# Computational Analysis of Language Use in Autism

Masoud Rouhizadeh

Presented to the Center for Spoken Language Understanding
within the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in
Computer Science & Engineering

October 2015

Center for Spoken Language Understanding
School of Medicine
Oregon Health & Science University

---

CERTIFICATE OF APPROVAL

---

This is to certify that the Ph.D. dissertation of
Masoud Rouhizadeh
has been approved.

---

Richard Sproat, Thesis Advisor
Research Scientist
Google, Inc.

---

Jan van Santen, Thesis Advisor
Professor
Oregon Health & Science University

---

Steven Bedrick
Assistant Professor
Oregon Health & Science University

---

Kemal Sonmez
Associate Professor
Oregon Health & Science University

---

Cecilia Ovesdotter Alm
Assistant Professor
Rochester Institute of Technology

# Acknowledgment

I would like to express my sincere gratitude to my advisors, Richard Sproat, and Jan van Santen, for their continuous support during my PhD study, for their immense knowledge, and their invaluable instruction. In addition, I would like to thank the other members in my committee: Steven Bedrick, Kemal Sonmez, and Cecilia Ovesdotter Alm for their interest in my work and their insightful discussion and comments.

My sincere thanks also goes to all of the current and former faculty at CSLU, especially Brian Roark, Alison Presmanes Hill, and Kyle Gorman for their important input and motivation. Peter Heeman and Patricia Dickerson deserve a special mention for their constant help in my educational pathway.

I cannot thank enough my fellow students, both past and present, for making CSLU a great place to work, especially Joel Adams, Meysam Asgari, Alireza Bayesteh, Russ Beckley, Nate Bodenstab, Shiran Dudy, Aaron Dunlop, Mahsa Elyasi, Andrew Fowler, Kristy Hollingshead Seitz, Rebecca Lunsford, Archana Machireddy, Meg Mitchell, Hamidreza Mohammadi, Eric Morley, Emily Prud'hommeaux, Golnar Sheikhshab, Guillaume Thibault, and Damien Zufferey.

Finally, this dissertation would not have been possible without the true love, support, and patience of: my wife, Mahsa, my mom, dad, my siblings and Mahsa's family.

This work was supported in part by NSF grant #BCS-0826654, and NIH NIDCD grants #R01-DC007129 and #1R01DC012033-01. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the NSF or the NIH.

# Contents

# List of Tables

# List of Figures

# Abstract

**Computational Analysis of Language Use in Autism**

Masoud Rouhizadeh

Doctor of Philosophy
the Center for Spoken Language Understanding within
the Oregon Health & Science University
School of Medicine

October 2015

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired social communication, and restricted, repetitive patterns of behavior and interest. These two core symptoms can appear at the language level and result in problems such as using inappropriate words, idiosyncratic language, topic repetition, and lack of conversational responsiveness. Existing methods for the evaluation of language in ASD are mainly based on subjective parental and clinical reports. In this thesis, we propose fast, objective, scalable, automatic analysis of these interrelated aspects of ASD language, utilizing computational methods for natural language processing based on unannotated verbatim transcripts of conversations. We first apply word ranking and distributional semantic models to automatically determine off-topic lexical content in children's narrative retellings. Our classification of unexpected words is sufficiently accurate to distinguish the retellings of children with autism from those with typical development (TD). Second, we utilize

semantic similarity measures to identify idiosyncratic topic digressions expressed in narratives. Our findings indicate that TD children tend to use similar words and semantic concepts when retelling the same narrative, while children with ASD use different words and concepts that are potentially related to their individual topics of interest. Third, we try to quantify restrictive and repetitive interests and topic repetition in spontaneous conversations of autistic children. Using various similarity measures, we show that the children with ASD have significantly higher ratio of semantically overlapping dialogue turns compared to their TD peers, as a result of higher topic perseveration in their conversations. Finally, we focus on social communication and interaction in children with ASD and we analyze their question responsiveness as well as the use of discourse markers and acknowledgments in various contexts. Our findings suggest that the ASD children are less responsive compared to TD children, and have problems in the appropriate use of conversational cues. Our proposed methods and results in this thesis underscore the potential of automated natural language processing techniques for improving the understanding of the prevalence and diagnostic significance of language use in ASD.

# Chapter 1

# Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by impaired communication and social behavior as well as by the presence of restrictive and repetitive behavior (RRB) (American Psychiatric Association, 2000, 2013). According to the Centers for Disease Control and Prevention (CDC), ASD affects one in 68 children in the United States as of 2014 (CDC, 2014), a dramatic increase from one in 88 in 2012 (CDC, 2012), and one in 152 in 2007 (CDC, 2007). This large increase is partially due to changes in diagnosis criteria and procedures, and increased access to diagnosis or treatment, although the exact cause is still unknown (Newschaffer et al., 2007). Nevertheless, there are likely many children with ASD conditions who remain undiagnosed due to lack of access to qualified clinical assessment. In addition, those families who are aware of their children's ASD conditions may have difficulty in accessing professional evaluation, monitoring, and treatment due to lack of financial resources, living in rural areas, or cultural and language barriers. Even if the parents have adequate access to professional help, limited-time clinical visits might not provide a "broad picture" of the language and behavioral issues of their child.

ASD is hard to diagnose. Currently there is no genetic or other physical or neuroimaging tests for the diagnosis of autism, and it is mostly diagnosed by clinical autism-specific behavioral assessments as well as parental reports (London, 2007). Although impaired language is not necessary for a diagnosis of ASD, it is nevertheless frequently observed in ASD, especially in the form of atypical pragmatic language (Bishop, 1989; Tager-Flusberg, 1999; Tager-Flusberg et al., 2005b). This is not surprising, given that atypical pragmatic

language is a plausible manifestation of a core deficit of ASD, impaired social communication. The evaluation of these characteristics is required in major ASD diagnostic instruments such as the Autism Diagnostic Observation Schedule (ADOS, Lord et al., 2002), Social Communication Questionnaire (SCQ, Rutter et al., 2003), and the Autism Diagnostic Interview-Revised (ADI-R, Lord et al., 1994).

Such assessments, however, are largely based on subjective, aggregate ratings. Objective evaluation requires analyzing natural language samples, which is not frequently done in clinical practice and research environments. A proper and comprehensive analysis of language generally involves manual coding which is labor-intensive and usually requires extensive training. Manual coding is also a subjective procedure, and it is difficult to implement a reliable and objective coding system across coders and different research groups. Moreover, coding protocols usually need to be adapted for new tasks, populations, and diagnosis criteria. Although language-based measures are useful and informative for tracking responses to language and communication intervention, the demand for human experts, time, and training, makes such measures hard to apply in research studies and clinical practice related to ASD treatment.

Automated methods from Natural Language Processing (NLP) could be of great utility in addressing these issues. First, they provide objective and quantitative evaluation of language samples, as opposed to qualitative impressions. Second, they merely rely on manual verbatim transcriptions, hence they are completely unbiased to the child's diagnosis even compared to manual analysis by the clinician. Third, once the methods are developed, they require only low-cost non-expert "raw" transcription with considerably less training than labor-intensive expert coding. Fourth, the automated methods are far more generalizable across different research and clinical sites as they provide robust and reliable analysis of language tasks. Fifth, they facilitate the analysis of larger, contextually more diverse language samples from a child, compared to short-term clinical visits. This analysis can provide a more illustrative picture of the child's language and aid in tracking his/her response to communication intervention. In addition, a large-scale analysis could result in detecting significant language behaviors (such as topic repetition) which

may not be easy to observe and capture by clinicians during administration of structured tasks.

## 1.1 Problem statement

Semantic and pragmatic language are generally atypical or impaired in autism spectrum disorder (Bishop, 1989; Tager-Flusberg, 1999; Tager-Flusberg et al., 2005b). It includes atypical language and topic digressions in narrative retellings, idiosyncratic topic perseveration in dialogue, and a lack of conversational reciprocity. Although these problems have been frequently observed in research studies on language use in ASD, their objective and quantitative evaluation is not well addressed in current autism diagnosis instruments, and their assessment is mostly limited to high-level question items, requiring impressionistic judgments from parents, teachers, and/or clinicians.

The in-depth analysis of pragmatic language use in clinical and research settings requires expert-level manual annotation. Such annotations are time-consuming, tedious, and prone to human error. In addition, there is little or no standard definition for many of these problems in the literature, and most researchers have designed their own guidelines for the annotation and characterization these phenomena.

Computational methods for natural language processing, and in particular distributional semantic models, could provide automated, fast, objective, large-scale, reliable, standard measurements for the analysis of semantic and pragmatic aspects of language. However, as we will see in later sections of this thesis, little or no previous work has addressed the automated analysis of these phenomena in autism.

## 1.2 Research objectives

In this thesis, we propose, implement, and evaluate the significance of automated measures for the analysis of pragmatic language use in autism. In particular, we focus on determining off-topic words and topic shifts in narrative retellings, verbal perseveration,

and lack of conversational responsiveness. More precisely, the six major objectives of the thesis are the following:

**Atypical language as a behavioral marker of autism.** We will demonstrate –by manual annotation– that atypical language and, in particular, idiosyncratic lexical items are significantly more frequent in narratives produced by children with ASD.

**Word ranking and semantic expansion for detecting atypical language.** We will show that distributional semantic models for word association and ranking, in addition to corpus-based and knowledge-based semantic expansion, can very accurately identify idiosyncratic lexical items in narrative retellings.

**Characterizing topic digression in narratives.** We will present our methods for pairwise semantic comparison of narratives of children children with ASD and typical development, and identifying idiosyncratic topic digressions in the retellings generated by children with ASD.

**Similarity measures for quantifying topic repetition in conversations.** We will describe the applications of distributional semantic similarity methods for deriving verbatim and/or purely semantic similarity scores between conversational turns at different distance windows in order to capture topic repetitions.

**Computational analysis of conversational reciprocity.** We will explain our methodology for question classification, and the analysis of appropriate use of discourse markers and acknowledgment terms in post-question and post-statement contexts.

**Statistical significance of our measures.** For each measure, we analyze their statistical power in differentiating children with ASD and typical development.

## 1.3 Organization of the dissertation

In the next two chapters we present the necessary background for our original work. In Chapter 2, we give an overview of the computational background for distributional semantic models and similarity measures at the lexical and document levels. Chapter

3 provides detailed information about the participants of this study, their recruitment, screening and diagnosis process, as well as their neuropsychological assessments. We also explain the two main data sets that we analyze in this thesis: children's retellings of the NEPSY Narrative memory (NMM), and conversation transcripts of the Autism Diagnostic Observation Schedule (ADOS).

In the later chapters we present the original research conducted for this thesis. Chapters 4 and 5 focus on semantic and pragmatic atypicality in narrative retellings. Chapter 4 explores the presence of atypical lexical items in narratives and our automated methods for detecting such items, and Chapter 5 describes our technique for characterizing idiosyncratic interest and topic deviations of children with ASD in the NEPSY narrative retelling task. Chapters 6 and 7 focus on the analysis of the semi-structured ADOS dialogues. Chapter 6 reports our methods for characterizing topic perseveration and restrictive repetitive patterns in conversations, and Chapter 7 discusses our analysis of conversational responsiveness and the use of appropriate conversational cues. Finally, in Chapter 8, we conclude the findings of the thesis and summarize our contributions. We propose future work, including suggestions for additional analysis based on our findings and methods, as well as applying the methods developed in the thesis on clinical data from other neuropsychological disorders.

| | |
|---|---|
| ADOS | Autism Diagnostic Observation Schedule |
| ALI | autism with language impairment |
| ALN | autism language normal (no language impairment) |
| ASD | autism spectrum disorder |
| CA | chronological age |
| Cos | cosine similarity score |
| IDF | inverse document frequency |
| IR | information retrieval |
| Jac | Jaccard similarity coefficient |
| LI | meeting criteria for a language impairment |
| LS | Lin's universal similarity |
| LSA | latent semantic analysis |
| NEPSY | A Developmental NEuroPSYchological Assessment |
| NLP | natural language processing |
| NNM | NEPSY Narrative Memory |
| NVIQ | nonverbal IQ |
| PLR | pure lexical relatedness |
| RFM | relative frequency measure |
| RRB | restrictive and repetitive behavior |
| SLI | specific language impairment |
| SOR | semantic overlap ratio |
| TD | typical development/typically developing |
| TF | term frequency |
| VIQ | verbal IQ |
| VLR | verbatim word overlap with lexical relatedness |
| WJC | weighted Jaccard similarity coefficient |
| WM | WordNet-based mutual similarity |
| WV | WordNet-based vector similarity |

Table 1.1: Acronyms and initialisms used in this thesis.

# Chapter 2

# Preliminaries and technical background

In this chapter we present the technical background for the original research that we conducted in later chapters. We start with the definition of vector space models and the way we use them for semantic representation of documents and language samples. We continue to describe distributional similarity measures, and their integrated term weighting, association measures, and knowledge-based lexical similarity metrics. In addition, we briefly describe the prerequisites for similarity methods we used, including $n$-gram language models, part-of-speech tagging, and text pre-processing.

## 2.1 Vector space model

The vector space model is an algebraic model for document representation in the form of vectors of words or terms . It is widely used in distributional semantics to represent the meaning of words, phrases, or documents, and measure the similarity between those. The idea is to represent each document in the corpus as a vector in vector space. Vectors that are closer together in the space use a noticeable proportion of the same words, and so are presumably semantically more similar.

Consider that we observe three children are talking about *climbing a branch of an oak tree*. We see that the first child mentioned the word "climb" five times, "tree" four times, "branch" one time, and "oak" three times in his or her conversation. More formally we represent the frequency of this set of four words as:

$d$ = {'climb': 5, 'tree': 4, 'branch': 1, 'oak': 3 }

We can illustrate this set of words or *document* $d$ in a vector space representation as:

$$\vec{d} = (5, 4, 1, 3)$$

where each dimension of $\vec{d}$ corresponds to words "climb", "tree", " branch", and "oak" respectively.

Now assume that the second child mentioned only the term "climb" from the above list and he repeated that four times, and a third child mentioned each word "climb" and "tree" only once. The vector space representation of these two cases will be:

$$\vec{d'} = (4, 0, 0, 0)$$
$$\vec{d''} = (1, 1, 0, 0)$$

It is easy to say that the first and the third children ($d$ and $d''$ respectively) have similar conversations since they both have words "climb" and "tree" in common. The second child, however, is different from the two since he or she only shares the word "climb" with them. This similarity can be illustrated in the plot of these three vectors based on the first two dimensions "climb" and "tree", where we can see the smaller angle between $d$ and $d''$ (Figure 2.1).

We use some terminology in the vector space model definition. A *document* refers to the unit of text, indexed in the system and could be used in retrieval. A *term* represents a

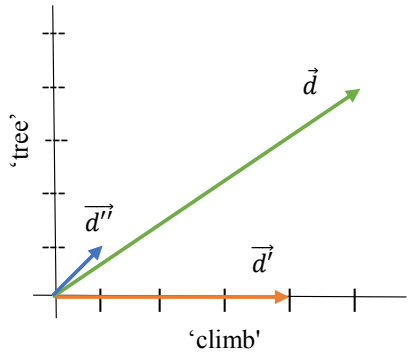



Figure 2.1: A graphical illustration of document similarity in a 2-dimensional vector space

lexical item or longer phrases, depending on the application. A *query* refers to the user's query in form of a set of terms. Documents and queries are represented as vectors $\vec{d}$ and $\vec{q}$. The vectors for a document $d$ and a query $q$ is represented as follows:

$$\vec{d} = (d_1, d_2, \ldots, d_n) \tag{2.1}$$
$$\vec{q} = (q_1, q_2, \ldots, q_n)$$

Each dimension of the vectors corresponds to a separate term. So the number of dimensions in the vector is the total number of terms in all of the documents. The value of each dimension is the term weight, which is usually a function of the term's frequency in the document. Several different ways of computing the term weight's values have been developed. The simplest form is the raw frequency counts of the terms in documents (Luhn, 1957). In Section 2.3, we describe several term weighing methods.

The vector space model has been used in many semantic tasks of natural language processing, such as in information retrieval, information extraction, topic modeling, text summarization, question answering, indexing, and relevance rankings. Some variants of the vector space model are used in Web search engines to retrieve relevant text documents to word-based queries of users (Manning et al., 2008). Other variants have been successfully applied as similarity metrics between words, sentences, and larger documents in various tasks. The main well-performing algorithms for calculating semantic relatedness between words utilize vector space models (Pantel and Lin, 2002; Rapp, 2003; Turney et al., 2003). The same is the case for major algorithms for characterizing lexical semantic relations (Lin and Pantel, 2001; Turney, 2006; Nakov and Hearst, 2008; Turney et al., 2010).

## 2.2 Similarity measures

By modeling the documents as vectors, we can measure the similarity of two documents by performing algebraic operations between their vectors. A large number of similarity measures between two vectors exist. This is historically due to the fact that different researchers have come up with various approaches to quantify similarity. Nevertheless,

this diversity reveals that similarity can be quantified and conceptualized in different ways, each highlighting different relations and association between vectors (Sepkoski Jr, 1974).

### 2.2.1 Jaccard similarity coefficient

Jaccard similarity coefficient, introduced by Jaccard (Jaccard, 1912), measures similarity between two sets by comparing the size of their overlap against the total size of the two sets, i.e., the similarity is defined as the size of intersection divided by the size of union of the sets. A natural generalization of Jaccard similarity to $n$-dimensional vectors is defined as follows (Dagan, 2000; Jurafsky and Martin, 2009; Grefenstette, 1994; Kendrick, 1964):

$$sim_{Jaccard}(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{N} min(q_i, d_i)}{\sum_{i=1}^{N} max(q_i, d_i)} \tag{2.2}$$

Intuitively, Jaccard similarity is the amount of features present in both vectors, divided by the amount of features present in either or both vectors.

### 2.2.2 Cosine similarity score

We can measure the distance of two vectors by the cosine of the angle between the two, as formulated in Equation 2.3. A cosine similarity score of 1.0 indicates that two documents have the same terms and in the same frequencies, whereas cosine of 0 means that they share no common terms.

$$sim_{cosine}(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{N} q_i \times d_i}{\sqrt{\sum_{i=1}^{N} q_i^2} \times \sqrt{\sum_{i=1}^{N} d_i^2}} \tag{2.3}$$

This metric measures the orientation of similarity and not the magnitude. If two vectors point to the points far from each other, they still could be similar if there is a small angle between the vectors. If the vectors contain raw frequencies, the cosine cannot be negative,

but term weighting can introduce negative values (see Section 2.3 for more information about term weighting).

A notational variant for the representation of cosine similarity is to divide the intersection (i.e. the dot product) of the document vector and the query vector by the multiplication of the norm (i.e. the length) of the two vectors, as formulated in Equation 2.4. This is also called normalized dot product.

$$sim_{cosine}(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}||\vec{d}|} \tag{2.4}$$

### 2.2.3 Relative frequency measure

Relative frequency measure was introduced by Hoad and Zobel (2003) as a similarity measure between two co-derivative documents. The main application of this measure is to identify plagiarism at the document level. A simple variation of this measure, presented by Metzler et al. (2005) is formulated as follows:

$$Sim_{RelFreq}(\vec{q}, \vec{d}) = \frac{1}{1 + \frac{max(|\vec{q}|,|\vec{d}|)}{min(|\vec{q}|,|\vec{d}|)}} \sum_{i=1}^{N} \frac{1}{1 + |q_i - d_i|} \tag{2.5}$$

The numerator is 1 in this variation but it can be replaced by a term weighting factor to assign higher weights to document-specific words. The denominator consists of two parts. The first part penalizes the difference between the lengths of the vectors, and the second part penalizes the differences between the term values.

### 2.2.4 Vector distance measures

Similarity can be computed using a measure of distance between the two vectors. Turney et al. (2010) suggest to convert a distance measure to a similarity measure by inversion or subtraction:

$$sim(\vec{q}, \vec{d}) = \frac{1}{dist(\vec{q}, \vec{d})} \tag{2.6}$$
$$sim(\vec{q}, \vec{d}) = 1 - dist(\vec{q}, \vec{d})$$

Two of the simplest vector distance measures are Levenshtein and Euclidean distance. The Levenshtein distance, also known as Manhattan distance or L1 norm, gives the Manhattan or "city block" distance between two vectors. This metric is defined as:

$$dist_{Manhattan}(\vec{q}, \vec{d}) = \sum_{1}^{N} |q_i - d_i| \tag{2.7}$$

The Euclidean distance metric, also known as L2 norm, defines the squared distance of two vectors as the sum of squared differences in their coordinates. The distance itself is the square root of the squared distance:

$$dist_{Euclidean}(\vec{q}, \vec{d}) = \sqrt{\sum_{1}^{N} (q_i - d_i)^2} \tag{2.8}$$

These two metrics are more popular in geometric measures than information retrieval and lexical similarity. Other common information theoretic distance metrics are Kullback-Leibler divergence or KL divergence (Kullback and Leibler, 1951), and Bhattacharya (Bullinaria and Levy, 2007) which measure the similarity of two vectors based on the similarity of their probability distributions.

### 2.2.5 BLEU score

BLEU (Papineni et al., 2002) is the most popular automated method for evaluation of machine translation (MT). It calculates a weighted average of the number of $n$-gram overlaps (see Subsection 2.6.1 for the definition of $n$-gram) between the MT output and one or more reference human translations. BLEU can be considered as a similarity metric between two documents, although it is not a variant of vector space model. It extends the

familiar precision metric to a *modified n-gram precision* in case of comparing the candidate translation against multiple reference translations. Simple precision is the number of candidate words (unigrams) in a reference translation, divided by the total number of words in the candidate. Modified unigram precision ensures that a reference word is considered exhausted after a matching candidate word is identified. To do this, the maximum number of occurrences of a word in any of the reference translations (*max_ref_count*) are counted, and then the count of each candidate word is clipped by its *max_ref_count*. The modified precision is similarly calculated for higher-order $n$-grams, often up to quadrigrams, as well.

The sentence-level modified $n$-gram precision is then extended over the entire test set. BLEU first adds the clipped $n$-gram counts for all the candidates, and then divides by the total number of candidate $n$-grams in the test set. The modifies precision score is thus:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count_{clip}(n\text{-}gram')} \tag{2.9}$$

A good candidate translation should be neither too long nor too short. Traditionally, precision is combined with recall to deal with such length-related problems. However, recall over multiple references is not a good measure, because a good candidate translation only recalls one of the references, not all. Thus, a multiplicative factor called *brevity penalty* is introduced in calculating the BLEU score. Brevity penalty (BP) is defined as:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-\frac{r}{c})} & \text{if } c \leq r \end{cases} \tag{2.10}$$

where $c$ is the length of the candidate translation and $r$ is the effective reference length computed by summing the best match lengths for each candidate. Finally, BLEU score is calculated as the geometric mean of the modified $n$-gram precisions, $p_n$, multiplied by the brevity penalty:

$$BLEU = BP \times \exp(\frac{1}{n}\sum_{i=1}^{n}\log p_i) \tag{2.11}$$

### 2.2.6 Application of similarity measures

In this thesis we apply various variations of the described similarity measures to analyze language samples of children with ASD and typical development. These samples include children's narrative retellings and conversational turns. Thus, in our case, units of comparisons, i.e. documents, are either narratives or dialogue turns. We don't have the notion of queries or 1-to-$n$ comparison, instead we compare documents against each other in an $n$-to-$n$ comparison schema. Moreover, following a common practice in information retrieval, we compute the similarity between two documents only based on the present words in those documents and not all the words in the corpus. This practice dramatically reduces the vector dimensions since it eliminates many zero values corresponding to the terms that do not appear in those two documents.

## 2.3 Term weighting and association measures

Term frequency (i.e., raw frequency of a term in a document) suggests that terms that occur more frequently are best terms for document content identification, compared to the other terms in the document. However, terms that frequently appear in many documents may not be useful to distinguish target documents. The best terms to identify target documents are the ones that are specific and mostly associated to those documents. To reflect how important a term is to a document in a corpus, several term weighting methods are available. We describe the *tf-idf* family of measures, *pointwise mutual information*, *log-likelihood ratio*, and *log-odds ratio*.

### 2.3.1 TF.IDF

The most widely-used family of measures for term weighting is obtained by the product of the term frequency or *tf* (Luhn, 1957), and the inverse document frequency or *idf* (Sparck Jones, 1972; Salton and Buckley, 1988). A term gets a high weight if it is frequent in the document (i.e., *tf* is high), but rare in other documents (i.e., *df* is low). The *idf* is computed as (the logarithm of) the number of the documents in the corpus or $D$, divided by the number of documents where the term $w$ appears, or $df_w$ (Sparck Jones, 1972):

$$idf_w = \log(\frac{D}{df_w}) \tag{2.12}$$

Let $tf_{wd}$ be the frequency of term $w$ in document $d$. Then $tf\text{-}idf_{wd}$ is calculated as (Manning et al., 2008):

$$tf\text{-}idf_{wd} = tf_{wd} \times idf_w \tag{2.13}$$

Similar to *idf*, $tf_{wd}$ is sometimes converted into log-domain as well.

### 2.3.2 Log likelihood ratio

Log likelihood ratio (Dunning, 1993a) is another well-performing method for term weighting and calculating word associations. It is defined as the ratio between probability of observing a term $w$ in document $d$ and the overall probability of observing $w$ in other documents than $d$ i.e. $\bar{d}$ (Manning and Schütze, 1999). The following two explanations are examined:

- Hypothesis 1. $P(w|d) = P(w|\bar{d})$
- Hypothesis 2. $P(w|d) \neq P(w|\bar{d})$

Hypothesis 1 expresses that the occurrence of word $w$ is *independent* of observing it in document $d$, and Hypothesis 2 says that occurrence of $w$ is *dependent* to seeing it in $d$ which is a good evidence that $w$ is an associated word to $d$.

We use the maximum likelihood estimate of the probabilities $p$, $p_1$, and $p_2$ based on the following formulation:

$$\begin{aligned} p &= P(w) = \frac{c_w}{N} \\ p_1 &= P(w|d) = \frac{tf_{wd}}{c_d} \\ p_2 &= P(w|\bar{d}) = \frac{c_w - tf_{wd}}{N - c_d} \end{aligned} \tag{2.14}$$

where $c_w$ is the total count of $w$ in our corpus, $c_d$ the number of words in document $d$, $tf_{wd}$ the count of word $w$ in document $d$, and $N$ total number of words in our corpus.

If we assume a binomial distribution:

$$b(k; n, x) = \binom{n}{k} x^k (1-x)^{(n-k)} \tag{2.15}$$

the likelihood of hypotheses 1 and 2, given the observed counts $c_w$, $c_d$, and $tf_{wd}$ is:

$$\begin{aligned} L(H_1) &= b(tf_{wd}; c_d, p) b(c_w - tf_{wd}; N - c_d, p) \\ L(H_2) &= b(tf_{wd}; c_d, p_1) b(c_w - tf_{wd}; N - c_d, p_2) \end{aligned} \tag{2.16}$$

The log likelihood ratio for the two hypotheses then will be calculated as follows:

$$\begin{aligned} \log \lambda &= \log \frac{(H_1)}{(H2)} \\ &= \log \frac{b(tf_{wd}; c_d, p) b(c_w - tf_{wd}; N - c_d, p)}{b(tf_{wd}; c_d, p_1) b(c_w - tf_{wd}; N - c_d, p_2)} \\ &= \log L(tf_{wd}, c_d, p) + \log L(c_w - tf_{wd}, N - c_d, p) \\ &\quad - \log L(tf_{wd}, c_d, p_1) - \log L(c_w - tf_{wd}, N - c_d, p_2) \end{aligned} \tag{2.17}$$

where $L(k, n, x) = x^k (1-x)^{(n-k)}$. Following Dunning (1993a) we compute $-2 \log \lambda$ which is close to $\chi^2$.

### 2.3.3 Log odds ratio

The next term weighting and association measure we discuss is the *log-odds ratio*. Assuming $p_s$ to be a probability of success, the odds ratio of $p_s$ is defined as:

$$\text{odds}(p_s) = \frac{p_s}{1 - p_s} \tag{2.18}$$

We then define odds ratio $\theta$ between a term $w$ and document $d$ as the ratio between the odds of seeing $w$ in $d$ or $P(w|d)$, and the odds of seeing $w$ in all other documents or $P(w|\bar{d})$. Given the counts and probabilities defined above, we can formulate $\theta$ as follows (Mladenic and Grobelnik, 1999):

$$\begin{aligned}
\theta &= \frac{\text{odds}(p_1)}{\text{odds}(p_2)} \\
&= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\
&= \frac{p_1(1-p_2)}{p_2(1-p_1)} \\
&= \frac{P(w|d)(1-P(w|\bar{d}))}{P(w|\bar{d})(1-P(w|d))} \\
&= \frac{tf_{wd}(N - c_d - c_w + tf_{wd})}{(c_w - tf_{wd})(c_d - tf_{wd})}
\end{aligned} \tag{2.19}$$

we then calculate the $\log\theta$. To smooth zero values, we add 0.5 to the each multiplier in the final equation above.

### 2.3.4 Pointwise Mutual Information

In addition to term weighting, we can calculate the association and relatedness between different terms in two document vectors. In this case, instead of looking for exact term matches, we consider weights of associations between terms. Pointwise Mutual Information (PMI, Fano and Hawkins, 1961; Church and Hanks, 1990), or point mutual information, is a measure of association used in information theory. PMI measures how much the probability of co-occurrence of two events differs from the expected probability of the individual

events, assuming they are independent. Given two terms $w_1$ and $w_2$, $PMI(w_1, w_2)$ is formalized as:

$$PMI(w_1, w_2) = \log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{2.20}$$

The numerator essentially says how many times we observe $w_1$ and $w_2$ together in a given context (e.g. document), and the denominator expresses the how often we expect $w_1$ and $w_2$ co-occur, assuming that their occurrence is independent of each other (Jurafsky and Martin, 2009).

## 2.4 WordNet lexical database

Lexical and knowledge resources such as FrameNet (Baker et al., 1998a), Cyc (Lenat, 1995), and WordNet (Fellbaum, 1998), are extensive sources for lexical semantic, or common sense knowledge. The most commonly used lexical resource for English NLP applications is WordNet[1], which is an online lexical database of English developed in 1995 in Princeton. In this thesis, we use WordNet in combination with our similarity and term weighting metrics for semantic expansion and extracting similarity between the lexical items.

WordNet includes three databases for nouns, verbs, as well as adjectives and adverbs. It does not include prepositions and other closed-class or function words. Each database includes a set of lemmas. An example lemma entry for 'cat' is shown in Figure 2.2. As can be seen, the lemma 'cat' is annotated with eight senses for the noun and two senses for the verb. Each sense indicates a specific concept, consisting of a set of synonym words (or sometimes a single word) called *synsets*, a definition of the concept called *gloss*, and sometimes an example of the usage of that concept in the form of a sentence.

[1] http://wordnet.princeton.edu/

**Noun**

1. cat[1], true cat[1] (feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats)
2. guy[1], cat[2], hombre[1], bozo[2], sod[4] (an informal term for a youth or man) "a nice guy"; "the guy's only doing it for some doll"; "the poor sod couldn't even buy a drink"
3. cat[3] (a spiteful woman gossip) "what a cat she is!"
4. kat[1], khat[1], qat[1], quat[1], cat[4], Arabian tea[1], African tea[1] (the leaves of the shrub Catha edulis which are chewed like tobacco or used to make tea; has the effect of a euphoric stimulant) "in Yemen kat is used daily by 85% of adults"
5. cat-o'-nine-tails[1], cat[5] (a whip with nine knotted cords) "British sailors feared the cat"
6. Caterpillar[2], cat[6] (a large tracked vehicle that is propelled by two endless metal belts; frequently used for moving earth in construction and farm work)
7. big cat[1], cat[7] (any of several large cats typically able to roar and living in the wild)
8. computerized tomography[1], computed tomography[1], CT[2], computerized axial tomography[1], computed axial tomography[1], CAT[8] (a method of examining body organs by scanning them with X rays and using a computer to construct a series of cross-sectional scans along a single axis)

**Verb**

1. cat[1] (beat with a cat-o'-nine-tails)
2. vomit[1], vomit up[1], purge[6], cast[11], sick[1], cat[2], be sick[1], disgorge[2], regorge[1], retch[1], puke[1], barf[1], spew[3], spue[2], chuck[4], upchuck[1], honk[4], regurgitate[4], throw up[1] (eject the contents of the stomach through the mouth) "After drinking too much, the students vomited"; "He purged continuously"; "The patient regurgitated the food we gave him last night"

Figure 2.2: WordNet 3.1 entry for 'cat'.

### 2.4.1 Synsets

Synsets or synonym sets are the most important semantic primitive in WordNet. Each concept is represented in WordNet as a list of cognitive synonyms or near synonyms instead of logical forms. As an example, the entry for lemma 'cat' includes synsets such as {*cat[1], true cat[1]*}, or {*computerized tomography[1], computed tomography[1], CT[2], [...], CAT[8]*}. Each of these lexical items can be used to refer to the specific concept of the synset. The number associated with each lexical item indicates the sense number of that item used in that specific synset.

| **Relation** | **Definition between X → Y** | **Example** |
|---|---|---|
| hypernym/superordinate | concept → superordinates | *breakfast*$^1$ → *meal*$^1$ |
| hyponym/subordinate | concept → subordinates | *meal*$^1$ → *lunch*$^1$ |
| coordinate/sister terms | two concepts share a hypernym | *breakfast*$^1$ — *lunch*$^1$ |
| meronym | whole → part | *table*$^2$ → *leg*$^3$ |
| holonym | part → whole | *window*$^1$ → *building*$^1$ |
| antonym | semantic opposition between concepts | *night*$^1$ ⇔ *date*$^4$ |

Table 2.1: Most frequent noun relations in WordNet

| **Relation** | **Definition between X → Y** | **Example** |
|---|---|---|
| hypernym | event → superordinate event | *walk*$^1$ → *move*$^1$ |
| troponom | event → subordinate event | *walk*$^1$ → *stroll*$^1$ |
| coordinate/sister terms | two events share a hypernym | *fly*$^9$ — *ride*$^2$ |
| entailment | verb/event → verb/event it entails | *divorce*$^2$ → *marry*$^1$ |
| antonym | semantic opposition between events | *increase*$^1$ ⇔ *decrease*$^1$ |

Table 2.2: Most frequent verb relations in WordNet

The WordNet 3.1 database contains 117,798 nouns with 82,115 synsets, 11,529 verbs with 13,767 synsets, 21,479 adjectives with 18,156 synsets, and 4,481 adverbs with 3,621 synsets. On average, each noun has 1.24 senses, each verb has 2.17 senses, and each adjective and adverb has 1.40, and 1.25 senses, respectively[2]. Verbs, are by far the most polysemous category.

### 2.4.2 Lexical relations in WordNet

Synsets are linked to other synsets by means of lexical sense relations. Tables 2.1 and 2.2 demonstrate the important relations between nouns and verbs in WordNet. The most important relation in organizing the meanings of adjectives and adverbs is *antonymy*, defined as symmetric semantic opposition between two concepts such as *wet*$^1$, and *dry*$^1$.

[2]http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

cat, true cat
⇒ feline, felid
⇒ carnivore
⇒ placental, placental mammal, eutherian, eutherian mammal
⇒ mammal, mammalian
⇒ vertebrate, craniate
⇒ chordate
⇒ animal, animate being, beast, brute, creature, fauna
⇒ organism, being
⇒ living thing, animate thing
⇒ whole, unit
⇒ object, physical object
⇒ physical entity
⇒ entity

Figure 2.3: WordNet hypernymy chain for the noun synset {*cat, true cat*}

### 2.4.3 Lexical hierarchies in WordNet

Nouns are the most common category in WordNet and they are mainly organized by hypernymy-hyponymy relations. Each synset is directly related to its immediately more general synset through direct hypernymy relation, and to its immediately more specific concept via hyponymy. There relations are transitive and can continue to produce longer concept chains on each side. Figure 2.3 demonstrates the hypernymy chain for the noun synset {*cat, true cat*}. The immediate hypernym of this synset is {*feline, felid*}, following the chain will lead to concepts like {*mammal, mammalian*}, and even further to {*animal, animate being*}. Eventually we get to the more abstract concepts such as {*object, physical object*}, and finally {*entity*} which is the top node of noun hierarchies in WordNet (also called root or unique beginner) (Jurafsky and Martin, 2009).

## 2.5 WordNet-based lexical similarity

Two words can be semantically similar to each other, whereas they might not be synonymous. Compared to synonymy, semantic similarity is a looser metric and requires some common semantic features or elements between two words. Lexical similarity can be

used in many NLP applications such as information retrieval (Hearst, 1994), information extraction (Spasic et al., 2005), text summarization (Erkan and Radev, 2004), question answering (Soubbotin and Soubbotin, 2001), plagiarism detection (Clough, 2000), and automatic essay grading (Mihalcea et al., 2006). In all of these applications lexical similarity helps to expand the semantic space of lexical items in document vectors by including similar terms between documents in addition to exact words. This facilitates capturing topically and semantically related words and documents.

A popular approach for computing lexical similarity is using knowledge-bases and lexical databases such as WordNet, in which similarity between words is defined based on the structure of the knowledge-base. In WordNet, in particular, lexical similarity is calculated based on the semantic relations defined above, mainly hypernymy-hyponymy relations. Since different parts-of-speech categories have separate databases in WordNet, most of the WordNet-based lexical similarity measures can only be calculated with the same part-of-speech. To assign part-of-speech tags to words in our corpus we use an HMM POS-tagger, as we describe in Subsection 2.6.2.

### 2.5.1 Simple path similarity

The simplest measure of similarity between two words is the length of the path between the two in the (hierarchy) graph of a knowledge-base. The intuition behind this approach is that words are more similar to their adjacent words in the hierarchy such as their sister, children, or parent terms. The more distant the words are in this particular representation of concepts, the less similar they are. We can formally define path length similarity as follows (Jurafsky and Martin, 2009):

$pathLen(c_1, c_2)$ = the number of edges in the shortest path in the thesaurus graph between the sense node $c_1$ and $c_2$.

Path length similarity is often defined in the log domain to reduce path count variations according to the following equation (Leacock and Chodorow, 1998):

$$Sim_{pathLen}(c_1, c_2) = -\log pathLen(c_1, c_2) \tag{2.21}$$

### 2.5.2 Information content lexical similarity

There are two problems with the path length similarity metric: First, it assumes that the hierarchy graph is uniformly distributed, or in other words all the edges in the hierarchy have similar values. However, in WordNet some concepts have a more fine-grained classification with longer hypernymy chain. Thus, a large number of edges in such hierarchies does not necessarily illustrate the semantic distance. Second, higher concepts in the hierarchy are very abstract. They connect many nodes in the hierarchy but such connections are not very informative in terms of semantic similarity.

Resnik (1995) suggested a probabilistic approach to calculate the *information content* of the nodes in hierarchy. He defined the frequency of concept $c$ or $freq(c)$ as the number of words in a corpus which are the hyponyms of the concept $c$. The probability of concept $c$, or $P(c)$ is then defined as $freq(c)$ normalized by the total number of words in the corpus, $N$:

$$freq(c) = \sum_{n \in words(c)} count(n) \tag{2.22}$$

where $words(c)$ is the set of hyponym words of $c$. Then:

$$P(c) = \frac{freq(c)}{N} \tag{2.23}$$

Following standard definitions in information theory, Resnik (1995) defines the information content of concept $c$ or $IC(c)$ as:

$$IC(c) = -\log P(c) \tag{2.24}$$

Concepts with high probability then have lower information content, hence the more abstract a concept, the lower its information content. Assuming the *lowest common subsumer* or *LCS* of two concepts $c_1$ and $c_2$ as the lowest node in WordNet hierarchy that is the hypernym of $c_1$ and $c_2$, Resnik similarity between the two concepts is defined as the information content of their lowest common subsumer, calculated by the following formula (Jurafsky and Martin, 2009):

$$Sim_{Resnik}(c_1, c_2) = -\log P(LCS(c_1, c_2)) \tag{2.25}$$

The intuition behind this similarity is that two words are more similar if they share a less abstract and more concrete subsumer, i.e. they have more *information* in common.

Lin (1998) extends this idea and defines similarity between two concepts as the relation between the amount of information needed to express their commonality and the information required to fully describe each of the two concepts by normalizing their shared information content with their probability. More formally, Lin's universal similarity is defined as (Lin, 1998):

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)} \tag{2.26}$$

An alternative formula, called Jiang-Conrath distance (Jiang and Conrath, 1997) derived from distance rather than similarity between two concepts:

$$Sim_{JC}(c_1, c_2) = 2 \times \log P(LCS(c_1, c_2)) - \log P(c_1) + \log P(c_2) \tag{2.27}$$

## 2.6 $N$-grams and part-of-speech tagging

### 2.6.1 $N$-grams

An $n$-gram is a sequence of $N$ tokens of text or speech. Tokens can be words, characters, phonemes, syllables, etc. according to the application. An $n$-gram of size 1, 2, and 3

are commonly called *unigram*, *bigram*, and *trigram* respectively. An $n$-gram model, computes the probability of the last token of an $n$-gram, given the previous tokens (Jurafsky and Martin, 2009). $N$-gram models are widely used in computational linguistics, computational biology, augmentative communication, and many speech and natural language processing tasks including part-of-speech tagging, machine translation, speech recognition, sentiment extraction, spelling correction and prediction, and lexical similarity.

The probability of a sequence of $k$ words $\mathbf{w} = w_1 w_2 \dots w_k$, using the chain rule of probabilities is estimated by multiplying conditional probabilities of each word given its history of previous words:

$$P(w_1 \dots w_k) = \prod_{i=1}^{k} P(w_i | w_0 \dots w_{i-1}) \tag{2.28}$$

However, computing the exact probability of a word given its entire long history is practically impossible. Thus, according to Markov assumption, an $n$-gram model limits the distance of dependencies by approximating the history to just the few last words. As a result, the estimation becomes:

$$P(w_1 \dots w_k) = \prod_{i=1}^{k} P(w_i | w_{i-n} \dots w_{i-1}) \tag{2.29}$$

The simplest method to compute the estimated $n$-gram probabiliy $P(w|h)$, for all words $w$ for any given history $h$ in a corpus, is to use maximum likelihood estimation. The relative frequency notation from Roark and Sproat (2007) for this estimation is:

$$\widehat{P} = \frac{c(hw)}{\sum'_w c(hw')} \tag{2.30}$$

where $c(hw)$ is the count of $hw$ in the corpus, and it is normalized by the sum of the counts of all $n$-grams that share the same history.

### 2.6.2 POS Tagging

Part-of-speech (POS) tagging is the process of assigning a POS tag to each word in a corpus. POS tagging is an important type of lexical disambiguation performed usually as a pre-processing step for the majority of computational linguistics tasks. POS tag of a word gives a lot of useful information for language processing about the word and its neighbors. It distinguishes the word class between eight main classes: noun, verb, pronoun, preposition, adverb, conjunction, participle, and article (Jurafsky and Martin, 2009). It also distinguishes subclasses such as different verb tenses, or whether a noun is singular or plural. For example, the Penn Treebank corpus (Marcus et al., 1994) has a tag set of size 45 POS tags or the CHILDES (MacWhinney, 2000) corpus has a tag set of size 31.

Several algorithms have been applied to POS tagging problem including rule-based methods and stochastic methods. Rule-based methods use hand-written rules to assign words' POS tags. One of the most common stochastic approaches to POS tagging is the Hidden Markov Model, or HMM. In an HMM model, POS tagging is a sequence classification task which, given the observation of a sequence of words (a sentence), aims to assign it the most likely hidden state sequence of POS tags. The main features used to train an HMM model are tag-transition and observation probabilities.

The HMM POS tagging model can be extended to a log-linear model to allow a variety of features to be used in the model (Roark and Sproat, 2007). Various information in the context, including $n$-gram of surrounding words, $n$-grams of surrounding tags, and additional orthographical features to tag rare and unknown words, can be captured in a log-linear model through linear combination of weighted features. The POS tagger we used in thesis uses a log-linear model (Yarmohammadi, 2014).

## 2.7 Pre-processing of the text

Recall that our overall goal is to apply NLP methods for the analysis of pragmatic language in autism. We pre-process our corpus of children's narrative retellings and conversations before applying vector space models and other measures for document and term similarity. The common pipeline for text pre-processing involves tokenization, lemmatization (or stemming), and removing stop words.

### Tokenization

Tokenization or word segmentation is the task of breaking and separating words (i.e. tokens) from a stream of text. Words in English are often separated by whitespace but this character is not always a good indicator for the tokenization boundary. In our tokenization process, we separate phrases like "I'll" into two terms "I" and "-'ll" and represent names like "Los Angeles" as "Los-Angeles"). We also separate utterance-final punctuations from the last word of the utterance.

### Lemmatization

The focus of the language analyses in this thesis is on the semantic and pragmatic level. Thus, similar to some other applications such as information retrieval, we try to reduce morphological variations of words that may cause data sparsity. A popular method for doing this is *stemming* which is the process for stripping off word endings like "-ing", "-ed", "-s", and "-es" to reduce the inflected word forms into their stem. A more complex alternative to stemming which we use in this thesis is *lemmatization*. In a simple lemmatization process, we map irregular inflectional variations as "went" or "sung" into the *lemma* forms "go" and "sing", using a look-up table.

### Stop words

Stop words are often high-frequency, closed-classed, function words such as “may”, “the”, or “of” which convey little meaning on their own. We follow the common practice in information retrieval and eliminate such words because of their low semantic weight and their high frequency which often makes the size of document vectors considerably larger. Our stop word list is created based on the popular list from Selkirk (1986).

# Chapter 3

# Participants and data

## 3.1 Participants

The total number of participants in the studies in this thesis is 115, from the Portland, OR metropolitan area, between the ages of four and eight years, including: 44 typically developing children (TD) (32 males), 20 children with Specific Language Impairment (SLI) (12 males), and 51 children with autism spectrum disorder (ASD) (45 males). The ASD group is divided into two groups: children with ASD and without language impairment (ALN; N = 25; 22 males), and children with ASD and language impairment (ALI; N = 26; 23 males). The reported information on the participants in this thesis are from the following publications: van Santen et al. (2013), Van Santen et al. (2010), and Hill et al. (2015).

### 3.1.1 Recruitment and screening procedure

Participants were originally recruited for an NIH-funded study on prosody in autism. The recruitment process for the ASD group was done through autism clinics, local healthcare specialists, parent groups, education service districts, and non-profit autism organizations. The recruitment process for the SLI group was done through local speech clinics, speech language pathologists, and the Oregon Speech and Hearing Association. In addition, the recruitment team put advertisements at "community tables" in local elementary schools and in local newspapers.

The minimum required full-scale IQ (FSIQ, the composite score of verbal and non-verbal IQ) score for the study participants was 70 (mean=104.6) indicating no mental retardation or intellectual disability. FSIQ was measured by the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-III) (Wechsler, 2002) for children at ages four to seven, and the Wechsler Intelligence Scale for Children (WISC-IV) (Wechsler, 2003) at ages seven and older. All participants were native speakers of English and they were "verbally fluent", requiring a mean length of utterance (MLU, measured in morphemes per utterance) of at least 3.

The exclusion criteria of the participant include any of the following: identified metabolic, neurological, or genetic disorder; brain lesion; gross sensory or motor impairment; orofacial abnormality (e.g., cleft palate); identified mental retardation; or bilinguality. In addition, during a clinical screening, a speech and language pathologist confirmed that each participant's expressive speech was acceptable for performing the speech tasks, and verified that the child did not have speech intelligibility impairment.

### 3.1.2 Diagnosis of ASD

Diagnosis of ASD was done by a team of clinicians with specific expertise in ASD (including two Clinical Psychologists, a Speech Language Pathologist, and an Occupational Therapist). The clinicians used the criteria for ASD in the DSM-IV-TR (American Psychiatric Association, 2000) to base their judgment. ASD diagnosis was determined through best-estimate clinical (BEC) judgment (Klin et al., 2000; Spitzer and Siegel, 1990). The ASD group in this study (ALI and ALN) includes only children which received consensus BEC diagnosis of ASD. BEC diagnosis was verified by scores above the established threshold for ASD on the Autism Diagnostic Observation Schedule (ADOS-G) (Gotham et al., 2007), as well as the suggested cut-off score of 12 (Lee et al., 2007) on the Social Communication Questionnaire (SCQ) (Rutter et al., 2003). The majority of children met ADOS-G criteria for autism (ALN=19; ALI=23) (Gotham et al., 2009); a total of eight children met ASD but not autism criteria (ALN=6; ALI=3).

### 3.1.3 Diagnosis of Language Impairment

Diagnosis of language impairment (ALI and SLI groups) was performed using the Clinical Evaluation of Language Fundamentals (CELF), a composite summary of receptive and expressive language abilities. A child was diagnosed with language impairment if he or she received a Core Language Scores (CLS) lower than one standard deviation below the mean (standard score $< 85$) on the CELF. CELF CLS was measured the CELF 4 (Semel et al., 2003) children with the age of six years or older, and by the CELF Preschool-2 (Semel et al., 2004) for children bellow the age of six. 26 children with ASD (52%) received the diagnosis of language impairment based on this criterion (the ALI group).

A child received a diagnosis of SLI if he or she met one of the following criteria: (1) a documented history of language delay and/or language deficits; and (2) BEC consensus judgment of language impairment but without ASD, considering all the available evidence, including: medical and family history, assessments performed in this study, locally-based assessments performed by others before this study, and information obtained from school. However, several children with BEC SLI diagnosis had above threshold scores for ASD on the ADOS-G ($n = 4$) or the SCQ ($n = 8$). Normally, it is not recommended to use only one of these measures (ADOS-G or SCQ) in isolation to identify the diagnostic status. If one of these measures is above the threshold, then BEC is considered to be the most accurate way to diagnose (Jones and Lord, 2013). Thus, in order to maintain a representative sample of the SLI population in this work, children with a BEC diagnosis of SLI are included if they exceeded thresholds on either ADOS-G or SCQ, but not both (van Santen et al., 2013).

We should note that the CELF CLS is a composite score of receptive and expressive language abilities including phonology, morphology, syntax, semantics, and vocabulary. Hence, SLI language problems in this research are not limited to morphosyntactic problems –as usually discussed in the literature– and may include deficits in any of the areas of language production and/or comprehension mentioned above. The SLI group, while

explicitly selected to not include children with ASD, is heterogeneous and can nevertheless be expected to be at risk for numerous neuropsychiatric issues (Beitchman, 1996). In addition, even though the children in the ALN group did not receive language impairment diagnosis based on the CELF Core Language Scores, their language development can not precisely be characterized as unimpaired or typical (Hill et al., 2015).

### 3.1.4 TD Children

Children who did not receive diagnosis of ASD and SLI were assigned to the TD group, but were excluded if they meet one of the following criteria: either a ADOS-G or a SCQ score above ASD threshold; a family member diagnosed with ASD or SLI; or a history of psychiatric disturbance (e.g., ADHD, Anxiety Disorder, Bipolar Disorder).

### 3.1.5 Measures

The ADOS-G and the SCQ was used for the evaluation of core ASD symptomatology. Language ability was measured by the CELF (Preschool-2 or 4). Depending on the child's age, WPPSI-III or the WISC-IV was used for measuring IQ. Receptive vocabulary of each child was evaluated by the Peabody Picture Vocabulary Test - 3rd Edition (PPVT-III) (Dunn and Dunn, 1997) based on established norms for children 2-6 through adulthood. Semantic verbal fluency was evaluated using subtests from the NEPSY (Korkman et al., 1998); the children were asked to list as many animals as possible in 60 seconds, and the same instruction was repeated for foods. The Nonword Repetition Task (NRT) was administered and scored based on Dollaghan and Campbell (1998). Table 3.1 shows the descriptive statistics of the four groups with the significance results of paired group contracts using the Welch's two-sample t-test: T-Statistics ($t$), P-Value ($p$), and Cohen's D for effect size ($d$). The reported measures include the following: chronological age (CA), FSIQ, nonverbal IQ (NVIQ), verbal IQ (VIQ), CLS, MLU, SCQ Total score, ADOS Total score, ADOS Restricted Repetitive Behaviors score (RRB), ADOS Social Affect (SA).

Table 3.1 shows the descriptive statistics of the four groups with the significance results of paired group contracts using the Welch's two-sample t-test: T-Statistics ($t$), P-Value ($p$), and Cohen's D for effect size ($d$). The reported measures include the following: chronological age (CA), FSIQ, nonverbal IQ (NVIQ), verbal IQ (VIQ), CLS, MLU, SCQ Total score, ADOS Total score, ADOS Restricted Repetitive Behaviors score (RRB), ADOS Social Affect (SA). We will analyze the narrative retellings (see Section 3.2) of a subset of the ALN and TD groups in Chapter 4, and 5. The descriptive statistics of these two groups are reported in Table 3.2.

## 3.2 NEPSY Narrative memory

The first data set we analyze in this thesis is the narrative retellings from the Narrative Memory subset of subtest of the NEPSY (A Developmental NEuroPSYchological Assessment) (Korkman et al., 1998). The NEPSY is a comprehensive series of neuropsychological tests in children from 3 to 16 years old in six functional domains: attention and executive functions, visuospatial functions, language and communication, memory and learning, sensorimotor functions, and social perception. The NEPSY Narrative Memory (NNM) subset (part of memory and learning domain) is a narrative retelling task in which the examiners reads a short story to the child, and he has to retell the story to the examiner.

### 3.2.1 Administration

In the first version of NNM (1998) which is the one used in our study, the examiner starts with "I am going to read you a story. Listen carefully so you can tell me the story when I am finished." then she reads the brief story shown in Figure 3.1. She then prompts the child to initiate his retelling by saying "Now you tell me the story.". If the child has difficulties in initiating his retelling, the examiner could help by saying "How did the story start?" In case the child stops before the end of the story, the examiner could prompt by saying "Then what happened?", or "Tell me more." up to three times. The child's response is expected to contain the essential information of the story. It is not required to be a

| *Contrast* | *Measure* | Mean (SD) *Group1* | Mean (SD) *Group 2* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| | | $n = 51$ | $n = 44$ | | | |
| | CA | 6.58 (1.2) | 6.2 (1.26) | 1.51 | 0.135 | 0.31 |
| | FSIQ | 98.29 (15.64) | 119.57 (11.72) | −7.56 | 0.000 | −1.56 |
| | NVIQ | 108.94 (17.22) | 118.43 (14.49) | −2.92 | 0.004 | −0.6 |
| ASD vs. TD | VIQ | 94.73 (17.86) | 119.45 (12.77) | −7.84 | 0.000 | −1.61 |
| | MLU | 4.2 (0.97) | 4.9 (0.96) | −3.53 | 0.001 | −0.73 |
| | SCQ Total | 20.68 (5.23) | 2.91 (2.6) | 21.39 | 0.0 | 4.4 |
| | ADOS Total | 7.41 (1.96) | 1.25 (0.53) | 21.53 | 0.0 | 4.43 |
| | ADOS RRB | 3.92 (1.82) | 0.39 (0.58) | 13.12 | 0.0 | 2.7 |
| | ADOS SA | 7.22 (2.15) | 1.34 (.71) | 18.39 | 0.0 | 3.78 |
| | | $n = 25$ | $n = 44$ | | | |
| | CA | 6.45 (1.3) | 6.2 (1.26) | 0.79 | 0.433 | 0.2 |
| | FSIQ | 107.64 (14.9) | 119.57 (11.72) | −3.44 | 0.001 | −0.86 |
| | NVIQ | 114.0 (16.99) | 118.43 (14.49) | −1.1 | 0.279 | −0.27 |
| ALN vs. TD | VIQ | 107.72 (14.81) | 119.45 (12.77) | −3.32 | 0.002 | −0.83 |
| | MLU | 4.53 (1.023) | 4.9 (0.96) | −1.49 | 0.144 | −0.37 |
| | SCQ Total | 20.41 (4.9641) | 2.91 (2.6) | 16.4 | 0.0 | 4.11 |
| | ADOS Total | 6.88 (2.02) | 1.25 (0.53) | 13.62 | 0.0 | 3.41 |
| | ADOS RRB | 3.44 (1.4457) | 0.39 (0.58) | 10.11 | 0.0 | 2.53 |
| | ADOS SA | 6.52 (2.4) | 1.34 (.71) | 10.53 | 0.0 | 2.64 |
| | | $n = 26$ | $n = 25$ | | | |
| | CA | 6.7 (1.11) | 6.45 (1.3) | 0.73 | 0.466 | 0.21 |
| | FSIQ | 89.31 (10.23) | 107.64 (14.9) | −5.1 | 0.000 | −1.43 |
| | NVIQ | 104.08 (16.29) | 114.0 (16.99) | −2.13 | 0.038 | −0.6 |
| ALI vs. ALN | VIQ | 82.23 (9.79) | 107.72 (14.81) | −7.22 | 0.000 | −2.02 |
| | CLS | 72.04 (12.08) | 106.92 (13.54) | −9.69 | 0.000 | −2.72 |
| | MLU | 3.87 (0.8036) | 4.53 (1.02) | −2.54 | 0.014 | −0.72 |
| | SCQ Total | 20.94 (5.559) | 20.41 (4.96) | .36 | .723 | .1 |
| | ADOS Total | 7.92 (1.78) | 6.88 (2.03) | 1.95 | 0.058 | 0.55 |
| | ADOS RRB | 4.38 (2.0411) | 3.44 (1.45) | 1.91 | 0.062 | 0.54 |
| | ADOS SA | 7.88 (1.6572) | 6.52 (2.4) | 2.35 | .023 | .66 |
| | | $n = 26$ | $n = 20$ | | | |
| | CA | 6.7 (1.11) | 6.96 (1.12) | −0.78 | 0.437 | −0.23 |
| | FSIQ | 89.31 (10.23) | 87.95 (8.15) | 0.5 | 0.619 | 0.15 |
| | NVIQ | 104.08 (16.29) | 100.85 (12.6) | 0.76 | 0.453 | 0.23 |
| ALI vs. SLI | VIQ | 82.23 (9.79) | 85.3 (6.39) | −1.28 | 0.207 | −0.38 |
| | CLS | 72.04 (12.08) | 73.5 (9.09) | −0.47 | 0.642 | −0.14 |
| | MLU | 3.87 (0.8036) | 4.07 (1.01) | −0.68 | 0.504 | −0.21 |
| | SCQ Total | 20.94 (5.559) | 11.61 (6.64) | 5.07 | 0.0 | 1.51 |
| | ADOS Total | 7.92 (1.78) | 2.85 (2.58) | 7.51 | 0.0 | 2.23 |
| | ADOS RRB | 4.38 (2.0411) | 1.6 (1.67) | 5.09 | 0.0 | 1.51 |
| | ADOS SA | 7.88 (1.6572) | 3.0 (2.22) | 8.22 | 0.0 | 2.45 |
| | | $n = 20$ | $n = 44$ | | | |
| | CA | 6.96 (1.11) | 6.2 (1.26) | 2.44 | 0.019 | 0.66 |
| | FSIQ | 87.95 (8.1465) | 119.57 (11.72) | −12.46 | 0.000 | −3.36 |
| | NVIQ | 100.85 (12.60) | 118.43 (14.49) | −4.93 | 0.000 | −1.33 |
| SLI vs. TD | VIQ | 85.3 (6.39) | 119.45 (12.77) | −14.24 | 0.000 | −3.84 |
| | MLU | 4.07 (1.0136) | 4.9 (0.96) | −2.93 | 0.007 | −0.84 |
| | SCQ Total | 11.61 (6.6376) | 2.91 (2.6) | 5.66 | 0.0 | 1.53 |
| | ADOS Total | 2.85 (2.58) | 1.25 (0.53) | 2.75 | 0.013 | 0.74 |
| | ADOS RRB | 1.6 (1.667) | 0.39 (0.58) | 3.17 | 0.005 | 0.85 |
| | ADOS SA | 3.0 (2.2243) | 1.34 (.71) | 3.26 | .004 | .88 |

Table 3.1: Descriptive statistics and paired group contrasts of all participants (van Santen et al., 2013; Van Santen et al., 2010; Hill et al., 2015)

| | Mean (SD) | | | | |
|---|---|---|---|---|---|
| *Measure* | *ALN* | *TD* | $t$ | $p$ | $d$ |
| | $n = 21$ | $n = 39$ | | | |
| CA | 6.51 (1.3256) | 6.2 (1.23) | 0.88 | 0.385 | 0.24 |
| FSIQ | 106.1 (15.2967) | 118.56 (11.63) | −3.26 | 0.003 | −0.88 |
| NVIQ | 112.71 (17.5503) | 117.08 (14.42) | −0.98 | 0.336 | −0.26 |
| VIQ | 105.76 (14.8859) | 118.9 (13.1) | −3.4 | 0.002 | −0.92 |
| SCQ TOT | 20.7 (5.2198) | 2.9 (2.57) | 14.69 | 0.0 | 3.98 |
| ADOS TOT | 12.48 (4.7605) | 1.36 (1.48) | 10.43 | 0.0 | 2.82 |
| ADOS RRB | 8.0 (1.4142) | 2.49 (2.04) | 12.28 | 0.0 | 3.32 |
| ADOS SA | 6.48 (2.3795) | 1.28 (0.65) | 9.81 | 0.0 | 2.66 |

Table 3.2: Descriptive statistics and group contrasts for the ALN and TD subsets of which we analyzed the narrative retellings

> Jim was a boy whose best friend was Pepper. Pepper was a big black dog. Jim liked to walk in the woods and climb the trees. Near Jim's house was a very tall oak tree with branches so high that he couldn't reach them. Jim always wanted to climb that tree, so one day he got a ladder from home and carried it to the oak tree. He climbed up, sat on a branch, and looked out over his neighborhood. When he started to get down, his foot slipped, his shoe fell off, and the ladder fell to the ground. Jim held onto a branch so he didn't fall, but he couldn't get down. Pepper sat below the tree and barked. Suddenly Pepper took Jim's shoe in his mouth and ran away. Jim felt sad. Didn't his friend want to stay with him when he was in trouble? Pepper took the shoe to Anna, Jim's sister. He barked and barked. Finally Anna understood that Jim was in trouble. She followed Pepper to the tree where Jim was stuck. Anna put the ladder up and rescued Jim. Wasn't Pepper a smart dog?

Figure 3.1: NNM narrative.

verbatim and this response will be counted towards the Free Recall score. If the child fails to provide information on some elements, then the examiner asks element-specific questions to determine the Cued Recall score. In this work we only analyze the narrative retellings from the Free Recall section of the test. Examples of such retelling from children in this study are shown in Figures 3.3 and 3.4.

| NNM Story Elements | |
|---|---|
| 1 | Jim |
| 2 | Pepper |
| 3 | big |
| 4 | black |
| 5 | liked to walk in the woods **or** climb trees |
| 6 | tree/oak with branches too high for Jim to reach |
| 7 | climbed the tree/oak |
| 8 | got a ladder **or** carried a ladder to the tree/oak |
| 9 | looked out over the neighborhood **or** looked around |
| 10 | slipped **or** shoe fell **or** ladder fell **or** got stuck **or** couldn't get down |
| 11 | Pepper ran for help **or** went to get help **or** ran away |
| 12 | Jim was sad **or** thought Pepper didn't want to stay |
| 13 | Anna |
| 14 | Jim's sister |
| 15 | took her Jim's shoe |
| 16 | barked and barked |
| 17 | Anna put the ladder back up **or** rescued Jim **or** helped Jim. |

Figure 3.2: Story element list in the Record Form for the NNM narrative.

> Pepper rescued Jim. Her sister put up the ladder and she and she rescued Jim. And Pepper barked and barked to tell her to tell Jim's sister. And that 's all. Oh, one one more thing. Jim was walking in the woods. I wanted to call it Pepper rescued Jim.

Figure 3.3: Sample retelling from a TD child (score = 6).

### 3.2.2 Scoring

Scoring is based on a Record Form, which is a score-sheet included in the NEPSY materials and consists of 17 story elements and paraphrases of some of them. The elements contained in the Record Form are presented in Figure 3.2. The NNM Free Recall score is calculated by counting how many from a set of 17 story elements were used by a child. The typical scoring procedure is in real time using the record from, however, the scores used in our study are obtained from a research work on narratives (Prud'hommeaux, 2012). Scores in that work were assigned by first identifying the words and phrases corresponding to

> I remember Jim went up a tree with a ladder. He lost his shoe, he got stuck, he hung from a from a branch. Pepper took his shoe. He showed it to his sister and she helped him down. And that's it. Let me look at this picture with my trusty vision gadget.

Figure 3.4: Sample retelling from an ALN child (score = 6).

each element in the Record Form, manually verifying those alignments, and deriving the summary Free Recall score. Agreement between scores in and those assigned by the examiner in real time was 0.97 as measured by the Pearson product-moment correlation coefficient.

## 3.3 The Autism Diagnostic Observation Schedule (ADOS)

The Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002) is a widely-used instrument for ASD diagnosis. It consists of a semi-structured series of spontaneous conversations and interactions between a participant and an examiner (usually 30 to 60 minutes long) in which the examiner asks questions and provides prompts that serve to bring out verbal and nonverbal behaviors indicative of ASD. It is designed for the assessment of communication, social interaction, play, imagination, and restrictive and repetitive behavior based on a schedule of activities including Picture Description, Play, and Wordless Picture Book Description.

### 3.3.1 Design and Administration

The ADOS consists of four overlapping modules, each of them is appropriate to different language-ability and developmental levels. Each module approximately takes 30 minutes to administer and only one module is used for an individual at a certain point of time (Lord et al., 2002).

Module 1 is intended for young children with no consistent use of spontaneous phrase speech. Module 2 is designed for non-verbally fluent children but with ability to generate some flexible phrases. Module 3 is destined for children who are verbally fluent, defined as the language-ability to (a) generate several sentence types (including multiple clauses) and syntactic structures, (b) speak about out-out-context events and objects, and (c) making logical connections in sentence boundaries. Playing with toys should also be age-appropriate for these children. Module 4 is designed for verbally-fluent adults and adolescents who are not willing to play with toys. It includes several tasks from Module

| Module 1 | Module 2 | Module 3 | Module 4 |
|---|---|---|---|
| Preverbal/ single words/ simple phrases | Flexible phrase speech | Fluent speech child/adolescent | Fluent speech adolescent/adult |
| Anticipation of a social routine | Construction task | Construction task | Construction task |
| | Make-believe play | Make-believe play | Current work/ school/daily living |
| Functional and symbolic imitation | Joint interactive play | Joint interactive play | Socioemotional questions: Plans and dreams |
| Free play Snack | Free play Snack | Break | Break |
| Response to name | Response to name | Cartoons | Cartoons |
| Response to joint attention | Response to joint attention | Socioemotional questions: Emotions | Socioemotional questions: Emotions |
| Birthday party | Birthday party | Socioemotional questions: Friends/loneliness/marriage | Socioemotional questions: Friends/loneliness/marriage |
| Bubble play | Bubble play | Socioemotional questions: Social difficulties/annoyance | Socioemotional questions: Social difficulties/annoyance |
| Anticipation of a routine with objects | Anticipation of a routine with objects | Creating a story | Creating a story |
| | Demonstration task | Demonstration task | Demonstration task |
| | Conversation | Conversation / reporting a nonroutine event | Conversation / reporting a nonroutine event |
| | Description of picture | Description of picture | Description of picture |
| | Looking at a book | Telling a story from a book | Telling a story from a book |

Table 3.3: ADOS Modules 1-4 activities (Lord et al., 2000)

3 in addition to conversation items about daily activities. The major difference between Modules 3 and 4 is whether play or a conversation interview is used by the examiner to acquire information on social-communication ability of the participant. Table 3.3 summarizes the required levels of language ability, as well as the list of activities for Modules 1 to 4.

### 3.3.2 Diagnostic algorithms and scoring

ADOS classifications are according to the subsets of items in each module listed in Table 3.4. These items are included in a scoring algorithm based on DSM-IV diagnostic criteria (American Psychiatric Association, 2000), resulting in ADOS Communication score,

ADOS Reciprocal Social Interaction score, and the ADOS Total score (the sum of the Communication and Reciprocal Social Interactions scores) (Lord et al., 2002).

Once calculated, the scores are compared against an algorithm threshold for autism and more broadly autism spectrum disorder on each of the above domains. If a child exceeds the specific threshold in all three domain, they are assumed as meeting the criteria for that specific classification (Lord et al., 2002).

## 3.4 Data collection procedure

All procedures in this study obtained the approval of Institutional Review Board (IRB) of Oregon Health & Science University. Participating families were completely informed about the study procedures and provided consent forms. Participants were asked to complete a set of experimental tasks as well as cognitive, language, and neuropsychological tests in approximately 6 sessions (each 2 - 3 hours long). ADOS and NNM sessions were recorded by a microphone connected to a computer. All ADOS and the majority of NNM sessions were also video-recorded.

Child and examiner speech were then transcribed at the word level, indicating utterance boundaries, by one or more annotators based on the 2004 EARS Official Annotation Guidelines for conversational speech (Strassel, 2004). The transcribers were blind to the study hypotheses and child diagnosis, cognitive level, and language ability. Unless noted otherwise, all of the automated analysis were performed on these unannotated, "raw", transcripts which can be obtained relatively quickly with no specific expertise or training. In a separate labeling process for the ADOS transcripts, mazes (false starts, revisions, and repetitions) were annotated and removed from our analysis.

## 3.5 Summary

In this chapter we provided a detailed overview of our study participants, their diagnosis processes, as well as their cognitive, language, and neuropsychological tests and scores.

| **Module 1** | **Module 2** | **Module 3** | **Module 4** |
|---|---|---|---|
| Preverbal/ single words/ simple phrases | Flexible phrase speech | Fluent speech child/adolescent | Fluent speech adolescent/adult |
| | | Algorithm items | |
| Stereotyped/idiosyncratic words or phrases | Stereotyped/idiosyncratic words or phrases | Stereotyped/idiosyncratic words or phrases | Stereotyped/idiosyncratic words or phrases |
| Gestures | Descriptive, conventional, instrumental gestures | Descriptive, conventional, instrumental gestures | Descriptive, conventional, instrumental gestures |
| Unusual eye contact | Unusual eye contact | Unusual eye contact | Unusual eye contact |
| Facial expressions directed to other | Facial expressions directed to other | Facial expressions directed to other | Facial expressions directed to other |
| Quality of social overtures | Quality of social overtures | Quality of social overtures | Quality of social overtures |
| Response to joint attention | Amount of reciprocal social communication | Amount of reciprocal social communication | Amount of reciprocal social communication |
| Shared enjoyment | Quality of social response | Quality of social response | Quality of social response |
| Use of other's body to communicate | Conversation | Conversation | Conversation |
| Pointing | Pointing to express interest | | Emphatic or emotional gestures |
| Showing | Overall quality of rapport | Overall quality of rapport | |
| Frequency of vocalization directed to others | Amount of social overtures | Insight | Empathy/comments on others' emotions |
| Spontaneous initiation of joint attention | Spontaneous initiation of joint attention | Reporting of events | Responsibility |
| | | Other items | |
| Immediate echoing | Immediate echoing | Immediate echoing | Immediate echoing |
| Speech abnormalities | Speech abnormalities | Speech abnormalities | Speech abnormalities |
| Imagination/functional play | Imagination/functional play | Imagination | Imagination |
| Mannerisms | Mannerisms | Mannerisms | Mannerisms |
| Unusual sensory behaviors | Unusual sensory behaviors | Unusual sensory behaviors | Unusual sensory behaviors |
| Repetitive interests and behaviors | Repetitive interests and behaviors | Excessive, specific interests | Excessive, specific interests |
| | | Rituals and compulsive behaviors | Rituals and compulsive behaviors |
| Overactivity | Overactivity | Overactivity | Overactivity |
| Negative behavior | Negative behavior | Negative behavior | Negative behavior |
| Anxiety | Anxiety | Anxiety | Anxiety |

Table 3.4: ADOS Modules 1-4 items for diagnosis of autism (Lord et al., 2000)

We discussed our four diagnosis groups and their descriptive statistics on various measures and the significance results of paired group contracts. We also described the two data sets that we analyze in this thesis: The NEPSY Narrative Memory (NNM) subset, and the Autism Diagnostic Observation Schedule (ADOS).

# Chapter 4

# Detecting semantically unexpected words in narratives

## 4.1 Introduction

Atypical and idiosyncratic language has been one of the core symptoms observed in verbal individuals with autism since Kanner first assigned a name to the disorder (Kanner et al., 1943). It is listed as a feature of autism in DSM-IV-TR and DSM-5 (American Psychiatric Association, 2000, 2013) and currently serves as a diagnostic criterion in many of the most widely used diagnostic instruments for ASD (Lord et al., 1994, 2002; Rutter et al., 2003). The phenomenon is especially marked in the areas of semantics and pragmatics (Tager-Flusberg, 2001; Volden and Lord, 1991).

Structured language assessment tools are often not sensitive to the ASD-specific semantic and pragmatic atypicality in language samples. Instead, the degree of language atypicality is often determined via subjective parental reports (e.g., asking a parent whether their child has ever used odd phrases (Rutter et al., 2003)) or general impressions during clinical examination (e.g., rating the child's degree of "stereotyped or idiosyncratic use of words or phrases" on a four-point scale (Lord et al., 2002)). This has led to a lack of reliable and objective information about the frequency of atypical language use and its precise nature in ASD.

### 4.1.1 Off-topic content in the narratives of individuals with ASD

Manual analysis of spontaneous language samples in children and young adults with autism has shown that they have significantly more irrelevant, bizarre, and idiosyncratic content in their narratives, compared to their typically developing peers within the same age range and with similar morphosyntactic language abilities.

In one of the earliest studies on narrative abilities in ASD, Loveland et al. (1990) compared retellings from 16 children and young adults with high-functioning autism (HFA), to 16 language- and aged-matched peers with Down Syndrome. The task was recalling the story from a short puppet show or video sketch to an examiner and answer follow-up questions. The two groups did not differ in producing understandable narratives, answering follow-up questions, and structural language use. However, individuals with HFA produced significantly more bizarre language and off-topic content both in their narratives, and in their responses to questions. They also had more problems grasping the story as a set of meaningful events and sometimes they recognized the story characters as objects instead of agents.

Landa, Martin, Minshew and Goldstein (1995, reported in Landa et al., 2000) examined narrative elicitation and story completion in adolescents and adults with HFA compared to a neurotypical group, matched on age, IQ, and gender. The authors found no significant difference in story length, although they observed more variability in story length in the HFA group. In addition, the individuals with HFA produced significantly less coherent narratives with more incomplete episodes and irrelevant content, failing to provide a clear gist of the story.

In another study, Capps et al. (1998) explored personal experience narratives in 15 children with autism compared to 15 children with developmental delay, matched on language and mental age. They showed that children with autism, in general, generated significantly fewer narratives on personal experience and their narratives were not always topically relevant.

Klin (2000) analyzed narrative productions in adolescents and adults with HFA, Asperger's Syndrome (AS), and typical development (TD). They had 20 subjects in each group, age- and IQ-matched. The task was to produce a narrative, describing the depicted events in short silent cartoon animations, and attribute social meaning to the cartoon. The author found that the HFA and AS groups mentioned significantly fewer social elements and almost 30% of those elements were irrelevant and unrelated to the social plot of the stories, a significant difference.

Losh and Capps (2003) explored the narrative abilities in 28 children with autism or AS compared to 22 TD matched on age and verbal IQ. They investigated narrative elicitations from a wordless picture book as well as personal experience narratives. They found no difference between the two groups in the narrative lengths and the range of topics in both tasks. The authors showed that the children with autism had more difficulties in personal experience narratives. In particular, they used less evaluation, less complex syntactic structures, and provided significantly more irrelevant and bizarre information. In addition, the children with autism had more problems in using casual relationships between the events in both narrative contexts.

In a more recent study, Diehl et al. (2006) investigated story recall and coherence of wordless picture book narratives from 17 children with HFA compared to 17 TD children, matched on age, gender, language, and cognitive ability. The authors did not observe any noticeable difference between the two groups in terms of narrative length, complexity, and number of elements recalled. There was also no difference in the use of the gist events of the story to assist recalling of the elements, as well as the use of important versus peripheral story elements. However, authors observed significantly more instances of idiosyncratic and unusual language in the narratives produced by children with ASD in addition to significantly fewer casual connections among events.

Goldman (2008) analyzed narratives of personal experience and events of 14 children with HFA compared to 12 children with developmental language disorder and 12 TD children, all matched on age and non-verbal IQ. They did not find any group difference in the narrative structure and conversation format, but they found that the HFA group required

significantly more support from the examiner to produce narratives. In addition, the HFA group had significantly higher ratio of off-topic narrative content and irrelevant answers. They also had problems in understanding and generating the narrative at the high-level as well as problems in specifying the goals of the story.

### 4.1.2 Goals of this study

Given the above research findings in the biomedical literature, our goal in this work is to automatically identify instances of contextually atypical language at the lexical level in narrative retellings produced by children with ASD and typically developing children. We expect that children with ASD use more idiosyncratic, unexpected, or off-topic words, and we manually determine those words in our corpus of narrative retellings. We then apply two word ranking methods and distributional semantic modeling in order to automatically identify these unexpected words.

In particular, we follow these three goals in this chapter:

1. We hypothesize more irrelevant, off-topic, unexpected, or off-topic lexical content in spontaneous narrative retellings of children with ASD.
2. We manually determine the instances of those words to test our hypothesis and create a gold standard to evaluate our automated measures.
3. We use word-ranking methods and semantic expansion to automatically detect these lexical items.

Our results indicate that, children with ASD produce more semantically unexpected and off-topic words in their narratives than typically developing children (as suggested by the previous findings). In addition, we demonstrate that our automated methods for identifying these words are accurate enough to serve as an adequate substitute for manual annotation. Although unexpected off-topic word use is only one aspect of the atypical language in ASD, our work in this chapter highlights the potential of computational language evaluation and analysis methods for improving our understanding of the linguistic

| | Mean (SD) | | | | |
|---|---|---|---|---|---|
| *Measure* | *ALN* | *TD* | $t$ | $p$ | $d$ |
| Narr. Length | 70.76 (48.17) | 86.03 (43.82) | −1.21 | .883 | −.33 |
| NDW | 39.39 (22.57) | 44.1 (17.1) | −.84 | .796 | −.23 |
| NDCR | 21.05 (13.47) | 24.64 (10.92) | −1.05 | .85 | −.28 |
| NNM FR | 4.9 (3.73) | 6.26 (3.43) | −1.38 | 0.176 | −0.37 |

NDW, number of distinct words; NDCR, number of distinct content word roots; NNM FR, NEPSY Narrative Memory Free Recall score.

Table 4.1: Descriptive statistics and group contrasts of NNM retellings scores

deficits associated with ASD. The methods and results in this chapter are in part based on Rouhizadeh et al. (2013).

## 4.2 Participants and data

Participants in this study are 39 children with typical development (TD) and 21 high-functioning children with ASD and without language impairment (ALN), ranging in age from 4 to 9 years. Table 3.2 in Chapter 3 reports descriptive statistics of these two groups.

The narrative retelling task analyzed here is the NEPSY Narrative Memory (NNM) subtest (Korkman et al., 1998), described in detail in Chapter 3, Section 3.2. The NNM was administered by a trained clinician to each study participant, and each participant's retelling was recorded, transcribed, and evaluated according to the published scoring guidelines.

Table 4.1 reports descriptive statistics of narrative retellings of our study participants in terms of overall narrative lengths, number of distinct words (NDW), number of distinct content word roots (NDCR), and the NEPSY Narrative Memory Free Recall score (NNM FR). As can be seen, there is no significant difference between the ALN and TD groups in these measures based on the Welch's two-sample t-test.

Using Amazon's Mechanical Turk (AMT), we also collected a corpus of 186 retellings from (assumingly) neurotypical adults. We describe how this corpus was used in Section 4.3, below. The task included listening (but not reading) a recording of the story and when finished, provided the story retelling in written form. We briefly reviewed the responses and removed about 20 empty, very short, or junk inputs.

Two annotators, blind to the diagnosis of the experimental subjects, identified every word in each retelling transcript that was unexpected (i.e. irrelevant, off-topic) given the larger context of the story. For instance, in the sentence *T-rex could smell things*, both *T-rex* and *smell* were marked as unexpected, since there is no mention of either concept in the story. In a seemingly more appropriate sentence, *the boy sat up off the bridge*, the word *bridge* is considered unexpected since the boy is trapped up in a tree rather than on a bridge.

## 4.3 Methods

We begin with the expectation that different retellings of the same source narrative will share a common lexical semantic space (see Chapter 3). The presence of words outside of this set of vocabulary or semantic space in a retelling could suggest that the speaker has digressed from the topic of the story. Our approach for automatically identifying these unexpected and off-topic words relies on the ranking of words according to the strength of their association with the target topic of the corpus of narrative retellings. Our word ranking methods are based on the frequency of a word in the child's retelling relative to the frequency of that word in other retellings in the larger corpus of retellings. These ranking methods are similar to those developed for the Information Retrieval (IR) task of topic modeling, where the objective is to identify topic-specific sets of words – i.e., words that appear frequently in only a subset of topically-related documents – in order to cluster together documents about a similar topic. Details about how we utilize and interpret these rankings for our particular goal are provided in the following sections.

Before calculating the word rankings, we first tokenize and downcase the transcripts and remove all punctuation. In addition, we lemmatize words to decrease the sparsity caused by inflectional variations. The pipeline for identifying the unexpected words in each retelling starts by assigning a rank score to each word in each retelling using two word ranking methods described below. A threshold over these scores is determined for each child using leave-one-out cross validation in order to select a set of potentially unexpected words (see Subsection 4.3.3). This set of potential unexpected words is then filtered using two semantic expansion approaches that allow us to eliminate words that were not used in other retellings but are semantically related to topic of the narrative. This final set of words is evaluated against the set of manually identified unexpected words in order to measure the performance accuracy of our automated off-topic word identification methods.

### 4.3.1 Word ranking methods

#### TF.IDF

The first word ranking method that we used is a variation of the *tf-idf*, or term frequency-inverse document frequency, a family of term weighting measures proposed by Luhn (1957); **?**. We use the following formulation to calculate $tf\text{-}idf_{w,m}$ for each word $w$ in each child's narrative retelling $m$ (Manning et al., 2008):

$$tf\text{-}idf_{w,m} = \begin{cases} (1 + \log(tf_{w,m})) \times log(\frac{N}{df_w}) & \text{if } tf_{w,m} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where $tf_{w,m}$ is the term frequency of word $w$ in narrative $m$; $df_w$ is the document frequency of $w$, or number of retellings from the full corpus of child and adult retellings containing the word $w$; and $N$ is the total number of retellings in the full corpus.

**Log odds ratio**

The *log-odds* ratio (van Rijsbergen et al., 1981), is another association measure and ranking method used in information retrieval and extraction tasks. It is defined as the ratio between the odds of a particular word, appearing in a child's retelling as estimated using its relative frequency in that retelling, and the odds of that word appearing in all other retellings, again estimated using its relative frequency in all other retellings. Let the probability of a word $w$ appearing in a retelling $m$ be $p_1$ and the probability of $w$ appearing in all other retellings $\bar{m}$ be $p_2$, i.e:

$$p_1 = P(w|m) \tag{4.2}$$

$$p_2 = P(w|\bar{m}) \tag{4.3}$$

We can express the odds ratio of $w$ in $m$ as follows (Mladenic and Grobelnik, 1999):

$$\text{odds ratio} = \frac{\text{odds}(p_1)}{\text{odds}(p_2)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \tag{4.4}$$

### 4.3.2 Application of word ranking methods

As previously mentioned, both of these word ranking methods are used in IR to cluster together documents about a similar target topic. In IR, words that appear only in a subset of documents from a large and varied corpus of documents will have high ranking scores, and the documents containing those words will likely be focused on the same topic.

In our task, however, we have a single cluster of documents focused on a single specific topic: the NNM narrative. Topic-specific words have to occur much more frequently than other words across all retellings. As a result, words with high *tf-idf* and *log-odds* scores are likely to be those *unrelated* to the topic of the NNM story. In other words, a high rank indicates that the word $w$ is very specific to the retelling $m$, which in turn suggests that the word might be unexpected or off-topic in the larger context of the NNM narrative.

If a child veers away from the topic of the NNM story and uses words that do not occur frequently in the retellings produced by neurotypical speakers, his or her retellings will contain more words with high ranking score. Thus we expect that this set of high-scoring words is likely to overlap significantly with the set of words identified by the manual annotators as unexpected or off-topic in the context of the NNM narrative.

### 4.3.3 Cross-validation and determining the operating points

Applying these word ranking methods to each word in each child's retelling provides a list of words from each retelling ranked in order of decreasing *tf-idf* or *log-odds* score. We use leave-one-out cross-validation to determine, for each method, the best operating point that maximizes the accuracy of unexpected word identification in terms of F-measure. For each child, the threshold is found using the data from all of the other children. This threshold is then applied to the ranked word list of the held-out child. All words above this threshold are potentially unexpected words, while all words below this threshold are considered to be expected and appropriate in the context of the NNM narrative. The next stage in the pipeline is therefore to use external resources to eliminate any semantically appropriate words from the set of potentially unexpected or off-topic words generated via thresholding on the *tf-idf* or *log-odds* score.

### 4.3.4 Semantic expansion of unexpected words

The corpus of retellings that we use to generate the word rankings described above, is very small. It is therefore possible that a child may have used an entirely appropriate word that by chance was never used by another child or one of the neurotypical adults. Thus we use a corpus-based and a knowledge-based approach to find semantically related words or synonyms of the potential unexpected words and check if either of those are present in the corpus of retellings.

### Corpus-based semantic expansion

We use the CHILDES corpus of transcripts of children's conversational speech (MacWhinney, 2000) to generate topic estimates for each potentially unexpected word. To do this, we first located every utterance in the CHILDES corpus containing that potentially unexpected word. We then measure the association of that word with every other open-class word collocated in an utterance, using the log likelihood ratio (Dunning, 1993b). The 20 words from the CHILDES corpus with the highest log likelihood ratio (i.e., the words most strongly associated with the potentially unexpected word), are assumed to collectively represent a particular topic (the number of 20 is rather an arbitrary choice but as we will see, it results in a very good performance). If more than two of the words in the vector of words representing this topic are also present in the NNM source narrative or the adult retellings, the word that generated that topic is eliminated from the set of unexpected words. Words that are not present in the CHILDES corpus are annotated as unexpected.

### Knowledge-based semantic expansion

Knowledge bases are semantically rich resources which can be used as an alternative resources for lexical semantic expansion. In particular, we use WordNet (Fellbaum, 1998), an online thesaurus of English, in which nouns, verbs, adjectives and adverbs are grouped into sets of synonyms called *synsets* (see Chapter 3). For each word in the set of potentially unexpected words, we locate the WordNet synset for that word. If any of the WordNet synonyms of the potentially unexpected word is present in the source narrative or in one of the adult retellings, that word is removed from the set of unexpected words. Same as above, words that are not present in WordNet are tagged as unexpected.

| *word* | **t-rex** | **smell** | **couch** | **meat** | ~~**guy**~~ | way | sense | far | great | rescue |
|---|---|---|---|---|---|---|---|---|---|---|
| *score* | 8.1 | 7.7 | 7.7 | 7.2 | 5.6 | 4.2 | 4.2 | 3.9 | 3.6 | 2.8 |

Table 4.2: Example of word raking based on *tf-idf* scores.

## 4.4 Results

Each of our word ranking methods (i.e. *tf-idf* or *log-odds*) can be applied separately, or in combination with corpus-based or knowledge-based semantic expansion procedures, forming six conditions overall. Table 4.2 shows an example of word ranking based on *tf-idf* scores. The highlighted words are the ones with score above threshold, hence annotated as potentially unexpected words. Since "boy" and "Jim" were words in the story and the word "guy" is a closely semantically associated to either, "guy" is crossed-out from the set of unexpected words.

It is important to note that the optimized threshold described in Subsection 4.3.3 is determined separately for each combination of term weighting and semantic expansion method. This may result in potentially a different threshold for each combination tested, hence we do not necessarily expect precision to increase and recall to decrease after semantic expansion. Rather, if the semantic expansion is effective, we expect F-measure to increase since the threshold is selected in order to optimize that measure.

### 4.4.1 Accuracy of automated methods

Table 4.3 reports the performance accuracy (precision, recall, and F-measure) of our methods in identifying unexpected words against the set of manually annotated off-topic words described in Subsection 4.2. We see that the *log-odds* method outperforms *tf-idf* both with and without semantic expansion. Semantic expansion improves F-measure under both word ranking schemes. Applying corpus-based semantic expansion yields higher precision at the expense of recall, and applying knowledge-based semantic expansion results in the noticeably highest F-measure with improvements in both precision and recall.

| *Word Ranking* | *Semantic Expansion* | *Precision* | *Recall* | *F-measure* |
|---|---|---|---|---|
| TF-IDF | NO | 75.31 | 80.26 | 77.71 |
| | Corpus-based | 78.67 | 77.63 | 78.15 |
| | Knowledge-based | 87.01 | 88.16 | 87.58 |
| Log-odds | NO | 75.29 | 84.21 | 79.5 |
| | Corpus-based | 79.22 | 80.26 | 79.74 |
| | Knowledge-based | 87.80 | 94.74 | 91.14 |

Table 4.3: Accuracy of unexpected word identification methods.

### 4.4.2 Significance of the group differences

As explained in Section 4.3, the purpose of identifying these unexpected words was to determine whether children with autism produce unexpected and off-topic words at a higher rate than children with typical development. To test the significance of difference in rate of unexpected word use between the two groups, we perform a Welch's two-sample one-tailed t-test, assuming that our groups have unequal variances and unequal sample sizes. For each word ranking method in combination with the semantic expansion techniques, we report the group average rate of unexpected word use as well as the significance test results in Table 4.4 and Figure 4.1.

Our expectation of higher rate of unexpected words in the ALN group appears to be true in our manually annotated data. On average, 13.37% of the words types produced by children with ALN were marked as unexpected, while only 4.28% of words produced by children with TD were marked as unexpected, a significant difference ($p < 0.05$). This significant between-group difference in rate of unexpected word use holds even when using the automated methods of unexpected word identification, with the best performing unexpected word identification method estimating a mean of 12.78% in the ALN group and 4.69% in the TD group ($p < 0.05$).

| *Word Ranking* | *Semantic Expansion* | *ALN* | *TD* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| TF-IDF | - | .1419 (.2266) | .0427 (.0603) | 1.97 | .031 | .52 |
| | Corpus-based | .1314 (.2216) | .0396 (.0559) | 1.87 | .038 | .51 |
| | Knowledge-based | .1207 (.1900) | .0437 (.0618) | 1.81 | .042 | .49 |
| Log-odds | - | .1599 (.2172) | .0578 (.0870) | 2.07 | .025 | .56 |
| | Corpus-based | .1331 (.1857) | .0521 (.0773) | 1.91 | .034 | .52 |
| | Knowledge-based | .1278 (.1958) | .0469 (.0690) | 1.83 | .04 | .50 |
| Manual annotation | | .1337 (.1979) | .0428 (.0656) | 2.04 | .026 | .55 |

Table 4.4: Significance of the group differences in the average rate of unexpected word use using automated unexpected word identification methods

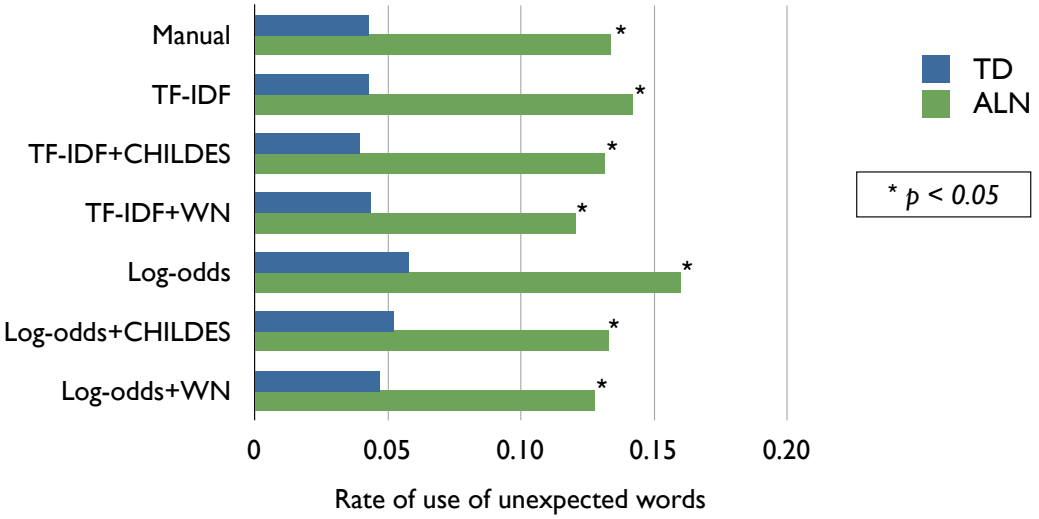



Figure 4.1: Plot of group comparison based on the average rate of unexpected word use

## 4.5 Conclusions and future work

In this chapter we presented automated methods for examining our hypothesis that children with autism have used significantly more off-topic and unexpected words in their narrative retellings compared to their typically developing peers. We first manually annotated the unexpected words in our narrative corpus and then we utilized automated methods for ranking and filtering unexpected words according to their distributions in different corpora, which are adapted from techniques originally developed for topic modeling in the context of information retrieval and extraction tasks.

Based on manual annotation we illustrated our hypothesis was true. In addition, we have demonstrated the utility of automated approaches for the analysis of semantics and pragmatics and we were able to use these methods to identify unexpected or off-topic words with high enough accuracy to replicate the patterns of unexpected word use manually observed in our two diagnostic groups.

In future work, we plan to use a development set to determine the optimal number of topical words (that could be larger or smaller than 20) and achieve a better topic estimate in the corpus-based semantic expansion stage of the pipeline in order to maintain improvements in precision without a loss in recall. We would also like to investigate using part-of-speech, word sense, and parse information to improve our approaches for both semantic expansion and topic estimation. Although the rate of unexpected word use alone is unlikely to provide sufficient power to classify the two diagnostic groups investigated here, we expect that it can serve as one feature in an array of features that capture the broad range of semantic and pragmatic atypicalities observed in the spoken language of children with autism. Finally, we plan to apply these same methods to identify the confabulations and topic shifts often observed in the narrative retellings of the elderly with neurodegenerative conditions.

## 4.6 Summary

Atypical semantic and pragmatic expression is frequently reported in the language of children with autism. Although this atypicality often manifests itself in the use of unusual or unexpected words and phrases, the rate of use of such unexpected words is rarely directly measured or quantified. In this chapter, we used distributional semantic models to automatically identify unexpected words in narrative retellings by children with autism. The classification of unexpected words is sufficiently accurate to distinguish the retellings of children with autism from those with typical development.

# Chapter 5

# Identifying idiosyncratic topic digressions in narratives

## 5.1 Introduction

One of the core features of Autism Spectrum Disorder (ASD) is preoccupation with specific restricted interests and activities (American Psychiatric Association, 2000, 2013). In verbal individuals with ASD, such a preoccupation can manifest itself as a tendency to fixate on a particular idiosyncratic topic, and a narrow range of interests.

As we explained in detail in Chapter 4, manual analysis of conversations and narratives of individuals with ASD by trained coders has shown that children and teenagers with autism include significantly more bizarre and irrelevant content in their narratives (Loveland et al., 1990; Losh and Capps, 2003) and introduce more abrupt topic digressions in their conversations (Lam et al., 2012) than their typically developing peers. In the previous chapter, we have explored automatic detection of idiosyncratic lexical items, using techniques originally developed for information extraction. There has been little work, however, in annotating the precise direction of the departure from a target topic. Thus, it is not clear whether children with ASD are initiating similar topic digressions or pursuing idiosyncratic directions in their narratives and conversations consistent with their restricted interests.

### 5.1.1 Goals of this study

In this chapter, we attempt to automatically identify topic digressions in the narrative retellings of children with ALN (ASD, with no language impairment) and to determine whether these digressions are influenced by their idiosyncratic or restricted interests. We utilize several measures designed to capture different aspects of semantic similarity between a pair of retellings. We compare all pairs of narratives from TD children, all pairs of narratives from ALN children, and each pair of narratives from TD and ALN children.

To clarify, we follow these two goals in this chapter:

1. We hypothesize less similarity and more topic digressions in the narrative retellings of children with ALN.
2. We use distributional semantic similarity measures to test our hypothesis and automatically identify topic digressions expressed in children's narrative retellings.

We find that TD children share the greatest semantic similarity with one another, while children with ALN have significantly less word similarity with TD children and even less similarity with other ALN children. These results suggest that TD children tend to adhere to the target topic in the narrative retellings, while children with ALN often stray from the target topic and introduce topic digressions. Furthermore, the fact that the lexical semantic space of an individual child with ALN seem not to resemble the lexical semantic space of other children with ALN, suggests that when a child with ALN chooses to abandon the target topic, he or she does so in an idiosyncratic way. The reported methodology and results in this chapter are based on Rouhizadeh et al. (2015a), and Rouhizadeh et al. (2014).

### 5.1.2 Related work

To the best of our knowledge, the only similar study on identifying semantic differences among the narratives of children with ASD and typical development is the work by Losh

and Gordon (2014). The authors used Latent Semantic Analysis (LSA, Landauer and Dumais, 1997) to compare the semantic space of narratives of 22 high-functioning individuals with autism and 26 typically-developing controls. The subjects were between 8-14 years old, matched on age, non-verbal, and verbal IQ. The narrative data include two different contexts: a narrative recall tasks (similar to NNM), and a wordless picture book narration task. For the story recall task, the authors calculated the LSA similarity of the retellings of each subject in each group with the original text of the story, and they found that this similarity is significantly lower in the ASD group compared to controls. They also observed a greater variability in similarity scores of the ASD group. In addition, for both story recall and picture book narration tasks, the most similar narratives from the control group were selected, and their LSA scores were averaged to generate an empirically derived standard, reflecting the 'average standard' topic space of each task. The authors then computed the LSA similarity between the retellings of the remaining participants and that average standard. They found that, compared to the controls, the story retelling of the ASD group is significantly less similar to the average standard, whereas their picture book narrations are not different.

A major caveat with the above work is the assumption that the most similar narratives from the control group generate an average standard semantic space. We think that this approach has the tendency to centralize the average standard topic space around the (randomly) repeated themes by selected participants from the control group. In addition, the authors did not provide specific information about their word ranking and term weighting methods when calculating the LSA similarities. As we have seen in the previous chapter, the off-topic words tend to have high rankings in specific narratives and these rankings will have an impact on the LSA score. Finally, the authors did not compare the narratives of each group to each other, providing comparisons only to the original story or the average standard. In our proposed method, all of these issues have been addressed, and in addition, we use multiple semantic similarity measures as opposed to single LSA similarity score. Note that the work by Losh and Gordon (2014) and the work we present in this

> Jim went up a tree with a ladder. He lost his shoe he got stuck he hung from a branch. Pepper took his shoe. He showed it to his sister and she helped him down. Let me look at this picture with my trusty vision gadget.
>
> The boy got stuck and someone rescued him and pepper was a really smart dog. Dogs have a great sense of smell too, like T-rex. T-rex could smell things that were really far away. T-rex could be over there and the meat could be way back there under the couch Well, that guy got stuck on the tree and then he, and then Pepper, his shoe fell out of the tree. Anna rescued it. Pepper brought his shoe back and Anna rescued them.

Figure 5.1: Two topically different NNM retellings with similar free recall scores (6 and 5, respectively).

chapter have been done independently and researchers on the both sides were unaware of each other's work.

## 5.2 Participants and data

Participants in this study are the same as Chapter 4 and include 39 TD children and 21 children with ALN, between 4 to 9 years old (for more details refer to Table 3.2, Chapter 3). Similar to Chapter 4, here we use also retellings from the NEPSY Narrative Memory (NNM). As explained in Chapter 3, under standard administration of the NNM, a retelling is scored according to how many story elements from a predetermined list it contains. The guidelines for scoring do not require verbatim recall for most elements and generally allow the use of synonyms and paraphrases. As is typically reported when comparing matched groups (Diehl et al., 2006), we observed no significant difference in the standard NNM free recall score (NNM FR) between the TD group and the ALN group (see Table 4.1 in Chapter 4). It might seem that a low similarity score between two retellings simply indicates that one retelling includes fewer story elements. However, given the equivalent number of story elements recalled by the two groups, we can assume that a low similarity score indicates a difference in the quality rather than the quantity of information in the retellings. Figure 5.1 shows two topically different NNM retellings with similar free recall scores.

## 5.3 Methods

### 5.3.1 Word weighting and normalization

We expect that two different retellings of the same narrative will lie in the same lexico-semantic space and will thus have high similarity scores. In this work we use well-known similarity measures with two modifications. As we discussed in Chapter 4 and Rouhizadeh et al. (2013), children with autism tend to use more off-topic and unexpected words. We have shown that such words always have high inverse document frequency (IDF) scores since they are very specific to a particular retelling. By including IDF weights, a similarity measure would be biased toward off-topic words rather than actual content words in the NNM story elements. Conventional IDF weights are therefore not useful for our particular purpose in the current chapter since we are interested to capture those similar narratives that convey the story gist. Nevertheless, we remove closed-class function words to avoid their bias in our similarity measures. In addition, we lemmatize our narrative corpus to reduce the sparsity due to inflectional variation.

### 5.3.2 Word overlap measures

As we explained, children talking about the same story are highly likely to use same or similar set of words. In this subsection we describe a family of semantic similarity measures that take overlapping words between two narratives into account.

#### Jaccard similarity coefficient

The Jaccard similarity coefficient ($Sim_{Jac}$, Jaccard, 1912) is a simple word overlap measure between a pair of narratives $n$ and $m$ defined as the size of intersection of the words in narratives $n$ and $m$, relative to the size of word union of $n$ and $m$:

$$Sim_{Jacc}(n, m) = \frac{|n \cap m|}{|n \cup m|} \tag{5.1}$$

### Cosine similarity score

Cosine similarity score $Sim_{Cos}$ is the similarity between two narratives $n$ and $m$ measured by taking the cosine of the angle between their term-frequency vectors. We use a non-weighted cosine similarity based on the following formula, where $tf_{w,n}$ is the term frequency of word $w$ in narrative $n$:

$$Sim_{Cos}(n,m) = \frac{\sum\limits_{w \in n \cap m} tf_{w,n} \times tf_{w,m}}{\sqrt{\sum\limits_{w_i \in n} (tf_{w_i,n})^2} \sqrt{\sum\limits_{w_j \in m} (tf_{w_j,m})^2}} \tag{5.2}$$

### Relative frequency measure

Relative frequency measure ($Sim_{RF}$, Hoad and Zobel, 2003) is an author identity measure for identifying plagiarism at the document level. This measure normalizes the frequency of the words appearing in both narratives $n$ and $m$ by the overall length of the two narratives, as well as the relative frequency of the words common to the two narratives. We used a simplified variation of this measure, described by Metzler et al. (2005) and formulated as follows:

$$Sim_{RF}(n,m) = \frac{1}{1 + \frac{max(|n|,|m|)}{min(|m|,|m|)}} \times \sum_{w \in n \cap m} \frac{1}{1 + |tf_{w,n} - tf_{w,m}|} \tag{5.3}$$

### BLEU

BLEU (Papineni et al., 2002) is commonly used measure of n-gram overlap for automatically evaluating machine translation output (see Chapter 2 for more details). Since BLEU is inherently an overlap measure, we could use that as similarity metric between two narratives. Because BLEU is a precision metric, the BLEU score for any pair of narratives $n$ and $m$ will depend on which narrative is considered the "reference". To create a single BLEU-based overlap score for each pair of narratives, we calculate $Sim_{BLEU(n,m)}$ as the avarage of $BLEU(m,n)$ and $BLEU(n,m)$.

### 5.3.3 Knowledge-based measures

It is reasonable to expect people to use synonyms or semantically similar words in their narrative retellings. It is therefore possible that children with autism are discussing the appropriate topic but choosing unusual words within that topic space in their retellings which could be consistent with the type of atypical language often observed in children with ASD. Based on this, we use a set of measures that consider the semantic overlap of two narratives using WordNet (Fellbaum, 1998) similarities (Achananuparp et al., 2008), in addition to simple word overlap, in order to distinguish instances of atypical but semantically appropriate language from true examples of poor topic maintenance.

Most WordNet-based similarity measures can only use word pairs with the same part-of-speech (POS). For automatic POS tagging of the narrative corpus, we trained a multi-class classifier (Yarmohammadi, 2014) from labeled training data from the CHILDES corpus of transcripts of children's conversational speech (MacWhinney, 2000). The classifier uses a discriminative linear model, learning the model parameters with the averaged perceptron algorithm (Collins, 2002). The feature set includes bigrams of surrounding words, a window of size 2 of the next and previous words, and the POS-tag of the previous word. An additional orthographical feature set is used to tag rare and unknown words. This feature set includes prefixes and suffixes of the words (up to 4 characters), and presence of a hyphen, digit, or an uppercase character. See Chapter 3, Subsection 2.6.2 for more information about part-of-speech tagging.

#### WordNet-based vector similarity

In a modified version of WordNet-based vector similarity ($Sim_{WV}$, Li et al., 2006) we first create vectors $v_n$ and $v_m$ for each narrative $n$ and $m$, where each element corresponds to a word in the type union of $n$ and $m$. We assign values to each element $e$ in $v_n$ using the following formulation:

$$S(e,n) = \max_{w_i \in n} LS(e, w_i) \tag{5.4}$$

where $LS$ is Lin's universal similarity (Lin, 1998). Lin's universal similarity of two identical words is equal to 1. Thus if the element $e$ is present in $n$, $S(e,n)$ will be 1. If not, the most similar word to $e$ will be chosen from words in $n$ using Lin's universal similarity and $S(e,n)$ will be that maximum score. The same procedure is applied to $v_m$, and finally the similarity score between $n$ and $m$ is derived from the cosine score between $v_n$ and $v_m$.

#### WordNet-based mutual similarity

In a modified version of WordNet-based mutual similarity ($Sim_{WM}$, Mihalcea et al., 2006), we find the maximum similarity score $S(w_i, m)$ for each word $w_i$ in narrative $n$ with words in narrative $m$ as described in Equation 5.4. The same procedure is applied to narrative $m$, and $Sim_{WM}$ is calculated as follows:

$$Sim_{WM}(n,m) = \frac{1}{2}\left(\frac{\sum_{w_i \in n} S(w_i, m)}{|n|} + \frac{\sum_{w_j \in m} S(w_j, n)}{|m|}\right) \tag{5.5}$$

### 5.3.4 Analytical plan

#### Pairwise similarity matrix

For each of the semantic similarity measures, we build a similarity matrix comparing every possible pair of children. Because this pairwise similarity matrix is diagonally symmetrical, we need only consider above-diagonal section of the matrix in our analyses.

Our hypothesis is that we observe less similarity in the narrative retellings of children with ALN and our goal is to explore the degree of similarity, as measured by the semantic

overlap measures, within and across diagnostic groups. With this in mind, we consider the following three sub-matrices for of each similarity matrix: one in which each TD child is compared with every other TD child (the TD.TD sub-matrix); one in which each ALN child is compared with every other ALN child (the ALN.ALN sub-matrix); and one in which each child is compared with the children in the diagnostic group to which he does not belong (the TD.ALN sub-matrix).

**Monte Carlo permutation**

We may not have enough information and a priori reason to assume that the pairwise similarity measures of all children are from any particular distribution. In order to calculate the statistical significance of these between-group differences, we therefore apply the Monte Carlo permutation method, a non-parametric procedure widely used in non-standard significance testing situations (see (Kroese et al., 2014) for more details on this method and its applications in various scientific fields).

For each pair of sub-matrices (e.g., TD.TD vs ALN.ALN) described above, we first calculate two statistics that compare the cells in one sub-matrix with the cells in other sub-matrices: the *t-statistic*, using the Welch Two Sample t-test; and the *w-statistic*, using the Wilcoxon rank sum test. We call these observed values *observed-t* and *observed-w* respectively. We next take a large random sample with replacement from all possible permutations of the data by shuffling the diagnosis labels of the children for a relatively large number of times (say 1000). We then calculate *t-statistic* and *w-statistic* for each shuffle and count the number of times the *observed-t* and *observed-w* exceed the values generated by the 1000 shuffles.

| Group | Top 10 overlapping words |
|---|---|
| TD.TD | *shoe, tree, climb, ladder, fall, Pepper, Jim, dog, sister, branch* |
| TD.ALN | *shoe, tree, Jim, climb, dog, ladder, Pepper, fall, branch, sister* |
| ALN.ALN | *shoe, tree, Jim, dog, climb, Pepper, ladder, branch, boy, run* |

Table 5.1: Top 10 overlapping words between the groups

| Group | Examples of non-overlapping words |
|---|---|
| TD.TD | *book, class, coconut, couch, lie, picture, spike, stuff, tight, watch* |
| TD.ALN | *arm, bottom, cousin, doctor, eat, fruit, giant, meat, push, sense* |
| ALN.ALN | *bite, bridge, crunch, donut, gadget, lizard, microphone, sell, table, vision* |

Table 5.2: Examples of non-overlapping words between the groups

## 5.4 Results

### 5.4.1 Overlap score between group pairs

Table 5.3 shows the average semantic overlap scores between the narratives for each of the three sub-matrices described above. We see that for both the word-overlap and the knowledge-based semantic similarity measures described in Section 5.3.1, on-average, TD children are most similar to other TD children. ALN children are less similar to TD children than TD children are to one another; and children with ALN are even less similar to other ALN children than to TD children. Some examples of overlapping and non-overlapping terms between the groups are provided in Tables 5.1 and 5.2 respectively. The non-overlapping terms appear to be topically diverse and at the same time irrelevant lexical items to the context of the NEPSY narrative.

### 5.4.2 Significance of group differences

We calculate the statistical significance of the following comparisons, using the Monte Carlo permutation method (see Subsection 5.3.4): TD.TD vs ALN.ALN; TD.TD vs

| | TD.TD | TD.ALN | ALN.ALN |
|---|---|---|---|
| $Sim_{Jac}$ | 0.19 | 0.14 | 0.11 |
| $Sim_{Cos}$ | 0.42 | 0.34 | 0.28 |
| $Sim_{RF}$ | 2.07 | 1.52 | 1.08 |
| $Sim_{BLEU}$ | 0.36 | 0.29 | 0.24 |
| $Sim_{WV}$ | 0.54 | 0.47 | 0.42 |
| $Sim_{WM}$ | 0.80 | 0.69 | 0.59 |

Table 5.3: Average semantic overlap scores for each group.

| *measure* | *statistic* | ***p-values*** | | |
|---|---|---|---|---|
| | | TD.TD vs ALN.ALN | TD.TD vs TD.ALN | TD.ALN vs ALN.ALN |
| $Sim_{Jac}$ | t | .014* | .022* | .022* |
| | w | .012* | .002* | .002* |
| $Sim_{Cos}$ | t | .025* | .043* | .027* |
| | w | .025* | .001* | .001* |
| $Sim_{RF}$ | t | .056 | .072 | .046* |
| | w | .012* | .002* | .002* |
| $Sim_{BLEU}$ | t | .032* | .039* | .034* |
| | w | .036* | .002* | .002* |
| $Sim_{WV}$ | t | .014* | .008* | .028* |
| | w | .01* | .01* | .01* |
| $Sim_{WM}$ | t | .018* | .007* | .042* |
| | w | .018* | .002* | .002* |

*: $p < .05$; t: t-test statistic; w: Wilcoxon rank sum test statistic

Table 5.4: Monte Carlo significance test p-values for each similarity measure

TD.ALN; and TD.ALN vs ALN.ALN. Table 5.4 summarizes the results of these significance tests. In all cases, the differences are significant at $p<0.05$ except for the first two comparisons in the t-test permutation of $Sim_{RF}$, which narrowly eluded significance.

### 5.4.3 Non-metric multidimensional scaling

In order to visualize the (dis-)similarities we observed in our statistical analysis between the TD and ALN groups, we utilize non-metric multidimensional scaling (MDS) (Borg and Groenen, 2005) of the pairwise similarity matrices. MDS finds Euclidean distances between the items in a pairwise similarity matrix, as well as the non-parametric monotonic

relationship between their dissimilarities using isotonic regression. These computations will result in finding the location of each item in a low-dimensional (here 2D) space.

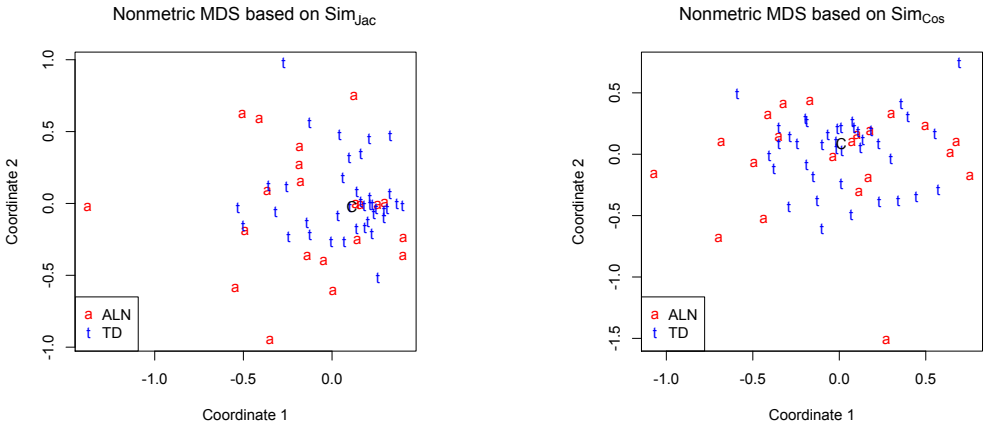
Figures 5.2-a-f visualize the location of semantic similarity of each narrative of the two groups in a 2-D MDS representation based on our similarity measures. As we can see in these figures, the TD group are more clustered around the center, whereas the ALN groups are more scattered, showing their dissimilarities with themselves and the TD group.

## 5.5 Conclusions and future work

Low lexical overlap similarity measures by themselves might indicate that children with ALN are using semantically appropriate but infrequent or sophisticated words that were not used by other children. We note, however, that the knowledge-based overlap measures ($Sim_{WV}$ and $Sim_{WM}$) follow the same pattern as the purely lexical overlap measures ($Sim_{Jac}$, $Sim_{Cos}$, $Sim_{RF}$, and $Sim_{BLEU}$). This suggests that it is not the case that children with ALN are simply using rare synonyms of the more common words used by TD children. Instead, it seems that the children with ALN are moving away from the target topic and following their own individual and idiosyncratic semantic paths. These findings provide additional quantitative evidence not only for the common qualitative observation that young children with ALN have difficulty with topic maintenance, but also for the more general behavioral symptom of idiosyncratic and restricted interests.

A focus of our future work will be to manually annotate the narrative retellings to determine the frequency of topic departures and the nature of these departures. Given the semantic differences seen here, we expect to find not only that children with ALN are departing from the topic of the source narrative more frequently than children with typical development, but also pursue topics of their own individual restricted interests.

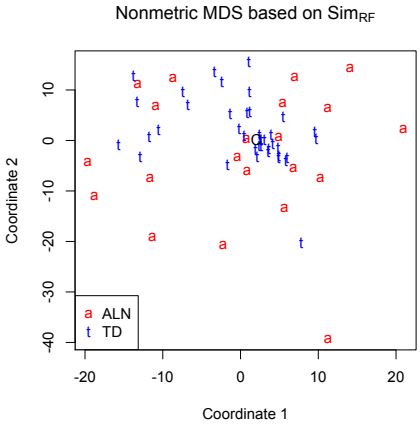
We are also interested in applying the analysis described above to a set of retellings from seniors with and without mild cognitive impairment, a frequent stage prior to dementia. Similar to children with ASD, seniors with dementia are also more likely to include irrelevant information in their narrative retellings. This irrelevant extra information, however,



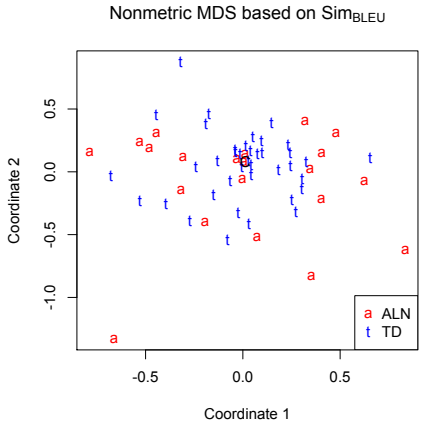

(a)



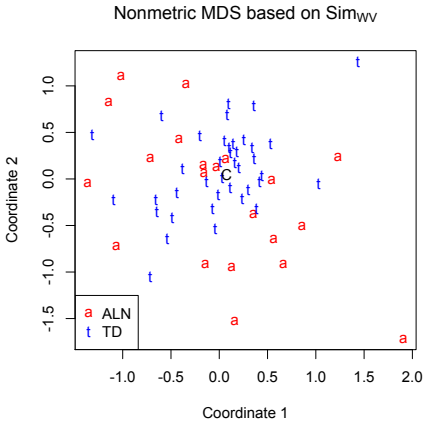

(b)



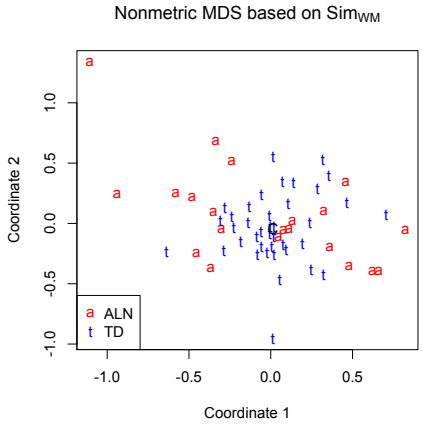

(c)




(d)




(e)




(f)

Figure 5.2: Plots of non-metric multidimensional scaling of semantic similarities between narratives of the ALN and TD groups

is often based on real-world knowledge, and thus may not result in a decrease in semantic overlap with narratives produced by unimpaired individuals.

## 5.6 Summary

A defining symptom of autism spectrum disorder is the presence of restricted and repetitive activities and interests, which can surface in language as a perseverative focus on idiosyncratic topics. In this chapter, we use semantic similarity measures to identify such idiosyncratic topics in narratives produced by children with TD and ALN. We find that neurotypical children tend to use the same words and semantic concepts when retelling the same narrative, while children with ALN, even when producing accurate retellings, use different words and concepts relative not only to neurotypical children but also to other children with ALN. Our results indicate that children with ALN not only stray from the target topic but do so in different individual ways.

# Chapter 6

# Quantifying topic repetition in spontaneous conversations

## 6.1 Introduction

An important characteristics of Autism Spectrum Disorder (ASD) is the presence of restrictive and repetitive behavior (RRB) (American Psychiatric Association, 2000, 2013). RRB is grouped into two major categories: (1) lower-order repetitive behaviors including sensory and motor movements (Prior and Macmillan, 1973; Turner, 1999). Examples of these behaviors are repetitively flapping the arms, and spending hours lining up toy bricks. Such behaviors are usually observed in younger and lower functioning children (Leekam et al., 2011), and have been the most widely studied form of repetitive behavior in autism (Turner, 1999); (2) higher-order cognitive behaviors including circumscribed interests, insistence on sameness, and topic repetition which are observed in older, higher functioning children even with good language ability (Boyd et al., 2012; Szatmari et al., 2006; Turner, 1999; Kanner et al., 1943).

Higher-order behaviors have been less systematically studied, but there is evidence for their presence in autism. Kanner et al. (1943) listed stereotypy, routines, echolalia and "obsessive desire for sameness" as main characteristic of individuals with ASD from the early days he conceptualized the disorder (Kanner et al., 1943). Wing and Gould (1979) claimed that the existence of detailed and elaborate routines is a clear distinguishing factor in individual with and without a history of autism (See Turner, 1999, and Leekam et al.,

2011 for an overview and discussion). Higher-order RRBs are manifested at the language level as well, and they affect linguistic processing, that involve semantics, pragmatics, and formulations of discourse (Rapin and Dunn, 1997).

### 6.1.1 Topic repetition in individuals with ASD

Topic repetition (or perseveration) at the language level is a specific item of higher-order RRB, listed under "highly restricted, fixated interests that are abnormal in intensity or focus" in DSM-IV-TR and DSM-5 (American Psychiatric Association, 2000, 2013). It is defined as fixation during conversation on a specific and sometimes idiosyncratic topic or theme (Sudhalter et al., 1990) and is referred to as patterns of circumscribed interest in the autism literature (Robinson and Vitale, 1954). Topic repetition is often observed as fixed and repetitive "information flow" throughout a conversation (Arora, 2008; Robinson and Vitale, 1954). It involves using words and phrases related to an individual's circumscribed interest (Rein, 1984), as well as frequently (re-)introducing tangental and repetitive conversation topics (Sudhalter et al., 1990). Loveland and Tunali-Kotoski (2005) define it as "limited topics of interest to no one but the child". Not surprisingly, it interferes with the ability to engage in a responsive and meaningful conversation (Arora, 2008; Rapin and Katzman, 1998).

Although widely reported as a characteristic of ASD individuals (Baker et al., 1998b; Baker, 2000; Koegel and Frea, 1993; Ross, 2002; Tager-Flusberg et al., 2005a), there has been little work on systematic and quantitative analysis of this phenomenon and evaluations of topic repetition has been based on subjective observations.

Rein (1984) explored whether individuals with ASD use verbal perseverations with communicative intent. He studied 15 subjects with autism, from 9.1 to 26.0 years old, through four school or work days. Defining verbal perseveration as talking excessively about a particular topic of interest, he coded conversation utterances based on their function, setting, response of listeners, and whether or not it was a verbal perseveration. He found that approximately 17% of the total number of utterances observed were verbal perseverations

and the majority of utterances were used with communicative intent. However, he did not report any comparison with a control group in his study.

Rehfeldt and Chambers (2003) examined perseverative speech for a 23-year-old man with ASD. Results showed that topic perseverations were not changed when the experimenter used social attention (e.g. eye contact). The authors found that interventions consisting of differential reinforcement of alternative verbal responses, and extinction of topic perseverations, were effective in decreasing verbal perseverations and increasing on-topic and appropriate verbal responses.

Stribling et al. (2009) studied topic repetition in conversations and interactions between a boy with ASD and a researcher, a school staff member and a mobile robot. The authors analyzed sequences of interaction and counted instances of repeated occurrences of a topical theme. Although they did not have any control group, the authors found repetitions of the same topic in the boy's conversations and they argue for the analysis of topic perseveration considering sequential context and interactional aspects of conversations.

Nadig et al. (2010) conducted an analysis of verbal exchange and eye gaze based on conversations of children with autism and typically developing children with an adult. The participants include 20 children with high-functioning autism, and 17 Typically Developing (TD) children. Groups were matched in terms of age, gender, language level, and Performance IQ. Subjects participated in a conversation with an adult about a circumscribed interest or favorite hobby, and a generic topic. Authors found that speaking about circumscribed interest negatively affected the reciprocal conversation and could cause more one-sided interaction in children with autism compared to TD. However, Boyd et al. (2007), Baker et al. (1998b), and Vismara and Lyons (2007) found the opposite effect of embedding circumscribed interest on the social behaviors of children with autism. Each of these studies examined three children diagnosed with autism and reported that the intervention of the perseverative interests improves the quality of interaction between the children and caregivers. In other words, they found that the caregivers had better interactions with the children if they allow the children talk about their restricted interests. Nadig et al. (2010) relate this difference to impact of circumscribed interests on social

communication at different levels of development on the autism spectrum, and suggests further research in that area.

Other previous work studied verbal perseveration, including topic repetition and excessive self-repetition, in other neurodevelopmental disorders. Martin et al. (2012) analyzed verbal perseveration and topic repetition of boys with Fragile X syndrome with and without ASD, Down syndrome and typical development. For each subject they manually coded their ADOS conversation for topic repetition. If a child repeated a topic or theme excessively and more than two times, it was assumed an instance of topic perseveration. The authors found significantly more topic preservation in boys that had both Fragile X syndrome and ASD, compared to all other groups and they suggest that autism status does affect topic preservation in boys with Fragile X syndrome. In another study Murphy and Abbeduto (2007) examined gender differences for verbal perseveration of individuals with fragile X syndrome. Narration and conversation contexts were designed to collect spontaneous language samples from 16 adult male and 18 adult female participants. Language transcripts were coded for utterance-level (i.e. words or phrases) repetition, topic repetition, and conversational device (i.e. rote phrases or expressions) repetition. Male participants produced more conversational device repetition than did female participants independent of cognitive or linguistic ability. They also found that more topic repetition occurred in conversation than in narration regardless of gender.

As we can see, assessment of topic repetition has been mainly based on (subjective) clinical observations and, in particular, conversation tasks. Robust evaluation of topic repetition usually requires topic annotation across all the dialogue, hence few studies have systematically addressed this phenomena in autism. In addition, some of the existing studies did not clearly distinguish verbatim versus topic repetition in their analysis. Some other only include one or few ASD participants and did compare their findings with a control group. Moreover, lack of scalable and reliable methods for measuring RRB and topic repetition in clinical interventions remains to be a challenge based on current research work (Leekam et al., 2011). Finally, to the best of our knowledge, no study have investigated topic repetition in autism, using an automated, quantitative, objective approach.

In a previous study, van Santen et al. (2013) reported an automated method for identifying and quantifying two types of repetitive speech in ASD: repetitions of what the child him- or herself said(*intra-speaker repetitions*), and repetitions of what the conversation partner said (*inter-speaker repetitions*, or *echolalia*). The authors found robust differences in echolalia between children with ASD and typical development but no significant difference in self-repetition. The focus of that study was on verbatim repeats of word n-grams at short turn distances. To the best of our knowledge, there is no study for computational assessment of topic repetition in spontaneous conversations of children with ASD.

### 6.1.2 Goals of this study

In this study, we attempt to automatically assess the presence of topic repetition (i.e. RRB at the semantic level) in children's conversation with an adult examiner during a semi-structured dialogue. We hypothesize a significantly higher semantic overlap ratio (SOR) (see Subsection 6.3.1) between dialogue turns in children with ASD compared to (a) those with typical development (TD), and (b) those with language impairment (LI). We compute the SOR for each child by averaging the similarity of every turn pair in four distance windows. To do so, we utilize multiple semantic similarity metrics – as we did in the previous chapter– based on verbatim word overlap and lexical relatedness, weighted by child specificity scores.

In summary, we explore these three goals in this chapter:

1. We automatically assess the presence of topic repetition in children's conversation.
2. We hypothesize more SOR in children with ASD than their TD and LI peers.
3. We use distributional semantic similarity measures to analyze our hypothesis.

Our analysis indicates that, based on different similarity metrics, the ASD group had a significantly higher SOR than the TD or LI groups in most of the distance windows. These results support our hypothesis. Thus, patterns of semantic similarity between a child's turns could provide an automated and robust ASD-specific behavioral marker.

The reported methodology and results in this chapter are partially based on Rouhizadeh et al. (2015b).

## 6.2 Participants and data

Participants in this study include 44 children with TD, 25 children with ASD without language impairment (ALN), 26 children with ASD and language impairment (ALI), and 20 children diagnosed with Specific Language Impairment (SLI), as described in detail in Chapter 3 and in particular, Section 3.1.

Our automated methods in this chapter are applied to unannotated ADOS conversation transcripts (see Chapter 3, Section 3.3). We especially assume that the ADOS conversations are well-suited for the analysis of topic repetitions. On one hand, the ADOS tries to minimize the regular variability of children in daily conversations, since it is a common, semi-structured context enhances comparability of the natural language samples across children and groups. On the other hand, it is designed to elicit behaviors that are indicative of ASD, including topic repetitions, and there are well-established procedures (via the ADOS manual and research training) for presentation of activities. Although all of the children are exposed to and encouraged to engage in the similar set of conversation topics by the examiner, we expect children with ASD to have more topic repetition and be less willing to collaborate with the examiner and to stick with their own topics of interest.

## 6.3 Methods

### 6.3.1 Measuring the semantic overlap ratio

In order to calculate the semantic similarity at different turn intervals, for each child, we calculate the similarity score between every turn pair $I$ and $J$ in the following turn distance windows $w$:

a) $0 < w \leq 3$: $J$ is between 1 to 3 turns after $I$ (short-distance window),

b) $3 < w \leq 9$ (mid-distance window),

c) $9 < w \leq 27$ (long-distance window),

d) $0 < w \leq l$, where $l$ is the length of conversation (overall conversation).

Then we compute the child's semantic overlap ratio (SOR) for a given window $w$ by averaging the similarity scores of turn pairs in $w$. Window $d$ ($0 < w \leq l$) provides the child's *overall SOR*.

For comparing turn pair $I$ and $J$, we explore three families of semantic similarly measures: a) measures that compute verbatim word overlap, b) measures that integrate verbatim word overlap with lexical relatedness, and c) measures that exclude verbatim overlap but consider lexical relatedness. We explain each of these measures in this section.

Because of MLU differences between groups (see Table 3.1), there may also be turn length differences between groups. Thus, it is in principle possible that differences in average turn length could bias the proposed similarity measures. To address this issue, we use a radical method in which we create completely artificial turns, as follows: all utterances of each child are concatenated and then chunked to strings of $n$ words. This was done under the assumption that each child's ADOS is a sequence of term distributions representing topics, and topic repetitions can be modeled by calculating the similarity of chunks of $n$ words. Throughout this study $n = 10$, and the word *turn* refers to strings of 10 consequent words from the child's utterances (we plan to explore the effect of applying different sizes of $n$ other than 10 in the future).

### 6.3.2 Child-specificity weight factor

We expect children with ASD to use more specific terms, relevant to their particular and often idiosyncratic interest due to their restrictive behavior. Therefore, all of our measures include a child-specificity weight factor. To assign a higher weight to words specific to a particular child, and a lower weight to the words used frequently by a large number of children, we apply an inverse document frequency (*idf*) term weight using the standard

definition of *idf* in Information Retrieval (IR) (Sparck Jones, 1972; Salton and Buckley, 1988; Manning et al., 2008):

$$idf_w = log\Big(\frac{N}{df_w}\Big) \tag{6.1}$$

where $N$ is the total number of participants and $df_w$ is the number of children who used the word $w$. We also lemmatize our corpus (see Section 2.7 in Chapter 2) to reduce the sparsity (hence higher *idf* weights) caused by inflectional variations of the same lexeme.

### 6.3.3 Measuring verbatim word overlap

#### Weighted Jaccard Similarity Coefficient

The weighted Jaccard similarity coefficient ($WJC$, Jaccard, 1912) is a word overlap measure between a pair of turns $I$ and $J$ defined as the sum of the minimum term frequency of each overlapping word $w$ in $I$ and $J$ weighted by $idf_w$ normalized by the sum of the maximum term frequency of each word in either turn:

$$WJC(I,J) = \frac{\sum\limits_{w \in I \cap J} \min(tf_{w,I}, tf_{w,J}) \times idf_w}{\sum\limits_{w \in I \cup J} \max(tf_{w,I}, tf_{w,J})} \tag{6.2}$$

where $tf_{w,I}$ is the term frequency of $w$ in $I$ (number of times $w$ occurs in $I$), and $tf_{w,J}$ is the term frequency of $w$ in $J$.

#### Relative Frequency measure

Relative frequency measure ($RFM$, Hoad and Zobel, 2003) is a similarity measure between co-derivative documents and sentences. Here we use this metric to measure the similarity of the two turns $I$ and $J$ by the weighted relative frequency of their common words (Metzler et al., 2005):

$$RFM(I,J) = \sum_{w \in I \cap J} \frac{idf_w}{1 + |tf_{w,I} - tf_{w,J}|} \tag{6.3}$$

### 6.3.4 Integrating verbatim word overlap with lexical relatedness

It is a valid assumption that children are discussing the same or similar topics using synonymous or semantically similar words. Specifically, children with autism often use idiosyncratic yet relevant words within a particular topic space. Based on this assumption, we use a WordNet-based similarity measure that integrates verbatim word overlap with lexical relatedness (*VLR*, Mihalcea et al., 2006).

We begin with finding the maximum lexical similarity score $S_{VLR}(w_i, J)$ for each word $w_i$ in turn $I$ with words in turn $J$ using the following formulation (see Subsection 2.5.2 in Chapter 2):

$$S_{VLR}(w_i, J) = \max_{w_j \in J} LS(w_i, w_j) \times idf_{w_i} \tag{6.4}$$

where $LS$ is Lin's universal similarity (Lin, 1998).

So, if $w_i \in J$, $\max_{w_j \in J} LS(w_i, w_j)$ is equal to 1, hence $S_{VLR}(w_i, J) = idf_{w_i}$. If $w_i \notin J$, the most similar word to $w_i$ will be chosen from words in $J$ using Lin's universal similarity and $S_{VLR}(w_i, J)$ will be that maximum score multiplied by $idf_{w_i}$. The same procedure is applied to the words in $J$, and finally the $VLR$ similarity between $I$ and $J$ is calculated using the following formulation:

$$VLR(I,J) = \frac{1}{2}\left(\sum_{w_i \in I} S_{VLR}(w_i, J) + \sum_{w_j \in J} S_{VLR}(w_j, I)\right) \tag{6.5}$$

Most WordNet-based similarity measures can only use word pairs with the same part-of-speech (POS). We therefore POS-tagged our narrative corpus using a perceptron tagger (Yarmohammadi, 2014). See Chapter 3, Subsection 2.6.2 for more information about part-of-speech tagging.

Lin's universal similarity can only be applied to word pairs with the same part-of-speech (POS). For POS tagging of the ADOS corpus, we used a multi-class classifier (Yarmohammadi, 2014). See Chapter 5, Subsection 5.3.3, and Chapter 3, Subsection 2.6.2 for more information about this tagger and part-of-speech tagging in general.

### 6.3.5 Pure lexical relatedness measure

To completely eliminate the effect of verbatim word overlap and exact repeats, we modified our WordNet-based similarity measure such that it takes account of lexical relatedness but precludes verbatim word overlap. This pure lexical relatedness similarity measure ($PLR$) is calculated based on the following formulation:

$$S_{PLR}(w_i, J) = \begin{cases} 0 & \text{if } w_i \in J \\ \max\limits_{w_j \in J} LS(w_i, w_j) \times idf_{w_i} & \text{otherwise} \end{cases} \tag{6.6}$$

The only difference of Equation 6.6 compared to Equation 6.4 is that $S_{PLR}(w_i, J)$ will be 0 if the word $w_i$ is present in $J$. If not, we apply the same procedure as in Equation 6.4 and choose the most similar word to $w_i$ from words in $J$. Similarly, words in $J$ are compared to $I$, and the $PLR$ similarity between $I$ and $J$ is calculated by Equation 6.5, replacing $S_{VLR}$ with $S_{PLR}$. In other words, the pure lexical relatedness similarity measure between $I$ and $J$ is derived only from the cumulative similarity scores of their most similar words.

## 6.4 Results

As described in Subsection 6.3.1, we calculate window-specific and overall SOR of each child using three families of similarity measures. To test the significance of difference in SOR between our diagnosis groups, we perform a two-tailed Welch's two-sample t-test, assuming that our groups have unequal variances and unequal sample sizes. For each distance window, and each similarity measure, we report the means of SOR in each group (with standard deviation) as well as the significance test results: t-Statistics ($t$), P-Value

| *Contrast* | *Similarity* | Mean SOR (SD) *Group1* | Mean SOR (SD) *Group 2* | $t$ | $p$ | $d$ |
|---|---|---|---|---|---|---|
| ASD vs. TD | *WJC* | .0109 (.0041) | .0081 (.0025) | 4.14 | 0.0* | .85 |
| | *RFM* | .1476 (.0458) | .118 (.0305) | 3.74 | 0.0* | .77 |
| | *VLR* | .5316 (.1747) | .4527 (.1133) | 2.65 | .01* | .54 |
| | *PLR* | .1971 (.0562) | .1824 (.0441) | 1.43 | .155 | .29 |
| ALN vs. TD | *WJC* | .0108 (.0046) | .0081 (.0025) | 2.81 | .008* | .7 |
| | *RFM* | .1475 (.0462) | .118 (.0305) | 2.86 | .007* | .72 |
| | *VLR* | .5803 (.1931) | .4527 (.1133) | 3.02 | .005* | .76 |
| | *PLR* | .2096 (.0508) | .1824 (.0441) | 2.24 | .03* | .56 |
| ALI vs. ALN | *WJC* | .011 (.0036) | .0109 (.0046) | .09 | .926 | .03 |
| | *RFM* | .1476 (.0464) | .1475 (.0462) | .01 | .988 | 0.0 |
| | *VLR* | .4848 (.1435) | .5803 (.1931) | −2.0 | .052 | −.56 |
| | *PLR* | .1852 (.0595) | .2096 (.0508) | −1.58 | .121 | −.44 |
| ALI vs. SLI | *WJC* | .011 (.0036) | .008 (.0024) | 3.35 | .002* | 1.0 |
| | *RFM* | .1476 (.0464) | .1088 (.0306) | 3.41 | .001* | 1.01 |
| | *VLR* | .4848 (.1435) | .4005 (.1069) | 2.28 | .027* | .68 |
| | *PLR* | .1852 (.0595) | .1613 (.0384) | 1.65 | .107 | .49 |
| SLI vs. TD | *WJC* | .008 (.0024) | .0081 (.0025) | −.13 | .9 | −.03 |
| | *RFM* | .1088 (.0306) | .118 (.0305) | −1.11 | .273 | −.3 |
| | *VLR* | .4005 (.1069) | .4527 (.1133) | −1.77 | .084 | −.48 |
| | *PLR* | .1613 (.0384) | .1824 (.0441) | −1.94 | .059 | −.52 |

Table 6.1: Results of paired-group contrasts in distance window $0 < w \leq 3$

($p$), and Cohen's D for effect size ($d$). These results are reported in Tables 6.1, 6.2, 6.3, and 6.4 and Figures 6.1, 6.2, 6.3, and 6.4.

Results are consistent and qualitatively similar in different combinations of distance windows and similarity measures. In all groups and in all measures, we observe a higher SOR in shorter distances that decreases as we increase the distance. In all combinations (including the overall SOR), we observe a higher SOR in ASD compared to TD; ALN compared to TD; and ALI compared to SLI. In the majority of cases these differences are statistically significant ($p < .05$). However, we did not observe significant difference between the following groups in any combination: ALI compared to ALN, and SLI compared to TD.

In the shortest distance window ($0 < w \leq 3$) we observe significantly higher SOR in ASD versus non-ASD groups using *WJC*, *RFM*, and *VLR*. However, using *PLR* (pure lexical relatedness similarity measure) we see a non-significantly higher SOR in ASD compared to

| *Contrast* | *Similarity* | Mean SOR (SD) *Group1* | Mean SOR (SD) *Group 2* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| ASD vs. TD | *WJC* | .0038 (.0019) | .0028 (.001) | 3.21 | .002* | .66 |
| | *RFM* | .0551 (.0251) | .0435 (.0145) | 2.79 | .007* | .57 |
| | *VLR* | .3074 (.0976) | .2615 (.0535) | 2.89 | .005* | .59 |
| | *PLR* | .1887 (.0609) | .1626 (.037) | 2.56 | .012* | .53 |
| ALN vs. TD | *WJC* | .0037 (.0015) | .0028 (.001) | 2.56 | .015* | .64 |
| | *RFM* | .0536 (.0184) | .0435 (.0145) | 2.36 | .023* | .59 |
| | *VLR* | .3209 (.0764) | .2615 (.0535) | 3.44 | .001* | .86 |
| | *PLR* | .1973 (.0508) | .1626 (.037) | 3.0 | .005* | .75 |
| ALI vs. ALN | *WJC* | .0039 (.0022) | .0037 (.0015) | .36 | .722 | .1 |
| | *RFM* | .0564 (.0305) | .0536 (.0184) | .4 | .692 | .11 |
| | *VLR* | .2943 (.1144) | .3209 (.0764) | −.98 | .333 | −.27 |
| | *PLR* | .1804 (.0691) | .1973 (.0508) | −1.0 | .322 | −.28 |
| ALI vs. SLI | *WJC* | .0039 (.0022) | .0026 (.0013) | 2.53 | .015* | .75 |
| | *RFM* | .0564 (.0305) | .0376 (.0184) | 2.59 | .013* | .77 |
| | *VLR* | .2943 (.1144) | .2329 (.073) | 2.21 | .032* | .66 |
| | *PLR* | .1804 (.0691) | .1512 (.0469) | 1.7 | .096 | .51 |
| SLI vs. TD | *WJC* | .0026 (.0013) | .0028 (.001) | −.79 | .436 | −.21 |
| | *RFM* | .0376 (.0184) | .0435 (.0145) | −1.27 | .214 | −.34 |
| | *VLR* | .2329 (.073) | .2615 (.0535) | −1.57 | .128 | −.42 |
| | *PLR* | .1512 (.0469) | .1626 (.037) | −.96 | .344 | −.26 |

Table 6.2: Results of paired-group contrasts in distance window $3 < w \leq 9$

| *Contrast* | *Similarity* | Mean SOR (SD) *Group1* | Mean SOR (SD) *Group 2* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| ASD vs. TD | *WJC* | .0019 (.001) | .0014 (.0005) | 3.07 | .003* | .63 |
| | *RFM* | .028 (.0141) | .0218 (.0082) | 2.68 | .009* | .55 |
| | *VLR* | .2428 (.073) | .2064 (.0418) | 3.03 | .003* | .62 |
| | *PLR* | .1793 (.0566) | .1579 (.0365) | 2.22 | .029* | .46 |
| ALN vs. TD | *WJC* | .0019 (.001) | .0014 (.0005) | 2.44 | .02* | .61 |
| | *RFM* | .0285 (.0137) | .0218 (.0082) | 2.23 | .032* | .56 |
| | *VLR* | .2565 (.0682) | .2064 (.0418) | 3.33 | .002* | .83 |
| | *PLR* | .1867 (.0469) | .1579 (.0365) | 2.65 | .011* | .66 |
| ALI vs. ALN | *WJC* | .0019 (.001) | .0019 (.001) | −.29 | .773 | −.08 |
| | *RFM* | .0275 (.0147) | .0285 (.0137) | −.24 | .808 | −.07 |
| | *VLR* | .2297 (.0764) | .2565 (.0682) | −1.32 | .192 | −.37 |
| | *PLR* | .1721 (.0647) | .1867 (.0469) | −.93 | .358 | −.26 |
| ALI vs. SLI | *WJC* | .0019 (.001) | .0013 (.0004) | 2.56 | .015* | .76 |
| | *RFM* | .0275 (.0147) | .0192 (.0053) | 2.68 | .011* | .8 |
| | *VLR* | .2297 (.0764) | .1836 (.0523) | 2.43 | .02* | .72 |
| | *PLR* | .1721 (.0647) | .1402 (.0436) | 2.0 | .052 | .59 |
| SLI vs. TD | *WJC* | .0013 (.0004) | .0014 (.0005) | −.86 | .395 | −.23 |
| | *RFM* | .0192 (.0053) | .0218 (.0082) | −1.51 | .137 | −.41 |
| | *VLR* | .1836 (.0523) | .2064 (.0418) | −1.72 | .096 | −.46 |
| | *PLR* | .1402 (.0436) | .1579 (.0365) | −1.58 | .123 | −.43 |

Table 6.3: Results of paired-group contrasts in distance window $9 < w \leq 27$

| *Contrast* | *Similarity* | Mean SOR (SD) *Group1* | Mean SOR (SD) *Group 2* | *t* | *p* | *d* |
|---|---|---|---|---|---|---|
| ASD vs. TD | *WJC* | .0018 (.0009) | .0011 (.0002) | 4.85 | 0.0* | 1.0 |
| | *RFM* | .0253 (.0113) | .0173 (.0034) | 4.75 | 0.0* | .98 |
| | *VLR* | .2331 (.0769) | .1889 (.0373) | 3.64 | 0.0* | .75 |
| | *PLR* | .1755 (.0558) | .1506 (.0355) | 2.63 | .01* | .54 |
| ALN vs. TD | *WJC* | .0016 (.0008) | .0011 (.0002) | 2.79 | .01* | .7 |
| | *RFM* | .0227 (.0093) | .0173 (.0034) | 2.78 | .01* | .7 |
| | *VLR* | .2356 (.0668) | .1889 (.0373) | 3.22 | .003* | .81 |
| | *PLR* | .1807 (.0483) | .1506 (.0355) | 2.72 | .01* | .68 |
| ALI vs. ALN | *WJC* | .002 (.001) | .0016 (.0008) | 1.41 | .165 | .39 |
| | *RFM* | .0277 (.0127) | .0227 (.0093) | 1.63 | .111 | .46 |
| | *VLR* | .2307 (.0867) | .2356 (.0668) | −.23 | .821 | −.06 |
| | *PLR* | .1706 (.0627) | .1807 (.0483) | −.65 | .521 | −.18 |
| ALI vs. SLI | *WJC* | .002 (.001) | .0012 (.0005) | 3.08 | .004* | .92 |
| | *RFM* | .0277 (.0127) | .0182 (.0059) | 3.39 | .002* | 1.01 |
| | *VLR* | .2307 (.0867) | .1812 (.0606) | 2.28 | .028* | .68 |
| | *PLR* | .1706 (.0627) | .1394 (.0433) | 1.99 | .053 | .59 |
| SLI vs. TD | *WJC* | .0012 (.0005) | .0011 (.0002) | 1.09 | .286 | .29 |
| | *RFM* | .0182 (.0059) | .0173 (.0034) | .61 | .551 | .16 |
| | *VLR* | .1812 (.0606) | .1889 (.0373) | −.52 | .606 | −.14 |
| | *PLR* | .1394 (.0433) | .1506 (.0355) | −1.01 | .319 | −.27 |

Table 6.4: Results of paired-group contrasts based on Overall SOR

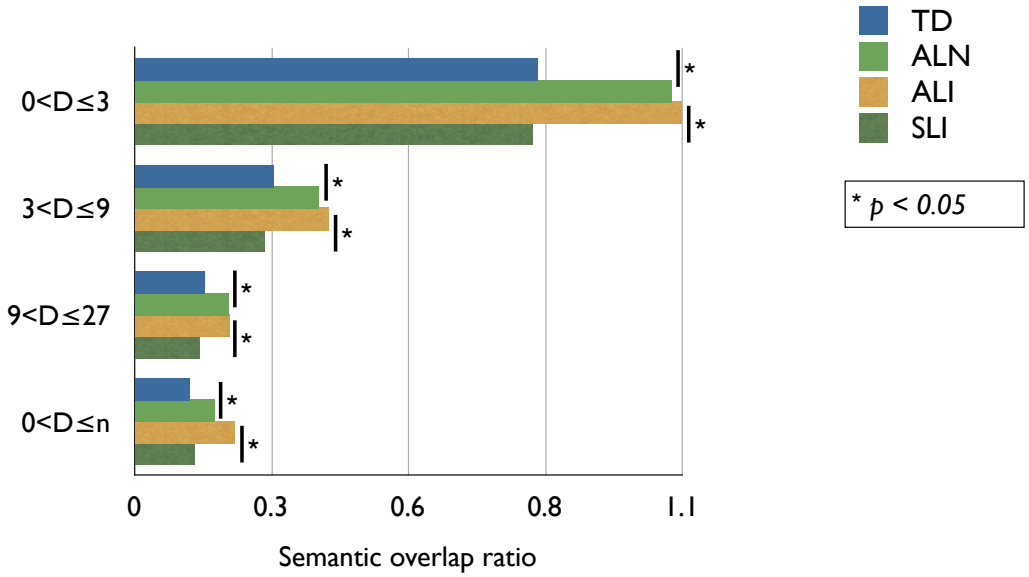



Figure 6.1: Plot of group comparison based on weighted Jaccard similarity coefficient

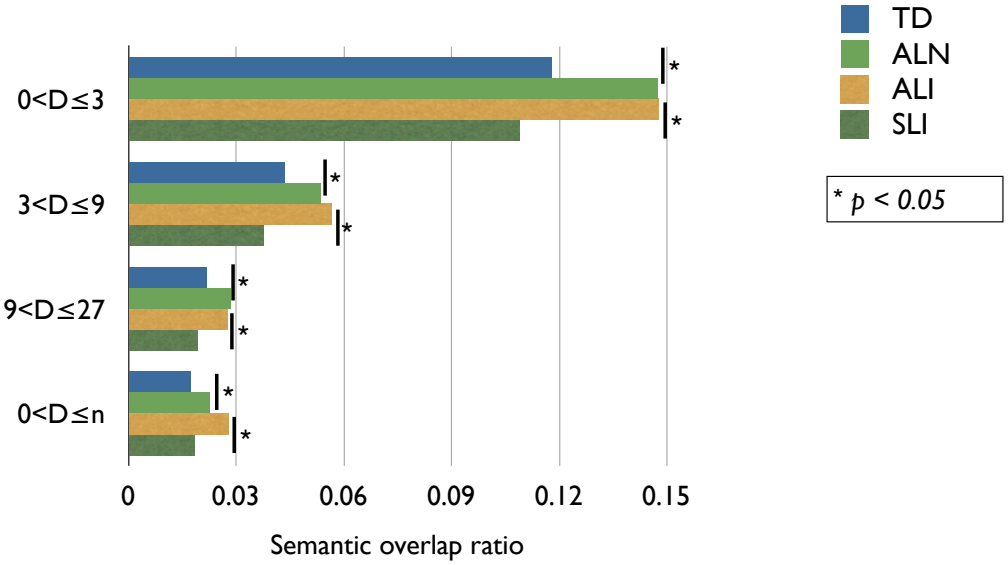



Figure 6.2: Plot of group comparison based on relative frequency measure

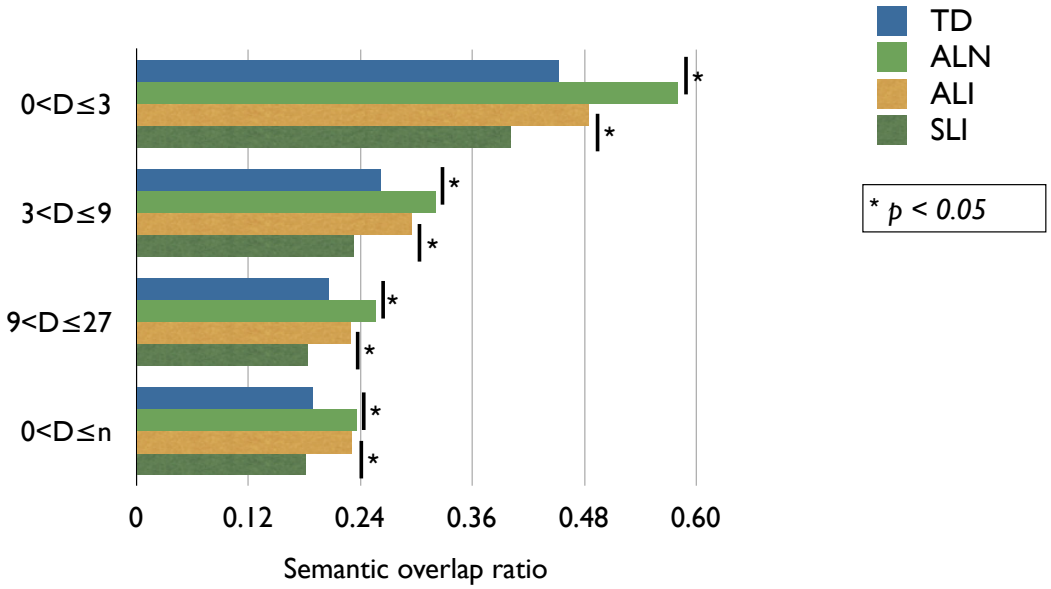



Figure 6.3: Plot of group comparison based on verbatim overlap & lexical relatedness

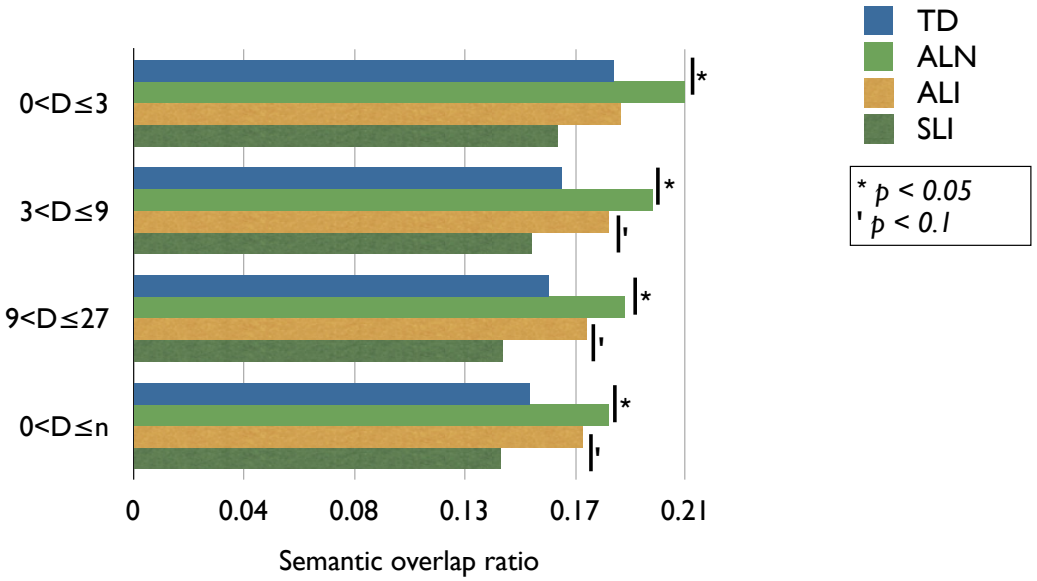



Figure 6.4: Plot of group comparison based on pure lexical relatedness measure

TD and ALI compared to SLI. However, the SOR is significantly higher in ALN compared to TD. An explanation for this observation could be that LI groups tend to use exact sets of words in shorter turn distances as opposed to topically related ones and this may be due to their lack of language ability and lexical diversity (Mainela-Arnold et al., 2010).

In the bigger distance windows we see the consistent pattern of significantly higher SOR in ASD compared to non-ASD groups with the exception of ALI versus SLI, using *PLR* which narrowly missed the standard significance levels in a two-tailed t-test. We do not observe any significant difference in LI compared to non-LI group. More importantly we observe the exact same pattern of group differences in the overall SOR (calculated for each child from their entire transcript) as well.

We can see that *VLR*, and *PLR* –which take into account lexical similarity with and without verbatim word overlap respectively– indicate similar patterns as verbatim word overlap measures (*WJC*, and *RFM*). This observation is reasonably consistent with our expectations that children may use synonyms and semantically similar words (rather than the exact set of words) within the same topic space.

| | *diff* | *lwr* | *upr* | *p-adj* |
|---|---|---|---|---|
| ALN-ALI | 0.301 | -0.97 | 1.568 | 0.926 |
| SLI-ALI | -1.77 | -3.12 | -0.43 | 0.004 |
| TD-ALI | -1.17 | -2.28 | -0.05 | 0.037 |
| SLI-ALN | -2.07 | -3.43 | -0.72 | 0.001 |
| TD-ALN | -1.47 | -2.60 | -0.34 | 0.005 |
| TD-SLI | 0.606 | -0.61 | 1.826 | 0.567 |

diff, the difference in the observed means; lwr, the lower end point of the interval; upr, the upper end point of the interval; *p*-adj, the p-value after adjustment for the multiple comparisons.

Table 6.5: Results of multiple group comparisons based on SOR using Tukey HSD test

Since we are doing multiple comparisons, we also perform Tukey's HSD (Honest Significant Difference) test (Tukey, 1949) that compares all possible pairs of group means based on studentized range distribution. The group-means are based on average –per subject– log scores across multiple windows D and across similarity measures leading into one total score for each group. The results are reported in Table 6.5 and presented in Figure 6.5 and they are again consistent with what we observe in individual windows using various measures.

For children with autism we examined correlations of their overall SOR with chronological age (CA), full-scale IQ (FSIQ), nonverbal IQ (NVIQ), verbal IQ (VIQ), CELF Core Language Score (CLS), ADOS total, and ADOS RRB with Bonferroni Correction method for multiple tests. Results are reported in Table 6.6 and indicate small to moderate (and rarely significant) negative correlations between SOR and CA, FSIQ, NVIQ, VIQ, and CLS. There is a moderate ($r > .30$) correlation between ADOS total and RRB scores (See Sections 3.1,and 3.3 in Chapter 3) and the overall SOR. This correlation is but significant ($p < .05$) for the ADOS RRB score based on verbatim overlap measures ($WJC$, $RFM$). The correlations are lower when we only apply lexical relatedness ($VLR$, $PLR$) and marginally missed the significance level in most cases.

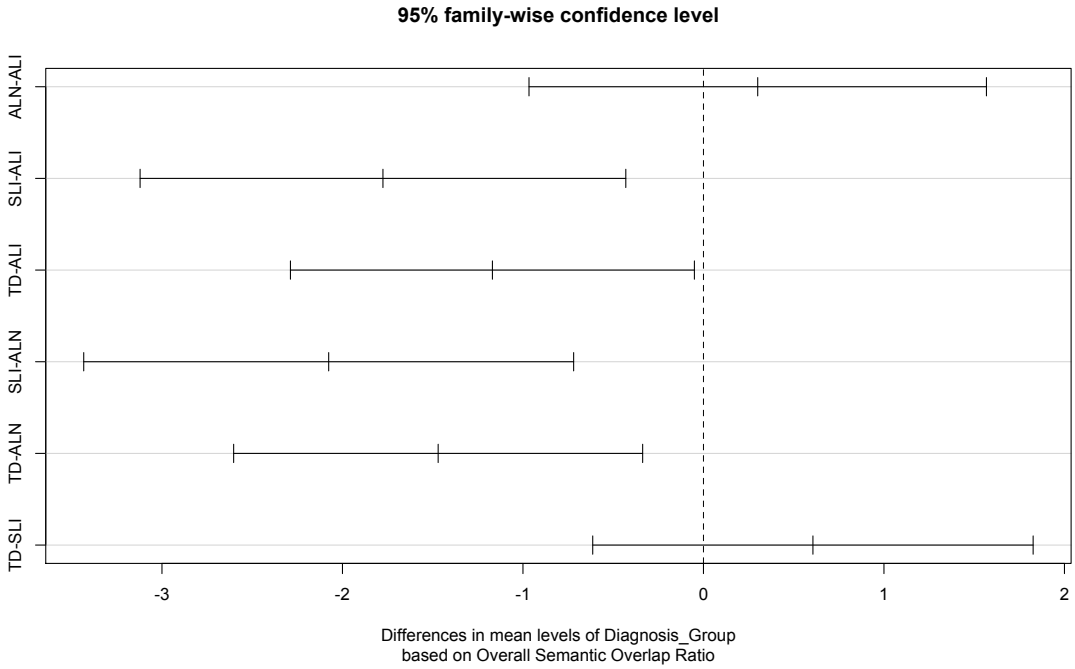



Figure 6.5: Plot of multiple group comparisons based on SOR using Tukey HSD test

| | *WJC* | *RFM* | *VLR* | *PLR* |
|---|---|---|---|---|
| CA | −.26 | −.22 | −.30 | −.26 |
| FSIQ | −.26 | −.28 | −.16 | −.10 |
| NVIQ | −.23 | −.25 | −.25 | −.10 |
| VIQ | −.27 | −.27 | −.30 | −.10 |
| CLS | −.27 | −.30 | −.12 | −.04 |
| ADOS Total | .31 | .33 | .27 | .20 |
| ADOS RRB | .35* | .36* | .31 | .25 |

*$p < .05$

CA, chronological age; FSIQ, full-scale IQ; NVIQ, non-verbal IQ; VIQ, verbal IQ; CLS, CELF Core Language Score.

Table 6.6: Correlations of SOR with cognitive and language measures for children with ASD

## 6.5 Conclusions and future work

In this study we presented automated methods for testing our hypothesis that children with autism have higher semantic overlap between their spontaneous conversational turns compared to their TD and LI peers. Our results support our hypothesis, and are consistent

even when we control for (a) verbatim versus semantically related repeats; and (b) specific language impairment. In addition, we observe moderate but significant correlations between our results and autism severity scores.

The first thing we want to explore in the future work will be examining the impact of changing the lengths of conversational chunks (i.e. $n$) in the obtained results. Given our assumption that each child's ADOS conversation is a sequence of terms representing topics, we expect to achieve similar results to what we presented using various sizes of $n$ other than 10.

A second area to investigate in the future is determining the children's conversation topics, especially the ones that are repeated. One way to do this is to combine the child specificity scores such as *idf* with the highly overlapping lexical items across different turns. To evaluate, we could use manual annotation and clinical impression to determine if a child has a particular (idiosyncratic) topic of interest. We could then compare these annotations with the findings from our automated measures.

A major focus of our future work will be to automatically detect the topics introduced by the examiner to the child. The main assumption of this work is that children with ASD return to a set of topics during their conversation, no matter if they or the examiner initiated the topic. Given the high semantic overlap ratio seen here, we expect that children with autism contribute in conversations related to their particular topic of interest, rather than collaborating with the examiner in a dialogue. This will provide a conversational responsiveness measure based on semantic relevance of the child's responses.

## 6.6 Summary

Restrictive and repetitive behavior is a core symptom of autism spectrum disorder (ASD), and one that can manifest at the language level. Based on this, we expect that children with autism to talk about fewer topics more repeatedly during their conversations. To measure this phenomenon quantitatively, we define the Semantic Overlap Ratio, and we hypothesize a higher semantic overlap ratio between dialogue turns in children with autism

compared to those with typical development (TD) and language impairment (LI). We apply several semantic similarity metrics on the children's dialogues with an examiner during a semi-structured ASD assessment conversation task. We find that the children with ASD have significantly more semantically overlapping dialogue turns compared to those with TD and LI across different turn intervals. These results support our hypothesis, and could provide a convenient and robust ASD-specific behavioral marker.

# Chapter 7

# Quantifying the use of conversational cues

## 7.1 Introduction

Impaired reciprocal social interaction has long been an important facet of Autism Spectrum Disorder (ASD). It is represented at the language level and includes problems in discourse maintenance, and responsiveness to an interlocutor's contribution in dialogue (Baron-Cohen, 2001; Rice et al., 2005). In this chapter we focus on conversational reciprocity in ASD and in particular, we investigate question responsiveness, use of agreement/disagreement responses to questions (in particular yes/no questions), as well as the use of acknowledgments and discourse markers in their appropriate contexts. In the following subsections, we first define these conversational cues, and then give an overview of the existing literature that addresses these phenomena in ASD. As we will show, existing research provides reliable evidence for our analysis from a clinical and biomedical standpoint. However, most existing work has been done manually, requiring tedious annotation.

### 7.1.1 Question responsiveness

Previous research reveals that children with ASD have problems responding to questions addressed to them. Research has shown that they either do not respond, or provide irrelevant, off-topic responses significantly more often than children without autism (Capps

et al., 1998). Responding to questions requires awareness that a) the conversation partner is asking a question, and b) the question requires a (relevant) response. Children with ASD could have problems in one of these or both (Fay, 1982).

A study by Capps et al. (1998) analyzed the question responsiveness of 15 children with ASD within the context of semi-structured conversations and compared them to 15 peers with Developmental Delay, matched on language ability and mental age. The authors found that children with autism were significantly less responsive, often ignoring the questions that were directed to them. Kremer-Sadlik (2004) explored linguistic performance of individuals with high-functioning autism spectrum disorder (HFASD) and Asperger Syndrome (AS, Wing, 1981), when answering everyday questions from their family members. She analyzed 64 hours of video recorded interaction of 16 HFASD and AS children aged 8 to 12 years old, and showed than 15 percent of the time the children ignored the questions addressed to them (no results for TD control are reported in this study). Jones and Schwartz (2009) examined family dinner conversations in 20 HFASD and 10 typically developing (TD) children. They video-recorded and transcribed the conversations between children and their family members. Then they coded the social communicative interactions, including question responsiveness of the children. The children with HFASD were significantly less responsive to questions and, in general, significantly less responsive to any communication bids from family members compared to the TD group.

### 7.1.2 Acknowledgments

Acknowledgments (ACKs) are words, such as 'yes', 'no', 'hmm', 'ok', 'right', or 'uh-huh', that are used to signal agreement, understanding, continued attention, interest, or the start of a new topic. Their appropriate usage requires awareness of the other person's desire to ensure mutual understanding (Jurafsky et al., 1998; Gravano et al., 2012; Heeman et al., 2010; Heeman and Allen, 1999). Acknowledgments are mainly used in after-statement utterances (a post-statement context), but a sub-class of them called **agreement/disagreement** (AD), are used after yes/no (YN) questions (a post-question

context). Their function is to provide either an affirmative or negative answer to the speaker's questions (Jurafsky et al., 1998; Gravano et al., 2012).

Few studies, have analyzed the use of post-statement ACKs in individuals with ASD. Using automated methods, Heeman et al. (2010) studied the use of ACKs in children's ADOS conversations (for more information about the ADOS see Subsection 7.2 in this chapter and Section 3.3 in Chapter 3). Their participants include 22 TD children and 26 with ASD, ranging in age from 4 to 8 years old. The authors did not provide further information about the subjects such as language impairment status in children with ASD, and possible group differences in age, nonverbal, and verbal IQ. They found that TD children used ACKs in 17.42% of their post-statement turns, while the children with ASD did this only 13.39% of the time, a statistically significant difference.

From the early days of conceptualizing autism, Kanner et al. (1943) noticed difficulties in using the word 'yes' as an agreement response to questions. In his later work (Kanner, 1946), he also reported very limited use of the word 'no' by a child with ASD. Researchers have subsequently shown that children with ASD use echolalia or exact repetition of the questions as their agreement or conformation response as opposed to the word 'yes', its variations, or even head nodding (See Fay, 1982, for a discussion).

Huang and Oi (2013) analyzed the responsiveness of HFASD Taiwanese children to several question types (including YN questions) in their conversations with their mothers. The authors examined the childrens response-adequacy to questions in a semi-structured setting. They compared responses from 12 children with HFASD, between the ages of 7.1 to 14.9 years, with 12 typically developing (TD) peers who were matched on age, sex, IQ and mean length of utterance in syllable (MLUs). They found that children with HFASD were less responsive to YN-questions compared to TD children.

### 7.1.3 Discourse markers

Discourse markers (DMs) are words, such as 'and', 'then', 'but', 'well', 'oh', 'so', 'wait', or 'actually', that signal connections between the current utterance and the discourse context. They are linguistic elements with procedural, and not conceptual or literal meaning that is interpreted by the dialogue context. Their main roles are creating a natural and fluent conversation, guiding a listener through a speaker's chain of thought, maintaining a listener's understanding of the conversation, and signaling an explicit relationship between two dialogue utterances (Schiffrin, 1988; Fox Tree and Schrock, 1999; Flowerdew and Tauroza, 1995; Jucker, 1993; Schourup, 1999; Bolden, 2006). Appropriate use of DMs contributes to a successful conversation.

DMs have been extensively studied in typical language, however few studies have focused on their analysis in the language of ASD individuals. A study by Solomon (2001) shows that children with HFASD use discourse markers such as 'and', 'but', and 'so', but not in their appropriate context. They often use them to connect topically irrelevant parts of the dialogue.

Morrison (2014) reports a quantitative (but not automated) analysis of pragmatic language including the use of DMs in adults with HFASD. The analysis was performed on annotated transcripts of guided discussions with 8 HFASD and 8 TD participants. The author found that the TD group used significantly more DMs (specifically the word 'like') than the HFASD but there was no difference in the use of DM 'you know'.

The most similar study to our current work is by Heeman et al. (2010) (see 7.1.2 for more information about the study). They described an automatic method for quantifying the interactional aspects of dialogue in children with ASD, including the use of DMs. They found significantly higher use of DMs in TD children compared to their ASD peers.

### 7.1.4 Measuring the use of conversational cues

Conversational responsiveness in ASD individuals is often measured using unstructured, semi-structured, or fully structured conversational tasks. Unlike topic repetition and idiosyncratic language, measures for conversational responsiveness and conversational cues in ASD are usually objective and not qualitative (as we have seen above). However, most of these studies are not automated, and they require manual coding by trained coders and time-consuming annotations. As a result, few of them (such as Heeman et al. (2010)) have addressed large-scale comprehensive conversational analysis in ASD.

Computational methods for conversational analysis could be very helpful to automate this resource-demanding annotation process. There has been a lot of work on automated analysis of conversations in typical populations (Hirschberg and Litman, 1993; Gravano et al., 2007, 2012; Jurafsky et al., 1998; Heeman et al., 2010; Heeman and Allen, 1999; Traum and Hinkelman, 1992; Delaherche et al., 2013). Most of these studies can potentially be extended and applied to the conversations of individuals with ASD. The algorithms in these studies provide fast, objective and quantitative assessments, and they can be applied to unannotated raw transcripts.

### 7.1.5 Goals of this study

In this study we plan to automatically quantify the appropriate use of conversational cues in a semi-structured interactive dialogue task between children and an adult examiner. We analyze the overall question responsiveness of a child along with the use of acknowledgement and discourse markers in various post-question, and post-statement contexts. Based on the current research in the autism literature that presented above, we expect children with ASD to be less responsive to an examiner's questions, and use the conversational cues less appropriately in their dialogue when compared to those with typical development. We expect to see such problems in children with ASD regardless of their language impairment status (i.e. autism with and without language impairment).

In particular, we follow these three goals in this chapter:

1. Automatically detecting the use of ACKs and DMs in conversational contexts.
2. Automatically classifying examiner's questions into yes/no or WH-questions.
3. Applying our tools to our corpus of ADOS transcripts to test our hypothesis that children with ASD use conversational cues less appropriately than TDs.

While this work is inspired by the Heeman et al. (2010) study, we will extend their research in various ways. Our study includes more subjects and more diagnostic groups, including children with ASD with language impairment, and children with SLI. In addition we will study question responsiveness, and the use of ACKs and DMs in more elaborate contexts (in particular post-question). We also present a rule-based classifier for identifying yes/no and WH-questions.

## 7.2 Participants and data

The participants of this study are the same as Chapter 6 and include 44 children with TD, 25 children with ASD without language impairment (ALN), 26 children with ASD and language impairment (ALI), and 20 children diagnosed with Specific Language Impairment (SLI). For more information about the participants, the diagnosis processes, and other measures refer to Chapter 3.

We analyze each participant's conversation responsiveness in the context of the ADOS (see Chapter 3, Section 3.3), where the examiner provides the same opportunity for the ASD and TD children to participate in the dialogue and respond to her various conversational interactions and reciprocal actions. Our expectations are that the children with ASD –compared to their TD peers– will be less engaged in the conversations, have fewer interactions with the examiner, and use improper conversational cue words.

| | |
|---|---|
| Acknowledgement (ACK) | a) One of the words 'yeah', 'yep', 'yes', 'no', 'nope', 'mmhmm', 'mm', 'ok', 'kay', 'right', 'hm', 'hmm', 'uhhuh'<br>b) Does not follow a question by the examiner. |
| Agree/disagree (AD) | a) Same set of words as acknowledgements<br>b) Follows a yes-no-question by the examiner. |
| Discourse marker (DM) | a) One of the words 'and', 'then', 'but', 'well', 'oh', 'so', 'wait', 'actually'.<br>b) Follows a question or a statement by the examiner. |

Table 7.1: Analytical criteria for conversational cue

## 7.3 Methods

In this section we describe our analysis methods for the ADOS transcripts. We explain the criteria for conversational cues and our analytical contexts, and provide a brief description of our question classifier.

### 7.3.1 Analytical criteria for conversational cues

We presented the definition for ACKs, AD question responses, and DMs in Section 7.1. We define the criteria for each of these conversational cues in Table 7.1 based on the work of Heeman and Allen (1999). All conversational cues defined here, must be the first word in the child's turn (i.e. conversational turn consisting one to several utterances).

### 7.3.2 Question classification

In order to analyze the use of conversational cues in more specific contexts, we classify the examiner's question into one of two types of **YN** (yes/no, e.g. "Have you ever been to Texas?") and **WH** questions (e.g. "What are the kinds of things that make you feel sad?". To do this, we built a rule-based question classifier that considers the occurrence and the position of the following classes of words in an utterance: a) WH or interrogative words such as 'what', 'who', and 'how', b) copulas such as 'am', 'is', and 'are'; c) modals such as

'can', 'may', and 'should'; d) auxiliaries such as 'do', 'does', and 'did'. The classifier can differentiate YN and WH questions with an accuracy of 97.57%. The reported accuracy is based on a gold standard set of 200 randomly selected questions (around 1% of all questions) manually annotated by two human annotators

### 7.3.3 Analytical contexts

As summarized in Table 7.2, we define multiple conversational contexts in order to analyze the appropriate use of conversational cues in different positions in the dialogue.

#### Post-question context

The first context we analyze is the post-question context. We begin with the analysis of the ratio of unanswered questions: the ratio of the number of questions unanswered by the child, divided by the total number of questions. As we classify the examiner's question types, we report the ratio of unanswered YN, WH, and both types of questions.

We then analyze the ratio of agree/disagree responses in the post-question context. This is the number of the child's agree/disagree responses, divided by the total number of questions the child responded to. In particular we are interested in the post YN-question responses since they require either an affirmative or a negative answer. Nevertheless we also report the ratio for WH and all questions combined.

The third analysis in the post-question context is the ratio of DM responses. This ratio is the number of child's responses to questions initiated with a discourse marker, divided by total number of questions responded to by the child. As above we report the ratios for different question types.

#### Post-statement context

Since, by definition, ACKs and DMs occur in post-statement context (see Subsections 7.1.2 and 7.1.3), our final analysis will focus on the ratio (or proportion) of ACK and DM usage

| | | |
|---|---|---|
| **Post-question context** | | |
| No response | WH | **E:** What do you do when you feel angry?<br>**C:** [No response]. |
| | YN | **E:** Are you an astronaut?<br>**C:** [No response]. |
| Agree/disagree | WH | **E:** What's going to happen?<br>**C:** Yeah, yeah, I can. |
| | YN | **E:** Have you ever been up to Mount Hood?<br>**C:** No, I've been to Cape Kiwanda and Cape Lookout. |
| Discourse marker | WH | **E:** What happens to it when it breaks open?<br>**C:** Well, it falls. |
| | YN | **E:** Are you thinking of another conestoga?<br>**C:** Actually, it's a pool. |
| **Post-statement context** | | |
| Acknowledgement | | **E:** He has a parachute on.<br>**C:** Yeah. |
| Discourse marker | | **E:** There's a man telling the story.<br>**C:** And there's the cat from the grandma's house. |

Table 7.2: Analytical contexts for the quantifying conversational reciprocity

in such context. For ACKs, this ratio is defined as number of the child's turns starting with an acknowledgement following the examiner's statement turns, divided by the total number of the child's turns following the examiner's statements. Similarly, for DMs, it is the number of the child's turns starting with a discourse marker following the examiner's non-question turns, divided by the total number of the child's turn following the examiner's statements.

## 7.4 Results

As explained in the previous section, we calculate the ratios for using conversational cues in different analytical contexts in different diagnosis groups. We then test the significance of between-group differences by performing Welch's two-sample t-test, with the assumption that our groups have unequal variances and unequal sample sizes. For each analytical

context, and each conversational cue, we report the mean ratio for each group (with standard deviation) as well as the significance test results: T-Statistics ($t$), P-Value ($p$), and Cohen's D for effect size ($d$).

### 7.4.1 Conversational responsiveness in post-question context

#### Overall question responsiveness

As we can see in Table 7.3 and Figure7.1, the ASD group responds to an examiner's questions significantly less than the TD group (i.e. ASD has higher unanswered question ratio). The results are similar regardless of question type (AllQ, WH, or YN). We see the same pattern in the ALN group compared to TD, although the difference narrowly missed the significance level in YN questions. This indicates that children with ASD are generally less responsive to the examiner's questions regardless of their language impairment status. However, we did not observe any significant difference between ALI compared to ALN or ALI compared to SLI. Given the heterogeneity of language impairment diagnosis, these findings should be interpreted with caution (See Chapter 3 for more discussion on the language impairment diagnosis).

#### Agreement/disagreement responses

Table 7.4 and Figure7.2 demonstrate the usage of agreement/disagreement (AD) responses. The ASD (and ALN) group used significantly fewer AD responses to YN-questions compared to the TD group ($p < .01$). As YN questions are acknowledgment-soliciting, responses should begin with an AD term. The ASD group failed to do this. On the other hand, no significant difference can be seen between ASD (and ALN) compared to TD in the number of AD responses to the examiner's WH-questions. This lack of differentiation is potentially due to the fact that WH-questions are asking for information and not for affirmation or negations which is the only focus of this analysis. In other words, we do not have enough information to investigate the group differences in post WH-question contexts.

| | | Mean Unanswered Q Ratio (SD) | | | | |
|---|---|---|---|---|---|---|
| *Contrast* | *Similarity* | *Group1* | *Group 2* | *t* | *p* | *d* |
| ASD vs TD | AllQ | .4005 (.0942) | .3343 (.0643) | 4.04 | 0.0 | .83 |
| | WH | .3867 (.0935) | .3139 (.07) | 4.33 | 0.0 | .89 |
| | YN | .4103 (.1066) | .3483 (.0779) | 3.27 | .002 | .67 |
| ALN vs TD | AllQ | .3861 (.087) | .3343 (.0643) | 2.6 | .013 | .65 |
| | WH | .3685 (.0714) | .3139 (.07) | 3.08 | .003 | .77 |
| | YN | .3982 (.1106) | .3483 (.0779) | 1.99 | .053 | .5 |
| ALI vs. ALN | AllQ | .4143 (.1004) | .3861 (.087) | 1.07 | .289 | .3 |
| | WH | .4041 (.1092) | .3685 (.0714) | 1.38 | .173 | .39 |
| | YN | .422 (.1035) | .3982 (.1106) | .79 | .432 | .22 |
| ALI vs. SLI | AllQ | .4143 (.1004) | .3819 (.1018) | 1.07 | .289 | .32 |
| | WH | .4041 (.1092) | .3755 (.0799) | 1.03 | .31 | .31 |
| | YN | .422 (.1035) | .3853 (.1266) | 1.05 | .3 | .31 |
| SLI vs. TD | AllQ | .3819 (.1018) | .3343 (.0643) | 1.93 | .065 | .52 |
| | WH | .3755 (.0799) | .3139 (.07) | 2.97 | .006 | .8 |
| | YN | .3853 (.1266) | .3483 (.0779) | 1.21 | .238 | .33 |

Table 7.3: Results of paired-group contrasts in unanswered question ratio

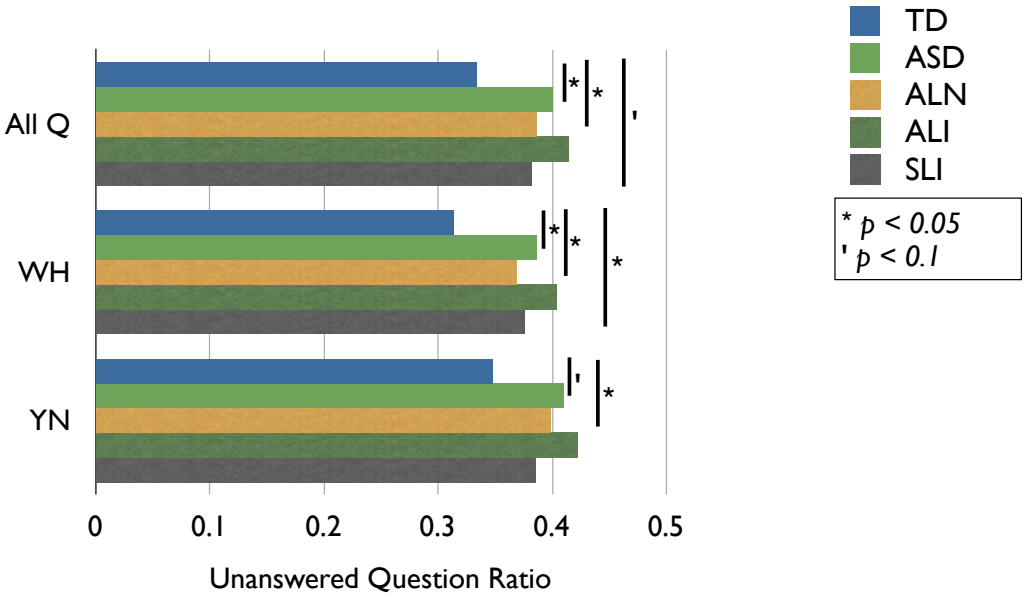



Figure 7.1: Plot of group comparison in unanswered question ratio

| | | Mean Post-Q AD Ratio (SD) | | | | |
|---|---|---|---|---|---|---|
| *Contrast* | *Similarity* | *Group1* | *Group 2* | *t* | *p* | *d* |
| ASD vs TD | AllQ | .3146 (.0787) | .364 (.0613) | −3.43 | .001 | −.71 |
| | WH | .1051 (.0443) | .1139 (.0556) | −.84 | .402 | −.17 |
| | YN | .4806 (.1175) | .5563 (.0861) | −3.61 | 0.0 | −.74 |
| ALN vs TD | AllQ | .3063 (.0891) | .364 (.0613) | −2.87 | .007 | −.72 |
| | WH | .1009 (.0439) | .1139 (.0556) | −1.07 | .288 | −.27 |
| | YN | .4713 (.1317) | .5563 (.0861) | −2.89 | .006 | −.72 |
| ALI vs. ALN | AllQ | .3226 (.0681) | .3063 (.0891) | .73 | .469 | .2 |
| | WH | .1092 (.0451) | .1009 (.0439) | .67 | .508 | .19 |
| | YN | .4895 (.1038) | .4713 (.1317) | .55 | .588 | .15 |
| ALI vs. SLI | AllQ | .3226 (.0681) | .3348 (.0713) | −.59 | .559 | −.18 |
| | WH | .1092 (.0451) | .1021 (.049) | .5 | .617 | .15 |
| | YN | .4895 (.1038) | .5135 (.1143) | −.73 | .468 | −.22 |
| SLI vs. TD | AllQ | .3348 (.0713) | .364 (.0613) | −1.58 | .124 | −.43 |
| | WH | .1021 (.049) | .1139 (.0556) | −.86 | .396 | −.23 |
| | YN | .5135 (.1143) | .5563 (.0861) | −1.49 | .146 | −.4 |

Table 7.4: Results of paired-group contrasts in agreement/disagreement response

There is no significant difference in ALI versus ALN, ALI versus SLI, or SLI versus TD in the number of agree/disagree responses to any type of examiner's question (AllQ, WH, and YN). This reveals that only children with ASD (but not SLI) have problems in providing AD responses to YN-questions.

**Discourse markers**

As we see in Table 7.5, there is no significant difference in all group comparisons in the use of discourse markers in response to the questions of any type. The only exception is that the SLI group has significantly fewer DMs after Wh-questions as compared to TD. Overall, we do not see any performance difference in post-question DM use. Almost all groups perform similarly in this particular context.

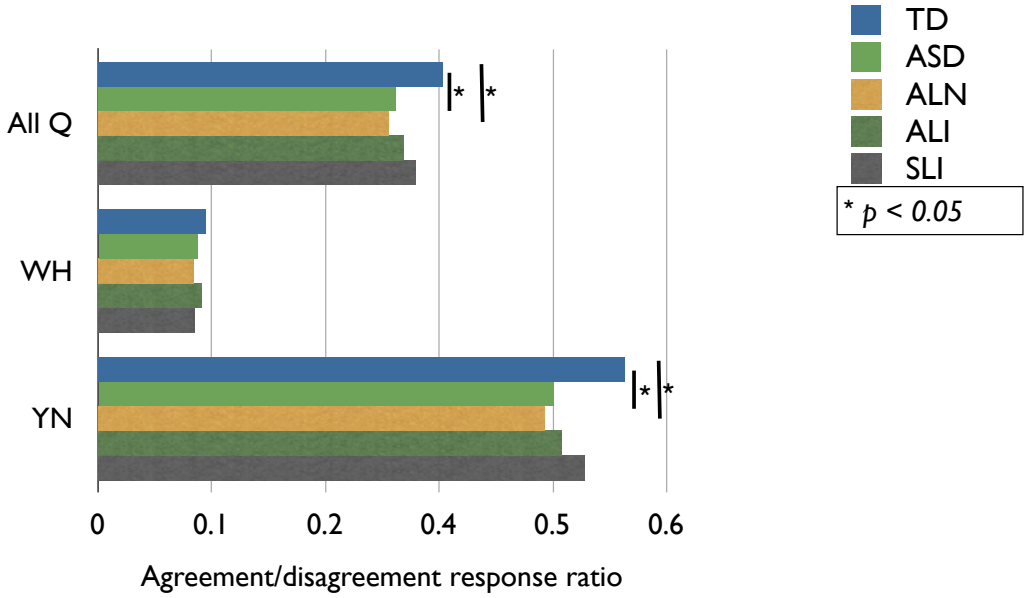



Figure 7.2: Plot of group comparison in agreement/disagreement response

| | | Mean Post-Q DM Use Ratio (SD) | | | | |
|---|---|---|---|---|---|---|
| *Contrast* | *Similarity* | *Group1* | *Group 2* | *t* | *p* | *d* |
| ASD vs TD | AllQ | .0829 (.0631) | .0809 (.044) | .18 | .855 | .04 |
| | WH | .0944 (.0745) | .1001 (.064) | −.4 | .688 | −.08 |
| | YN | .0739 (.0633) | .0657 (.0434) | .75 | .458 | .15 |
| ALN vs TD | AllQ | .0978 (.0725) | .0809 (.044) | 1.06 | .296 | .27 |
| | WH | .109 (.087) | .1001 (.064) | .45 | .656 | .11 |
| | YN | .0893 (.071) | .0657 (.0434) | 1.51 | .141 | .38 |
| ALI vs. ALN | AllQ | .0686 (.05) | .0978 (.0725) | −1.67 | .103 | −.47 |
| | WH | .0803 (.0586) | .109 (.087) | −1.38 | .176 | −.39 |
| | YN | .0592 (.052) | .0893 (.071) | −1.72 | .092 | −.48 |
| ALI vs. SLI | AllQ | .0686 (.05) | .0709 (.0378) | −.17 | .862 | −.05 |
| | WH | .0803 (.0586) | .0708 (.0407) | .65 | .52 | .19 |
| | YN | .0592 (.052) | .0698 (.0449) | −.74 | .462 | −.22 |
| SLI vs. TD | AllQ | .0709 (.0378) | .0809 (.044) | −.93 | .356 | −.25 |
| | WH | .0708 (.0407) | .1001 (.064) | −2.21 | .031 | −.6 |
| | YN | .0698 (.0449) | .0657 (.0434) | .34 | .735 | .09 |

Table 7.5: Results of paired-group contrasts in post-questions DM use

| | | Mean ACK and DM Use (SD) | | | | |
|---|---|---|---|---|---|---|
| *Contrast* | *Conv Cue* | *Group1* | *Group 2* | $t$ | $p$ | $d$ |
| ASD vs TD | ACK | .1491 (.0825) | .2046 (.0815) | −3.29 | .001 | −.68 |
| | DM | .1712 (.084) | .229 (.0798) | −3.43 | .001 | −.71 |
| ALN vs TD | ACK | .1536 (.0832) | .2046 (.0815) | −2.47 | .017 | −.62 |
| | DM | .1797 (.0857) | .229 (.0798) | −2.35 | .023 | −.59 |
| ALI vs. ALN | ACK | .1449 (.0832) | .1536 (.0832) | −.37 | .712 | −.1 |
| | DM | .1630 (.0832) | .1797 (.0857) | −.71 | .483 | −.2 |
| ALI vs. SLI | ACK | .1449 (.0832) | .1415 (.0733) | .15 | .885 | .04 |
| | DM | .163 (.0832) | .1745 (.0803) | −.47 | .638 | −.14 |
| SLI vs. TD | ACK | .1415 (.0733) | .2046 (.0815) | −3.08 | .004 | −.83 |
| | DM | .1745 (.0803) | .229 (.0798) | −2.52 | .016 | −.68 |

Table 7.6: Results of paired-group contrasts in post-statement ACK and DM use

### 7.4.2 Responsiveness in post-statement context

Table 7.6, and Figure7.3 show the pair-wise group contrasts in the use of ACKs and DMs in the post-statement context. We see that the ASD and the ALN groups were significantly less likely to use ACKs and DMs after examiner's statements compared to the TD group. No significant difference exists between ALI versus ALN, and ALI versus TD. However, SLI groups used significantly less ACKs and DMs than the TD group. These findings suggest that children with autism have problems in (a) signaling agreement due to lack of use of ACKs, and (b) making connections between dialogue turns due to problems in using DMs. These problems are not dependent to the language ability in the ASD groups. Not surprisingly, we observe similar difficulties in children with SLI as well. As the previous literature suggests, SLI children have problems in pragmatic aspects of language due to their problems with structural and semantic aspects of language (Rice et al., 2005).

### 7.4.3 Multiple comparisons with various scores

In addition to comparing groups pairs based on individual scores, we perform multiple comparisons using Tukey's HSD (Honest Significant Difference) test (Tukey, 1949). This

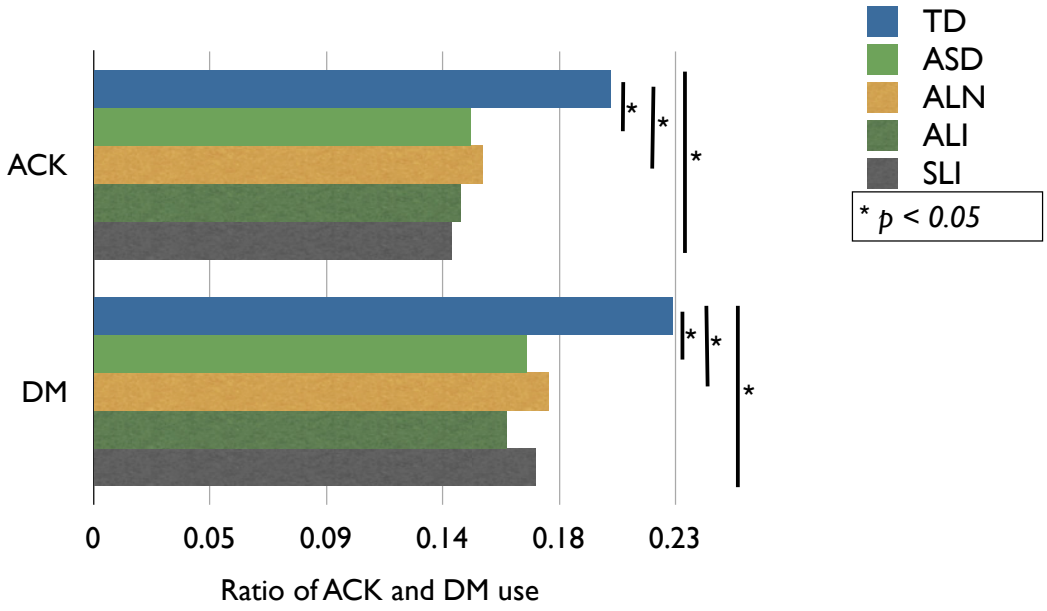



Figure 7.3: Plot of group comparison in post-statement ACK and DM use

test compares all possible pairs of group means according to studentized range distribution. The group-means are calculated using average –per subject– log scores across multiple conversational responsiveness measures resulting in one total score for each group. The results are reported in Table 7.8 and presented in Figure 7.4.

### 7.4.4 Correlations with cognitive and language measures

For children with ASD we examined correlations of their conversational responsiveness ratios –including no response to questions, agreement/disagreement after yes/no questions, post-questions discourse markers, acknowledgements use, and discourse marker use– with CA, FSIQ, NVIQ, VIQ, CLS, SCQ total, SCQ SRI, ADOS total, and ADOS SA score. We applied Bonferroni Correction method for $n = 45$ tests. Hence, the cutoff at a corrected $p$-value of 0.05 is 0.42 one-tailed.

Results are shown in Table 7.7. We can see small-to-moderate significant negative correlation between the unanswered question ratio and VIQ. However, the correlation is positive

| | No Res Q | AD-YNQ | Post-Q DM | ACK | DM |
|---|---|---|---|---|---|
| CA | −.30 | −.35 | .42* | .35 | .24 |
| FSIQ | −.26 | .00 | .21 | .08 | .14 |
| NVIQ | −.07 | −.11 | .16 | .08 | .10 |
| VIQ | −.40 | .01 | .24 | .09 | .19 |
| CLS | −.34 | .07 | .17 | .17 | .11 |
| SCQ Total | .09 | −.08 | .12 | .07 | −.12 |
| SCQ SRI | .10 | −.03 | .10 | .07 | −.10 |
| ADOS Total | .47* | −.20 | −.30 | −.26 | −.28 |
| ADOS SA | .28 | −.10 | −.30 | −.08 | −.21 |

*$p < .05$

No Res Q, no response to questions; AD-YNQ, agreement/disagreement after yes/no questions; Post-Q DM, post-questions discourse marker use; ACK, acknowledgements use ratio; DM, discourse markers use ratio; CA, chronological age; FSIQ, full-scale IQ; NVIQ, nonverbal IQ; VIQ, verbal IQ; CLS, CELF Core Language Score; SCQ SRI, SCQ social reciprocal interaction domain sum score; ADOS SA, ADOS social affect score.

Table 7.7: Correlations of conversational responsiveness ratios with cognitive and language measures children with ASD

| | *diff* | *lwr* | *upr* | *p-adj* |
|---|---|---|---|---|
| ALN-ALI | -0.216 | -0.872 | 0.440 | 0.825 |
| SLI-ALI | -0.295 | -0.992 | 0.401 | 0.687 |
| TD-ALI | -0.965 | -1.545 | -0.386 | 0.000 |
| SLI-ALN | -0.079 | -0.782 | 0.624 | 0.991 |
| TD-ALN | -0.749 | -1.336 | -0.162 | 0.006 |
| TD-SLI | -0.670 | -1.302 | -0.038 | 0.033 |

diff, the difference in the observed means; lwr, the lower end point of the interval; upr, the upper end point of the interval; *p*-adj, the p-value after adjustment for the multiple comparisons.

Table 7.8: Results of multiple group comparisons based on conversational responsiveness measures using Tukey HSD test

and significant between unanswered question ratio and the ADOS total score. The agreement/disagreement responses after YN questions have no significant negative correlation with any score. The post-question DM use shows positive significant correlation with CA and non-significant negative correlation with the ADOS total and the ADOS SA scores. The use of ACKs and DMs are not significantly correlated with any scores.

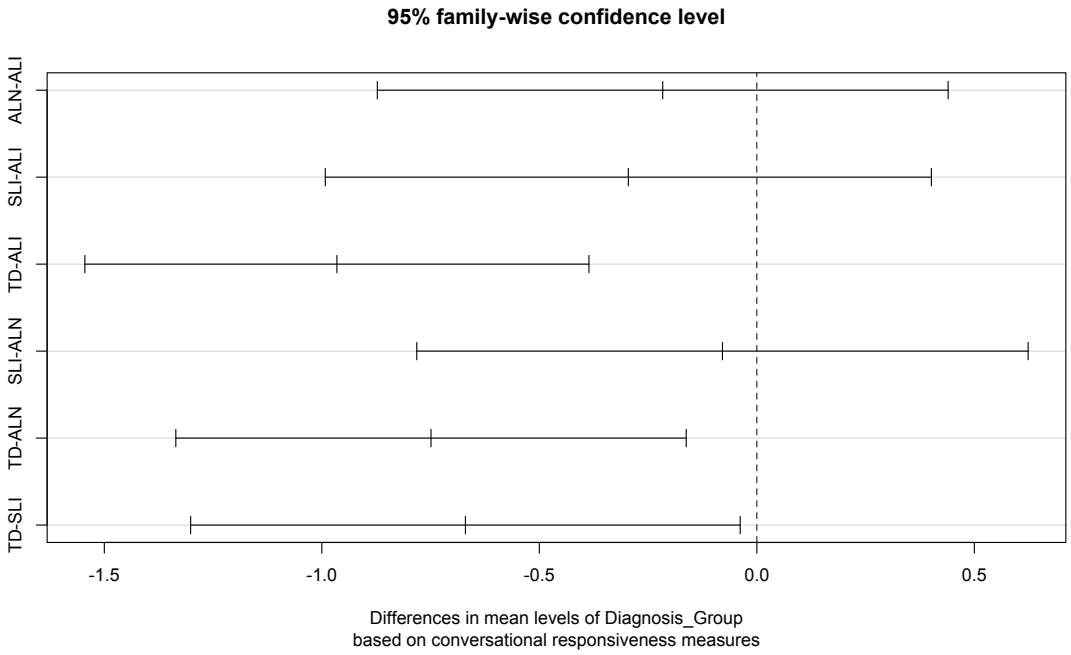



Figure 7.4: Plot of multiple group comparisons based on conversational responsiveness measures using Tukey HSD test

## 7.5 Conclusions and future work

In this study we provided an automated and quantitative analysis for testing the hypothesis that children with autism have difficulties in using conversational cues, in particular question responsiveness, use of agreement/disagreement, acknowledgements and discourse markers. We applied our methods to transcripts of semi-structured conversational tasks from typically developing children, children with and without ASD and language impairment.

The obtained results show patterns such as high ratio of unanswered questions, lack of agreement/disagreement responses, and inappropriate use of ACKs and DMs in children with ASD regardless of their language impairment status. While we observed similar patterns such as problems with using ACKs and DMs in children with SLI, it is suggested in the literature that such problems with discourse and pragmatics are usually secondary to deficits in semantic and syntax in this group (Rice et al., 2005).

Similar to our previous findings is this thesis, the methods and results presented in this chapter demonstrate the utility of computational methods for the evaluation of pragmatic language use in autism. These methods –in combination with our other findings– can be applied to provide automated, fast, objective, reliable and easy-to-use diagnostic tools for autistic language.

Future work will focus on disambiguating sentential and discourse uses of conversational cues, as well as identifying their dialogue acts. Hirschberg and Litman (1993) investigated text-based and prosodic features such as pitch-accent and prosodic phrasing for sense disambiguation of lexical items, such as 'well', that could be used as an explanatory comments or for initiating a response to the previous topic. Jurafsky et al. (1998) discussed the utility of lexical, syntactic and prosodic features for identifying subtype dialogue acts of acknowledgements, including continuers, assessments, incipient-speakership, and agreements. Such sense disambiguation and dialogue act classification of conversational cues could provide a more sophisticated analysis of the way that children with ASD and typical development use conversational cues in their dialogues. We are also interested to explore the use of conversational cues in other contexts than turn-initial.

Another area to investigate is the analysis of response types to questions. Previous research has shown that children with ASD may only repeat a YN-question as an affirmative response, rather than using the word 'yes' (Fay, 1982). To identify such echolalic affirmative responses, we can use the work by van Santen et al. (2013) on automated analysis of echolalia and self repeats. Furthermore, previous research has indicated that children with ASD provide significantly high numbers of off-topic and irrelevant responses to questions (Capps et al., 1998). The work on topic repetition in Chapter 6 could be useful for identifying semantic relatedness of children's responses to the examiner's questions. In addition, in order to distinguish echolalic responses from the topically related ones, we can utilize frequency-based term weighting and knowledge-based lexical similarity measures as discussed in Chapter 6.

## 7.6 Summary

Children with autism have problems in social communication and conversational reciprocity. These problems have rarely been investigated on natural language samples in the autism literature. In this chapter we applied computational text analysis tools for fast, quantitative, objective characterization of conversational reciprocity in ASD. Our findings support our hypothesis that children with ASD are less responsive to questions compared to TD children, have problems in agreement/disagreement responses to yes/no questions, and have difficulties in context-appropriate use of acknowledgments and discourse markers. The results also show these problems in the ASD group are not due to their language impairment status.

# Chapter 8

# Conclusions

## 8.1 Summary

The primary goal of this thesis was to explore the utility of automated natural language processing techniques in providing fast, objective, and scalable methods for the analysis and characterization of language use in autism. We have shown applications of our methods on different aspects of pragmatic language use in ASD: detecting idiosyncratic words and semantic digression in narratives, and characterizing topic repetition and conversational reciprocity in semi-structured dialogues.

We first presented a method to automatically identify irrelevant content and unexpected words in narrative retellings by children with autism. Our methods utilized common word ranking and lexical association techniques used in the information retrieval and information extraction tasks of document clustering and topic modeling. In these tasks, words used with relatively higher frequency in a given topic, such as "sports", are ranked higher in documents related to that topic. More common words, such as function words, are ranked lower because of their frequent occurrences across different documents and topics. In our narrative corpus, however, we only have a single topic: the short NEPSY Narrative Memory (NNM) story. Words related to that topic should appear frequently in all narrative retellings by different children. Thus, words that are not frequently used in retellings, achieve a higher relative frequency rank, and are good candidates to be unexpected or inappropriate in the context of the NNM story. It is also likely that children use different set of words that are nevertheless topically relevant to the NNM story. In

order to filter those words from the list of potentially unexpected words, we expand their meaning using corpus-based and knowledge-based approaches, and then look for their associated or synonymous words in our retelling corpus. We evaluated our methods against a set of manually identified unexpected words. Our automated methods for identifying the unexpected words are very accurate, especially when we combine word ranking with knowledge-based semantic expansion (87.80% precision, 94.74% recall, and 91.14% F-measure). Both manual and automated methods indicated a significantly higher ratio of irrelevant and idiosyncratic words in narratives from children with ASD compared to those from typically developing ones.

We then focused on topic digressions in narrative retellings of children with ASD. Given the large amount of off-topic content in their narratives, we investigate whether the direction of this departure from the target topic is similar across all children with ASD, or whether each of them chooses an idiosyncratic and different direction influenced by their restricted interests. We compared the narratives of children with ASD to one another, and with narratives of typically developing children. As a control, we also compared the narratives of TD children with one another. Between each pair of retellings, we extracted several measures designed to capture different facets of semantic similarity, including lexical overlap (such as Cosine similarity and BLEU score), and knowledge-based measures with semantic expansion (such as WordNet-based mutual similarity). Applying the Monte Carlo permutation method, we found that for all types of similarity measures, narratives of typically developing controls are the most similar to each other. We found significantly less similarity between children with ASD and typical development, and even less similarity among children with ASD when compared to each other. These observed differences suggest that children with ASD depart from the narrative topic following their own directions.

Next, we analyzed the presence of restrictive and repetitive behavior and, in particular, topic repetition in conversations of children with ASD. We expected a higher semantic overlap ratio between dialogue turns in children with autism compared to those with typical development in semi-structured ADOS conversations. In order to calculate the

semantic similarity at different turn intervals, for each child we compared every turn pair at four different distance windows, and averaged the turn similarities at each window to calculate a window-specific semantic overlap ratio. To compare turn pairs, we utilized three types of turn similarity measures: verbatim word overlap, verbatim word overlap integrated with lexical relatedness, and lexical relatedness without verbatim overlap. In all measures, we assigned higher weights to child-specific words and lower weights to frequent words, using inverse document frequency term weights. We obtained consistent results using different combinations of turn distance windows and similarity measures. In all combinations we observed a higher semantic overlap ratio in children with ASD (with and without language impairment) compared to their typically developing peers, and in the majority of cases this difference was significant. These findings indicate significantly more topic repetition and instances of sameness in the ASD group, regardless of the their language impairment status.

Finally, we investigated conversational reciprocity of children with ASD. We started with building a very accurate rule-based system to classify questions into yes/no and WH, and then we measured the question responsiveness of our participants. We found that children in the ASD group, on average, were significantly less likely to respond to an examiner's question. These results were the same for both YN and WH question types. We then explored the use of agreement/disagreement responses to examiner's YN questions and, as expected, we found that children with autism were less likely to provide such responses than typically developing children. Next, we focused on the use of acknowledgments (terms such as "yes", "no", "hmm", "right"), and discourse markers (terms such as "and", "then", "well") in post-statement context and our results showed patterns of atypical use of such conversational cues in children with ASD. The same results were observed in both groups of children with ASD, i.e. those with and without language impairment.

An important aspect of our results in the analysis of topic repetition and conversational reciprocity is that the observed differences between children with ASD and typical development could not be attributed to group differences in chronological age, structural

language abilities, or Performance IQ. This suggests that our findings are likely to be ASD-specific diagnostic language features.

## 8.2 Future work

### 8.2.1 Extensions

The first major focus of our future work will be to manually annotate our data to objectively validate our methods and results. While it is true that we achieved similar and significant results using various approaches and measures, validating our results against human-generated gold-standard data will provide stronger evidence for their performance and reliability. We are interested particularly in manual annotation of topic digressions in narratives (Chapter 5), topic repetition in conversations (Chapter 6), and different functions of conversational cues (Chapter 7).

Second, we plan to apply all of our methods to the output of an automatic speech recognition (ASR) system rather than manual transcripts. The ASR output is likely to contain word errors, especially in children's spontaneous speech. However, the fact that our distributional semantic models focus mainly on content words may make them robust to the sorts of function word recognition errors typically produced by ASR systems (Fitzgerald, 2009). The analysis of conversational cues will be an interesting challenge given that most of those terms are turn-initial functions words.

Third, we hope to explore the combination of methods presented in this thesis with the analysis of prosody in autism. Many individuals with autism show patterns of atypical prosody in their speech (McCann and Peppé, 2003) and existing methods for characterizing this atypicality are performing well-above chance (Van Santen et al., 2010; Kiss et al., 2012). An interesting research question will be to determine the degree to which instances of atypical language and atypical prosody co-occur and overlap with each other. Additionally, we can investigate the intonation and pitch variations of different question types in conversational speech of children with ASD. In the neurotypical population we

generally expect YN- and WH-questions to have rising and falling intonation, respectively. To examine question intonation in children with autism, we can combine our highly accurate question classification methods with presented methods by Kiss et al. (2012) for characterizing intonation at the utterance level.

We have applied various measures in the analysis of semantic similarity between words, narrative retellings and conversational turns. Nevertheless, it is reasonable to try additional measures. Since most of our lexical similarity measures have been knowledge-based, the fourth extension of our work will be applying corpus-driven lexical similarity and lexical relatedness measures such as Pointwise Mutual Information (PMI, Church and Hanks, 1990). The main advantage of this family of measures over the knowledge-based methods is that they are superior in capturing semantic 'relatedness' of lexical items in addition to semantic similarity. Two lexical items such as 'ball' and 'play' may be topically well-related, but such instances of relatedness are not usually represented in WordNet lexical relations. We also plan to apply other data-driven sentential/document similarity measures such as Latent Semantic Analysis (LSA, Landauer and Dumais, 1997) and Latent Dirichlet Allocation (LDA, Blei et al., 2003) for comparing narratives and conversation turns.

### 8.2.2 Applications in research and clinical practice

The automated methods we have developed in this thesis have significant potential to be applied in clinical practice and research settings. For instance, our methods for detecting idiosyncratic off-topic lexical content, and topic digression can easily be extended for the semantic analysis of other narrative elicitations from commonly used wordless picture books by clinicians and researchers such as "Frog, where are you?" (Mayer et al., 1969), "Tuesday" (Wiesner, 1991), and "Good night, gorilla" (Rathmann, 1994).

In addition, we have been working on combining our pragmatic analysis of narratives with (a) word alignment methods for narrative scoring (Prud'hommeaux, 2012; Prud'hommeaux and Roark, 2015 in press), (b) automated methods for measuring the Index of Productive

Syntax (IPSyn, Scarborough, 1990), (Sproat et al., 2011), (c) syntactic complexity measures (Prud'hommeaux et al., 2011; Roark et al., 2011), and (d) methods for detecting disfluencies such as revisions, repetitions, false starts, and filled pauses (Morley et al., 2014; Heeman et al., 2010). Our goal is to build an integrated narrative assessment system that can evaluate (a) narrative performance, (b) syntactic complexity, (c) presence of disfluencies, (d) use of off-topic words, (e) narrative coherence, and (f) topic shifts from the gist of the story. We are planning to make this system available for researchers and at a later stage for clinicians with a more user-friendly interface. Initial results for such comprehensive analysis have been presented in Prud'hommeaux et al. (2014).

The methods developed for the analysis of the ADOS conversations have substantial capacity to be utilized in research and clinical domains as well. We plan to integrate our techniques for characterizing a child's topic perseveration (Chapter 7) with methods introduced by van Santen et al. (2013) for identifying verbatim and exact repeats including the child's self-repeats, and repetitions of others (*echolalia*). Given that all of these analyses can be performed on raw transcripts, such integrated system could provide large-scale extensive multi-aspect analysis of verbal perseveration at any given part of the dialogue, on any number of conversation samples. In addition, the developed techniques for quantifying conversational responsiveness can be incorporated with methods for detecting disfluency (Morley et al., 2014; Heeman et al., 2010) to provide a bigger picture of conversational reciprocity in children with autism.

We should note, however, that a clinician must consider a variety of social, communication, and behavioral criteria when making a diagnosis of ASD, and hence it is unlikely that language features alone could reliably predict a diagnosis of ASD. The more significant potential in our approaches, systems and tools lies in the area of pragmatic language deficit detection and remediation. Carefully designed tools using objective generalizable methods could provide helpful information to clinicians and therapists working with children with ASD to screen, detect, and track changes in their atypical language, as well as researchers exploring the specific linguistic and behavioral deficits associated with ASD.

Another obvious extension of this thesis would be to apply our methods for the analysis of language use to other types of neuropsychological disorders such as mild cognitive impairment (MCI), dementia, Alzheimer's disease, and Fragile X syndrome. First, given that repetitive speech and topic repetition is a frequently observed symptom in most of these disorders and, we can straightforwardly apply our methods to quantifying this phenomenon in conversations of individuals with these conditions. Next, we are interested in applying our techniques for the analysis of narratives to a set of retellings from seniors with and without MCI, a frequent precursor to dementia. Much like children with ASD, seniors with dementia are also more likely to include irrelevant information in their narrative retellings. These intrusions, however, are often informed by real-world knowledge and thus may not result in a decrease in measures of word overlap with narratives produced by unimpaired individuals. Once developed and evaluated, these methods can be combined with existing automated measures for the analysis of language use in MCI, including spoken and syntactic complexity measures (Roark et al., 2007, 2011), discriminative joint modeling of lexical variation and acoustic confusion (Lehr et al., 2013), and narrative assessment (Prud'hommeaux, 2012; Lehr et al., 2012). The extracted measures could be used for statistical analysis or as features for machine learning classification in order to provide a more representative picture of language behavior in this disorder.

# Bibliography


Achananuparp, P., Hu, X. and Shen, X. 2008. The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery*, pages 305–316, Springer.

American Psychiatric Association. 2000. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR (4th ed.)*. American Psychiatric Publishing, Inc.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. Washington, DC: American Psychiatric Publishing.

Arora, T. 2008. *Perseveration in young children with autism and the association with joint attention*. ProQuest.

Baker, C., Fillmore, C. and Lowe, J. 1998a. The Berkeley Framenet Project. In *Proceedings of the 17th international conference on computational linguistics*, pages 86–90.

Baker, M. J. 2000. Incorporating the thematic ritualistic behaviors of children with autism into games increasing social play interactions with siblings. *Journal of Positive Behavior Interventions* 2(2), 66–84.

Baker, M. J., Koegel, R. L. and Koegel, L. K. 1998b. Increasing the social behavior of young children with autism using their obsessive behaviors. *Research and Practice for Persons with Severe Disabilities* 23(4), 300–308.

Baron-Cohen, S. 2001. Theory of mind in normal development and autism. *Prisme* 34(1), 74–183.

Beitchman, J. H. 1996. *Language, Learning, and Behavior Disorders: Developmental, Biological, and Clinical Perspectives*. Cambridge University Press.

Bishop, D. V. 1989. Autism, Asperger's syndrome and semantic-pragmatic disorder: where are the boundaries? *International Journal of Language & Communication Disorders* 24(2), 107–121.

Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 3, 993–1022.

Bolden, G. B. 2006. Little words that matter: Discourse markers so and oh and the doing of other-attentiveness in social interaction. *Journal of Communication* 56(4), 661–688.

Borg, I. and Groenen, P. J. 2005. *Modern Multidimensional Scaling: Theory and Applications*. Springer Science & Business Media.

Boyd, B. A., Conroy, M. A., Mancil, G. R., Nakao, T. and Alter, P. J. 2007. Effects of circumscribed interests on the social behaviors of children with autism spectrum disorders. *Journal of Autism and Developmental Disorders* 37(8), 1550–1561.

Boyd, B. A., McDonough, S. G. and Bodfish, J. W. 2012. Evidence-based behavioral interventions for repetitive behaviors in autism. *Journal of Autism and Developmental Disorders* 42(6), 1236–1248.

Bullinaria, J. A. and Levy, J. P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39(3), 510–526.

Capps, L., Kehres, J. and Sigman, M. 1998. Conversational abilities among children with autism and children with developmental delays. *Autism* 2(4), 325–344.

CDC. 2007. Prevalence of autism spectrum disorders—Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2002. *Morbidity and Mortality Weekly Report* 56(1).

CDC. 2012. Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. *Morbidity and Mortality Weekly Report* 61(3).

CDC. 2014. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *Morbidity and Mortality Weekly Report* 63(2).

Church, K. W. and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1), 22–29.

Clough, P. 2000. Plagiarism in natural and programming languages: an overview of current tools and technologies. *Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK* pages 1–31.

Collins, M. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on*

*Empirical Methods in Natural Language Processing-Volume 10*, pages 1–8, Association for Computational Linguistics.

Dagan, I. 2000. Contextual word similarity. *Handbook of Natural Language Processing* pages 459–475.

Delaherche, E., Chetouani, M., Bigouret, F., Xavier, J., Plaza, M. and Cohen, D. 2013. Assessment of the communicative and coordination skills of children with Autism Spectrum Disorders and typically developing children using social signal processing. *Research in Autism Spectrum Disorders* 7(6), 741–756.

Diehl, J. J., Bennetto, L. and Young, E. C. 2006. Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of Abnormal Child Psychology* 34(1), 87–102.

Dollaghan, C. and Campbell, T. F. 1998. Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research* 41(5), 1136–1146.

Dunn, L. M. and Dunn, L. M. 1997. *Peabody Picture Vocabulary Test-Third Edition*. Circle Pines, MN: American Guidance Service.

Dunning, T. 1993a. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1), 61–74.

Dunning, T. 1993b. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1), 61–74.

Erkan, G. and Radev, D. R. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research* pages 457–479.

Fano, R. M. and Hawkins, D. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics* 29(11), 793–794.

Fay, W. H. 1982. The development of yes and no answers in autistic children. *Topics in Language Disorders* 3(1), 24–32.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Fitzgerald, E. C. 2009. *Reconstructing Spontaneous Speech*. ProQuest.

Flowerdew, J. and Tauroza, S. 1995. The effect of discourse markers on second language lecture comprehension. *Studies in Second Language Acquisition* 17(04), 435–458.

Fox Tree, J. E. and Schrock, J. C. 1999. Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language* 40(2), 280–295.

Goldman, S. 2008. Brief report: Narratives of personal events in children with autism and developmental language disorders: Unshared memories. *Journal of Autism and developmental disorders* 38(10), 1982–1988.

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.

Gotham, K., Risi, S., Pickles, A. and Lord, C. 2007. The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders* 37(4), 613–627.

Gravano, A., Benus, S., Hirschberg, J., Mitchell, S. and Vovsha, I. 2007. Classification of discourse functions of affirmative words in spoken dialogue. In *Interspeech*, pages 1613–1616.

Gravano, A., Hirschberg, J. and Beňuš, Š. 2012. Affirmative cue words in task-oriented dialogue. *Computational Linguistics* 38(1), 1–39.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Springer Science & Business Media.

Hearst, M. A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, Association for Computational Linguistics.

Heeman, P. A. and Allen, J. F. 1999. Speech repairs, intonational phrases, and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics* 25(4), 527–571.

Heeman, P. A., Lunsford, R., Selfridge, E., Black, L. and van Santen, J. 2010. Autism and interactional aspects of dialogue. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 249–252, Association for Computational Linguistics.

Hill, A. P., van Santen, J., Gorman, K., Langhorst, B. H. and Fombonne, E. 2015. Memory in language-impaired children with and without autism. *Journal of neurodevelopmental disorders* 7(1), 19.

Hirschberg, J. and Litman, D. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics* 19(3), 501–530.

Hoad, T. C. and Zobel, J. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American Society for Information Science and Technology* 54(3), 203–215.

Huang, S.-F. and Oi, M. 2013. Responses to Wh-, Yes/No-, A-not-A, and choice questions in Taiwanese children with high-functioning autism spectrum disorder. *Clinical Linguistics & Phonetics* 27(12), 969–985.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. *New Phytologist* 11.

Jiang, J. J. and Conrath, D. W. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *ROCLING 97* .

Jones, C. D. and Schwartz, I. S. 2009. When asking questions is not enough: An observational study of social communication differences in high functioning children with autism. *Journal of Autism and Developmental Disorders* 39(3), 432–443.

Jones, R. M. and Lord, C. 2013. Diagnosing autism in neurobiological research studies. *Behavioural Brain Research* 251, 113–124.

Jucker, A. H. 1993. The discourse marker well: A relevance-theoretical account. *Journal of Pragmatics* 19(5), 435–452.

Jurafsky, D. and Martin, J. H. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall, second edition.

Jurafsky, D., Shriberg, E., Fox, B. and Curl, T. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120.

Kanner, L. 1946. Irrelevant and metaphorical language in early infantile autism. *American Journal of Psychiatry* 103(2), 242–246.

Kanner, L. et al. 1943. *Autistic Disturbances of Affective Contact*. Publisher not identified.

Kendrick, B. 1964. *Quantitative Characters in Computer Taxonomy*. Phenetic and Phylogenetic Classification.

Kiss, G., van Santen, J. P., Prud'hommeaux, E. T. and Black, L. M. 2012. Quantitative analysis of pitch in speech of children with neurodevelopmental disorders. In *Interspeech*.

Klin, A. 2000. Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The social attribution task. *Journal of Child psychology and Psychiatry* 41(07), 831–846.

Klin, A., Lang, J., Cicchetti, D. V. and Volkmar, F. R. 2000. Brief report: Interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: Results of the DSM-IV autism field trial. *Journal of Autism and Developmental Disorders* 30(2), 163–167.

Koegel, R. L. and Frea, W. D. 1993. Treatment of social behavior in autism through the modification of pivotal social skills. *Journal of Applied Behavior Analysis* 26(3), 369–377.

Korkman, M., Kirk, U. and Kemp, S. 1998. *NEPSY: A Developmental Neuropsychological Assessment*. San Antonio: The Psychological Corporation.

Kremer-Sadlik, T. 2004. How children with autism and Asperger Syndrome respond to questions: A naturalistictheory of mind task. *Discourse Studies* 6(2), 185–206.

Kroese, D. P., Brereton, T., Taimre, T. and Botev, Z. I. 2014. Why the Monte Carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics* 6(6), 386–392.

Kullback, S. and Leibler, R. A. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* pages 79–86.

Lam, Y. G., Sze, S. and Yeung, S. 2012. Towards a convergent account of pragmatic language deficits in children with high-functioning autism: Depicting the phenotype using the Pragmatic Rating Scale. *Research in Autism Spectrum Disorders* 6(2), 792–797.

Landa, R., Martin, M., Minshew, N. and Goldstein, G. 1995. Discourse and abstract language ability in non-retarded individuals with autism. In *Biennial Meeting of the Society for Research in Child Development, Indianapolis, IN*.

Landa, R., Klin, A., Volkmar, F. et al. 2000. Social language use in Asperger syndrome and high-functioning autism. *Asperger syndrome* pages 125–155.

Landauer, T. K. and Dumais, S. T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104(2), 211.

Leacock, C. and Chodorow, M. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database* 49(2), 265–283.

Lee, L.-C., David, A. B., Rusyniak, J., Landa, R. and Newschaffer, C. J. 2007. Performance of the Social Communication Questionnaire in children receiving preschool special education services. *Research in Autism Spectrum Disorders* 1(2), 126–138.

Leekam, S. R., Prior, M. R. and Uljarevic, M. 2011. Restricted and repetitive behaviors in autism spectrum disorders: a review of research in the last decade. *Psychological bulletin* 137(4), 562.

Lehr, M., Prud'hommeaux, E. T., Shafran, I. and Roark, B. 2012. Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In *Interspeech*.

Lehr, M., Shafran, I., Prud'hommeaux, E. T. and Roark, B. 2013. Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment. In *HLT-NAACL*, pages 211–220.

Lenat, D. B. 1995. CYC: a large-scale investment in knowledge infrastructure. *Communications of the ACM* 38, 33–38.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D. and Crockett, K. 2006. Sentence similarity based on semantic nets and corpus statistics. *Knowledge and Data Engineering, IEEE Transactions on* 18(8), 1138–1150.

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304.

Lin, D. and Pantel, P. 2001. DIRT – Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*, pages 323–328.

London, E. 2007. The role of the neurobiologist in redefining the diagnosis of autism. *Brain Pathology* 17(4), 408–411.

Lord, C., Risi, S., Lambrecht, L., Cook Jr, 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.

Lord, C., Rutter, M., DiLavore, P. and Risi, S. 2002. *Autism Diagnostic Observation Schedule: ADOS*. Western Psychological Services Los Angeles, CA.

Lord, C., Rutter, M. and LeCouteur, A. 1994. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders* 24, 659–685.

Losh, M. and Capps, L. 2003. Narrative ability in high-functioning children with autism or Asperger's syndrome. *Journal of Autism and Developmental Disorders* 33(3), 239–251.

Losh, M. and Gordon, P. C. 2014. Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence. *Journal of Autism and Developmental Disorders* 44(12), 3016–3025.

Loveland, K., McEvoy, R. and Tunali, B. 1990. Narrative story telling in autism and Down's syndrome. *British Journal of Developmental Psychology* 8(1), 9–23.

Loveland, K. A. and Tunali-Kotoski, B. 2005. The school-age child with an autistic spectrum disorder. *Handbook of Autism and Pervasive Developmental Disorders, Volume 1, Third Edition* pages 247–287.

Luhn, H. P. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1(4), 309–317.

MacWhinney, B. 2000. *The CHILDES project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mainela-Arnold, E., Evans, J. L. and Coady, J. A. 2010. Explaining lexical–semantic deficits in specific language impairment: The role of phonological similarity, phonological working memory, and lexical competition. *Journal of Speech, Language, and Hearing Research* 53(6), 1742–1756.

Manning, C. D., Raghavan, P. and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. 1994. The Penn Treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119, Association for Computational Linguistics.

Martin, G. E., Roberts, J. E., Helm-Estabrooks, N., Sideris, J., Vanderbilt, J. and Moskowitz, L. 2012. Perseveration in the connected speech of boys with fragile X syndrome with and without autism spectrum disorder. *American Journal on Intellectual and Developmental Disabilities* 117(5), 384–399.

Mayer, M., Ormond, J., Murray, S., Templeton, G. and Osborn, S. 1969. *Frog, where are you?*. Dial Press New York.

McCann, J. and Peppé, S. 2003. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders* 38(4), 325–350.

Metzler, D., Bernstein, Y., Croft, W. B., Moffat, A. and Zobel, J. 2005. Similarity measures for tracking information flow. In *Proceedings of the 14th ACM international conference on Information and Knowledge Management*, pages 517–524, ACM.

Mihalcea, R., Corley, C. and Strapparava, C. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Mladenic, D. and Grobelnik, M. 1999. Feature selection for unbalanced class distribution and naive bayes. In *ICML*, volume 99, pages 258–267.

Morley, E., Hallin, A. E. and Roark, B. 2014. Challenges in automating maze detection. In *Proceedings of the First Workshop on Computational Linguistics and Clinical Psychology*, pages 69–77.

Morrison, K. E. 2014. *A Quantitative Analysis of Pragmatic Language in Adults with High-Functioning Autism*. Masters Thesis, The Ohio State University.

Murphy, M. and Abbeduto, L. 2007. Gender differences in repetitive language in fragile X syndrome. *Journal of Intellectual Disability Research* 51(5), 387–400.

Nadig, A., Lee, I., Singh, L., Bosshart, K. and Ozonoff, S. 2010. How does the topic of conversation affect verbal exchange and eye gaze? A comparison between typical development and high-functioning autism. *Neuropsychologia* 48(9), 2730–2739.

Nakov, P. and Hearst, M. A. 2008. Solving relational similarity problems using the web as a corpus. In *ACL*, pages 452–460.

Newschaffer, C. J., Croen, L. A., Daniels, J., Giarelli, E., Grether, J. K., Levy, S. E., Mandell, D. S., Miller, L. A., Pinto-Martin, J., Reaven, J. et al. 2007. The epidemiology of autism spectrum disorders. *Annu. Rev. Public Health* 28, 235–258.

Pantel, P. and Lin, D. 2002. Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, pages 613–619, ACM.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational :inguistics*, pages 311–318.

Prior, M. and Macmillan, M. B. 1973. Maintenance of sameness in children with Kanner's syndrome. *Journal of Autism and Childhood Schizophrenia* 3(2), 154–167.

Prud'hommeaux, E. 2012. *Alignment of Narrative Retellings for Automated Neuropsychological Assessment*. Ph. D.thesis, Oregon Health & Science University.

Prud'hommeaux, E., Morley, E., Rouhizadeh, M., Silverman, L., van Santeny, J., Roark, B., Sproat, R., Kauper, S. and DeLaHunta, R. 2014. Computational analysis of trajectories of linguistic development in autism. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 266–271, IEEE.

Prud'hommeaux, E. and Roark, B. 2015 in press. Graph-based word alignment for clinical language evaluation. *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, Association for Computational Linguistics.

Rapin, I. and Dunn, M. 1997. Language disorders in children with autism. In *Seminars in Pediatric Neurology*, volume 4, pages 86–92, Elsevier.

Rapin, I. and Katzman, R. 1998. Neurobiology of autism. *Annals of Neurology* 43(1), 7–14.

Rapp, R. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit*, pages 315–322.

Rathmann, P. 1994. *Good Night, Gorilla*. Penguin.

Rehfeldt, R. A. and Chambers, M. R. 2003. Functional analysis and treatment of verbal perseverations displayed by an adult with autism. *Journal of Applied Behavior Analysis* 36(2), 259–261.

Rein, R. P. 1984. *Observational Study of the Use of Verbal Perseverations by Persons with Autism*. Ph. D.thesis, University of California, Los Angeles.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* .

Rice, M. L., Warren, S. F. and Betz, S. K. 2005. Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile X, specific language impairment, and Williams syndrome. *Applied Psycholinguistics* 26(01), 7–27.

Roark, B., Mitchell, M. and Hollingshead, K. 2007. Syntactic complexity measures for detecting mild cognitive impairment. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 1–8, Association for Computational Linguistics.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K. and Kaye, J. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(7), 2081–2090.

Roark, B. and Sproat, R. 2007. *Computational Approach to Morphology and Syntax*. Oxford University Press.

Robinson, J. F. and Vitale, L. J. 1954. Children with circumscribed interest patterns. *American Journal of Orthopsychiatry* 24(4), 755.

Ross, D. E. 2002. Replacing faulty conversational exchanges for children with autism by establishing a functionally equivalent alternative response. *Education and Training in Mental Retardation and Developmental Disabilities* pages 343–362.

Rouhizadeh, M., Prud'hommeaux, E., Roark, B. and van Santen, J. 2013. Distributional semantic models for the evaluation of disordered language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Rouhizadeh, M., Prud'hommeaux, E., van Santen, J. and Sproat, R. 2014. Detecting linguistic idiosyncratic interests in autism using distributional semantic models. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–50, Association for Computational Linguistics.

Rouhizadeh, M., Prud'hommeaux, E., van Santen, J. and Sproat, R. 2015a. Measuring idiosyncratic interests in children with autism spectrum disorder. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, Association for Computational Linguistics.

Rouhizadeh, M., Sproat, R. and van Santen, J. 2015b. 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, Association for Computational Linguistics.

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

Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing and Management* 24(5), 513–523.

Scarborough, H. S. 1990. Index of Productive Syntax. *Applied psycholinguistics* 11(01), 1–22.

Schiffrin, D. 1988. *Discourse Markers*. Cambridge University Press.

Schourup, L. 1999. Discourse markers. *Lingua* 107(3), 227–265.

Selkirk, E. O. 1986. *Phonology and syntax: the relationship between sound and structure*. MIT press.

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

Semel, E., Wiig, E. and Secord, W. 2004. *Clinical Evaluation of Language Fundamentals Preschool - Second Edition*. San Antonio, TX: The Psychological Corporation.

Sepkoski Jr, J. J. 1974. Quantified coefficients of association and measurement of similarity. *Journal of the International Association for Mathematical Geology* 6(2), 135–152.

Solomon, O. 2001. Discourse markers in everyday conversation of high-functioning autistic children with family members. In *American Association for Applied Linguistics Annual Conference, St Louis, MO*.

Soubbotin, M. M. and Soubbotin, S. M. 2001. Patterns of Potential Answer Expressions as Clues to the Right Answers. In *TREC*.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21.

Spasic, I., Ananiadou, S., McNaught, J. and Kumar, A. 2005. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in Bioinformatics* 6(3), 239–251.

Spitzer, R. and Siegel, B. 1990. The DSM-III-R field trial of pervasive developmental disorders. *Journal of the American Academy of Child & Adolescent Psychiatry* 29(6), 855 – 862.

Sproat, R., Black, L., Prud'hommeaux, E., van Santen, J. and Roark, B. 2011. Automated analysis of natural language samples: comparison of children with autism spectrum disorders, developmental language disorders, and typical development. In *International Meeting for Autism Research (IMFAR)*.

Strassel, S. 2004. *Simple Metadata Annotation Specification v6.2*. Linguistic Data Consortium.

Stribling, P., Rae, J. and Dickerson, P. 2009. Using conversation analysis to explore the recurrence of a topic in the talk of a boy with an autism spectrum disorder. *Clinical linguistics & phonetics* 23(8), 555–582.

Sudhalter, V., Cohen, I. L., Silverman, W. and Wolf-Schein, E. G. 1990. Conversational analyses of males with fragile X, Down syndrome, and autism: comparison of the emergence of deviant language. *American Journal on Mental Retardation* .

Szatmari, P., Georgiades, S., Bryson, S., Zwaigenbaum, L., Roberts, W., Mahoney, W., Goldberg, J. and Tuff, L. 2006. Investigating the structure of the restricted, repetitive behaviours and interests domain of autism. *Journal of Child Psychology and Psychiatry* 47(6), 582–590.

Tager-Flusberg, H. 1999. A psychological approach to understanding the social and language impairments in autism. *International Review of Psychiatry* 11(4), 325–334.

Tager-Flusberg, H. 2001. Understanding the language and communicative impairments in autism. *International Review of Research in Mental Retardation* 23, 185–205.

Tager-Flusberg, H., Paul, R. and Lord, C. 2005a. *Language and Communication in Autism*, pages 335–364. John Wiley & Sons, Inc.

Tager-Flusberg, H., Paul, R., Lord, C. et al. 2005b. Language and communication in autism. *Handbook of Autism and Pervasive Developmental Disorders* 1, 335–364.

Traum, D. R. and Hinkelman, E. A. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence* 8(3), 575–599.

Tukey, J. W. 1949. Comparing individual means in the analysis of variance. *Biometrics* pages 99–114.

Turner, M. 1999. Annotation: Repetitive behaviour in autism: A review of psychological research. *Journal of Child Psychology and Psychiatry* 40(06), 839–849.

Turney, P., Littman, M., Bigham, J. and Shnayder, V. 2003. Combining Independent Modules to Solve Multiple-choice Synonym and Analogy Problems. In *Proceedings International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489, Borovets, Bulgaria.

Turney, P. D. 2006. Similarity of Semantic Relations. *Computational Linguistics* 32(3), 379–416.

Turney, P. D., Pantel, P. et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1), 141–188.

van Rijsbergen, C., Harper, D. and Porter, M. 1981. The selection of good search terms. *Information Processing and Management* 17(2), 77–91.

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

Van Santen, J. P., Prud'hommeaux, E. T., Black, L. M. and Mitchell, M. 2010. Computational prosodic markers for autism. *Autism* 14, 215–236.

Vismara, L. A. and Lyons, G. L. 2007. Using Perseverative Interests to Elicit Joint Attention Behaviors in Young Children With Autism Theoretical and Clinical Implications for Understanding Motivation. *Journal of Positive Behavior Interventions* 9(4), 214–228.

Volden, J. and Lord, C. 1991. Neologisms and idiosyncratic language in autistic speakers. *Journal of Autism and Developmental Disorders* 21, 109–130.

Wechsler, D. 2002. *Wechsler Primary and Preschool Scale of Intelligence - Third edition (WPPSI-III)*. San Antonio: Harcourt Assessment.

Wechsler, D. 2003. *Wechsler Intelligence Scales for Children - Fourth Edition (WISC-IV)*. San Antonio: The Psychological Corporation.

Wiesner, D. 1991. *Tuesday*. Houghton Mifflin Harcourt.

Wing, L. 1981. Asperger's syndrome: a clinical account. *Psychological Medicine* 11(01), 115–129.

Wing, L. and Gould, J. 1979. Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders* 9(1), 11–29.

Yarmohammadi, M. 2014. Discriminative training with perceptron algorithm for POS tagging task. Technical Report CSLU-2014-001, Center for Spoken Language Understanding, Oregon Health & Science University.