychologist, all of whom were trained to research reliability level. Sessions were videotaped and then transcribed at a later date, and all sessions were scored using the revised algorithms (Gotham et al., 2009). Previous work with ADOS language samples (Salem et al., 2021; MacFarlane et al., 2023) has shown that computational methods are able to capture a variety of differences in the language used by autistic children from such dialogue samples.

Of the fourteen activities in Module 3 of the ADOS, four were chosen for our analyses:

1. *Emotions*
2. *Social Difficulties and Annoyance*
3. *Friends, Relationships, and Marriage*
4. *Loneliness*

We chose these activities in particular because of their conversational structure and naturalistic dialogue. Additionally, these are also the only activities in the ADOS-2 Module 3 where the examiner is provided a list of open-ended, interview questions to ask. In the following sections, we briefly describe the goals for these four activities and provide any other relevant information.

Emotions

In the *Emotions* activity, the examiner is asked to observe how the participant describes their own and others' emotions as well as assess if the participant displays insight into the social situations and relationships from which these emotion may arise. Table 3.6 lists the interview questions given to the examiners. The examiners do not need to ask all of the questions provided or in the order provided. Once the examiner has adequate responses for two emotions they can conclude the activity.

Social Difficulties and Annoyance

In this activity, examiners assess the participant's insight into personal social difficulties as well as their sense of responsibility for their own actions. The interview questions are listed in Table 3.7. Examiners are instructed to ask the questions in the order provided. They are allowed to ask older adolescents about work instead of or in addition to school if appropriate.

Friends, Relationships, and Marriage

For this activity, the examiner evaluate the participant's understanding of the concept of friendship, marriage, and other social relationships. The interview questions for this activity are shown in Table 3.8. Examiners are instructed to ask all of the interview questions regardless of the participant's age. However, these questions do not need to be asked verbatim and examiners are allowed to make modifications based on the participant's developmental level. For example, the question "Why do you think some people get married?" could be modified to "Do you know anyone who is married?". The instructions also state that the question "Do you have a girlfriend or boyfriend?" should be asked as worded to participants of either sex.

Loneliness

In this activity, the examiner is evaluating the participant's understanding of the concept of loneliness and how it pertains to themselves or to others. Table 3.9 lists the interview questions for this activity. The examiners are instructed to ask the questions in the order given.

3.2.1 Transcription

Transcription of the ADOS sessions was completed by a team of trained transcribers. All transcribers were blind to the participants' diagnostic status and intellectual abilities. Sessions were transcribed according to modified Systematic Analysis of Language Transcripts (SALT) guidelines (Miller and Iglesias, 2012). Both child and examiner utterances were transcribed, with child utterances prefixed with 'C:' and examiner utterances prefixed with 'E:'. After transcription, sessions were manually partitioned into activities according to established guidelines in order to ensure consistency across transcripts. Details of the particular text preprocessing steps performed for each set of analyses are described in Chapters 4, 5, and 6. For all analyses, all utterances were initially tokenized, converted to lowercase, and lemmatized (e.g., *troubling* and *troubles* both become *trouble*).

C-unit

During transcription, all speech was segmented into Communication-units (C-units) according to SALT guidelines (Miller and Iglesias, 2012). A C-unit is an independent clause with its associated modifiers, including dependent clauses. Throughout this dissertation, we use the term "utterance" as a proxy for C-unit. A selection of utterances and their corresponding c-units are shown in Table 3.2.

Utterance	C-unit(s)
"The frog was sitting on a lily-pad, and it got a tan."	*C: The frog was sitting on a lily-pad.* *C: And it got a tan.*
"The frog was eating biscuits and gravy."	*C: The frog was eating biscuits and gravy.*
"The frog was tan because it sat it the sun."	*C: The frog was tan because it sat in the sun.*
"The frog jumped in the water and then it got wet and then it wanted to get out and then it got out and the boy looked at it."	*C: The frog jumped in the water.* *C: And then it got wet.* *C: And then it wanted to get out.* *C: And then it got out.* *C: And the boy looked at it.*
"Well here is a dolphin, a big alligator, a boy playing with a beach ball, a blue house."	*C: Well here is a dolphin.* *C: A big alligator.* *C: A boy playing with a beach ball.* *C: A blue house.*
"I like candy, popsicles, and stuff."	*C: I like candy.* *C: Popsicles.* *C: And stuff.*

Table 3.2: Examples of transcribed utterances segmented into C-units.

Final punctuation

At the end of each utterance, one of the following five punctuation marks was included: . ? ! > ^. Table 3.3 shows examples of these punctuation marks in practice and describes their intended usage.

Symbol	Usage	Example
.	Most natural utterances	*C: The dinosaur breathes fire.*
!	Exclamations	*C: I like this book!*
?	Questions	*E: What happens next?*
>	Abandoned utterances	*C: Yesterday I went to>*
^	Interrupted utterances	*E: What do you^*

Table 3.3: Transcription guidelines for final punctuation and example usage.

Overlapping speech

For instances where the examiner and the child both spoke at the same time, spans of overlapping speech were surrounded by angled brackets. Examples of this in practice are shown in Table 3.4, with “XX” representing a segment of unintelligible speech.

1.	*E: What happens <next>?* *C: <Then the> boy left.*
2.	*E: What <do you>^* *C: <I like this> book!*
3.	*E: How about <XX>.* *C: <This is> fun.*

Table 3.4: Examples of transcribed overlapping speech.

Mazes

Any instances of mazes (i.e., intervals of disfluent speech) – including repetitions, false starts, revisions, and fillers – were manually marked by transcribers. The corresponding word or words were surrounded by parentheses, with each set of parentheses only containing a single maze. Table 3.5 contains examples of transcribed mazes.

Type of Maze	Example
Repetition	*C: I (want to) (want to) want to eat.*
False Start	*C: (I hate the) I really don't like the highway.*
Revision	*C: I (like) love butter.*
Filler	*C: He threw it (uh) so fast it went out of the park!*

Table 3.5: Examples of transcribed mazes.

Emotions
1. “What do you like doing that makes you feel happy and cheerful?”
2. “What kinds of things make you feel this way? How *do* you feel when you’re happy? Can you describe it?”
3. “What about things that you’re afraid of?”
4. “What makes you feel frightened or anxious? How does it feel? What do you do?”
5. “What about feeling angry?”
6. “What kinds of things make you feel that way? How do you feel ‘inside’ when you’re angry?”
7. “Most people have times when they feel sad. What kinds of things make you feel that way?”
8. “How *do* you feel when you’re sad? What is it like when you’re sad? Can you describe that?”
9. “How about feeling relaxed or content? What kinds of things make you feel that way?”

Table 3.6: Interview questions provided for the *Emotions* activity in the ADOS-2, Module 3.

Social Difficulties and Annoyance
1. “Have you ever had problems getting along with people at school? How about at home with your family? Do you ever get in trouble? Why? What for?”
2. “Are there things that other people do that irritate or annoy you? What are they?”
3. “What about things you do that annoy others? (If no response, ask: “What about your brother(s) or sister(s) or parents?”)
4. “Have you ever been teased or bullied? Why, do you think?”
5. “Have you ever tried to change these things? Have you ever done anything so that others wouldn’t tease you? How has it worked?”
6. “Are there other kids/people you know who get teased or bullied?”

Table 3.7: Interview questions provided for the *Social Difficulties and Annoyance* activity in the ADOS-2, Module 3.

Friends, Relationships, and Marriage
1. "Do you have some friends? Can you tell me about them?"
2. "What do you like doing together? How did you get to know them? How often do you get together?"
3. "What does being a friend mean to you? How do you know someone is your friend?'
5. "Do you have a girlfriend or boyfriend? What is her/his name? How old is she/her?"
6. "When do you see her/him last?"
7. "What is she/he like? What do you like to do together?"
8. "How do you know she/he is your girlfriend/boyfriend?"
9. "Where do you want to live when you get older? What kind of place (apartment, house, condo)?"
10. "Whom do you think you would like to live with? Your family, a roommate(s), by yourself?"
11. "Do you ever think about having a long-term relationship or getting married (when you are older)?"
12. "Why do you think some people get married or live with a girlfriend or boyfriend when they grow up?"
13. "What would be nice about it? What might be difficult about being married or living with a girlfriend or boyfriend? Or living with a roommate?"

Table 3.8: Interview questions provided for the *Friends, Relationships, and Marriage* activity in the ADOS-2, Module 3.

Loneliness
1. "Do you ever feel lonely?"
2. "Do you think other kids/people your age ever feel lonely?"
3. "Are there things that you do to help yourself feel better? What about things other people do to help themselves feel better when they're lonely?"

Table 3.9: Interview questions provided for the *Loneliness* activity in the ADOS-2, Module 3.

Chapter 4

Filler Usage, *Um* and *Uh*

Although sometimes overlooked, disfluencies are a common and enduring feature of spoken language (Wieling et al., 2016). A disfluency is a break or interruption that upsets the flow of speech. Common types of disfluencies include repetitions, repairs, false starts, silent pauses, and fillers. Also known as filled pauses, fillers can be sounds, words, non-words, and phrases. Two of the most frequently used fillers in English are *um* and *uh*[1]. Both are thought to have important pragmatic contributions to a conversation. Clark and Fox Tree (2002) posit that they are used by the speaker to signal to the listener that they are anticipating a delay or pause in their speech caused by a planning or production problem. Furthermore, the length of the expected pause varies based on which filler is used, with *um* signaling a major delay and *uh* signaling a minor delay. By using *um* or *uh*, the speaker is able to hold their turn in the conversation while they quickly address the source of the delay (e.g., recall a forgotten word or formulate a response to a prior question). These fillers fill what would otherwise be an abrupt, unexplained, and perhaps silent pause that the listener might interpret as confusing.

Thus far, the studies investigating filler usage in the language of autistic individuals support Clark and Fox Tree's (2002) view that fillers are intentionally used by the speaker to benefit the listener (Engelhardt et al., 2017; Engelhardt, 2019; Gorman et al., 2016; Irvine et al., 2016; Lake, 2008; Lake et al., 2011; Lunsford et al., 2010; Parish-Morris et al., 2017). These studies hypothesize that because ASD is characterized by deficits in conversational reciprocity and social communication, individuals with this disorder would be less likely to exhibit language and behavior that benefits their conversational partner, such as *um* and *uh*. Engelhardt (2019) further hypothesizes that higher rates of disfluencies, including *um* and *uh*, would be associated with higher executive

The contents of this chapter was previously published as a journal article in the Journal of Autism and Developmental Disorders (JADD; Lawley et al., 2022) and as an abstract at the International Society for Autism Research Annual Meeting (INSAR; Lawley et al., 2021).

[1] While there are many words besides *um* and *uh* that are considered fillers in the English language, from this point forward we will use "filler" to refer to *um* and *uh* exclusively.

functioning, higher intelligence, and better social communication abilities. If *um* and *uh* are intentionally used to manage listeners' expectations, then we would expect that autistic individuals would use these fillers in a conversational setting significantly less often than Typically Developing (TD) individuals. Moreover, we would expect that among autistic individuals, filler usage rates would correlate with the autistic individual's executive functioning, intelligence, and social communication abilities.

While there are many who support Clark and Fox Tree's (2002) view that *um* and *uh* are intentionally used to support listener comprehension, others question whether the fillers are simply byproducts of a planning or production difficulty that happen to also benefit listener comprehension (Finlayson and Corley, 2012). To our knowledge, there have not been any studies on the language of autistic individuals that explore this latter hypothesis in depth. If *um* and *uh* are merely byproducts of a problem with language planning and production, then their usage may be associated with an autistic individual's structural or general language ability and not with their social communication and pragmatic language ability.

Previous studies have shown that autistic children and young adults use *um* significantly less than their TD peers. This result has held true across multiple samples and a variety of language sampling contexts. Preliminary analyses indicate that lower *um* usage (overall and relative to *uh* usage) is associated with the lower social communication abilities characteristic of ASD and not with general language ability, executive functioning, or intelligence, but these analyses would benefit from a multivariate approach that controls for potential confounding factors such as age and sex. Furthermore, whether measures of pragmatic language ability or structural language ability (i.e., planning and production) are associated with filler usage among autistic children remains to be explored. Lastly, previous studies were limited by small sample sizes and few female participants.

4.1 Objectives

The specific objectives of the analyses in this chapter were to:

1. Ascertain filler usage differences between ASD and TD groups in a large, well-characterized sample.

2. Test the robustness of diagnostic group differences (ASD; TD) of filler usage by controlling for various participant-level variables such as age, sex, and IQ.

3. Examine, within the ASD group, whether filler usage is associated with the difficulties with

social communication, pragmatic language, and structural language characteristic of ASD while controlling for participant age, sex, and IQ.

4.2 Previous studies

Previous studies have reported that autistic children and young adults use *um* significantly less frequently than the TD controls, while the two groups do not differ in *uh* usage (Gorman et al., 2016; Irvine et al., 2016; McGregor and Hadden, 2020; Salem et al., 2021).[2] Irvine et al. (2016) analyzed spontaneous speech samples gathered from a painting description activity for a small sample of 24 autistic participants (21 male) and 16 TD participants (14 male) aged 8 to 21 years old. They found that the ASD group said *um* at a significantly lower rate than the TD group, while the two groups did not differ significantly in *uh* usage. Gorman et al. (2016) compared transcribed Autism Diagnostic Observation Schedule (ADOS) sessions for a sample of 50 autistic children (45 male) and 43 TD children (31 male), 4 to 8 years of age. After controlling for participant age, intelligence, ADOS activity type, and the position of the filler (i.e., initial token in utterance or non-initial), the autistic children still used *um* significantly less than TD children, while not differing in *uh* usage. In addition to general *um* and *uh* frequency, Gorman et al. (2016) also compared participant *um-ratio*, the ratio of *ums* and *uhs* produced by each participant which was calculated as total *um* / (total *um* + total *uh*). After controlling for the same participant-level variables as before, they found that the autistic participants had significantly lower *um-ratios* than TD participants. McGregor and Hadden (2020) compared spontaneous speech samples gathered from a structured, interview-style task of 31 autistic children (29 male) and 32 TD children (16 male) aged 7 to 15. Like Gorman et al. (2016) and Irvine et al. (2016), they found that the ASD group used *um* less frequently than the TD group, with no difference in *uh* frequency. Salem et al. (2021) analyzed transcribed ADOS sessions for a sample of 96 autistic children (80 male) and 45 TD children (31 male), ages 7 to 17. They compared *um-ratio* (which they refer to as *um proportion*) and also found that autistic participants had lower *um-ratios* than TD participants.

While previous results support the hypothesis that autistic children use *um* and *uh* differently than TD children, with autistic children specifically using *um* at much lower rates, the underlying processes that cause this difference remain unclear. This question has been addressed by some authors (Gorman et al., 2016; McGregor and Hadden, 2020; Irvine et al., 2016; Salem et al., 2021) with variable results. Irvine et al. (2016) found that within the ASD group, *um* usage was not

[2] The participant population analyzed in this chapter partially overlaps with both the sample analyzed by Gorman et al. (2016) and the sample analyzed by Salem et al. (2021).

correlated with measures of executive functioning, intelligence, or general language ability, but *um* usage was negatively correlated with Social Communication Questionnaire (SCQ; Rutter et al., 2003) scores (a caregiver-reported measure of social communication impairment, where a higher score indicates a more pronounced difficulties). In other words, the autistic participants who had less severe social communication difficulties also used *um* more frequently. Similarly, Gorman et al. (2016) reported that among the autistic participants, *um* usage relative to *uh* usage (i.e., *um-ratio*) was also negatively correlated with SCQ scores, but was also not correlated with measures of executive functioning, intelligence, or general language ability. Surprisingly, although McGregor and Hadden (2020) successfully replicated the primary findings of Gorman et al. (2016) and Irvine et al. (2016) (i.e., no difference in *uh* usage, ASD group had lower *um* usage than TD group), they did not find an association between SCQ scores and either measure of *um* usage. In contrast, Salem et al. (2021) found that across all participants (including TD participants), *um-ratio* was positively correlated with structural, pragmatic, and overall language ability as measured by the Children's Communication Checklist, 2nd edition (CCC-2; Bishop).

4.3 Methods

4.3.1 Data

Analyses were performed on transcribed ADOS, Module 3 sessions for 117 autistic children (98 male) and 65 TD children (37 male) aged 4 to 15 years old. Participants came from two larger studies, both conducted at Oregon Health & Science University in Portland, OR, USA. For autistic participants, diagnosis was confirmed based on expert clinical judgment according to DSM-IV-TR criteria (American Psychiatric Association, 2000). All participants were native English speakers, had an IQ ≥ 70, had fluent speech, and had a mean length of utterance in morphemes (MLUM) ≥ 3.0. More information on the participants, language samples, and standardized measures can be found in Chapter 3.

4.3.2 Language samples

Analyses were performed on transcribed ADOS sessions (Lord et al., 2000). All participants were administered the ADOS-2, Module 3, which is designed for children and adolescents with fluent speech. Sessions were transcribed by a team of trained transcribers according to modified Systematic Analysis of Language Transcript (SALT) guidelines (Miller and Iglesias, 2012). Further details on the transcription process can be found in Chapter 3.2.

4.3.3 Filler usage measures

For each participant, we counted the total number of *um* tokens, total number of *uh* tokens, and total number of words overall. Following previous studies,[3] any immediate repeats of *um* or *uh* (e.g., *um um um yes*) were preserved. We then computed three measures of filler usage:

- *uh-rate* = total *uh* / total words
- *um-rate* = total *um* / total words
- *um-ratio* = total *um* / (total *um* + total *um*)

The *uh-rate* and *um-rate* measures represent how many *uhs* or *ums* a participant said relative to the total number of words they said, with a higher value indicating a greater *uh* or *um* usage, respectively. The *um-ratio* measure corresponds to how many of the fillers a participant said were *um* (rather than *uh*). A higher value indicates that more *ums* compared to *uhs* were said while a lower value indicates that more *uhs* compared to *ums* were said. An *um-ratio* of 0.75 for example, would mean that 75% of the fillers said were *ums* while the remaining 25% were *uhs*.

Of the 182 participants in the sample, three autistic participants never used *um* or *uh*. These three participants were omitted from the *um-ratio* analyses but were included in the *uh-rate* and *um-rate* analyses.

4.3.4 Statistical analyses

Our first objective was to ascertain differences in filler usage between ASD and TD groups in both samples. As all three filler usage measures (*uh-rate*; *um-rate*; *um-ratio*) failed to satisfy standard tests of normality (Shapiro-Wilk Normality test; $p < 0.001$), nonparametric Wilcoxon-Mann-Whitney tests were used to compare groups with *uh-rate*, *um-rate*, or *um-ratio* as the dependent variable and diagnostic group (ASD; TD) as the independent variable. To determine the magnitude of the effect of diagnostic group, if any, effect sizes were estimated as rank-biserial correlations (r_{rb}). We used the following ranges to interpret the resulting value: small = 0.10 - 0.29, medium = 0.30 - 0.49; large = 0.50 - 1.0) (Cureton, 1956; Wendt, 1972).

Our second objective was to investigate group differences of *uh-rate*, *um-rate*, and *um-ratio* while also controlling for individual difference variables. Looking to the inferential analyses detailed in Lunsford et al. (2012) and Gorman et al. (2016) for guidance, we fit a mixed effects logistic

[3] Irvine et al. (2016) and McGregor and Hadden (2020) did not specify whether immediate repeats were removed, so we assumed that any repeats were left as is. While Gorman et al. (2016) did remove repeats, they repeated all statistical analyses with repeats preserved and found no difference in the results obtained.

regression model for each filler usage measure. Since the data consisted of all tokens said by the participants (i.e., one token per row), a per-participant random intercept was included in all three models in order to group the tokens accordingly. Our first model estimated participant *uh-rate*. To do so, we coded each *uh* token as a "hit" and every other token as a "miss". We used the resulting variable as the binary response variable in the model. The primary predictor variable was participant-level diagnosis, with participant-level sex, age, and IQ also included as predictor variables. Continuous variables were transformed into z-scores and categorical variables were encoded using sum coding. Our second model estimated participant *um-rate*. The structure of this model was identical to our first model, with the exception of the binary response variable: each *um* token was a "hit" and every other token was a "miss". Lastly, our third model estimated participant *um-ratio*. For this model, all tokens that were not *um* or *uh* were excluded from the input data. For the binary response variable, each *um* token was coded as a "hit" and each *uh* token as a "miss". All other structural elements were the same as the previous models. Table 4.1 shows the R formulas used to create these three models.

Measure	R formula syntax
uh-rate	`uh_or_not ~ dx + sex + age + iq + (1|study_id)`
um-rate	`um_or_not ~ dx + sex + age + iq + (1|study_id)`
um-ratio	`um_or_uh ~ dx + sex + age + iq + (1|study_id)`

Table 4.1: R formulas used to model *uh-rate*, *um-rate*, and *um-ratio* for Objective 2.

Lastly, our third objective was to investigate whether filler usage differences within the ASD group were associated with the social communication, pragmatic language, and structural language difficulties associated with ASD while controlling for participant age, sex, and IQ. Since the TD participants were the control group for this dataset, they did not have meaningful scores on measures of autism symptom severity. As such, we restricted the following analysis to only the autistic participants. For each of our three filler usage measures (*uh-rate*; *um-rate*; *um-ratio*), we fit mixed effects logistic regression models that were adjusted for participant-level age, IQ, and sex. All models had the same input data, per-participant random intercept, and binary response variables as our second objective models. The ADOS Social Affect (SA), CCC-2 Pragmatic, and CCC-2 Structural scores were included as predictors. After fitting separate models for each of these scores, we fit an "omnibus" model: a fully-adjusted model that included all three of the scores as predictors. Table 4.2 shows the R formulas used to create the models described above.

We chose these particular measures from the ADOS and the CCC-2 to represent the social

Model	Measure	R formula syntax
A	*uh-rate*	`uh_or_not ~ ados_sa + sex + age + iq + (1|study_id)`
	um-rate	`um_or_not ~ ados_sa + sex + age + iq + (1|study_id)`
	um-ratio	`um_or_uh ~ ados_sa + sex + age + iq + (1|study_id)`
B	*uh-rate*	`uh_or_not ~ ccc2_prag + sex + age + iq + (1|study_id)`
	um-rate	`um_or_not ~ ccc2_prag + sex + age + iq + (1|study_id)`
	um-ratio	`um_or_uh ~ ccc2_prag + sex + age + iq + (1|study_id)`
C	*uh-rate*	`uh_or_not ~ ccc2_struct + sex + age + iq + (1|study_id)`
	um-rate	`um_or_not ~ ccc2_struct + sex + age + iq + (1|study_id)`
	um-ratio	`um_or_uh ~ ccc2_struct + sex + age + iq + (1|study_id)`
D	*uh-rate*	`uh_or_not ~ ados_sa + ccc2_prag + ccc2_struct + sex + age + iq + (1|study_id)`
	um-rate	`um_or_not ~ ados_sa + ccc2_prag + ccc2_struct + sex + age + iq + (1|study_id)`
	um-ratio	`um_or_uh ~ ados_sa + ccc2_prag + ccc2_struct + sex + age + iq + (1|study_id)`

Table 4.2: R formulas used to model *uh-rate*, *um-rate*, and *um-ratio* for Objective 3.

communication difficulties associated with ASD for two reasons: (1) Both the CCC-2 and the ADOS are widely used tools that are used when assessing and diagnosing ASD; (2) The ADOS SA score is clinician-derived and is calculated following 30-60 minutes of direct observation and interaction with the subject; the CCC-2 Structural and Pragmatic scores both are caregiver-derived and represent the participants' current language abilities in everyday, naturalistic contexts.

As previous studies have found that *uh-rate* does not distinguish autistic children and TD children (Gorman et al., 2016; Irvine et al., 2016; McGregor and Hadden, 2020), we were not expecting to find an association between *uh-rate* and any of the ADOS and CCC-2 scores. Instead, we were expecting to find an association between both *um-rate* and *um-ratio* with these scores. For both the CCC-2 Pragmatic and Structural scores, a higher score is indicative of better language skills. Therefore, with the assumption that the ASD group would have lower *um-rates* and *um-ratios* than TD group, we expected that higher CCC-2 scores within the ASD group would be associated with higher *um-rates* and *um-ratios*. In contrast, as higher ADOS SA score is indicative of more profound differences in social communication, we were expecting a negative association with *um-rate* and *um-ratio*. In other words, we anticipated that higher ADOS SA scores would be associated with lower *um-rate* and *um-ratio* values.

All analyses were completed using the statistical programming language R (R Core Team, 2020) using the `lme4` package (Bates et al., 2015) to create the mixed effects logistic regression models.

4.4 Results

4.4.1 Objective 1

Our first objective was to ascertain filler usage differences between the ASD and TD groups. The median and interquartile range (IQR) values for *uh-rate*, *um-rate*, and *um-ratio* for the ASD and TD groups are shown in Table 4.3. There was an unexpected but significant difference in *uh-rate*

	ASD	TD	U	p	r_{rb}
uh-rate	0.007 [0.003, 0.013]	0.004 [0.002, 0.008]	4793.0	<0.01	0.260
um-rate	0.005 [0.001, 0.014]	0.017 [0.008, 0.028]	2213.5	<0.001	0.418
um-ratio	0.468 [0.167, 0.751]	0.806 [0.611, 0.900]	2036.0	<0.001	0.450

Table 4.3: Filler usage frequency by diagnostic group.

($p < 0.01$; small effect size: $r_{rb} = 0.260$; Table 4.3). The ASD participants used *uh* more frequently than the TD participants (ASD = 0.007 [0.003, 0.013] > TD = 0.004 [0.002, 0.008]; median [IQR]). Consistent with findings from previous studies, we also found a significant difference in *um-rate* between diagnostic groups ($p < 0.001$; medium effect size: $r_{rb} = 0.418$), with autistic participants using *um* less frequently than TD participants (ASD = 0.005 [0.001, 0.014] < TD = 0.017 [0.008, 0.028]). Lastly, there was a significant difference in *um-ratio* between diagnostic groups ($p < 0.001$; medium effect size: $r_{rb} = 0.450$). The ASD participants had a lower *um-ratio* compared to the TD participants (ASD = 0.468 [0.167, 0.751] < TD = 0.806 [0.611, 0.900]). In other words, when producing a filler (i.e., *um* or *uh*), TD participants said *um* 80.6% of the time while ASD participants said *um* only 46.8% of the time.

4.4.2 Objective 2

Our second objective was to investigate filler usage differences between the ASD and TD groups while controlling for age, sex, and IQ. Results for the models of *uh-rate*, *um-rate*, and *um-ratio* are shown in Table 4.4.

After adjusting for participant age, sex, and IQ, diagnostic group no longer had a significant effect in predicting *uh-rate*. In contrast, diagnostic group still had a significant effect in both the

	Uh-rate				*Um-rate*				*Um-ratio*			
Predictor	Log-odds	S.E.	χ^2	$P(\chi^2)$	Log-odds	S.E.	χ^2	$P(\chi^2)$	Log-odds	S.E	χ^2	$P(\chi^2)$
Intercept	-5.35	0.098	—	—	-4.67	0.109	—	—	0.723	0.150	—	—
DX	—	—	1.237	0.266	—	—	16.317	<.001	—	—	13.347	<.001
ASD	0.114	0.102	—	—	-0.475	0.116	—	—	-0.598	0.161	—	—
TD	-0.114	—	—	—	0.475	—	—	—	0.598	—	—	—
Sex	—	—	3.903	0.048	—	—	1.470	0.225	—	—	4.695	0.030
Male	0.198	0.099	—	—	-0.135	0.111	—	—	-0.338	0.155	—	—
Female	-0.198	—	—	—	0.135	—	—	—	—	0.338	—	—
Age	0.124	0.087	2.045	0.153	0.179	0.099	3.230	0.072	0.023	0.137	0.027	0.869
IQ	-0.107	0.092	1.366	0.243	-0.007	0.106	0.004	0.949	0.162	0.146	1.235	0.266

Table 4.4: Effect of diagnostic group on filler usage rates after adjusting for sex, age, and IQ.

um-rate model ($P(\chi^2) < 0.001$; $\chi^2 = 16.317$; Table 4.4) and the *um-ratio* model ($P(\chi^2) < 0.001$; $\chi^2 = 13.347$), with the ASD participants having significantly lower *um-rate* and *um-ratio* than the TD participants. Sex was significant in the *uh-rate* model ($P(\chi^2) = 0.048$; $\chi^2 = 3.903$), with female participants using *uh* less frequently than male participants, regardless of diagnostic group. The significant effect of diagnostic group on *uh-rate* found in the unadjusted analyses for Objective 1 was likely a reflection of the higher proportion of male participants in the ASD group compared to the TD group. After adjusting for sex in the *uh-rate* model, diagnostic group was no longer significant. There was no significant effect of sex in the *um-rate* model. However, sex was significantly associated with *um-ratio* ($P(\chi^2) = 0.030$; $\chi^2 = 4.695$), with male participants having a lower *um-ratio* compared to the female participants. Age and IQ were not associated with *um-rate*, *uh-rate*, or *um-ratio*.

4.4.3 Objective 3

Our third analysis examined whether filler usage rates were associated with the difficulties with social communication, pragmatic language, and structural language observed in ASD. To this effect, we restricted analyses to only the ASD sample. Results for the models are summarized in Table 4.5; the full set of results are reported at the end of this chapter in Table 4.6. All three models were adjusted for participant age, sex, and IQ. As anticipated, all *uh-rate* models showed no significant

Model	Predictor	*Uh-rate* OR	*Uh-rate* p	*Um-rate* OR	*Um-rate* p	*Um-ratio* OR	*Um-ratio* p
A	ADOS SA	1.21	0.282	1.04	0.861	0.80	0.484
B	CCC-2 Pragmatic	1.06	0.804	1.56	0.152	1.56	0.268
C	CCC-2 Structural	1.08	0.595	1.68	0.009	1.61	0.064
D	ADOS SA	1.22	0.275	1.05	0.840	0.81	0.497
	CCC-2 Pragmatic	0.99	0.966	0.94	0.862	0.96	0.936
	CCC-2 Structural	1.10	0.635	1.73	0.028	1.63	0.133

Table 4.5: Effect of three measures of social communication on filler usage rates within ASD group after controlling for sex, age, and IQ – summarized results.

association between *uh-rate* and any of the autism symptom severity measures, whether clinician-based (ADOS SA) or caregiver-derived (CCC-2 Pragmatic, CCC-2 Structural). This was the case both when the scores were modeled individually (Models A-C; Table 4.5) and in the omnibus model (Model D). Contrary to our initial hypotheses for the *um-rate* and *um-ratio* models, the clinician-based measure (ADOS SA) was not significantly associated with participant *um-rate* or *um-ratio* when modeled in isolation (Model A) or after controlling for the two other social communication scores (Model D). In regards to the two caregiver-derived scores, the CCC-2 Pragmatic score was not significant in any of the *um-rate* or *um-ratio* models (Models B-D). In contrast, the CCC-2 Structural score had a significant effect on participant *um-rate* when modeled on its own (OR = 1.68, $p = 0.009$; Model C). This significant association remained after adjusting for the CCC-2 Pragmatic and ADOS SA scores (OR = 1.73, $p = 0.028$; Model D). An odds ratio value of 1.73 means that an increase of 1 standard deviation in the CCC-2 Structural score corresponds to a 73% increase in *um-rate* (to be discussed further below). For the *um-ratio* models, the CCC-2 Structural score was not significant in either the individual (Model C) or combined (Model D) models. However, even though they did not reach significance, it is noteworthy that the odds ratios for the CCC-2 Structural in *um-ratio* Model C (OR = 1.61) and Model D (OR = 1.63) were of similar magnitude as the odds ratios for the same score in the *um-rate* Model C (OR = 1.68) and Model D (OR = 1.73).

For the three additional covariates (age, sex, and IQ), there were only two significant effects: sex in *um-ratio* Model C (OR = 0.61, $p = 0.037$; Table 4.6); age in *um-rate* Model D (OR = 1.33, $p = 0.048$). Given that a total of 66 statistical comparisons were performed in this analysis, it is likely that these two results were due to Type I error. As before, both IQ and age did not significantly predict filler usage rate in any of the models.

4.5 Discussion

We analyzed *um* and *uh* usage in a large, well-characterized sample of autistic and TD children and investigated whether usage patterns observed within the ASD group were associated with differences in social communication and language skills. Before controlling for participant-level age, sex, and IQ, the ASD group had significantly higher *uh-rate* than the TD group as well as significantly lower *um-rate*, and *um-ratio* values. The difference in *uh-rate* is likely due to the high number of female participants in the sample (see below). Once we adjusted for age, sex, and IQ, the two groups no longer significantly differed in *uh-rate* but the ASD group still had significantly lower *um-rate* and *um-ratio* values than the TD group. These results were in line with previous studies on filler usage differences on autistic children and TD children. Within the ASD group, social communication and pragmatic language ability did not predict *uh-rate*, *um-rate* or *um-ratio*, regardless of age, sex, and IQ. Furthermore, while structural language ability was not predictive of *uh-rate* or *um-ratio* among autistic participants, it did significantly predict *um-rate*, with a better structural language ability corresponding to a greater *um-rate*. Structural language ability was predictive of *um-rate* within the ASD group regardless of participant's age, sex, and IQ, and remained predictive after social communication and pragmatic language ability were accounted for. In summary, we found that after controlling for age, sex, and IQ, lower "um" usage by autistic children was associated with lower structural language ability and not with lower social communication or lower pragmatic language ability. This suggests that lower *um* usage by autistic children may be a byproduct of a planning or language production difficulty (Finlayson and Corley, 2012) instead a difficulty with taking the listener perspective into consideration (Clark and Fox Tree, 2002; Engelhardt, 2019).

The significant difference in *uh-rate* between diagnostic groups found in our Objective 1 analysis is likely due to our sample having a greater proportion of female participants (26%) than previous similar studies, such as Gorman et al. (2016) and Irvine et al. (2016), which had 17.7% and 12.5% female participants overall, respectively. Sex-based differences in *um* and *uh* usage by autistic children have been previously explored by Parish-Morris et al. (2017) in a study that specifically focused on differences between female and male individuals with ASD. They analyzed filler usage in ADOS samples for 65 autistic children and 17 children with TD, aged 6 to 17 years old. They found that within both the ASD and TD groups, female participants used less *uhs* than their male counterparts. Lower *uh* usage by female participants further resulted in the females having higher *um-ratio* values than the males. Given Parish-Morris et al.'s (2017) findings and *uh-rate* no longer distinguishing the ASD and TD groups after we adjusted for sex, age, and IQ, it is likely the initial *uh-rate* finding is derived from the higher proportion of females in our study sample.

In line with our initial hypotheses and results from prior studies, the ASD group had significantly lower *um-rate* and *um-ratio* values than the TD group, irrespective of participant age, sex, and IQ. This finding suggests that *um-rate* and *um-ratio* differences between autistic children and TD children are persistent and observable over a wide range of ages (4 to 15 years old). These differences are also robust to measures of intelligence, further supporting the hypothesis that the difference is due to language ability and not individual difference variables. As such, *um-rate* and *um-ratio* could be useful indicators of social communication language ability for a wide variety of autistic children who have fluent speech, irrespective of their IQ and age. Currently, disfluencies usage patterns, such as *um* and *uh*, are not used in conventional assessments of social communication ability, such as the CCC-2. If confirmed, these measures could be included as an additional language feature in screening instruments or as an outcome measure in treatment studies.

Contrary to our initial hypotheses, the clinician-derived measure of social communication skills (ADOS SA) and the caregiver-reported measure of pragmatic language ability (CCC-2 Pragmatic) were not predictive of *um* usage among the autistic participants. Surprisingly, only the caregiver-reported measure of structural language ability (CCC-2 Structural) predicted *um* usage within the ASD group, with a higher CCC-2 Structural score (i.e., better structural language ability) being associated with a higher *um-rate* (*ums* relative to total words). Our result suggests that filler usage, specifically *um* usage, is due to a language production or planning difficulty and not by a failure to account for the listener's needs during a conversation. This conflicts with previous findings: Irvine et al. (2016) and Gorman et al. (2016) found that social communication ability (as measured by the SCQ total score) predicted *um* usage (the former found it to predict *um-rate* while the latter found it to predict *um-ratio*). One possible explanation for why our results differed from the results of Irvine et al. (2016) and Gorman et al. (2016) is that our sample was considerably larger than both of the previous samples and also spanned a larger age range. Alternatively, the difference may be due to different social communication measurements used across studies (ADOS-2, CCC-2, and SCQ). We did not have SCQ scores for our sample, so we were not able to investigate the effect of the SCQ total score on *um* usage.

Our finding that the *um-ratio* of autistic males was significantly lower than autistic females was consistent with the results of Parish-Morris et al. (2017). Even though the current study included more female participants than usual, there was still a marked male preponderance in our study sample. Future studies that include more female participants, allowing for well-powered analyses stratified by sex, are needed before any definite conclusions can be drawn. Historically, a majority of research on ASD has focused on males (Halladay et al., 2015; Happé and Frith, 2020). There is limited data available on autistic, resulting in few studies on sex differences. However, there has been a recent effort to investigate the language of autistic girls (Parish-Morris et al., 2017; Song et al., 2020; Wood-Downie et al., 2021). As more time and resources are devoted, we are optimistic for a better representation of autistic children in the research community and literature.

Using NLP approaches to measure pragmatic language problems, such as disfluency usage patterns, in transcribed language samples is an exciting and promising alternative to conventional language assessments. Although transcription of language samples is costly, we anticipate future improvements to voice-to-text technology that will make the process more feasible. With the use of NLP, robust measures of pragmatic language features, including "um" and "uh" usage, as well as other pragmatic features, can be obtained in an automated, reliable, and relatively inexpensive fashion.

4.5.1 Limitations

There were several limitations in this analysis that we would like to mention. Our findings on disfluency usage patterns cannot be extended to minimally-verbal autistic children, those with intellectual disability, autistic preschoolers, or autistic adults. Additionally, the language samples analyzed were collected using a semi-structured clinical instrument (ADOS-2). While using the ADOS to assess spontaneous expressive language has been recommended and reviewed closely (Kover et al., 2014), some argue that language samples collected by caregivers during naturalistic interactions are ideal candidates to use when deriving language outcome measures (Barokova and Tager-Flusberg, 2020). As such, whether our findings generalize to everyday conversational contexts remains to be examined.

4.6 Summary

In this chapter, we analyzed *um* and *uh* in a large, well-characterized sample of autistic and TD children and investigated whether usage patterns observed within the ASD group were associated with differences in social communication and language skills. In line prior studies, the ASD group had significantly lower *um-rate* and *um-ratio* values than the TD group, irrespective of participant age, sex, and IQ. This finding suggests that *um-rate* and *um-ratio* differences between autistic children and TD children are persistent and observable over a wide range of ages (4 to 15 years old). These differences are also robust to measures of intelligence, further supporting the hypothesis that the difference is due to language ability and not individual difference variables. As such, *um-rate* and *um-ratio* could be useful indicators of social communication language ability for a wide variety of autistic children who have fluent speech, irrespective of their IQ and age. Currently, disfluencies usage patterns, such as *um* and *uh*, are not used in conventional assessments of social communication ability, such as the CCC-2. If confirmed, these measures could be included as an additional language feature in screening instruments or as an outcome measure in treatment studies.

We further found that after controlling for age, sex, and IQ, lower *um* usage by autistic children was associated with lower structural language ability and not with lower social communication or lower pragmatic language ability. This suggests that lower *um* usage by autistic children may be a byproduct of a planning or language production difficulty (Finlayson and Corley, 2012) instead a difficulty with taking the listener perspective into consideration (Clark and Fox Tree, 2002; Engelhardt, 2019).

		Uh-rate		*Um-rate*		*Um-ratio*	
Model	Predictor	OR	p	OR	p	OR	p
A	*(Intercept)*	0.01	< 0.001	0.01	< 0.001	1.27	0.377
	Sex	1.14	0.371	0.75	0.116	0.66	0.088
	Age	1.14	0.258	1.25	0.131	1.06	0.777
	IQ	1.03	0.840	1.04	0.800	1.04	0.864
	ADOS SA	1.21	0.282	1.04	0.861	0.80	0.484
B	*(Intercept)*	0.01	< 0.001	0.01	< 0.001	1.56	0.212
	Sex	1.17	0.273	0.75	0.115	0.64	0.067
	Age	1.13	0.295	1.26	0.115	1.08	0.703
	IQ	0.95	0.664	1.00	0.979	1.10	0.611
	CCC-2 Pragmatic	1.06	0.804	1.56	0.152	1.56	0.268
C	*(Intercept)*	0.01	< 0.001	0.01	< 0.001	1.50	0.143
	Sex	1.16	0.306	0.71	0.053	0.61	0.037
	Age	1.13	0.269	1.33	0.050	1.13	0.539
	IQ	0.93	0.561	0.88	0.409	0.98	0.922
	CCC-2 Structural	1.08	0.595	1.68	0.009	1.61	0.064
D	*(Intercept)*	0.01	< 0.001	0.01	< 0.001	1.60	0.206
	Sex	1.13	0.417	0.70	0.051	0.62	0.052
	Age	1.15	0.230	1.33	0.048	1.11	0.573
	IQ	1.00	1.000	0.90	0.525	0.91	0.668
	ADOS SA	1.22	0.275	1.05	0.840	0.81	0.497
	CCC-2 Pragmatic	0.99	0.966	0.94	0.862	0.96	0.936
	CCC-2 Structural	1.10	0.635	1.73	0.028	1.63	0.133

Table 4.6: Effect of three measures of social communication on filler usage rates within ASD group after controlling for sex, age, and IQ – extended results.

Chapter 5

Topic Maintenance

Throughout the course of a dialogue many different topics are traversed with varying frequencies, and many analytical tasks in Natural Language Processing (NLP) depend on the ability to meaningfully quantify or otherwise characterize these patterns. For example, a system designed to automatically summarize meetings might need to detect when a new topic has been introduced; in a clinical context, we might want to characterize the topics discussed during a patient visit to facilitate a downstream analysis involving clustering or classification. Topic modeling techniques such as Latent Dirichlet Allocation (LDA; Blei et al., 2003) allow us to capture and quantify the topic distributions across a collection of language samples.

The social communication difficulties that are characteristic of Autism Spectrum Disorder (ASD) sometimes include problems with topic maintenance (Baltaxe and D'Angiola, 1992; Paul et al., 2009), with autistic children having more difficulty staying on topic than TD children. This difference may result in a signal that could be captured by a topic model as the autistic and TD children would have different proportions of their speech assigned to different topics. While there are many existing topic modeling approaches from the NLP field that can be leveraged to investigate topic maintenance in this context, the accompanying evaluation metrics are frequently designed with NLP applications in mind. Typical methods for evaluating the resulting topic distributions use intrinsic metrics such as within-topic coherence; however, to our knowledge there remains a shortage of methods for statistically comparing the topic distributions produced by a model.

The application of topic modeling methods in clinical research has become more common in recent years (Hagg et al., 2022; Boyd-Graber et al., 2017; Jelodar et al., 2019). While topic modeling approaches have advanced significantly over the last twenty years, evaluation methods have lagged behind (see Hoyle et al., 2021 for a recent survey of methods). Current metrics tend to focus on

The contents of this chapter was previously published as a long paper at the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL; Lawley et al., 2023b).

intrinsically assessing model performance (via perplexity on held-out data) or on attempting to measure the quality of the topics that a model produces using metrics based on constructs such as human interpretability of the topics themselves (sometimes referred to as "coherence"). In a clinical research setting, however, the topic distributions produced by a model are themselves often meant for use in meaningfully quantifying differences between clinical populations. In such a scenario, usefully evaluating the quality of a topic model's "fit", or comparing that "fit" to that of another model (perhaps trained via a different algorithm, or with a different choice of hyperparameters) becomes a question of *extrinsic* evaluation, as intrinsic metrics such as perplexity or coherence are unlikely to be sufficient. Additionally, in clinical research, topic models are typically one piece of a larger analytical puzzle, one which often depends on traditional hypothesis-driven inferential statistical approaches (rather than stand-alone evaluation or use, as is more typical with topic models in machine learning scenarios).

In this chapter, we present our novel statistical method for evaluating group differences between topic distributions using the topic modeling method known as LDA. Our proposed method allows for a robust and statistically meaningful evaluation of the output of a topic model in both clinical and non-clinical contexts.

5.1 Objectives

The specific objectives of the accompanying analyses in this chapter were to:

1. Validate our method on the *20Newsgroup* corpus, a widely-used reference corpus for developing and evaluating topic modeling algorithms (Mitchell, 1997), by comparing topic distributions between groups of documents that we expect to be similar and groups that we expect to be different.

2. Apply our method on a corpus of transcribed Autism Diagnostic Observation Schedule (ADOS) sessions of autistic and Typically Developing (TD) children and compare the topic distribution vectors between the ASD and TD groups for both child speech and examiner speech separately.

For the child speech, we hypothesize that autistic children will have different topic distributions than the TD children (i.e., talk about different topics than the TD children), as prior evidence suggests that autistic children struggle with staying on topic more than their TD peers. For the examiner speech, we hypothesize that the examiners will have similar topic distributions regardless of whether they are talking with autistic or TD children, as the ADOS is designed and examiner are trained to ensure uniform assessment irrespective of the participant's diagnostic status.

5.2 Background

5.2.1 LDA

LDA is a unsupervised, generative probabilistic model that is used on a corpus of text documents to model each document as a finite mixture over k topics (Blei et al., 2003). Each document is treated as a bag-of-words (i.e., order does not matter) and is represented as a set of words and their associated frequencies. Given M documents and an integer k, LDA produces a $M \times k$ document-topic matrix (θ). LDA also produces a $k \times V$ topic-word matrix (β), where V is the total number of unique words across the entire corpus of documents.[1] In the $M \times k$ document-topic matrix, θ, each row represents a single document and each column represents one topic.

The elements are the estimated proportion of words in a document that were generated by a topic.

$$\theta = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \dots & \theta_{1,k} \\ \theta_{2,1} & \theta_{2,2} & \dots & \theta_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{M,1} & \theta_{M,2} & \dots & \theta_{M,k} \end{bmatrix} \tag{5.1}$$

From this matrix, each document can now be represented as a k-dimensional topic distribution vector. For example, given a document, m, its corresponding topic distribution vector would be:

$$\begin{bmatrix} \theta_{m,1} & \theta_{m,2} & \dots & \theta_{m,k} \end{bmatrix} \tag{5.2}$$

These LDA-derived topic distribution vectors often serve as useful document representations for downstream analyses, such as a feature vectors for documentation classification or clustering. They are also commonly used as proxies for document content in more qualitative analyses of the composition of text corpora.

[1] Since we will not be using the topic-word matrix in this analysis, from this point forward, we will use the phrases "LDA model" and "document-topic matrix" interchangeably.

5.2.2 Compositional data

Now that we have a way to represent each document as a distribution of topics, next we wanted to compare whether two groups of topic distributions (e.g, ASD vs. TD) were significantly different from one another. To our knowledge, a statistical method for comparing topic distribution vectors between groups of documents has not yet been proposed. One reason for this is due to the numerical properties of the resulting topic distribution vectors (each component θ_i is bounded between $\{0, 1\}$ with the further constraint of $\sum_{i=1}^{k} \theta_i = 1$), which render them unsuitable for use with many parametric statistical methods. This is an important limitation, because as previously mentioned, as the applications of topic modeling methods expand in clinical and behavioral research, the need for statistically based evaluation tools grows.

As we were unable to turn to existing methods, we began to search for other options. We realized that since the components in a topic distribution vector are proportions and all sum to one, they meet the definition of "compositional" data as formalized by Aitchison (1982), who also proposed a family of statistical approaches for such data.

Compositional data are vectors of positive numbers that together represent parts of some whole: e.g., the demographic profile of a city or the mineral compositions of rocks. There are three linear transformations currently defined for compositional data: additive logratio (ALR), center logratio (CLR), and isometric logratio (ILR) transformation. The transformation that best suits our purposes is the ILR transformation. This transformation was introduced by Egozcue et al. (2003) in an effort to broaden the range of statistical methods that can be applied to compositional data by mapping compositonal data into real space. This transformation maps a composition from its original sample space (the D-part simplex) to the $D-1$ Euclidean space with all metric properties preserved:

$$\text{ILR} : S^D \to \mathbb{R}^{D-1} \tag{5.3}$$

We can use the ILR transformation to map the topic distribution vectors from their original sample space (the k-part simplex) to the $k-1$ Euclidean space. It is important to note here that by performing this transformation we will lose one dimension from our topic distributions vectors. This loss of dimension is not an issue in our case since we are not assigning specific meaning to each k individually. However, it is still important to take note of this change in dimension. After the ILR transformation, the topic distribution vector for document m as shown in Equation 5.2 would now be:

$$\begin{bmatrix} \theta_{m,1} & \theta_{m,2} & \dots & \theta_{m,k-1} \end{bmatrix} \tag{5.4}$$

Once the compositions are in $\mathbb{R}^{D-1}$, we are able to use classical multivariate analysis tools such as multivariate analysis of variance (MANOVA) to explore group differences (Egozcue et al., 2003; van den Boogaart et al., 2023). It is important to note that our ability to use MANOVA here is contingent on statistical assumptions that must be met before proceeding. These assumptions are discussed in more detail in the following section.

5.2.3 MANOVA

MANOVA is used to compare multivariate sample means and examines the effect of one discrete, independent variable on multiple continuous, dependent variables. For the analyses described in this chapter, the independent variable is the pre-labeled grouping variable for the documents (e.g., topic category for the *20Newsgroups* corpus; diagnosis (ASD, TD) for the clinical corpus). The dependent variables in both analyses are the various topic distribution probabilities in the document-topic matrix created by LDA: $\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,k-1}$ where $i = 1, 2 \ldots, M$.

It is important to note that a different discrete variable can be used as the independent variable, as long as it separates the documents into groups (e.g., author if modeling a corpus of newspaper articles); if one wished to incorporate multiple independent variables, one could could instead use multivariate analysis of covariance (MANCOVA).

Assumptions

Before proceeding further with MANOVA, there are multiple statistical assumptions that must be met (Tabachnick and Fidell, 2013). These assumptions are as follows:

1. Each combination of independent and dependent variables should be multivariate normally distributed.
2. Dependent variables should have a linear relationship with each group of the independent variable.
3. Variance-covariance matrices for dependent variables should be equal across groups.
4. There should be no extreme outliers in the dependent variables.

Since there are more than 20 observations for each dependent $\times$ independent variable combination the multivariate central limit theorem holds and so we can assume the multivariate normality assumption also holds.

The second assumption was not initially met since each of the original topic distribution vectors summed to 1. However after performing the ILR transformation described earlier in Section 5.2.2, this is no longer the case and the second assumption is met.

The third assumption can be tested using Box's M test (Box, 1949), which tests the null hypothesis that the matrices are equal. If the resulting p-values are > 0.001 for each comparison made, we accept the null hypothesis and the third assumption is met. However, if the resulting p values are all < 0.001, we are still able to proceed since MANOVA is robust to unequal covariance matrices when Pillai's criterion is used (Tabachnick and Fidell, 2013; Pillai, 1955).

For the fourth assumption, extreme outliers can be identified by calculating the Mahalanobis distance for each observation and then performing a chi-squared test (using $df = k - 1$) to calculate the corresponding p-values. The null hypothesis here is that the observation is not an outlier.

Why MANOVA?

MANOVA is a compelling choice for this analysis for several reasons. As detailed above, it enables us to statistically determine whether the topic distributions learned by our topic model are significantly associated with our other variables of interest (group membership, etc.) under a conventional hypothesis-testing framework. Second, MANOVA allows us to calculate interpretable measurements of effect size, which in turn facilitate comparison between different models (even if they are trained using different modeling algorithms). Third, this framework enables us to incorporate additional covariates as independent variables (via upgrading to MANCOVA), in a way that a more traditional classification-centric downstream task would not. Lastly, MANOVA is a well-characterized and well-established statistical method and as such has numerous useful extensions; for example, it can be combined with post-hoc Roy–Bargmann stepdown procedure (Tabachnick and Fidell, 2013) which enables detailed statistical analysis of the relationship between individual topics (or combinations of topics) and our independent variable, thereby facilitating a far richer quantitative interpretation of our topic model's output than other methods. Note, however, that this would be slightly complicated under our protocol due to our use of ILR, which results in the loss of a dimension into a new feature space that is decoupled from the original topics learned by the model (but which preserves important semantic properties of the original feature space). In this work, we explore only the first two points mentioned, leaving the rest for future work.

Effect size

MANOVA is commonly followed up by calculating effect size with partial eta-squared (η^2). Partial η^2 tell us what proportion of variance of the linear combination of the topics can be explained by the independent variable (Tabachnick and Fidell, 2013). In other words, the effect size tells us the magnitude of the effect the independent variable has on the dependent variables.

5.3 Novel statistical approach

5.3.1 Summary

In summary, our statistical approach is performed as follows:

1. Given a corpus of M documents, fit LDA for k topics.
2. Transform the resulting $M \times k$ document-topic matrix (θ) using the ILR transformation (ILR: $S^k \to \mathbb{R}^{k-1}$).
3. Verify that all assumptions for MANOVA are satisfied (described in Section 5.2.3).
4. Perform MANOVA, with the grouping variable (e.g., topic label, diagnosis, author, etc.) as the independent variable and the topic distribution probabilities in θ as the dependent variables.
5. Calculate effect size using partial η^2.

We have also included an illustrated example of our approach as shown in Figure 5.1. In this figure, there are a total of twelve documents divided into two groups, each with six documents. The two groups represent our discrete grouping variable (e.g., topic label, diagnosis, author, etc.) for the documents. A value of $k = 5$ is used due to size constraints.

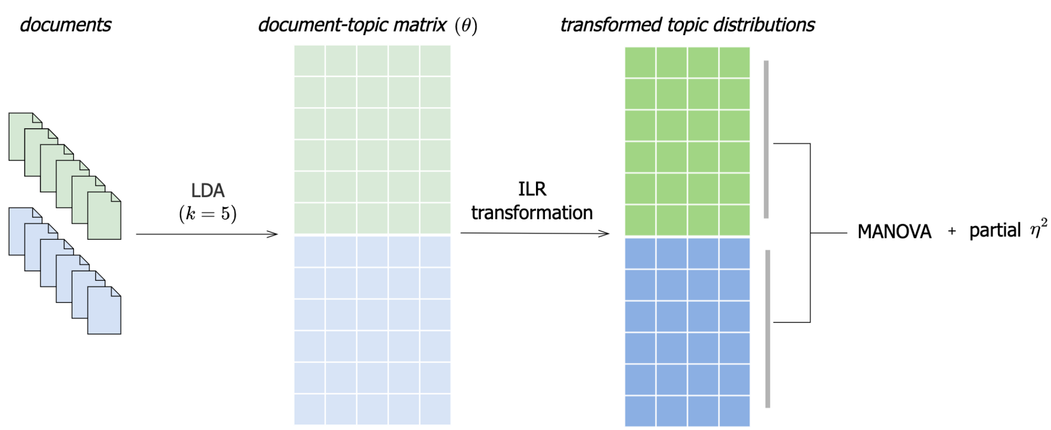

Figure 5.1: Example workflow for the described statistical approach (using $k = 5$) to quantify group differences in topic distributions captured by topic models.

5.4 *20Newsgroup* corpus

5.4.1 Data

The *20Newsgroups* corpus is a collection of approximately 18,000 posts from twenty different Usenet newsgroups,[2] and is a classic and widely-used dataset for text classification and analysis (Mitchell, 1997). We used the version of the *20Newsgroups* corpus that is available through the Python library `scikit-learn` (Pedregosa et al., 2011). Of the twenty topics available, we used documents from the following four topics: *comp.sys.ibm.pc.hardware*, *comp.sys.mac.hardware*, *rec.sport.baseball*, and *rec.sport.hockey*. During preprocessing, we omitted any documents that contained 500 characters or less. All utterances were tokenized, converted to lowercase, and lemmatized (e.g., *troubling* and *troubles* both become *trouble*). Using the lexicon of stop words provided in the `tidytext` package (Silge and Robinson, 2016), we removed all stop words that appeared in the documents. We also removed the following fillers: *uhhuh*, *mmhmm*, *hmm*, *ah*, *yep*, *wow*, *huh*, *mm*, and *alright*.

All analyses were completed using the statistical programming language R (R Core Team, 2020). LDA models were estimated using the the `topicmodels` package (Grün and Hornik, 2011). The ILR transformation was performed using the `compositions` package (van den Boogaart et al., 2023). Box's M Test was performed using the `heplots` package (Friendly et al., 2022) and partial η^2 was calculated using the `effectsize` package (Ben-Shachar et al., 2020). Our code for the *20Newsgroup* analysis was made available part of our previous publication of this work (Lawley et al., 2023b) and is publicly available on GitHub.[3]

5.4.2 Statistical plan

Using the documents from four different topics, we fit a single LDA model with a k value of 20. After transforming the topic distribution vectors using the ILR transformation, we performed seven MANOVA tests. For each of these MANOVA tests, the independent variable is the topic label and the dependent variables are the topic probability values from the document-topic vectors. The null hypothesis is that the multivariate means of the categories are equal. We will now describe the different comparisons performed.

First, we compared the topic distributions between the broader *comp.sys.** and *rec.sport.** categories. Our hypothesis was that the topic distributions between these broader categories would be very different. The *comp.sys.** category was composed of documents from the *comp.sys.ibm.pc.hardware*

[2] Usenet was an early internet-based network of hierarchically-organized discussion groups where users could post messages about a given topic.

[3] https://github.com/gracelawley/lawley-etal-sigdial-2023

and *comp.sys.mac.hardware* categories. The *rec.sport.** category was composed of documents from the *rec.sport.baseball* and *rec.sport.hockey*. We performed one MANOVA test for this first comparison:

1. *comp.sys.** vs. *rec.sport.**

Second, we compared topic distributions between subcategories with our hypothesis being that these groups will also be different, but not as different as the previous comparison (as captured by effect size). We performed two MANOVA tests for this comparison:

2. *comp.sys.ibm.pc.hardware* vs. *comp.sys.mac.hardware*
3. *rec.sport.baseball* vs. *rec.sport.hockey*

Third, we compared the topic distributions within each of the four topics by randomly splitting each topic into two. Each category were split into two by randomly assigning each document to 1 or 2. By splitting the documents in each category this way, the size of the resulting subcategories may not be the same. We hypothesized that since the documents are originally from the same category, we hypothesized that there will be no difference between the topic distributions. We performed a total of four MANOVA tests for this final set of comparisons:

4. *comp.sys.ibm.pc.hardware.1* vs. *comp.sys.ibm.pc.hardware.2*
5. *comp.sys.mac.hardware.1* vs. *comp.sys.mac.hardware.2*
6. *rec.sport.baseball.1* vs. *rec.sport.baseball.2*
7. *rec.sport.hockey.1* vs. *rec.sport.hockey.2*

MANOVA assumptions

Before proceeding further with MANOVA, there are multiple assumptions that must be met (Tabachnick and Fidell, 2013). First, each combination of independent and dependent variables should be multivariate normally distributed. Since there are more than 20 observations for each dependent $\times$ independent variable combination the multivariate central limit theorem holds and so we can assume the multivariate normality assumption also holds.

Second, dependent variables should have a linear relationship with each group of the independent variable. This assumption was initially not met since each topic distribution vector summed to 1. However after performing the ILR transformation described in Section 5.2.2, this is no longer the case.

Third, variance-covariance matrices for dependent variables should be equal across groups. This can be tested using Box's M test (Box, 1949), which tests the null hypothesis that the matrices are equal. For our data, Box's M test yielded p-values of $p < 0.001$ for each topic for the *20Newsgroups* documents and also for each conversation activity for both child and examiner speech, and thus this assumption (of equal covariance matrices) was not met. However, MANOVA is robust to unequal covariance matrices when Pillai's criterion is used (Tabachnick and Fidell, 2013; Pillai, 1955), and as such we are able to proceed .

Lastly, there should be no extreme outliers in the dependent variables. Extreme outliers can be identified by calculating the Mahalanobis distance for each observation and then performing a chi-squared test (using $df = k - 1$) to calculate the corresponding p-values. The null hypothesis is that the observation is not an outlier. We repeated analyses with identified outliers excluded and saw no difference in results. The results presented here are with these outliers included.

5.4.3 Results

Our first objective in this chapter was to demonstrate the application of our approach on the *20Newsgroup* corpus, a popular corpus for topic modeling. The results for the MANOVA tests are reported in Table 5.1. There was a significant difference between the topic distributions from the *comp.sys.** and *rec.sport.** categories, $F(19, 1710) = 414.240$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.82$. Between the *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware* subcategories, topic distributions were significantly different, $F(19, 795) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.26$. Topic distributions were also significantly different between the *rec.sport.baseball* and *rec.sport.hockey* subcategories, $F(19, 895) = 15.008$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.57$. When comparing topic distributions within each topic (by randomly splitting the documents into two groups), there were no significant differences found.

5.4.4 Discussion

The results for all of the topic distribution comparisons on the *20Newsgroups* documents were in line with our hypotheses. The topic distributions were significantly different between documents from broader categories (*comp.sys.** vs. *rec.sport.**) with the largest effect size (partial $\eta^2 = 0.82$) out of all the comparisons. In addition, the topic distributions were significantly different between documents from related but distinct subcategories: (*comp.sys.ibm.pc.hardware* vs. *comp.sys.mac.hardware*; *rec.sport.baseball* vs. *rec.sport.hockey*). The effect sizes for the subcategory comparisons were lower than the effect size for the broader category comparison, with the

Topic	n	df	Pillai	approx. F	df_1	df_2	p	partial η^2
*comp.sys.**	815	1	0.822	414.240	19	1710	<0.001	0.82
*rec.sport.**	915							
comp.sys.ibm.pc.hardware	447	1	0.264	15.008	19	795	<0.001	0.26
comp.sys.mac.hardware	368							
rec.sport.baseball	423	1	0.571	62.722	19	895	<0.001	0.57
rec.sport.hockey	492							
comp.sys.ibm.pc.hardware.1	219	1	0.020	0.460	19	427	0.976	0.02
comp.sys.ibm.pc.hardware.2	228							
comp.sys.mac.hardware.1	198	1	0.044	0.840	19	348	0.659	0.04
comp.sys.mac.hardware.2	170							
rec.sport.baseball.1	206	1	0.041	0.903	19	403	0.579	0.04
rec.sport.baseball.2	217							
rec.sport.hockey.1	247	1	0.029	0.738	19	472	0.780	0.03
rec.sport.hockey.2	245							

Table 5.1: *20Newsgroups*, comparison of LDA topic distribution vectors between and within topics.

comp.sys.ibm.pc.hardware vs.*comp.sys.mac.hardware* comparison having a smaller effect size (partial $\eta^2 = 0.26$) than the *rec.sport.baseball* vs. *rec.sport.hockey* comparison (partial $\eta^2 = 0.57$). Recall from earlier that partial η^2 tells us what proportion of variance of the linear combination of topics can be explained by the independent variable (i.e., category). In other words, it measures how much the independent variable effects the dependent variables (i.e., topic distributions). We can interpret the *rec.sport.baseball* vs. *rec.sport.hockey* comparison having a larger effect size than *comp.sys.ibm.pc.hardware* vs.*comp.sys.mac.hardware* to mean that the baseball and hockey documents are more different from each other than the PC hardware and Macintosh hardware documents. This result makes sense when considering that the terminology associated with PC hardware and Macintosh hardware overlaps quite a bit (e.g., CPU, RAM, motherboard, graphics card, etc.). On other hand, the terminology for baseball and hockey do not overlap as much (e.g., baseball: ball, bat, single, double, pitcher, catcher, first baseman, etc.; hockey: puck, stick, goal, center, goalie, right winger, left winger, etc.).[4] Our final set of comparisons for the documents from the *20Newsgroups* corpus were the comparisons between documents that were from the same category. For all four of these comparisons, there was no significant difference in topic distributions and the accompanying partial η^2 values were all close to zero.

5.5 Clinical corpus

5.5.1 Data

Our clinical corpus for this analysis consisted of transcribed ADOS, Module 3 sessions for 117 autistic children (98 male) and 65 TD children (37 male) between the ages of 4 to 15 years old. Participants were originally recruited for two separate studies conducted at Oregon Health & Science University. All participants were native English speakers, had fluent speech, and had an IQ $\geq$ 70. Intellectual ability was estimated using the the Wechsler Preschool and Primary Scale of Intelligence – Third Edition (WPPSI-III; Wechsler, 2002) for participants younger than 7 years old or a short form of the Wechsler Intelligence Scale for Children – Fourth Edition (WISC-IV; Wechsler, 2003) for participants ages 7 and older.

The language samples analyzed consisted of four conversation activities from the ADOS-2, Module 3:

1. *Emotions*

2. *Social Difficulties and Annoyance*

3. *Friends, Relationships, and Marriage*

4. *Loneliness*

Transcription was done by a team of trained transcribers in our group. All transcribers were blind to the participants' diagnostic status and intellectual abilities. Sessions were transcribed according to modified Systematic Analysis of Language Transcripts (SALT) guidelines (Miller and Iglesias, 2012). Both child and examiner utterances were transcribed, with child utterances prefixed with 'C:' and examiner utterances prefixed with 'E:'. Sessions were manually partitioned into activities after transcription according to established guidelines in order to ensure consistency across transcripts. Additional information on the participants, language samples, and transcription process can be found in Chapter 3.

[4]The author would like to note that her knowledge of baseball and hockey is limited and that these similarities and differences in terminology were confirmed by her colleagues.

In preparation for the analysis described in this chapter, for each transcript, all utterances were tokenized, converted to lowercase, and lemmatized (e.g., *troubling* and *troubles* both become *trouble*). During this process, all punctuation was removed. Using the lexicon of stop words provided in the `tidytext` package (Silge and Robinson, 2016), we removed all stop words that appeared in the documents. We also removed the following fillers: *uhhuh*, *mmhmm*, *hmm*, *ah*, *yep*, *wow*, *huh*, *mm*, and *alright*.

As before, all analyses were completed using the statistical programming language R (R Core Team, 2020). LDA models were estimated using the the `topicmodels` package (Grün and Hornik, 2011). The ILR transformation was performed using the `compositions` package (van den Boogaart et al., 2023). Box's M Test was performed using the `heplots` package (Friendly et al., 2022) and partial η^2 was calculated using the `effectsize` package (Ben-Shachar et al., 2020).

5.5.2 Statistical plan

Since our plan involved analyzing the child and examiner speech separately, we created two separate LDA models: one containing only the child speech and one containing only the examiner speech. In both models, we define a document as all words said by a speaker during a single ADOS conversation activity. Each child-examiner conversation is associated with four, distinct documents: one for each of the four activity types (*Emotions*; *Social*; *Friends*; *Marriage*). As LDA does not require that documents be independent from one another, this does not cause any issues. We used a k value of 20 for both the child-speech and examiner-speech models. This decision was informed by prior knowledge of the type and quantity of questions the examiners are instructed to ask during the ADOS conversation activities. Hyperparameter estimation was done using the variational expectation-maximization (VEM) algorithm with a starting α value of $50/k$ (Grün and Hornik, 2011; Griffiths and Steyvers, 2004).

After transforming the resulting topic distribution vectors, we performed a total of eight MANOVA tests.

1. Child-speech, *Emotions*, ASD vs. TD
2. Child-speech, *Social*, ASD vs. TD
3. Child-speech, *Friends*, ASD vs. TD
4. Child-speech, *Marriage*, ASD vs. TD
5. Examiner-speech, *Emotions*, ASD vs. TD

6. Examiner-speech, *Social*, ASD vs. TD
7. Examiner-speech, *Friends*, ASD vs. TD
8. Examiner-speech, *Marriage*, ASD vs. TD

For each of these MANOVA tests, the independent variable is diagnosis (either ASD or TD) and the dependent variables are the topic probability values from the document-topic vectors. Since we used a k of 20 in our analysis and one dimension was lost during the ILR transformation there were 19 dependent variables. The null hypothesis is that the multivariate means of the ASD and TD groups are equal.

MANOVA assumptions

For the first assumption of multivariate normality, sine there are more than 20 observations for each dependent $\times$ independent variable combination the multivariate central limit theorem holds and so we can assume the multivariate normality assumption also holds. The second assumption is met due to the ILR transformation. For the third assumption, Box's M test yielded p-values of $p < 0.001$ each conversation activity for both child and examiner speech, and thus this assumption (of equal covariance matrices) was not met. However, as previously stated, MANOVA is robust to unequal covariance matrices when Pillai's criterion is used (Tabachnick and Fidell, 2013; Pillai, 1955), and as such we are able to proceed . For the fourth, we repeated analyses with identified outliers excluded and saw no difference in results. The results presented here are with these outliers included.

5.5.3 Results

Our second objective involved comparing the topic distributions between diagnostic groups (ASD; TD) for the child-speech and the examiner-speech. We will first present the results for child speech.

Child speech

The results of the MANOVA tests for each ADOS conversation activity for child speech are reported in Table 5.2. The children's topic distributions were significantly different between the autistic and TD children within the *Social Difficulties and Annoyance* activity, $F(19, 169) = 2.055$, $p = 0.0083$, with a large effect size, partial $\eta^2 = 0.19$. There was no significant group difference in topic distributions within the other three conversation activities (*Emotions*; *Friends, Relationships, and Marriage*; *Loneliness*). To address potential Type I error from multiple comparisons, p-values can be evaluated using a Bonferroni adjusted α of 0.0125. When evaluating the results

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.093	0.941	19	175	0.5334	0.09
Social	dx	1	0.188	2.055	19	169	0.0083	0.19
Friends	dx	1	0.131	1.388	19	175	0.1381	0.13
Loneliness	dx	1	0.135	1.275	19	156	0.207	0.13

Table 5.2: Child speech, comparison of LDA topic distribution vectors between ASD and TD groups.

using the adjusted α of 0.0125, the significant result within the *Social Difficulties and Annoyance* conversation activity remains.

Examiner

The results of the statistical analyses performed on the examiner speech are reported in Table 5.3. The examiners' topic distributions differed significantly between ASD and TD groups within three

		df	Pillai	approx. F	df_1	df_2	p	partial η^2
Emotions	dx	1	0.195	2.235	19	175	0.0035	0.20
Social	dx	1	0.296	3.858	19	174	<0.001	0.30
Friends	dx	1	0.165	1.833	19	176	0.0224	0.17
Loneliness	dx	1	0.151	1.557	19	167	0.0726	0.15

Table 5.3: Examiner speech, comparison of LDA topic distribution vectors between ASD and TD groups.

of the four conversation activities examined: *Emotions*, $F(19, 175) = 2.235$, $p = 0.0035$, with a large effect size, partial $\eta^2 = 0.20$; *Social Difficulties and Annoyance*, $F(19, 174) = 3.858$, $p < 0.001$, with a large effect size, partial $\eta^2 = 0.30$; *Friends, Relationships, and Marriage*, $F(19, 176) = 1.833$, $p = 0.0224$, with a large effect size, partial $\eta^2 = 0.17$. There was no significant difference between groups for the *Loneliness* conversation activity. A Bonferroni adjusted α of 0.0125 can be used to address potential Type I error from multiple comparisons. With this adjusted α, a significant group difference within the *Emotions* and *Social Difficulties and Annoyance* activities remains; however, the previous group difference within *Friends, Relationships, and Marriage* is no longer significant.

5.5.4 Discussion

The autistic children and TD children had significantly different topic distributions for one of the four conversation analyzed: *Social Difficulties and Annoyance*. We expected to observe a group difference in all four of the conversation activities instead of only one. Incorporating additional participant-level information such as IQ and age or examining other measures of conversational reciprocity such as the length and complexity of utterances may help shed some light as to why a group difference was only seen in one of the four activities analyzed. In addition, further investigation into sampling context differences between the conversation activities is needed before conclusions can be drawn. This finding illustrates the value of our proposed statistical approach, in that we have numerous ways we could incorporate these additional covariates into our analysis in quantitatively useful ways within the same statistical framework.

The examiners' topic distributions differed significantly between the ASD and TD groups for two of the four activities: *Emotions*; *Social Difficulties and Annoyance*. This is surprising as our initial hypothesis was there would not be any significant group differences for the examiners' topic distributions. ADOS examiners are instructed to cover the same questions for each child, regardless of diagnosis, and are trained to a high standard of consistency and repeatability, as the assessment is meant for clinical use. Since one goal of the conversation activities is to foster a dialogue, the examiner would likely avoid actions that could discourage the child from conversing and sharing their interests. It may be the case that the examiners are mirroring the topics introduced by the children during the activities and those topics are being picked up by the topic distributions created by LDA. Examiners have been found to adjust their conversational patterns when speaking to patients with other cognitive conditions, such as Alzheimer's disease (Nasreen et al., 2021). Perhaps a similar adjustment our results are due to a similar adjustment occurring.

5.6 Conclusion

There are a few points about the statistical approach outlined in this chapter we would like to highlight. Although we demonstrate this method using the document-topic distribution matrix created by LDA, this method can be extended to any topic modeling algorithm that outputs a topic distribution that can be treated as a composition. We decided to use LDA here as it is a well-established technique that has been extended and built upon many times over since it was first introduced by Blei et al. (2003). Another important point to highlight is that, although not shown in here, this analysis has the potential to be extended further with a post-hoc Roy-Bargmann step down procedure to explore how much each topic (or combination of topics) contributes to the significant effect of the independent variable (Tabachnick and Fidell, 2013). However, as previously mentioned, the loss of a dimension during the ILR transformation would need to be addressed first. Overall, the statistical approach presented in this chapter represents a very promising direction for methods of making topic models more interpretable in a quantitative way, beyond human inspection of topics.

As the application of topic modeling methods continues to grow into areas such as clinical and behavioral research, so does the need for statistically based methods for evaluation and comparison. Our hope is that the statistical approach described in this chapter contributes to bridging that gap by focusing on improving evaluation metrics for existing topic modeling methods.

5.6.1 Limitations

There are several limitations of this analysis that should be mentioned. First, the decision to set k to 20 was specific to the particular clinical discourse corpus used. Our decision was informed by of the type and quantity of questions the examiners are instructed to ask during the ADOS conversation activities; however, it may not always be possible to choose a value for k using existing knowledge of the corpus. Second, as mentioned in Section 5.2.2, after performing the ILR transformation we lose one dimension from our original topic model's output and go from k to $k-1$ elements in each vector. A consequence of this is that there is no direct mapping between dimensions of the ILR-transformed $\mathbb{R}^{k-1}$ vector and the original k topics after the transformation, though the new dimensions retain the information contained in the original data (as shown by their ability to be used via MANOVA). Depending on the nature of the analysis that one is conducting, this may or may not be an issue; it was not during the present analysis, since we were interested in the overall topic distributions of each document (rather than in specific document-topic associations) but this may not always be the case. A possible direction for future work would be to draw further upon statistical methods from compositional spaces to assist with this issue.

5.7 Summary

In this chapter, we presented a novel application of existing statistical methods to evaluate the document-topic distribution vectors created by topic models in order to investigate group differences. By treating the document-topic distribution vectors as compositional data (Aitchison, 1982), we are able to use the ILR transformation (Egozcue et al., 2003) to map the vectors from their original sample sample, the k-part simplex, into the $k-1$ Euclidean space (ILR: $S^k \to \mathbb{R}^{k-1}$). Once in $\mathbb{R}^{k-1}$, we are able to use classical multivariate analysis tools such as MANOVA (Egozcue et al., 2003).

When applied to an LDA model fitted to the *20Newsgroups* corpus, our method successfully identified that the topic distributions for documents from different categories (computer hardware vs. sports) and also documents from related subcategories (PC hardware vs. Macintosh hardware; baseball vs. hockey) were significantly different. The effect size, measured with partial η^2, also varied across these comparisons, with the effect size being the largest when comparing computer hardware vs. sports and smallest when comparing Macintosh vs. PC hardware. Furthermore, our method did not find that topic distributions are significantly different when comparing groups of documents from the same category.

We also demonstrate the application of this method using LDA and a corpus of child-examiner dialogues of autistic and TD children, where prior clinical research gave us reason to expect to find group differences. We found that the topic distributions of autistic and TD children were significantly different during one of the four ADOS conversation activities examined. This result aligns with prior clinical research that autistic children often have difficulties with topic maintenance in a conversational context. Interestingly, we also found that examiners' topic distributions were significantly different whether they were conversing with an autistic child or a TD child for two of the four ADOS conversation activities examined. This may indicate that although the examiners are trained to ask the same set of questions irrespective of diagnosis status, tangential topics introduced by the child during the conversation may be mirrored by the examiner and thus are reflected in the associated topic distributions.

Chapter 6

Backchanneling Patterns

Providing feedback to your conversational partner in the form of backchanneling is a pervasive component of verbal communication. A backchannel is a short utterance (e.g., *mmhmm*, *yes*, *uhhuh*) said by person A while person B continues to have the floor (Levinson and Torreira, 2015). Although these brief utterances do not contribute new meaning to the dialogue, they still contribute important pragmatic information; by using a backchannel, a person is signaling that they are engaged and following along but that they also understand the other person is not ready to yield the floor. Backchannels sometimes (but not always) overlap other utterances (Levinson and Torreira, 2015). Deficits in backchanneling ability could lead to miscommunications or problems related to turn-taking. An extended pause before a backchannel could cause the backchannel to be interpreted as negative rather than positive (e.g., an excessive pause before saying *okay*). Starting a backchannel too close to the end of the other speaker's utterance could be interpreted as an attempt to take the floor (Schegloff, 2000).

Difficulties with social communication is one of the key characteristics of Autism Spectrum Disorder (ASD). One way that these communication difficulties could manifest is in difference in backchanneling usage. To the best of our knowledge, very limited work has been done examining backchanneling usage autistic individuals. Prior work by our group found that autistic children backchannel differently and less often overall than their Typically Developing (TD) peers (Heeman et al., 2010; Lunsford et al., 2012). However, both studies have been limited by small sample size, few female participants, and lack of controlling for participant-level variables such as age, sex, and intellectual ability. Furthermore, to our knowledge, previous work has not examined backchanneling in combination with overlap length. A study by Wehrle et al. (2024) on small sample of German speaking adults (14 with ASD and 14 without ASD) found that autistic adults used backchannels at a significant lower rate non-autistic adults.

The contents of this chapter was previously published as a short paper at the Connecting Multiple Disciplines to AI Techniques in Interaction-centric Autism Research and Diagnosis (ICARD) workshop (Lawley et al., 2023a).

In this chapter, we investigate whether autistic children use backchannels at different rates their TD peers using a multivariate approach that allows us to control for potential confounding participant-level variables such as age, sex, and IQ. Since autistic children frequently have difficulties with pragmatic language skills, we hypothesize that the ASD group will use less backchannels overall than the TD group. We also investigate whether group differences in backchannel rates are affected by whether the backchannel is an overlapping utterance and the length of the overlap (if any). Assuming that producing an overlapping-backchannel requires better turn-taking abilities than producing a backchannel that does not overlap, we hypothesize that the ASD group will produce less overlapping-backchannels and the ones they do produce will have a shorter overlap length.

6.1 Objectives

The specific objectives of the analyses presented in this chapter were to:

1. Compare overall usage rates of backchannel and overlapping-backchannels between diagnostic group differences (ASD; TD) without adjusting for differences in age, sex, and IQ.

2. Using mixed effects logistic regression models, investigate whether group differences (ASD; TD) in backchannel rates are robust to participant-level age, sex, and IQ as well as utterance-level overlap length (if any).

6.2 Methods

6.2.1 Data

The analyses in this chapter were performed on transcribed Autism Diagnostic Observation Schedule (ADOS), Module 3 sessions for 117 autistic children (98 male) and 65 TD children (37 male) between 4 to 15 years old. This sample is comprised of participants from two larger studies, both conducted at Oregon Health & Science University. All participants were native English speakers, had an IQ ≥ 70, and had fluent speech. Additional information regarding the participants and the original studies can be found in Chapter 3.

Language samples

A full description of the transcription process for the language samples were described in detail in Section 3.2; only the information relevant to the analyses in this chapter is included below. Four

ADOS activities were selected for this analysis: *Emotions*; *Social Difficulties and Annoyance*; *Friends, Relationships and Marriage*; *Loneliness*. These activities were chosen because of their conversational structure and similarities to naturalistic dialogue. Before analyses, all transcribed utterances were converted to lowercase and all punctuation was removed.

6.2.2 Backchannels

For each child, we calculated the total number of utterances that were backchannels. We considered an utterance to be a backchannel if it:

1. Appeared in the following, predefined list: *mmhmm*, *yes*, *ok*, *uhhuh*, *right*, *yeah*, *yep*.

2. Was not the first utterance of the transcript.

3. Did not follow a question (i.e., its predecessor utterance, as defined in Section 6.2.3, was not a question).

Creation of the predefined list of backchannels was informed by spelling conventions followed by our transcription team for words that commonly occur in natural conversation. These spelling conventions were strictly followed during transcription. We omitted utterances that immediately followed a question to avoid catching instances where words such as *yes* and *yeah* were used as an affirmative reply to a question. There was a total of 32,235 utterances overall, of which, 1,189 were backchannels.

There were a total of 1,837 utterances that satisfied criteria (1) and (2) but not (3). To test the validity of our rule of omitting these utterances, we took a random sample of 200 utterances from this subset (100 for each diagnostic group) and manually checked each. Of the 200 random utterances, 2 were false positives (i.e., were backchannels) and 198 were true negatives (i.e., were not backchannels), giving us a false positive rate of 0.01.

6.2.3 Overlap length

For a given utterance, we defined the overlap length as the amount (in seconds) that it overlaps with its predecessor utterance. Before we proceed with calculating the overlap length we will first explain predecessor utterances. Following Lunsford et al. (2016), we defined the predecessor of a given utterance as follows:

> Given an utterance u said by speaker A, let u' be the previous utterance said by the same speaker. Let w be the most recent utterance said by speaker B, such that the

start time of $w <$ the start time of u. Whichever of u' and w has the later end time is the predecessor of u.

When following this definition, not every utterance is a predecessor utterance and a single utterance can be the predecessor for multiple utterances. Additionally, the initial utterance in a transcript will not have a predecessor. An example of a dialogue with all predecessor utterances marked is shown in Figure 6.1. In this figure, the arrows point towards the predecessor of a given utterance.

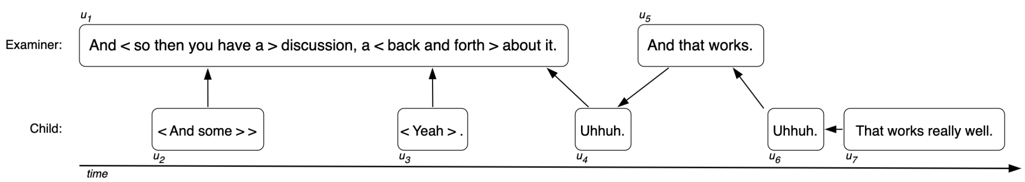

Figure 6.1: Example of predecessor utterances.

To find the predecessor of the utterance u_7, for example, our two potential options are u_6 and u_5. Both u_6 and u_5 have start times earlier than the start time of u_7, with u_6 being the most recent utterance said by the same speaker (i.e., Child) and u_5 being the most recent utterance said by the other speaker (i.e., Examiner). Since u_6 has a later end time than u_5, u_6 is the predecessor of u_7.

After identifying the predecessors for each utterance, we can proceed with calculating the overlap length (if any) of each utterance and its predecessor. We first subtracted the end time of the predecessor from the start time of the utterance. If the resulting value is positive (i.e., a pause or gap), the overlap length is 0. If instead the resulting value is negative (i.e., an overlap), the overlap length is the absolute value of this number.

6.2.4 Overlapping-backchannels

We defined an overlapping-backchannel as an utterance that is:

1. A backchannel (as defined earlier in Section 6.2.2).

2. Overlaps its predecessor utterance by longer than 0 ms.

Note that according to this definition, overlapping-backchannels are a subset of backchannels. As the sessions were recorded using a single audio channel, our options for identifying points where overlapping speech began and ended were limited. During the transcription process, the transcribers manually marked the start and end of every overlap in the transcription software used. We were able to use the timestamps created during this process to identify the overlap boundaries.

In an attempt to investigate whether the overlap could be attributed to a reaction time delay – that is, the time it takes for a person to process speech – we grouped the identified overlapping-backchannels into four categories:

- length of overlap > 0 ms
- length of overlap ≥ 100 ms
- length of overlap ≥ 200 ms
- length of overlap ≥ 300 ms

We chose these amounts in particular since 200 ms is thought to be how long it takes for a person to process speech (Fry, 1975; Levinson and Torreira, 2015).

6.2.5 Statistical analysis

We first compared the rates of both backchannels (total backchannels / total utterances) and overlapping-backchannels (total overlapping-backchannels / total utterances) between the ASD and TD groups without incorporating additional participant-level variables. Since normality assumptions were not met (Shapiro-Wilk Normality test; $p < 0.001$), we used the nonparametric Wilcoxon-Mann-Whitney test to compare groups. The dependent variable was backchannel or overlapping-backchannel rate and the independent variable was diagnosis (ASD; TD). After completing the Wilcoxon-Mann-Whitney tests, we calculated effect sizes using rank-biserial correlations to determine the magnitude of the effect of diagnostic group, if any. As in prior chapters, we used the following ranges to interpret the resulting value: small = 0.10 - 0.29, medium = 0.30 - 0.49; large = 0.50 - 1.0 (Cureton, 1956; Wendt, 1972).

Next, to investigate group differences in backchannel rates while also taking into account the participants' age, sex, and IQ as well as overlap length, we fit a mixed effects logistic regression model. The binary response variable was created as follows: with the data formatted as one utterance per row, each backchannel was coded as 1 and every other utterance was coded as 0. A per-participant random intercept was included since each participant was associated with multiple utterances. The primary predictor variable was diagnosis (ASD; TD). The other predictor variables included were participants' age, sex, and IQ and the utterance overlap length. Additionally, an interaction term between diagnosis and overlap length was included. All continuous variables were transformed into z-scores prior to model estimation.

Lastly, we repeated our second experiment but for just the overlapping-backchannels. To create the binary response variable, each overlapping-backchannel was coded as 1 and every other utterance was coded as 0. We did not include a diagnosis and overlap length interaction term in this model since the results of analysis of variance (ANOVA) showed that the inclusion of an interaction term did not significantly contribute to the model.

All analyses were performed using the statistical programming language R (R Core Team, 2020). The `lme4` package (Bates et al., 2015) was used to create the mixed effects logistic regression models.

6.3 Results

Figure 6.2 shows the distribution of both the backchannel and overlapping-backchannel rates within each diagnostic group. In this figure, the x-axis (shared by both plots) is the propor-

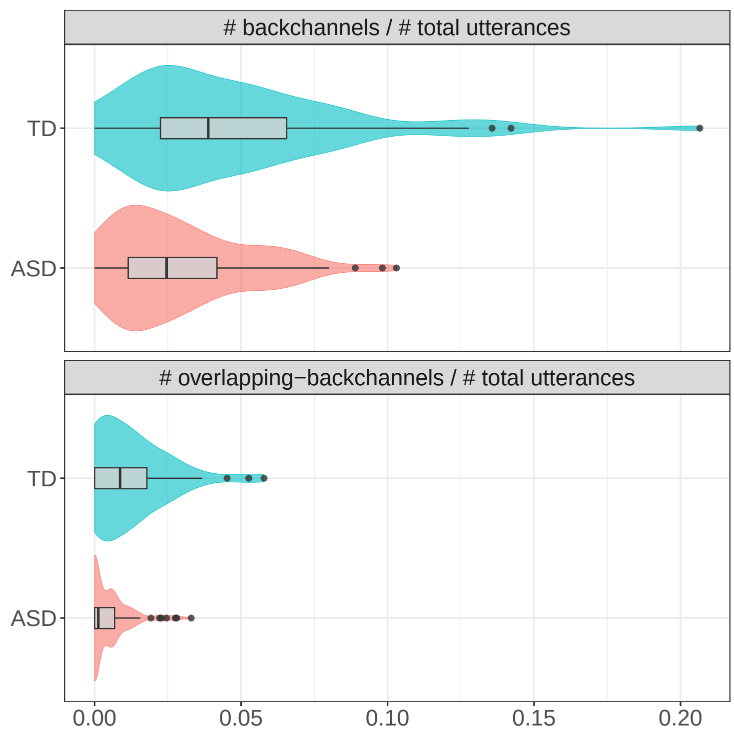

Figure 6.2: Distributions of backchannel and overlapping-backchannel rates by diagnosis.

tion of backchannels or overlapping-backchannels said by a child. Behind the boxplots are violin plots. Violin plots are mirrored kernel density plots, where wider areas correspond to a higher density of observations. The widths of the violin plots are comparable within each subplot but not overall. We can see that both the median backchannel and overlapping-backchannel rates for the ASD group are less than that of the TD group. Furthermore, the distributions of both rates for the TD group have a wider range than those of the ASD group.

The median and interquartile range (IQR) values and the results from the Wilcoxon-Mann-Whitney tests for backchannel and overlapping-backchannel rates are reported in Table 6.1. There

	ASD	TD	U	p	r_{rb}
backchannels	0.025 [0.011, 0.042]	0.039 [0.022, 0.067]	2651.5	0.001	0.303
overlapping-backchannels ($>$ 0 ms)	0.004 [0.000, 0.009]	0.011 [0.004, 0.025]	2350.0	$<$ 0.001	0.382
overlapping-backchannels ($\geq$ 100 ms)	0.004 [0.000, 0.008]	0.011 [0.004, 0.023]	2289.5	$<$ 0.001	0.398
overlapping-backchannels ($\geq$ 200 ms)	0.000 [0.000, 0.007]	0.008 [0.000, 0.018]	2291.0	$<$ 0.001	0.398
overlapping-backchannels ($\geq$ 300 ms)	0.000 [0.000, 0.007]	0.008 [0.000, 0.018]	2385.5	$<$ 0.001	0.373

Table 6.1: Backchannel and overlapping-backchannel usage rates by diagnostic group.

was a significant difference in backchannel usage between the ASD and TD groups ($p = 0.001$; small effect size: $r_{rb} = 0.303$). The ASD group used less backchannels than the TD group overall (ASD = 0.025 [0.011, 0.042] < TD = 0.039 [0.022, 0.067]). For overlapping-backchannels with an overlap length greater than 0 ms, there was also a significant group difference ($p < 0.001$; medium effect size: $r_{rb} = 0.382$), with the ASD group producing less overlapping-backchannels than the TD group (ASD = 0.004 [0.000, 0.009] < TD = 0.011 [0.004, 0.025]). This difference was still significant even when comparing rates of overlapping-backchannels with overlap lengths that were greater than or equal to 100 ms, 200 ms, and 300 ms.

Next, the results of the mixed effects logistic regression model for backchannel usage are reported in Table 6.2. A significant group difference in backchannel usage was still found after adjusting for age, sex, IQ, and overlap length ($\chi^2 = -3.164$, $P = 0.002$, Table 6.2). As before, the ASD had a lower backchannel rate than the TD group. There was no significant effect on backchannel rate of participant age, sex, or IQ. Overlap length significantly contributed to backchannel rate ($\chi^2 = 10.329$, $P < 0.001$), with overlap length increasing the likelihood that an utterance is a backchannel. There was also a significant interaction between diagnosis and overlap length ($\chi^2 = -2.420$, $P = 0.016$), with the ASD group being less likely to produce a backchannel as the overlap length increases.

Lastly, the results of the mixed effects logistic regression model for overlapping-backchannels are reported in Table 6.3. For this model, inclusion of an interaction term between diagnosis and overlap length did not significantly contribute to the model so the interaction was left out. After controlling for age, sex, IQ, and overlap length, a significant group difference in overlapping-backchannel usage remained ($\chi^2 = -4.131$, $P < 0.001$), with the ASD group again using less

Predictor	Log-odds	S.E.	χ^2	$P(\chi^2)$
(Intercept)	-3.235	0.140	—	—
DX	—	—	-3.164	0.002
ASD	-0.490	0.155	—	—
TD	0.490	—	—	—
Sex	—	—	-0.044	0.965
Male	-0.007	0.149	—	—
Female	0.007	—	—	—
Age	0.029	0.067	0.430	0.667
IQ	-0.023	0.072	-0.321	0.748
Overlap	0.227	0.022	10.329	< 0.001
DX:Overlap	—	—	-2.420	0.016
ASD:Overlap	-0.087	0.036	—	—

Table 6.2: Mixed effects logistic regression model predicting likelihood of a backchannel utterance.

backchannels than the TD group. The age, sex, and IQ of the participants had no significant effect on overlapping-backchannel rate. The overlap length significantly effected the likelihood that an utterance was an overlapping-backchannel ($\chi^2 = 19.591$, $P < 0.001$), irrespective of participant's age, sex, IQ, or diagnosis. In other words, the longer the overlap, the more likely that an utterance was an overlapping-backchannel.

6.4 Discussion

When comparing rates of backchannels and overlapping backchannels in isolation (i.e., without controlling for participant age, sex, IQ or utterance overlap length), we found that the ASD group had significantly lower rates of backchannels and overlapping-backchannels than their TD peers. These differences persisted after adjusting for difference in participant age, sex, and IQ, with the ASD group again using backchannels and overlapping-backchannels at a significantly lower rate than the TD group. Furthermore, utterances were more likely to be backchannels the more they overlapped with their corresponding predecessor utterance. The diagnostic group and overlap length interaction significantly effected the likelihood an utterance would be a backchannel, with the ASD group being less likely than the TD group to produce a backchannel with a greater overlap length.

Predictor	Log-odds	S.E.	χ^2	$P(\chi^2)$
(Intercept)	-4.625	0.203	—	—
DX	—	—	-4.131	< 0.001
ASD	-0.945	0.229	—	—
TD	0.945	—	—	—
Sex	—	—	-0.694	0.448
Male	-0.153	0.221	—	—
Female	0.153	—	—	—
Age	-0.017	0.103	-0.165	0.869
IQ	-0.001	0.112	-0.016	0.987
Overlap	0.405	0.021	19.591	< 0.001

Table 6.3: Mixed effects logistic regression model predicting likelihood of an overlapping-backchannel utterance.

These results suggest that autistic children use backchannels less than TD children and that this difference is affected by whether the backchannel overlaps and how long the overlap is. This could indicate that the TD group is more skilled at timing backchannels since they produced more overlapping utterances than the ASD group.

6.5 Summary

In this chapter, we investigated differences in backchannel usage in language samples of autistic and TD children and examined whether differences were associated with participant age, sex, IQ, or overlap length. We found that irrespective of participant age, sex, and IQ, autistic children used backchannels and overlapping-backchannels at significantly lower rates than their TD peers. Utterances were more likely to be a backchannel than not a backchannel is they overlapped their predecessor utterance. Although further investigation is needed, these results suggest that lower backchanneling may be a pragmatic language feature that could be used alongside other metrics to assess social communication ability in autistic children.

Chapter 7

Conclusion

7.1 Summary

In this dissertation, we presented our work on capturing and quantifying pragmatic language differences in the language of autistic children by using methods from the fields of Natural Language Processing (NLP) and statistics. The pragmatic language areas investigated include differences in the usage of the fillers *um* and *uh*, differences in topic maintenance ability as represented by topic distributions, and differences in the usage of backchannels such as *uhhuh*, *right*, and *okay*.

In Chapter 4, we analyzed *um* and *uh* usage in a large, well-characterized sample of autistic children and TD children and investigated whether usage patterns observed within the ASD group were associated with differences in social communication and language skills. Although the groups did not differ in *uh* usage, the ASD group used fewer *ums* than the TD group. This held true after controlling for age, sex, and IQ. Within ASD, social affect and pragmatic language scores did not predict filler usage; however, structural language scores predicted *um* usage. Lower *um* rates among children with ASD may reflect problems with planning or production rather than pragmatic language.

In Chapter 5, we presented a novel statistical approach for investigating group difference in the document-topic distribution vectors created by Latent Dirichlet Allocation (LDA). Topic distribution matrices created by topic models are typically used for document classification or as features in a separate machine learning algorithm. Existing methods for evaluating these topic distributions include metrics such as coherence and perplexity; however, there is a lack of statistically grounded evaluation tools. Our approach involves transforming topic distribution vectors using Aitchison geometry, followed by using multivariate analysis of variance (MANOVA) to compare sample means and calculate effect size using partial eta-squared. We report the results of successfully validating this method on a subset of the *20Newsgroup* corpus. When applying our approach to our clinical corpus, we found that the topic distributions of autistic children differed from those of TD children when responding to questions about social difficulties. We also applied our approach to the examiner speech in our clinical corpus. When doing so, we found that the examiners' topic distributions differed between the autistic and TD groups when discussing emotions and social difficulties. Our results support the use of topic modeling in studying clinically relevant features of social communication such as topic maintenance.

In Chapter 6, we investigated rates of backchanneling and overlapping-backchannels in the language of autistic children. Backchanneling (e.g., *right*, *okay*, *uhhuh*) during a dialogue signals that a person is engaged and following along with what is being said. Although backchannels often overlap with other utterances, they are not interpreted as an attempt to take the floor when used successfully. Limited work has been done on investigating the frequency and overlap length of backchannels in the language of autistic children. After controlling for age, sex, and IQ, we found that autistic children used significantly less backchannels than their TD peers. Additionally, we found that autistic children were less likely than TD children to produce a backchannel with a greater overlap length.

7.2 Applications and future work

Using NLP approaches to measure pragmatic language problems in transcribed language samples is an exciting and promising augmentation to conventional language assessments. With the use of NLP, robust measures of pragmatic language features, including *um* and *uh* usage, topic distributions, backchannels, as well as other pragmatic features, can be obtained in an automated, reliable, and relatively inexpensive fashion. We can also use statistical methods to better understand which underlying processes, if any, influence various pragmatic language differences. Creation of such measures pragmatic language features is first step towards building clinically informative automated language assessments that can be used to augment the standardized assessments used by clinicians such as the ADOS-2 and CCC-2 (Lord et al., 2000; Bishop, 2003).

Furthermore, work on measuring pragmatic language problems may provide additional context and information about the language patterns of autistic individuals. This insight can be used to improve training materials for teachers and support materials for caregivers of autistic children. Some theorize that the social communication difficulties that autistic individuals is caused in part by a communication breakdown between both sides of a conversation (Zamzow, 2021). By improving our understanding of how autistic individuals use language, we can in turn learn how to become better conversational partners ourselves. At the time of writing this dissertation, the potential impact of large language models such as GPT-4 on healthcare has been at the forefront of conversations (Gallifant et al., 2024). Every individual uses language in a unique way, and a one-size-fits-all approach when designing language models may not be best. NLP-driven pragmatic language metrics can be used to help inform the design of future language tools so that they can accommodate for pragmatic language differences.

Bibliography

Adams, C. (2002). Practitioner Review: The assessment of language pragmatics. *Journal of Child Psychology and Psychiatry*, 43(8):973–987.

Adams, J. R., Salem, A. C., MacFarlane, H., Ingham, R., Bedrick, S. D., Fombonne, E., Dolata, J. K., Hill, A. P., and van Santen, J. P. H. (2021). A Pseudo-Value Approach to Analyze the Semantic Similarity of the Speech of Children With and Without Autism Spectrum Disorder. *Frontiers in Psychology*, 12:3089.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–177.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders*. 4th ed., text rev. edition.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. 5th ed. edition.

Baltaxe, C. A. M. and D'Angiola, N. (1992). Cohesion in the discourse interaction of autistic, specifically language-impaired, and normal children. *Journal of Autism and Developmental Disorders*, 22(1):1–21.

Barokova, M. and Tager-Flusberg, H. (2020). Commentary: Measuring Language Change Through Natural Language Samples. *Journal of Autism and Developmental Disorders*, 50(7):2287–2306.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Ben-Shachar, M. S., Lüdecke, D., and Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, 5(56):2815.

Bishop, D. V. (2003). *The Children's Communication Checklist, version 2 (CCC-2)*. Pearson, London.

Bishop, D. V. (2013). Children's Communication Checklist (CCC-2). In Volkmar, F. R., editor, *Encyclopedia of Autism Spectrum Disorders*, pages 614–618. Springer New York, New York, NY.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993—1022.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4):317–346.

Boyd-Graber, J., Hu, Y., and Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Brown, L. (n.d.). Identity-First Language. Autistic Self Advocacy Network (ASAN). `https://web.archive.org/web/20230119074031/https://autisticadvocacy.org/about-asan/identity-first-language/`. Accessed: 2023-01-19.

Capps, L., Kehres, J., and Sigman, M. (1998). Conversational Abilities Among Children with Autism and Children with Developmental Delays. *Autism*, 2(4):325–344.

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

Cureton, E. E. (1956). Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

DaWalt, L. S., Usher, L. V., Greenberg, J. S., and Mailick, M. R. (2019). Friendships and social participation as markers of quality of life of adolescents and adults with fragile X syndrome and autism. *Autism*, 23(2):383–393.

de Marchena, A. and Eigsti, I.-M. (2010). Conversational gestures in autism spectrum disorders: Asynchrony but not decreased frequency. *Autism Research*, 3(6):311–322.

Dolata, J. K., Suarez, S., Calamé, B., and Fombonne, E. (2022). Pragmatic language markers of autism diagnosis and severity. *Research in Autism Spectrum Disorders*, 94:101970.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C. (2003). Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35:279–300.

Engelhardt, P. E. (2019). Speaker Versus Listener-Oriented Disfluency. In Volkmar, F. R., editor, *Encyclopedia of Autism Spectrum Disorders*, pages 1–10. Springer New York, New York, NY.

Engelhardt, P. E., Alfridijanta, O., McMullon, M. E. G., and Corley, M. (2017). Speaker-Versus Listener-Oriented Disfluency: A Re-examination of Arguments and Assumptions from Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, 47(9):2885–2898.

Finlayson, I. R. and Corley, M. (2012). Disfluency in dialogue: an intentional signal from the speaker? *Psychonomic Bulletin & Review*, 19(5):921–928.

Friedman, L., Sterling, A., DaWalt, L. S., and Mailick, M. R. (2019). Conversational Language Is a Predictor of Vocational Independence and Friendships in Adults with ASD. *Journal of Autism and Developmental Disorders*, 49(10):4294–4305.

Friendly, M., Fox, J., and Monette, G. (2022). *heplots: Visualizing Tests in Multivariate Linear Models*. R package version 1.4-2.

Fry, D. (1975). Simple Reaction-Times to Speech and Non-Speech Stimuli. *Cortex*, 11(4):355–360.

Gallifant, J., Fiske, A., Levites Strekalova, Y. A., Osorio-Valencia, J. S., Parke, R., Mwavu, R., Martinez, N., Gichoya, J. W., Ghassemi, M., Demner-Fushman, D., McCoy, L. G., Celi, L. A., and Pierce, R. (2024). Peer review of GPT-4 technical report and systems card. *PLOS Digital Health*, 3(1):1–15.

Goodkind, A., Lee, M., Martin, G. E., Losh, M., and Bicknell, K. (2018). Detecting language impairments in autism: A computational analysis of semi-structured conversations with vector semantics. In *Proceedings of the Society for Computation in Linguistics*, volume 1, pages 12–22.

Gorman, K., Olson, L., Hill, A. P., Lunsford, R., Heeman, P. A., and van Santen, J. P. H. (2016). Uh and um in children with autism spectrum disorders or language impairment. *Autism Research*, 9(8):854–865.

Gotham, K., Pickles, A., and Lord, C. (2009). Standardizing ADOS Scores for a Measure of Severity in Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 39(5):693–705.

Griffiths, T. L. and Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35.

Grün, B. and Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13):1–30.

Hagg, L. J., Merkouris, S. S., O'Dea, G. A., Francis, L. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Westrupp, E. M., Macdonald, J. A., and Youssef, G. J. (2022). Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review. *Journal of Medical Internet Research*, 24(11):e33166.

Halladay, A. K., Bishop, S., Constantino, J. N., Daniels, A. M., Koenig, K., Palmer, K., Messinger, D., Pelphrey, K., Sanders, S. J., Singer, A. T., Taylor, J. L., and Szatmari, P. (2015). Sex and gender differences in autism spectrum disorder: summarizing evidence gaps and identifying emerging areas of priority. *Molecular Autism*, 6(1):36.

Happé, F. and Frith, U. (2020). Annual Research Review: Looking back to look forward – changes in the concept of autism and implications for future research. *Journal of Child Psychology and Psychiatry*, 61(3):218–232.

Heeman, P. A., Lunsford, R., Selfridge, E., Black, L., and van Santen, J. P. H. (2010). Autism and Interactional Aspects of Dialogue. In *Proceedings of the SIGDIAL 2010 Conference*, pages 249–252, Tokyo, Japan. Association for Computational Linguistics.

Hill, A. P., Santen, J. P. H. v., Gorman, K., Langhorst, B. H., and Fombonne, E. (2015). Memory in language-impaired children with and without autism. *Journal of Neurodevelopmental Disorders*, pages 1–13.

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., and Resnik, P. (2021). Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence.

Irvine, C. A., Eigsti, I.-M., and 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.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey. *Multimedia Tools and Applications*, 78(11):15169–15211.

Kover, S. T., Davidson, M. M., Sindberg, H. A., and Weismer, S. E. (2014). Use of the ADOS for Assessing Spontaneous Expressive Language in Young Children With ASD: A Comparison of Sampling Contexts. *Journal of Speech Language and Hearing Research*, 57(6):2221–13.

Lake, J. (2008). *Listener vs. Speaker-Oriented Speech: Studying the Speech of Individuals with Autism*. PhD Thesis, McMaster University.

Lake, J., Humphreys, K. R., and Cardy, S. (2011). Listener vs. speaker-oriented aspects of speech: Studying the disfluencies of individuals with autism spectrum disorders. *Psychonomic Bulletin & Review*, 18(1):135–140.

Landa, R. (2000). Social language use in Asperger syndrome and high-functioning autism. In Klin, A. M., Volkmar, F. R., and Sparrow, S. S., editors, *Asperger syndrome.*, pages 125–155. The Guilford Press, New York, NY, US.

Lawley, G. O., Bedrick, S., Dolata, J. K., and Fombonne, E. J. (2021). Investigation on Examiner "Um" and "Uh" Usage in ADOS-2 Sessions. In *International Society for Autism Research Annual Meeting (INSAR)*.

Lawley, G. O., Bedrick, S., MacFarlane, H., Dolata, J. K., Salem, A. C., and Fombonne, E. (2022). "Um" and "Uh" Usage Patterns in Children with Autism: Associations with Measures of Structural and Pragmatic Language Ability. *Journal of autism and developmental disorders*, pages 1–12.

Lawley, G. O., Heeman, P. A., and Bedrick, S. (2023a). Computational Analysis of Backchannel Usage and Overlap Length in Autistic Children. In *Proceedings of the First Workshop on Connecting Multiple Disciplines to AI Techniques in Interaction-centric Autism Research and Diagnosis (ICARD 2023)*, pages 17–23, Prague, Czechia. Association for Computational Linguistics.

Lawley, G. O., Heeman, P. A., Dolata, J. K., Fombonne, E., and Bedrick, S. (2023b). A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 55–65, Prague, Czechia. Association for Computational Linguistics.

Levinson, S. C. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731).

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The Autism Diagnostic Observation Schedule, Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223.

Lunsford, R., Heeman, P., Black, L., and van Santen, J. P. H. (2010). Autism and the use of fillers: Differences between 'um' and 'uh'. In *Proceedings of DiSS-LPSS Joint Workshop 2010*, pages 107–110, Tokyo, Japan.

Lunsford, R., Heeman, P. A., and Rennie, E. (2016). Measuring Turn-Taking Offsets in Human-Human Dialogues. In *Proc. Interspeech 2016*, pages 2895–2899.

Lunsford, R., Heeman, P. A., and van Santen, J. P. H. (2012). Interactions Between Turn-taking Gaps, Disfluencies and Social Obligation. In *Proc. Interspeech 2012*, pages 607–610.

MacFarlane, H., Gorman, K., Ingham, R., Hill, A. P., Papadakis, K., Kiss, G., and van Santen, J. P. H. (2017). Quantitative analysis of disfluency in children with autism spectrum disorder or language impairment. *PLOS ONE*, 12(3):1–20.

MacFarlane, H., Salem, A. C., Bedrick, S., Dolata, J. K., Wiedrick, J., Lawley, G. O., Finestack, L. H., Kover, S. T., Thurman, A. J., Abbeduto, L., and Fombonne, E. (2023). Consistency and reliability of automated language measures across expressive language samples in autism. *Autism Research*, 16(4):802–816.

Maenner, M. J., Warren, Z., Williams, A. R., and et al. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR Surveillance Summaries*, 72(2):1–14.

McGregor, K. K. and Hadden, R. R. (2020). Brief Report: "Um" Fillers Distinguish Children With and Without ASD. *Journal of Autism and Developmental Disorders*, 50(5):1816–1821.

Miller, J. and Iglesias, A. (2012). *SALT: Systematic analysis of language transcripts [Research version]*. SALT Software, Middleton, WI.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Nasreen, S., Rohanian, M., Hough, J., and Purver, M. (2021). Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features. *Frontiers in Computer Science*, 3.

Parish-Morris, J., Liberman, M. Y., Cieri, C., Herrington, J. D., Yerys, B. E., Bateman, L., Donaher, J., Ferguson, E., Pandey, J., and Schultz, R. T. (2017). Linguistic camouflage in girls with autism spectrum disorder. *Molecular Autism*, 8(1):48–60.

Paul, R., Orlovski, S. M., Marcinko, H. C., and Volkmar, F. R. (2009). Conversational Behaviors in Youth with High-functioning ASD and Asperger Syndrome. *Journal of Autism and Developmental Disorders*, 39(1):115–125.

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

Pillai, K. C. S. (1955). Some New Test Criteria in Multivariate Analysis. *The Annals of Mathematical Statistics*, 26(1):117–121.

Prud'hommeaux, E., van Santen, J., and Gliner, D. (2017). Vector space models for evaluating semantic fluency in autism. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 32–37, Vancouver, Canada. Association for Computational Linguistics.

Prud'hommeaux, E. T., Roark, B., Black, L. M., and van Santen, J. (2011). Classification of Atypical Language in Autism. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 88–96, Portland, Oregon, USA. Association for Computational Linguistics.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reimers, N. and Gurevych, I. (2017). Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Rouhizadeh, M., Sproat, R., and van Santen, J. (2015). Similarity Measures for Quantifying Restrictive and Repetitive Behavior in Conversations of Autistic Children. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 117–123, Denver, Colorado. Association for Computational Linguistics.

Rutter, M., Bailey, A., and Lord, C. (2003). *The Social Communication Questionnaire (SCQ)*. Western Psychological Services, Los Angeles, CA.

Salem, A. C., MacFarlane, H., Adams, J. R., Lawley, G. O., Dolata, J. K., Bedrick, S., and Fombonne, E. (2021). Evaluating atypical language in autism using automated language measures. *Scientific Reports*, 11(1):10968.

Sattler, J. M. and Dumont, R. (2004). *Assessment of Children: WISC-IV and WPPSI-III Supplement*. La Mesa, CA: Jerome M. Sattler, Publisher, Inc.

Schegloff, E. A. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29(1):1–63.

Semel, E., Wiig, E., and Secord, W. (2003). *Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4)*. The Psychological Corporation, San Antonio, TX.

Semel, E., Wiig, E. H., and Secord, W. (2004). *Clinical Evaluation of Language Fundamentals Preschool – Second Edition (CELF-Preschool-2)*. The Psychological Corporation, San Antonio, TX.

Silge, J. and Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, 1(3).

Song, A., Cola, M., Plate, S., Petrulla, V., Yankowitz, L., Pandey, J., Schultz, R. T., and Parish-Morris, J. (2020). Natural language markers of social phenotype in girls with autism. *Journal of Child Psychology and Psychiatry*.

Tabachnick, B. G. and Fidell, L. S. (2013). *Using Multivariate Statistics*. Pearson, 6th edition.

Tager-Flusberg, H. and Caronna, E. (2007). Language Disorders: Autism and Other Pervasive Developmental Disorders. *Language, Communication, and Literacy: Pathologies and Treatments*, 54(3):469–481.

Tager-Flusberg, H. and Kasari, C. (2013). Minimally Verbal School-Aged Children with Autism Spectrum Disorder: The Neglected End of the Spectrum. *Autism Research*, 6(6):468–478.

Tager-Flusberg, H., Paul, R., and Lord, C. (2005). Language and Communication in Autism. In *Handbook of autism and pervasive developmental disorders: Diagnosis, development, neurobiology, and behavior, Vol. 1, 3rd ed.*, pages 335–364. John Wiley & Sons Inc, Hoboken, NJ, US.

van den Boogaart, K. G., Tolosana-Delgado, R., and Bren, M. (2023). *compositions: Compositional Data Analysis*. R package version 2.0-6.

van Santen, J. P. H., Sproat, R. W., and Hill, A. P. (2013). Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383.

Volden, J., Coolican, J., Garon, N., White, J., and Bryson, S. (2008). Brief Report: Pragmatic Language in Autism Spectrum Disorder: Relationships to Measures of Ability and Disability. *Journal of autism and childhood schizophrenia*, 39(2):388.

Volden, J. and Phillips, L. (2010). Measuring pragmatic language in speakers with autism spectrum disorders: Comparing the children's communication checklist–2 and the test of pragmatic language. *American journal of speech-language pathology*, 19(3):204–212.

Wechsler, D. (2002). WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence - 3rd ed. *San Antonio, TX: Psychological Corporation.*

Wechsler, D. (2003). WISC-IV: Wechsler Intelligence Scale for Children. *San Antonio, TX: Psychological Corporation.*

Wehrle, S., Vogeley, K., and Grice, M. (2024). Backchannels in conversations between autistic adults are less frequent and less diverse prosodically and lexically. *Language and Cognition*, 16(1):108—-133.

Wendt, H. W. (1972). Dealing with a common problem in social science: A simplified rank-biserial coefficient of correlation based on the U statistic. *European Journal of Social Psychology*, 2(4):463–465.

Wieling, M., Grieve, J., Bouma, G., Fruehwald, J., Coleman, J., and Liberman, M. (2016). Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change*, 6(2):199–234.

Wittke, K., Mastergeorge, A. M., Ozonoff, S., Rogers, S. J., and Naigles, L. R. (2017). Grammatical Language Impairment in Autism Spectrum Disorder: Exploring Language Phenotypes Beyond Standardized Testing. *Frontiers in Psychology*, 8:532.

Wood-Downie, H., Wong, B., Kovshoff, H., Mandy, W., Hull, L., and Hadwin, J. A. (2021). Sex/Gender Differences in Camouflaging in Children and Adolescents with Autism. *Journal of Autism and Developmental Disorders*, 51(4):1353–1364.

Young, E. C., Diehl, J. J., Morris, D., Hyman, S. L., and Bennetto, L. (2005). The Use of Two Language Tests to Identify Pragmatic Language Problems in Children With Autism Spectrum Disorders. *Language, speech, and hearing services in schools*, 36(1):62–72.

Zamzow, R. (2021). Double empathy, explained. Spectrum. `https://doi.org/10.53053/MMNL2849`.