Similarity Measures for Quantifying Restrictive and Repetitive Behavior in Conversations of Autistic Children

Masoud Rouhizadeh[†], **Richard Sproat**[§], **Jan van Santen**[†]
[†]Center for Spoken Language Understanding, Oregon Health & Science University
[§] Google, Inc.
{rouhizad,vansantj}@ohsu.edu, rws@xoba.com

Abstract

Restrictive and repetitive behavior (RRB) is a core symptom of autism spectrum disorder (ASD) and are manifest in language. Based on this, we expect children with autism to talk about fewer topics, and more repeatedly, during their conversations. We thus hypothesize a higher semantic overlap ratio between dialogue turns in children with ASD compared to those with typical development (TD). Participants of this study include children ages 4-8, 44 with TD and 25 with ASD without language impairment. We apply several semantic similarity metrics to the children's dialogue turns in semi-structured conversations with examiners. We find that children with ASD have significantly more semantically overlapping turns than children with TD, across different turn intervals. These results support our hypothesis, and could provide a convenient and robust ASD-specific behavioral marker.

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by two broad groups of symptoms: impaired social communication and presence of restrictive and repetitive behavior (RRB) (American Psychiatric Association, 2000; American Psychiatric Association, 2013). RRB comprises both lower-order behaviors such as motor movements and higher-order cognitive behaviors such as circumscribed interests and insistence on sameness. Both of these are manifest in language as well. (Boyd et al., 2012; Szatmari et al., 2006; Turner, 1999; Kanner, 1943). All major ASD diagnostic instruments require the evaluation of RRB (Rutter et al., 2003; Lord et al., 2002; Lord et al., 1994). Individuals with ASD have significantly more RRB, stereotyped phrases, and idiosyncratic utterances in their conversations (Nadig et al., 2010; Capps et al., 1998; Volden and Lord, 1991).

However, such assessments are mostly qualitative, relying on clinical impressions or parental reports. There has been little work on quantitative or automated assessment methods for these behaviors in ASD, possibly due to the significant effort of detailed annotation of conversations that this would entail. Previous research in our group analyzed automatic detection of poor topic maintenance and use of off-topic words (Rouhizadeh et al., 2013; Prud'hommeaux and Rouhizadeh, 2012). We have also explored the different directions of departure from the target topic in ASD (rou, 2014; Prud'hommeaux et al., 2014).

In this paper, we attempt to automatically assess the presence of RRB in language, specifically at the semantic level, in children's conversation with an adult examiner during a semi-structured dialogue. We expect children with ASD to talk about fewer topics more repeatedly during their conversations. Specifically, we hypothesize a significantly higher semantic overlap ratio (SOR) between dialogue turns in children with ASD compared to those with typical development (TD). In order to calculate the SOR at different turn intervals for each child, we apply multiple semantic similarity metrics (weighted by child specificity scores) on every turn

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 117–123,
Denver, Colorado, June 5, 2015.

pair in four distance windows. We then compute the SOR for each child by averaging the similarity of every turn pair in the four distance windows. Our analysis indicates that, based on different similarity metrics, the ASD group had a significantly higher SOR than the TD group in most of the distance windows. These results support our hypothesis. Thus, patterns of semantic similarity between child's turns could provide an automated and robust ASD-specific behavioral marker.

In a previous study, van Santen and colleagues (van Santen et al., 2013) reported an automated method for identifying and quantifying two types of repetitive speech in ASD: repetitions of what child him or herself said (*intra-speaker repetitions*) and of what the conversation partner said (*inter-speaker repetitions*, or *echolalia*). The focus of this study was on verbatim repeats of word n-grams at short turn distances. The present study differs in several ways. (1) We focus on intra-child repetitions only. (2) We do so using bag-of-words similarity measures and lexical semantic expansion. (3) We consider short and long turn distance windows. (4) We use frequency weighting, assigning lower weights to frequent words.

2 Participants and data

Participants in this study include 44 children with TD and 25 children with ASD. ASD was diagnosed via clinical consensus according to the DSM-IV-TR criteria (American Psychiatric Association, 2000) and established threshold scores on two diagnostic instruments: the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002); and the Social Communication Questionnaire (Rutter et al., 2003). None of the ASD children in this study met criteria for a language impairment, defined as having a Core Language Score (CLS) on the CELF (Semel et al., 2003) of more than one standard deviation below the mean. The groups were well matched in age (6.41 vs. 5.91 years for the ASD and TD groups, respectively; $p>0.2$), and Nonverbal IQ (114.0 and 118.4; $p>0.25$), but not for nonverbal IQ (108 and 119; $p<0.0025$).

Each participant's ADOS session was recorded and the recordings were transcribed. The examiner and transcribers were unaware of the child's diagnostic status, the study hypothesis, and the computational methods. The automated methods in this paper are applied to these un-annotated raw transcripts.

The ADOS is a widely-used instrument for ASD diagnosis. It consists of a semi-structured series of spontaneous conversations and interactions between a child and a 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. The ADOS covers a broad range of conversational topics and activities, including Picture Description, Play, and Wordless Picture Book Description activities. Our expectation is that even though the activities, conversation topics, and actual questions are standardized, ASD children will tend to stick with their own topics of interest to a larger degree than children with TD.

3 Measuring the semantic overlap ratio (SOR)

For each child, we compute the semantic similarity score between every turn pair I and J in the following exponentially increasing distance windows, D:

a) $0<D\leq 3$: J is between 1 to 3 turns after I,
b) $3<D\leq 9$,
c) $9<D\leq 27$,
d) $27<D\leq 81$.

Then we compute the child's SOR for a given window D by averaging the similarity scores of turn pairs in D. We explored four semantic similarly measures which we describe in this section.

3.1 Semantic Similarity Measures

We expect ASD children to use more specific terms, relevant to their particular and often idiosyncratic interest due to their restrictive behavior. Therefore, we want our measures to be sensitive to how common or uncommon the words used by an individual child are. To assign lower weights to 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) (Manning et al., 2008):

$$idf_w = log\Big(\frac{N}{df_w}\Big) \quad (1)$$

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 to reduce the sparsity (hence higher IDF weights) caused by inflectional variations of the same lexeme.

3.1.1 Weighted Jaccard Similarity Coefficient

The weighted Jaccard similarity coefficient (Jac) (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, and then normalized by the sum of the maximum term frequency of each word in either turn:

$$Jac(I,J)=\frac{\sum\limits_{w\in I\cap J}\min(tf_{w,I},tf_{w,J})\times idf_w}{\sum\limits_{w\in I\cup J}\max(tf_{w,I},tf_{w,J})} \quad (2)$$

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

3.1.2 Cosine Similarity Score

The cosine similarity score (Cos) is a popular metric in IR to measure the similarity between the two turns I and J via the cosine of the angle between their vectors. We assign IDF weights to term frequencies, and then normalize the turn vectors by their length and the term weights:

$$Cos(I,J)=\frac{\sum\limits_{w\in I\cap J}tf_{w,I}\times tf_{w,J}\times(idf_w)^2}{\sqrt{\sum\limits_{w_i\in I}(tf_{w_i,I}\times idf_{w_i})^2}\times\sqrt{\sum\limits_{w_j\in j}(tf_{w_j,J}\times idf_{w_j})^2}} \quad (3)$$

3.1.3 Relative Frequency Measure

The relative frequency measure (RF) (Hoad and Zobel, 2003) is introduced as an author identity measure for detecting plagiarism at the document level. However, it has been shown to be applicable to the sentence level as well (Metzler et al., 2005). For this measure, we first normalize the differences in the turn lengths, and, second, we measures the similarity of the two turns I and J by the weighted relative frequency of their common words:

$$RF(I,J)=\frac{1}{1+||I|-|J||}\times\sum_{w\in I\cap J}\frac{idf_w}{1+|tf_{w,I}-tf_{w,J}|} \quad (4)$$

3.1.4 Knowledge-Based Similarity Measure

We now generalize our measures that are based on verbatim overlap to non-verbatim overlap. Toward this end, we use a knowledge-based turn similarity measure KBS that integrates verbatim word overlap with lexical relatedness (Mihalcea et al., 2006).

We begin with finding the maximum lexical similarity score $S(w_i,J)$ for each word w_i in turn I with words in turn J using the following formulation:

$$S(w_i,J)=\begin{cases}1\times idf_{w_i} & \text{if } w_i\in J\\ \max\limits_{w_j\in J}LS(w_i,w_j)\times idf_{w_i} & \text{otherwise}\end{cases} \quad (5)$$

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

In other words, if the word w_i is present in J, $S(w_i,J)$ will be 1 multiplied by idf_{w_i}. If not, the most similar word to w_i will be chosen from words in J using Lin's universal similarity and $S(w_i,J)$ will be that maximum score multiplied by idf_{w_i}. The same procedure is applied to the words in J, and finally the similarity between I and J is calculated :

$$KBS(I,J)=\frac{1}{2}\Big(\frac{\sum\limits_{w_i\in I}S(w_i,J)}{\sum\limits_{w_i\in I}idf_{w_i}}+\frac{\sum\limits_{w_j\in J}S(w_j,I)}{\sum\limits_{w_j\in J}idf_{w_j}}\Big) \quad (6)$$

Lin's universal similarity can only be applied to word pairs with the same part-of-speech (POS). For automatic POS tagging of the ADOS 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.

4 Results

As described in Section 3, we use our measures to calculate the similarity scores of all turn pairs for each distance window. Table 1 shows examples of similar turn pairs in the four distance windows based on the Weighted Jaccard Similarity Coefficient score.

We then calculate the SOR of each child in each given distance window by averaging the similarity scores of turn pairs in that window. Finally, we perform a two-tailed Mann-Whitney's U test, which is a non-parametric test of significance that does not assume that scores have a normal distribution. It evaluates the statistical difference between the SOR in ASD and TD children by comparing the medians of the two groups. For each similarity measure we report the medians of SOR in ASD and TD groups (with the group mean rank) as well as the significance test results: Mann-Whitney's U-Value (reported as W), P-Value (p), and the effect size (R).

Table 2 shows that both ASD and TD groups have a greater SOR in shorter distances with more significant difference and higher effect size. We see a decreasing trend in SOR by exponentially increasing the window size and distance. For each analysis, ASD group has a higher SOR than TD and the difference is statistically significant ($p<0.05$) in all short distances (up to $9<D\leq27$) and marginally missed the standard significance levels for the longest window ($p<0.1$ in $27<D\leq81$). We also investigated the effect of distance window on SOR in a different window set. The results are shown in Figure 1 using the KBS measure. We observe the exact same trend in these new windows as our main distance windows. All the differences between SOR in ASD and TD are statistically significant as well ($p<0.05$).

The comparison between various semantic similarity measures also indicates that KBS measure which takes into account lexical similarity in addition to word overlap, have more statistical power

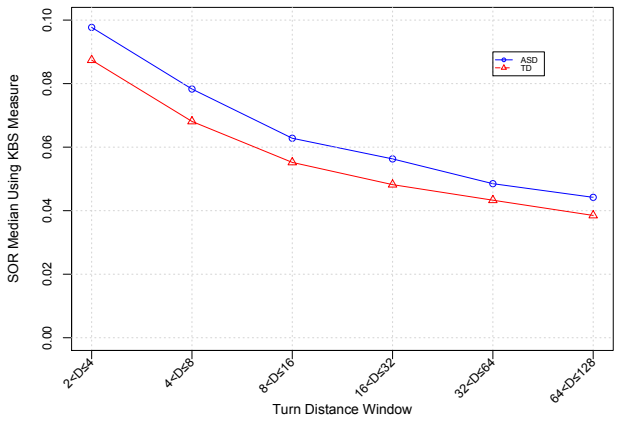

Figure 1: Semantic Overlap Ratio in ASD and TD at different turn distance windows using the KBS measure

to distinguish between ASD and TD groups in the longer windows ($9<D\leq27$ and $27<D\leq81$). 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 especially in the longer distances.

To address the possible confounding effect of verbal IQ, where a small but significant difference between the groups was found, we conducted two additional analyses. In one, we used analysis of covariance, with age, VIQ, and NVIQ as covariates; unlike W, there is no non-parametric equivalent of the analysis of covariance. In the other, we applied an algorithm that iteratively removes data until no significant group difference remains (at $p>0.15$) on age, VIQ, or NVIQ. Both analyses provided results that, while quantitatively different, were qualitatively the same.

5 Conclusions and future work

The results obtained with the methods presented here for measuring the semantic overlap between conversational turns in children with and without ASD in a spontaneous conversation indicate the utility of natural language processing for capturing diagnostically relevant information. The higher ratio of semantic overlap in children with ASD compared with TD children suggests that children with ASD are returning to specific topics more repeatedly. Thus, the findings support our hypothesis.

Window	*Example of turn pairs*
$0<D\leq3$	That is a crab with a humongous tail. Crab with a humongous tail is called a lobster.
$3<D\leq9$	So well, plus I got my and I got my magic carpets. You could use my magic carpet as a blanket.
$9<D\leq27$	Could you please get me some sports action figures? I just really want to play with sports action figures.
$27<D\leq81$	Yeah, just challenge him for one more duel. Alright, but first I challenge you for a duel.

Table 1: Examples of similar turns in four distance windows based on the Weighted Jaccard Similarity Coefficient

Similarity	*Window*	*ASD Mdn* (M Rank)*	*TD Mdn* (M Rank)*	W	p	r
Jac	$0<D\leq3$	.72 (43.68)	.59 (30.07)	333	.006	.33
	$3<D\leq9$	.25 (42.84)	.17 (30.55)	354	.014	.29
	$9<D\leq27$	.14 (42.44)	.09 (30.77)	364	.02	.28
	$27<D\leq81$	.08 (40.32)	.05 (31.98)	417	.09	.2
Cos	$0<D\leq3$	6.0 (45.28)	4.6 (29.16)	293	.001	.39
	$3<D\leq9$	2.2 (41.64)	1.8 (31.23)	384	.038	.25
	$9<D\leq27$	1.3 (42.32)	1.0 (30.84)	367	.022	.28
	$27<D\leq81$	.76 (40.6)	.53 (31.82)	410	.082	.21
RF	$0<D\leq3$	1.8 (44.48)	1.4 (29.61)	313	.003	.36
	$3<D\leq9$	.59 (45.2)	.41 (29.2)	295	.001	.38
	$9<D\leq27$	.31 (42.52)	.23 (30.73)	362	.018	.28
	$27<D\leq81$	.16 (40.68)	.13 (31.77)	408	.077	.21
KBS	$0<D\leq3$	15.0 (43.16)	12.0 (30.36)	346	.01	.31
	$3<D\leq9$	7.7 (41.64)	6.9 (31.23)	384	.038	.25
	$9<D\leq27$	5.9 (42.72)	5.0 (30.61)	357	.016	.29
	$27<D\leq81$	4.7 (43.76)	4.2 (30.02)	331	.006	.33

* *ASD and TD SOR Median values are multiplied by* 10^2.

Table 2: Significance Test Results of Semantic Overlap Ratio in ASD and TD groups at different turn distance windows, D

We are proposing a method of enabling measurement of a characteristic of language use in ASD that is currently "known" to be aberrant but is now ascertained only by impressionistic judgments rather than by quantification; and this is performed automatically on easy-to-obtain raw transcriptions of a clinical behavioral observation session (the ADOS) as opposed to requiring labor-intensive expert coding. To the best of our knowledge, this is the first time that verbal repetitiveness in natural language samples has been successfully measured — quantitatively, and automatically.

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.

A second area to investigate in the future is determining the children's conversation topics, especially the ones that are repeated. We could combine the child specificity scores such as IDF with the highly overlapping lexical items across different turns. We could also 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.

Third, we are also interested in trying additional similarity measures including BLEU (Papineni et al., 2002), ROUGE, (Lin, 2004), and Latent Semantic Analysis (Deerwester et al., 1990) to verify the robustness of our findings even further.

Finally, we plan to apply our methods to the output of Automatic Speech Recognition (ASR) systems to eliminate the transcription process. Measuring semantic similarity on ASR output will be an interesting challenge since it will likely contain word errors especially in children's spontaneous speech.

Acknowledgments

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.

References

American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, Washington, DC.

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

Brian A Boyd, Stephen G McDonough, and James W Bodfish. 2012. Evidence-based behavioral interventions for repetitive behaviors in autism. *Journal of autism and developmental disorders*, 42(6):1236–1248.

Lisa Capps, Jennifer Kehres, and Marian Sigman. 1998. Conversational abilities among children with autism and children with developmental delays. *Autism*, 2(4):325–344.

Michael Collins. 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.

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Timothy C Hoad and Justin Zobel. 2003. Methods for identifying versioned and plagiarized documents. *Journal of the American society for information science and technology*, 54(3):203–215.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Leo Kanner. 1943. Autistic disturbances of affective content. *Nervous Child*, 2:217–250.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Catherine Lord, Michael Rutter, and Anne LeCouteur. 1994. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disor-

ders. *Journal of Autism and Developmental Disorders*, 24:659–685.

Catherine Lord, Michael Rutter, Pamela DiLavore, and Susan Risi. 2002. *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services, Los Angeles.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Donald Metzler, Yaniv Bernstein, W Bruce Croft, Alistair Moffat, and Justin Zobel. 2005. Similarity measures for tracking information flow. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 517–524. ACM.

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

Aparna Nadig, Iris Lee, Leher Singh, Kyle Bosshart, and Sally Ozonoff. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Emily Prud'hommeaux and Masoud Rouhizadeh. 2012. Automatic detection of pragmatic deficits in children with autism. In *WOCCI*, pages 1–6.

Emily Prud'hommeaux, Eric Morley, Masoud Rouhizadeh, Laura Silverman, Jan van Santen, Brian Roark, Richard Sproat, Sarah Kauper, and Rachel DeLaHunta. 2014. Computational analysis of trajectories of linguistic development in autism. In *IEEE Spoken Language Technology Workshop (SLT 2014)*, South Lake Tahoe.

2014. Detecting linguistic idiosyncratic interests in autism using distributional semantic models. *ACL 2014*, page 46.

Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan van Santen. 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*.

Michael Rutter, Anthony Bailey, and Catherine Lord. 2003. *Social Communication Questionnaire (SCQ)*. Western Psychological Services, Los Angeles.

Eleanor Semel, Elisabeth Wiig, and Wayne Secord. 2003. *Clinical Evaluation of Language Fundamentals- Fourth Edition*. The Psychological Corporation, San Antonio, TX.

Peter Szatmari, Stelios Georgiades, Susan Bryson, Lonnie Zwaigenbaum, Wendy Roberts, William Mahoney, Jeremy Goldberg, and Lawrence Tuff. 2006. Investigating the structure of the restricted, repetitive behaviours and interests domain of autism. *Journal of Child Psychology and Psychiatry*, 47(6):582–590.

Michelle Turner. 1999. Annotation: Repetitive behaviour in autism: A review of psychological research. *Journal of child psychology and psychiatry*, 40(6):839–849.

Jan van Santen, Richard Sproat, and Alison Presmanes Hill. 2013. Quantifying repetitive speech in autism spectrum disorders and language impairment. *Autism Research*, 6(5):372–383.

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

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