Automated Discourse Analysis of Narrations by Adolescents with Autistic Spectrum Disorder

Michaela Regneri
IT Department
SPIEGEL-Verlag
Hamburg, Germany
michaela.regneri@spiegel.de

Diane King
National Foundation for
Educational Research (NFER)
London, United Kingdom
d.king@nfer.ac.uk

Abstract

We present a study about automated discourse analysis of oral narrative language in adolescents with autistic spectrum disorder (ASD). The basis of this evaluation is an existing dataset of fictional narrations of individuals with ASD and two matched comparison groups. We use three robust measures for quantifying different aspects of text cohesion on this corpus. These measures and several combinations of them correlate strongly with human cohesion annotations. Our evaluation will show which of these also distinguish the ASD group from the two comparison groups, which do not, and which differences are related to language competence rather than to factors specific to ASD.

1 Introduction

Language is, in many ways, a window to the mind. Written or spoken utterances convey much more than their content – they also provide information about the person who is writing or speaking the respective words. The research field of computational stylometry is concerned with the analysis of (written or transcribed) text and how it reveals information about the person who has produced this (see Daelemans (2013) for an overview). Typical applications, often with a focus on frequently updated websites and social media, are automated authorship attribution, gender distinction or forensic purposes.

A growing and very interesting subfield of computational stylometry is the detection of idiosyncratic language which may be found in individuals who have cognitive, affective or developmental disorders: while standard stylometry uses mostly focus on the pure identification of certain users or user groups, often with hardly interpretable features (like function word use), diagnostic analysis has the additional goal of making sense out of the actual features. The hope here is to gain more insight into the underlying disorder by analysing how it affects language. Additionally, there are also systems that automatically can identify or predict the onset of the condition in question.

Our focus is on the diagnostic analysis of oral narratives produced by adolescents with autistic spectrum disorder (ASD). ASD is a neurodevelopmental disorder characterised by impairment in social communication and restricted, repetitive and stereotyped patterns of behaviour (American Psychiatric Association, 2013). Although the social and communication difficulties of individuals with ASD have been well documented, little is known about narrative language in this population: whilst there has been a great deal of research on ASD by psychologists and neurologists, there are not many corpus analyses to support assumptions on language development and ASD. We are particularly interested in discourse cohesion, with cohesion being defined as the way in which devices are used to link together sentences, clauses and propositions. This includes the sequencing of and transitions between each event in a narrative. Although the production of a cohesive narrative is reported to be challenging for individuals with ASD, there is only limited work on systematic corpus analyses, mainly due to the lack suitable datasets.

Our work is based on a recently published dataset of fictional narratives told by young people with ASD (King et al., 2014). We expressly do not aim to just automatically identify stories from an ASD group, because that would be easily accomplished using crude features like story length. Our goal is instead to find meaningful cohesion-related features that distinguish the language of individuals wth ASD.

Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning, pages 1–9,
Berlin, Germany, August 11, 2016.

Our contribution is threefold: First, we present robust measures that allow the automated assessment of cohesion in short texts, and introduce skewness as a new measure for coreference chains. Second, we show which features of the text cohesion we measure are ASD-specific according to our data, and which are related to language competence. Lastly, we also show the correlation of our measures with human judgments of story cohesion.

2 Related Work

Many automated approaches to diagnostic analysis detect Alzheimer's and related forms of dementia: there are extensive studies on the specific language changes in people that develop dementia (Hirst and Wei Feng, 2012; Le et al., 2011), showing how the syntactic complexity of sentences declines with the disease's progress. Some classifiers are capable of automated diagnosis from continuous speech (Baldas et al., 2011), and, additionally, the "Nun study" resulted in a system that can predict whether or not an individual will develop Alzheimer's decades before the actual onset of cognitive decline (Riley et al., 2005).

Other systems recognize spontaneous speech by individuals with more general mild cognitive impairments, for adults (Roark et al., 2011) and also for children (Gabani et al., 2009). Hong et al. (2012) present an unusual study on the language of adult patients with schizophrenia.

Previous research on narratives of children with ASD has reported difficulties with both structural and evaluative language. Individuals with ASD struggle with expressing sentiment and make fewer references to mental states than their typically developing peers (Capps et al., 2000; Tager-Flusberg, 1996). However, other experiments show that, when carefully matched with comparison groups on cognitive and language ability, many of these differences are not evident.

More basic problems emanate from a general lower syntactic complexity (Tager-Flusberg and Sullivan, 1995) and difficulties in producing a coherent narrative. Karmiloff-Smith (1985) argues that the production of a coherent narrative is dependent on the integration of knowledge of both coherence and cohesion; coherence being defined as the structure of a story and cohesion as the devices used to link together sentences, clauses and propositions, thereby maintaining a common theme. Loveland and Tunali (1993) found that individuals with autism were less likely to tell a story as a coherent sequence and more likely to produce narratives that included bizarre, unrelated or inappropriate material. Diehl et al. (2006) also report that narratives produced by individuals with ASD were significantly less coherent than those of a comparison group. However, Tager-Flusberg and Sullivan (1995) found no significant differences in the use of lexical cohesion devices between three groups of children with autism, learning disabilities and typically developing, matched on verbal mental age.

Some of these language difficulties have been subject to automated analysis: Prud'hommeaux et al. (2011) analyzed data of very young children (6-7 years old). They built an automated classifier that distinguished sentences uttered by children with ASD from sentences of two control groups (one with children with a language-impairment, one with typically developing children). The authors themselves note some drawbacks of their underlying dataset, in particular that some children in the ASD group were also classified as language-impaired. In consequence, a clear distinction between these groups was impossible.

In two follow-up studies (Rouhizadeh et al., 2013; Rouhizadeh et al., 2015), the authors analysed whole narratives (retellings) told by children (mean age 6.4) with ASD compared to a typically developing control group (with the same average age and IQ). As discourse-related measures, they use the tf-idf measure (Luhn, 1957) and several measures of text similarity to identify idiosyncratic words and topics. The texts from the control group and some crowdsourced retellings from typically developing adults served as a basis for determining unusualness.

Regneri and King (2015) present a study on a much larger dataset with non-fictional stories about everyday scenarios (like *having a birthday* or *being angry*). Next to several shallow language features, they also evaluate tf-idf and show that this is actually more closely related to language competence than to ASD. However, they do not evaluate any other discourse-related features.

For our study, we use a dataset with fictional stories, and take the discourse-level investigation a step further: we present different measures of text cohesion, which quantify some actually ASD-specific difficulties with narrative.

FOREST				
Group	ASD	Lang.	Age	All
Cohesion	1.79	2.93	2.76	2.00
Sent. / Story	7.64	9.07	10.10	8.86
Words / Sent.	10.46	12.19	11.37	11.40

MOUNTAIN				
Group	ASD	Lang.	Age	All
Cohesion	2.21	2.96	3.04	2.49
Sent. / Story	10.00	9.90	12.39	10.62
Words / Sent.	10.59	10.57	10.98	10.71

ALL STORIES				
Group	ASD	Lang.	Age	All
Cohesion	2.00	2.95	2.90	2.24
Sent. / Story	8.88	9.49	11.24	9.76
Words / Sent.	10.54	11.33	11.15	11.01

Table 1: Manually assigned cohesion scores, average story and sentence lengths for the corpus.

3 Data

We base our analysis on a dataset collected by King et al. (2014), which we describe in more detail in the following. The corpus contains transcripts of fictional stories constructed by the children after one of two different prompts. Appendix A shows some examples from the story collection. King et al. also report extensive manual annotation of the narratives, parts of which we will use as a gold standard for our automated experiments.

3.1 Data collection

The participants were divided in three groups: 27 high functioning adolescents with ASD aged 11 to 14 years, one comparison group of 27 adolescents matched with the ASD group on chronological age and nonverbal ability, and a second comparison group of 27 children and adolescents aged between 7 and 14 years, who were individually matched with the ASD group on a measure of expressive language (Recalling Sentences subtest of the CELF IV (Semel et al., 2006)) and on nonverbal ability. All groups had average scores on non-verbal and verbal measures, as measured by the Matrices test of the BAS II (Elliot et al., 1996) and the BPVS II (Dunn and Dunn, 1997). There were no significant differences between the groups in measures of non-verbal ability, verbal ability or expressive language. The average age difference between the language-matched control group and the two other groups is 17 months.

Participants in all three groups were presented with two story stems and asked to continue the narrative. Each story stem was accompanied with a picture illustrating each prompt. The development of these materials was based on the work of Stein and Albro (1997), but adapted to be more suitable for the age group of this study. To prevent order effects, the presentation of the story stems was counterbalanced. After one practice story, each participant completed the following two story stems:

1. The "forest" story:

   The boy ran into the forest. He looked ahead of him and saw a little green man in a spaceship.

2. The "mountain" story:

   When the girl climbed up the mountain, she saw, hidden among the trees, a little wooden house covered in snow.

Overall, there were 54 stories per group, totaling 162 stories in the corpus. This corpus is particularly well suited to analyse difficulties with cohesion because it contains texts that were freely invented, without any structural guidance. Moreover, the inclusion of the language-matched and the age-matched control groups enables us to distinguish language development issues from ASD-specific difficulties.

3.2 Corpus annotations and statistics

The stories were recorded, transcribed and manually coded and scored according to the Narrative Scoring Scheme (Stein and Albro, 1997, NSS). The NSS rates stories on a 0-5 scale in several categories: introduction, character development, mental states, referencing, conflict/resolution, cohesion and conclusion. To ensure the reliability of the coding, 10% of the narratives (16) were also coded by an independent researcher. Inter-reliability was found to be high (0.87).

Because we are specifically interested in discourse structure, the NSS annotations for cohesion will serve as a gold standard for our own evaluation (cf. Section 5.2). We show these ratings along with some basic corpus figures in Table 1:

Cohesion refers to the manual cohesion annotations, *Sent. / Story* is the average number of utterances per story, and *Words / Sent.* quantifies the average number of words per sentence.

The ASD group has significantly lower cohesion scores than the two comparison groups (*Lang.* for the group matched by language competence, *Age* for the controls matched by chronological age). Between the two groups with neurotypical participants, there is no significant difference. The *mountain* story prompt resulted in longer, more cohesive stories, consisting of shorter sentences. This difference is particularly clear for the ASD group and the age-matched controls.

4 Measures for Story Cohesion

In a preprocessing step, we apply the coreference resolution module of Stanford CoreNLP (Manning et al., 2014) to the whole corpus. On this basis, we compute three coreference-related measures: the proportion of sentences with anaphoric references, the average length of coreference chains (normalized by story length) and *Skewness*, a measure we derive from statistics and apply to clusterings.

4.1 Sentences with anaphoric references

As a simple measure for cohesion in a text t, we define $anaphors(t)$ as the proportion of sentences that contain at least one anaphoric reference (with $sentences(t)$ being the set of sentences in t):

$$anaphors(t) = \frac{|(sentences\ w.\ anaphors\ in\ t)|}{|sentences(t)|}$$

4.2 Average length of coreference chains

The average length of coreference chains in a text is a common indicator for cohesion (the longer the chains, the stronger the cohesion). Computing this as an absolute number will also directly measure the average text length, which is always lower for the ASD group. In order to isolate the cohesion part, we divide the average coreference chain length by the number of sentences in the text. We compute $chain\_length(t)$ of a text t as follows (with C_t as the set of all coreference chains in t):

$$chain\_length(t) = \frac{\sum\limits_{c \in C_t} length(C)}{|C_t| * |sentences(t)|}$$

The average chain length for the same story will be higher if there are fewer coreference sets (and thus fewer characters and objects).

4.3 *Skewness* of coreference chains

As a third coherence measure, we introduce the notion of *Skewness* for coreference chains. Skewness is originally a measure for probability distributions, indicating (the lack of) uniformity.

We interpret this score as a geometric measure for a set partition: for mentions in different coreference chains, this measure shows whether the narrator has the tendency to devote equally long story parts to all participants (resembling a uniform distribution) or whether he or she focuses more on a few main characters or objects, with some supporting entities which are less frequently mentioned (skewed distribution). We thus interpret the distribution of mentions as a probability distribution Pr over a random variable x, with each value x_i in X corresponding to a coreference chain c_i in $C(t)$, and $Pr(X = x_i)$ being the number of mentions in c_i divided by the overall number of mentions. $skewness(t)$ is computed as follows (with E being the expectation operator, μ the mean of the distribution X, σ the standard deviation):

$$skewness(t) = abs\left(\mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]\right)$$

Skewness originally does not only indicate the strength, but also the *direction* using negative or positive values. Because we are only interested in the overall asymmetry of the coreference chains, we only note the absolute value of the result. According to Bulmer (1979), a result with an absolute value greater than 1 is considered to indicate strong skewness.

4.4 Measure combinations

In the final evaluation, we also use pairwise measure combinations, and the combination of all three together. *Combining* here means that we first make the measures comparable, and then average the results. To arrive at meaningful scores, we process the chain length and skewness as follows:

- The **coreference chain length** correlates negatively with human judgements (cf. Table 3), so we combine a "negative chain length" ($1 - chain\_length$) with the respective other measure.
- **Skewness** is normalized to a value between 0 and 1 before combination to match the value span of the other two measures.

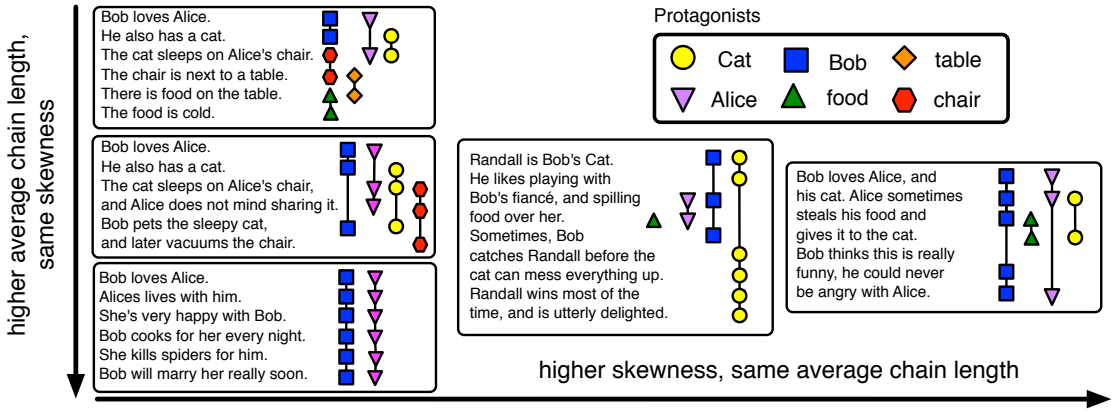

Figure 1: Constructed examples illustrating chain length vs. skewness

To better illustrate the relationship of skewness and chain length, Figure 1 shows examplary (made-up) stories with coreference chains of different lengths and skewness. The left column shows three stories with the same skewness (0), but different average chain length (2, 4 and 6, top to bottom, without normalization). An inverse case is sketched in the middle row: the chains all have the same average chain length (4), but different skewness (0,1.2 and1.4 respectively). Despite the equally trivial plots, the stories with 4 or more characters appear more readable when their coreference clusters are more skewed.

5 Experiments

In the following, we first compare the ASD group with the two control groups using our cohesion measures and their combinations (Section 5.1).

In a second step, we show the correlation of our measures with human coherence annotations reported by King et al. (2014) (Section 5.2).

5.1 Comparison of the three groups

The computed results for all measures and their combination is shown in Table 2.

Viewed in isolation, only the number of sentences with anaphors distinguishes the ASD group from the two control groups. While the (normalized) average coreference chain length is equal for all groups, skewness seems to be a matter of language competence rather than exposing anomalies from the ASD group. However, the picture is not entirely clear: when just considering the ”forest” stories, we see a tendency for the ASD group to have less skewed coreference chains.

The combination of anaphoric references plus the average chain length distinguishes the ASD group most clearly, showing no difference between the two neurotypical groups (even though they differ in general language competence). The same pattern is evident in the expert annotation with NSS scores: there is no difference between the control groups, but the stories from the ASD group are rated significantly as less cohesive.

All other combinations distinguish all three groups from each other, which means that the differences between the groups are related to both ASD and language competence. For the combination of anaphoric references with skewness, our data indicates that the group differences are more strongly related to factors specific to ASD: the alpha level for the significance of the difference between the two control groups is lower than for the remaining differences ($p < 0.05$, whereas all other significance levels are at $p < 0.01$).

The remaining two feature combinations (chain length with skewness, all three measures together) also distinguish the ASD group from the comparison groups, but (as expected) additionally bear components of language competency.

5.2 Correlation with manual annotations

The results of our automated evaluation were mixed with respect to ability to distinguish the ASD group from the comparison groups. From the isolated measures, only counting sentences with anaphoric references shows results specific for the ASD group. The combinations of different cohesion components gives a clearer picture, but only one combination shows the same pattern observed for manual annotations: the neurotypical groups are indistinguishable for the combination of anaphors and chain length.

To understand better what contributes most to

	FOREST			MOUNTAIN			ALL STORIES		
Measure	ASD	Lang	Age	ASD	Lang	Age	ASD	Lang	Age
anaphors (an)	0.83	0.93	0.95	0.80	0.92	0.91	0.81	0.93	0.93*
chain length (cl)	0.56	0.52	0.55	0.47	0.47	0.44	0.51*	*0.50*	*0.50**
skewness (skew)	0.79	0.91	1.04	0.95	0.97	1.22	0.90*	*1.04*	1.30*
an & cl	0.73	0.80	0.80	0.75	0.81	0.81	0.74	0.80	0.81*
an & skew	0.54	0.63	0.66	0.56	0.62	0.68	0.55	0.63	0.67
cl & skew	0.35	0.41	0.41	0.42	0.43	0.50	0.39	0.42	0.46
All Combined	0.57	0.64	0.66	0.61	0.65	0.69	0.59	0.65	0.67
NSS score	1.79	2.93	2.76	2.21	2.96	3.04	2.00	2.95	*2.90*

Table 2: Results of cohesion evaluation, along with the manually assigned cohesion scores. Significance is measured for the group of all stories only. *Emphasized* values have no significant difference ($p > 0.05$) to the ASD group, starred values $*$ have no significant difference to the language-matched group.

the cohesion perceptions of human experts, we calculate the correlation of our measures with the NSS scores assigned in our source dataset. To measure correlation, we use Spearman's rank correlation coefficient (ρ), a non-parametric test which is widely used for similar comparisons of system ratings with manually assigned scores (Mitchell and Lapata, 2008; Erk and McCarthy, 2009, among others). Spearman's ρ compares how similarly two measures rank the same set of samples (in our case, a sample is a story).

Table 3 shows the results. For the complete corpus, all measures show significant correlations.

The best isolated measure is skewness ($\rho = 0.44$), which shows the highest correlation for the mountain stories. However, combining skewness with any of the other measures does not result in a higher ρ value. A partial explanation lies in the differences in story length: We did not normalize skewness for story length, because "skewness per sentence" is not a meaningful measure. However, skewness of coreference chains is intuitively a more distinguishing feature if the stories are longer, simply because the possible skewness values have a higher range when there are more referring elements to distribute. In contrast, the other measures seem to be less suitable for longer texts, because skewness has the highest ρ values on the *mountain* stories, which are, on average, longer than the *forest* stories.

The average chain length has a significant negative correlation, i.e. cohesion is higher when the chains are shorter on average. (For combinations, we therefore use an inverted value, cf. Section 4).

Measure	FOREST	MOUNTAIN	ALL
anaphors (an)	0.23	0.17	0.19
chain length (cl)	-0.31	-0.22	-0.28
skewness (skew)	0.48	**0.38**	0.44
an & cl	0.55	0.36	**0.46**
an & skew	0.54	0.09 (ns)	0.31
cl & skew	0.47	0.13 (ns)	0.31
All Combined	**0.57**	0.21	0.40

Table 3: Correlation with manual evaluation (in Spearman's ρ). Values in italics are *not significant* ($p > 0.05$), **maxima** are in boldface.

The number of anaphoric references displays the lowest ρ-values (for the *mountain* sub-corpus, the correlation is not significant). When combined with coreference chain length, the fused measure has the highest overall correlation ($\rho = 0.46$), so these two features make different contributions to the overall cohesion.

The combination of all three measures has the second highest correlation with the manual annotations, and the highest ρ for the forest prompt.

Overall, our automated measures correlate strongly with the human annotations, but surprisingly much more so for the forest stories compared to the mountain stories. Skewness seems to be a very good measure in general, but the number of reference-bearing sentences combined with the average chain length obviously contributes similar information to the evaluation of short stories.

6 Discussion

The measures we evaluated are all coreference-based, quantifying different aspects of text cohesion: the proportion of sentences with anaphoric references reflects the sheer number of coreference links. The average coreference chain length (normalized over story length) mainly measures the number of protagonists and objects (cf. Figure 1). Skewness applied to coreference sets shows whether the protagonists differ in importance within the story, i.e. whether there is a recognizable main character (or a few of them) next to several supporting characters (or objects).

We succeeded in demonstrating that these measures strongly correlate with human assessment of cohesion, and that some combinations of them yield different results for the stories from the ASD group compared to the control groups. In particular, the measure combination that showed the strongest correlation with human judgements (chain length plus number of anaphoric references) seems to be directly influenced by ASD, and not just an indicator of general language competence: there was no difference between the two neurotypical control groups, but the score of the ASD group differed significantly from both.

Skewness, which we used as a new measure for quantifying the distribution of referring expressions into coreference sets, shows the highest correlation with human judgement as an isolated measure. However, skewness seems to work better for longer stories, which is intuitively clear: the possible variation of coreference set distribution is higher if there are more anaphoric references, and skewness becomes more distinguishing if the results show a higher variation.

Obviously, our measures cannot assess the whole spectrum of discourse features: they do not include any lexical features or semantic discourse relations. While we tried to compute such indicators, neither lexical chains nor discourse relations lead to a meaningful evaluation on our dataset. This is mostly due to the brevity of the stories, but is also because the setup of oral narration does not yield the discourse structure typically found in written language. Analyses with deeper discourse features would require a different dataset, which, however, might be difficult to create.

The most important outcome of this analysis is that an automatic evaluation of cohesion for diagnostic stylometry can be successfully used to validate theoretical claims. We also took important steps towards identifying cohesion-based measures to analyze unusual language traits in adolescents with autistic spectrum disorder. Our measures proved suitable for short stories, which is important because the participants we focus on have difficulties with producing longer texts. Further, our approach is robust enough to assess cohesion in transcripts of spoken narrations, which are more difficult to process with than written language. Future work needs to show how our ideas can be extended beyond this point, either with different measures, or with different datasets, or both.

7 Conclusion

We have presented an automatic evaluation showing differences in stories narrated by adolescents with autistic spectrum disorder in comparison with two control groups. For this purpose, we presented three robust measures applicable to the short story transcripts, namely the proportion of sentences with anaphors, the (normalized) average coreference chain length, and skewness as a new measure related to coreference set clusterings.

We showed that skewness is the measure that best correlates with manual cohesion annotation, and that it seems to be more meaningful for longer stories. Further we have shown that the combination of coreference chain length and the number of sentences with anaphors is sufficient to assess cohesion in shorter stories.

In future work, we would seek to find other measures of cohesion which could help to assess the difficulties of individuals with ASD compared to neurotypical controls, possibly on a different dataset with longer stories. Further, it would be interesting to establish whether the features that we found persist with age, and whether they are comparable to the effects reported for other disorders and diseases such as dementia.

References

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. APA, Washington, DC, 5th ed. edition.

Vassilis Baldas, Charalampos Lampiris, Christos Capsalis, and Dimitrios Koutsouris. 2011. Early Diagnosis of Alzheimer's Type Dementia Using Continuous Speech Recognition. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer Berlin Heidelberg.

M.G. Bulmer. 1979. *Principles of Statistics*. Dover Books on Mathematics Series. Dover Publications.

Lisa Capps, Molly Losh, and Christopher Thurber. 2000. "The Frog Ate the Bug and Made his Mouth Sad": Narrative Competence in Children with Autism. *Journal of Abnormal Child Psychology*, 28(2).

Walter Daelemans. 2013. Explanation in computational stylometry. In *Proc. of CICLing'13*.

Joshua J. Diehl, Loisa Bennetto, and Edna Carter Young. 2006. Story Recall and Narrative Coherence of High-Functioning Children with Autism Spectrum Disorders. *Journal of Abnormal Child Psychology*, 34(1).

Lloyd M. Dunn and Leota M. Dunn. 1997. *The British Picture Vocabulary Scale Second Edition (BPVS II)*. Windsor Berkshire: NFER-NELSON Publication Company.

Colin Elliot, Pauline Smith, and Kay McCullouch. 1996. *The British Ability Scales II (BASII)*. Windsor Berkshire: NFER-NELSON Publication Company.

Katrin Erk and Diana McCarthy. 2009. Graded word sense assignment. In *Proc. of EMNLP 2009*.

Keyur Gabani, Melissa Sherman, Thamar Solorio, Yang Liu, Lisa Bedore, and Elizabeth Peña. 2009. A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children. In *Proc. of NAACL-HLT 2009*.

Graeme Hirst and Vanessa Wei Feng. 2012. Changes in style in authors with alzheimer's disease. *English Studies*, 93(3).

Kai Hong, Christian G. Kohler, Mary E. March, Amber A. Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proc. of EMNLP-CoNLL 2012*.

Annette Karmiloff-Smith. 1985. Language and cognitive processes from a developmental perspective. *Language and Cognitive Processes*, 1(1).

Diane King, Julie Dockrell, and Morag Stuart. 2014. Constructing fictional stories: a study of story narratives by children with autistic spectrum disorder. *Research in developmental disabilities*, 35(10).

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three british novelists. *Literary and Linguistic Computing*, 26(4).

Katherine Loveland and Belgin Tunali. 1993. Narrative language in autism and the theory of mind hypothesis: a wider perspective. In *Understanding other minds: Perspectives from autism.* Oxford University Press.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, 1(4).

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proc. of ACL 2014: System Demonstrations*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proc. of ACL 2008*.

Emily T Prud'hommeaux, Brian Roark, Lois M Black, and Jan van Santen. 2011. Classification of atypical language in autism. *ACL HLT 2011*.

Michaela Regneri and Diane King. 2015. Automatically evaluating atypical language in narratives by children with autistic spectrum disorder. In *Proc. of NLPCS 2014*.

Kathryn P. Riley, David A. Snowdon, Mark F. Desrosiers, and William R. Markesbery. 2005. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the Nun Study. *Neurobiology of Aging*, 26(3).

Brian Roark, Margaret Mitchell, J Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7).

Masoud Rouhizadeh, Emily Prud'hommeaux, Brian Roark, and Jan van Santen. 2013. Distributional semantic models for the evaluation of disordered language. In *Proc. of NAACL-HLT 2013*.

Masoud Rouhizadeh, Emily Prud'Hommeaux, Jan Van Santen, and Richard Sproat. 2015. Measuring idiosyncratic interests in children with autism. In *Proc. of ACL 2015*.

Eleanor Semel, Elisabeth .H Wiig, and Wayne Secord. 2006. *Clinical Evaluation of Language Fundamentals (CELF-4 UK)*. Pearson Assessment, fourth edition uk edition.

Nancy L Stein and Elizabeth R Albro. 1997. Building complexity and coherence: Children's use of of goal-structured knowledge in telling stories. *Narrative development: Six approaches*, page 5.

Helen Tager-Flusberg and Kate Sullivan. 1995. Attributing mental states to story characters: A comparison of narratives produced by autistic and mentally retarded individuals. *Applied Psycholinguistics*, 16.

Helen Tager-Flusberg. 1996. Brief report: Current theory and research on language and communication in autism. *Journal of Autism and Developmental Disorders*, 26(2).

Appendix A Corpus Examples

The following shows some examples from our story corpus collected by King et al. (2014). For each story stem (repeated below), we show 2 examples from the ASD group, and one from each control group. For the sake of brevity, we do not show the manual annotations from the original corpus. Slashes (/) indicate utterance boundaries.

A.1 The *Forest* story

The boy ran into the forest. He looked ahead of him and saw a little green man in a spaceship.

A.1.1 ASD

Example 1: the spaceship was quite small. / And the alien was about the size of a small cat. / And it was friendly. / but it didn't really understand how humans said hello. / So it thought, to say 'hello' you had to vaporise the person in front of you. / and then the boy ran away, shut his door and then decided not to drink anymore whisky or beer.

Example 2: The green man had three eyes. / It had claws and fangs. / It looked at him and ran into the spaceship. / Out came three more green men carrying laser guns, dun dun dun.

A.1.2 Language-Matched Controls

He was shocked at first because he didn't know what it is. / So he walked up. / and he got suck/ed in by a tractor beam. / and he found himself in a UFO. / he was surround/ed by weird looking creatures like aliens. / and they started speaking like this unknown language to him. / and he couldn't understand a thing about them. / So he tried to escape. / he ran away. / but he couldn't cos—because wherever he went he was surround/ed by aliens. / and they eventually capture/ed him and took him off to Mars. /

A.1.3 Age-Matched Controls

The little green man waved at him and yelled at him to come and help him with his spaceship which had crashed into the forest after he had lost all his fuel. / The little boy went up to him and said that he didn't know what much he could do because he's obviously not very talented at fixing alien space craft. / so he ran out of the forest went to go get the nearest person he could see. / He then brought him back to the spaceship where the little green man was waiting. / and the man he'd brought back was absolutely amazed. / he went up to the little alien and started speaking with him. / And then after a while he persuaded it to come out of it's spaceship. / and it went to meet the locals who were very very amazed. / and then all the people joined together. / and they pulled his spaceship out of the mud. / and he flew back in it to mars.

A.2 The *Mountain* story

When the girl climbed up the mountain, she saw, hidden among the trees, a little wooden house covered in snow.

A.2.1 ASD

Example 1: She went up to the mountain to see the house. / She went inside and had a cup of tea. / After that she can't get out because the snow block/ed the door. / And the men came came in and broke it. / but snow came again. / and then she was stuck. / That's it.

Example 2: the snow house was was a zombie. / and the zombie / he went up to the door. / and the zombie scared him. / and the zombie went to chase the girl. / and the girl ran away to to her grandma that was climbing up the mountain. / and she screamed and jumped off the mountain

A.2.2 Language-Matched Controls

She saw it was abandoned. / so she went down to see what it was like. / She peeked inside the window. / and inside there was a pixie . / and then the pixie saw the girl and said 'go away from my window'. / and then he threw a bowl of soup over her. / the little girl went home and said: 'daddy there was a pixie who threw some soup over me'. / and then the dad said 'don't be silly'. / stop telling your little stories'.

A.2.3 Age-Matched Controls

She walked towards the house. / the house lit up. / lights switched on. / She knocked on the door. / she was cold. / she asked if she could come in. / There was a strange lady come to the door, pimples and spots all over her, mouldy ugly hair and very very small. / she went in. / the lady was actually a witch in disguise. / She grabbed the girl and threw her into the oven. / her friend had also came into the house five minutes later and seen her in the oven. / She had pushed the witch over, got her out and ran off. / they reported it all to the police. / The police came up the next day. / The house was not there.