Classification of atypical language in autism

Emily T. Prud'hommeaux, Brian Roark, Lois M. Black, and Jan van Santen
Center for Spoken Language Understanding
Oregon Health & Science University
20000 NW Walker Rd., Beaverton, Oregon 97006
{emily,roark,lmblack,vansanten}@cslu.ogi.edu

Abstract

Atypical or idiosyncratic language is a characteristic of autism spectrum disorder (ASD). In this paper, we discuss previous work identifying language errors associated with atypical language in ASD and describe a procedure for reproducing those results. We describe our data set, which consists of transcribed data from a widely used clinical diagnostic instrument (the ADOS) for children with autism, children with developmental language disorder, and typically developing children. We then present methods for automatically extracting lexical and syntactic features from transcripts of children's speech to 1) identify certain syntactic and semantic errors that have previously been found to distinguish ASD language from that of children with typical development; and 2) perform diagnostic classification. Our classifiers achieve results well above chance, demonstrating the potential for using NLP techniques to enhance neurodevelopmental diagnosis and atypical language analysis. We expect further improvement with additional data, features, and classification techniques.

1 Introduction

Atypical language and communication have been associated with autism spectrum disorder (ASD) since Kanner (1943) first gave the name *autism* to the disorder. The Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002) and other widely used diagnostic instruments include unusual word use as a diagnostic criterion. The broad and conflicting definitions used in diagnostic instruments for ASD, however, can lead to difficulty distinguishing the language peculiarities associated with autism.

The most recent and the most systematic study of unusual word use in ASD (Volden and Lord, 1991) found that certain types of atypical word use were significantly more prevalent in ASD speech than in the speech of children with typical development (TD). Although the results provided interesting information about unusual language in ASD, the process of coding these types of errors was laborious and required substantial linguistic and clinical expertise.

In this paper, we first use our own data to reproduce a subset of the results reported in Volden and Lord (1991). We then present a method of automatically identifying the types of errors associated with ASD using spoken language features and machine learning techniques. These same features are then used to differentiate subjects with ASD or a developmental language disorder (DLD) from those with TD. Although these linguistic features yield strong classification results, they also reveal a number of obstacles to distinguishing language characteristics associated with autism from those associated with language impairment.

2 Previous Work

Since it was first recognized as a neurodevelopmental disorder, autism has been associated with language described variously as: "seemingly nonsensical and irrelevant", "peculiar and out of place in ordinary conversation" (Kanner, 1946); "stereotyped", "metaphorical", "inappropriate" (Bartak et al., 1975); and characterized by "a lack of ease in

Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, pages 88–96,
Portland, Oregon, June 2011. ©2011 Association for Computational Linguistics

the use of words" (Rutter, 1965) and "the use of standard, familiar words or phrases in idiosyncratic but meaningful way" (Volden and Lord, 1991). The three most common instruments used in ASD diagnosis – the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002), the Autism Diagnostic Interview-Revised (ADI-R) (Lord et al., 1994), and the Social Communication Questionnaire (SCQ) (Rutter et al., 2003) – make reference to these language particularities in their scoring algorithms. Unfortunately, the guidelines for identifying this unusual language are often vague (SCQ: "odd", ADI-R: "idiosyncratic", ADOS: "unusual") and sometimes contradictory (ADOS: "appropriate" vs. ADI-R: "inappropriate"; ADOS: "phrases...they could not have heard" vs. SCQ: "phrases that he/she has heard other people use").

In what is one of the only studies focused specifically on unusual word use in ASD, Volden and Lord (1991) transcribed two 10-minute speech samples from the ADOS for 20 school-aged, high-functioning children with autism and 20 with typical development. Utterances containing non-English words or the unusual use of a word or phrase were flagged by student workers and then categorized by the authors into one of three classes according to the type of error:

- Developmental syntax error: a violation of a syntactic rule normally acquired in early childhood, such as the use of object pronoun in subject position or an overextension of a regular morphological rule, e.g., *What does cows do?*

- Non-developmental syntax error: a syntactic error not commonly observed in the speech of children acquiring language, e.g., *But in the car it's some.*

- Semantic error: a syntactically intact sentence with an odd or unexpected word given the context and intended meaning, e.g., *They're siding the table.*

The authors found that high-functioning children with ASD produced significantly more non-developmental and semantic errors than children with typical development. The number of developmental syntax errors was not significantly different between these two groups.

Although there has been virtually no previous work on automated analysis of unannotated transcripts of the speech of children with ASD, automatically extracted language features have shown promise in the identification of other neurological disorders such as language impairment and cognitive impairment. Gabani et al. (2009) used part-of-speech language models to derive perplexity scores for transcripts of the speech of children with and without language impairment. These scores offered significant diagnostic power, achieving an F1 measure of roughly 70% when used within an support vector machine (SVM) for classification. Roark et al. (in press) extracted a much larger set of language complexity features derived from syntactic parse trees from transcripts of narratives produced by elderly subjects for the diagnosis of mild cognitive impairment. Selecting a subset of these features for classification with an SVM yielded accuracy, as measured by the area under the receiver operating characteristic curve, of 0.73.

Language models have also been applied to the task of error identification, but primarily in writing samples of ESL learners. Gamon et al. (2008) used word-based language models to detect and correct common ESL errors, while Leacock and Chodorow (2003) used part-of-speech bigram language models to identify potentially ungrammatical two-word sequences in ESL essays. Although these tasks differ in a number of ways from our tasks, they demonstrate the utility of using both word and part-of-speech language models for error detection.

3 Data Collection

3.1 Subjects

Our first objective was to gather data in order reproduce the results reported in Volden and Lord (1991). As shown in Table 1, the participants in our study were 50 children ages 4 to 8 with a performance IQ greater than 80 and a diagnosis of either typical

Diagnosis	Count	Age (s.d.)	IQ (s.d.)
TD	17	6.24 (1.38)	125.7 (11.63)
ASD	20	6.38 (1.25)	108.9 (16.41)
DLD	13	7.01 (1.10)	100.6 (10.95)

Table 1: Count, mean age and IQ by subject group.

development (TD, n=17), autism spectrum disorder (ASD, n=20), or developmental language disorder (DLD, n=13).

Developmental language disorder (DLD), also sometimes known as specific language impairment (SLI), is generally defined as the delayed or impaired acquisition of language without accompanying comparable delays or deficits in hearing, cognition, and socio-emotional development (McCauley, 2001). The language impairments that characterize DLD are not related to articulation or "speech impediments" but rather are associated with more profound problems producing and often comprehending language in terms of its pragmatics, syntax, semantics, and phonology. The DSM-IV-TR (American Psychiatric Association, 2000) includes neither DLD nor SLI as a disorder, but for the purposes of this work, DLD corresponds to the DSM's designations *Expressive Language Disorder* and *Mixed Expressive-Receptive Language Disorder*.

For this study, a subject received a diagnosis of DLD if he or she met one of two commonly used criteria: 1) The Tomblin Epi-SLI criteria (Tomblin, et al., 1996), in which diagnosis of language impairment is indicated when scores in two out of five domains (vocabulary, grammar, narrative, receptive, and expressive) are greater than 1.25 standard deviations below the mean; and 2) The CELF-Preschool-2/CELF-4 criteria, in which diagnosis of language impairment is indicated when one out of three index scores and one out of three spontaneous language scores are more than one standard deviation below the mean.

A diagnosis of ASD required a previous medical, educational, or clinical diagnosis of ASD, which was then confirmed by our team of clinicians according to the criteria of the DSM-IV-TR (American Psychiatric Association, 2000), the revised algorithm of the ADOS (Lord et al., 2002), and the SCQ parental interview (Rutter et al., 2003). Fifteen of the 20 ASD subjects participating in this study also met at least one of the above described criteria for DLD.

3.2 Data Preparation

The ADOS (Lord et al., 2002), a semi-structured series of activities designed to reveal behaviors associated with autism, was administered to all 50 subjects. Five of the ADOS activities that require significant amounts spontaneous speech (Make-Believe Play, Joint Interactive Play, Description of a Picture, Telling a Story From a Book, and Conversation and Reporting) were then transcribed at the utterance level for all 50 speakers. All utterances from the transcripts longer than four words (11,244) were presented to individuals blind to the purposes of the study, who were asked to flag any sentence with atypical or unusual word use. Those sentences were then classified by the authors as having no errors or one of the three error types described in Volden and Lord. Examples from our data are given in Table 2.

3.3 Reproducing Previous Results

In order to compare our results to those reported in Volden and Lord, we calculated the rates of the three types of errors for each subject, as shown in Table 2. With a two-sample (TD v. ASD) t-test, the rates of nondevelopmental and semantic errors were significantly higher in the ASD group than in the TD group, while there was no significant difference in developmental errors between the two groups. These results reflect the same trends observed in Volden and Lord, in which the raw counts of both developmental and semantic errors were higher in the ASD group.

Using ANOVA for significance testing over all three diagnostic groups, we found that the rate of developmental errors was significantly higher in the DLD group than in the other groups. The difference in semantic error rate between TD and ASD using the t-test was preserved, but the difference in nondevelopmental error rate was lost when comparing all three diagnostic groups with ANOVA, as shown in Figure 1.

Error	Example
Dev.	I have a games. The baby drinked it. The frogs was watching TV.
Nondev.	He locked him all of out. Would you like to be fall down? He got so the ball went each way.
Sem.	Something makes my eyes poke. It smells like it's falling on your head. All the fish are leaving in the air.

Table 2: Examples of error types.

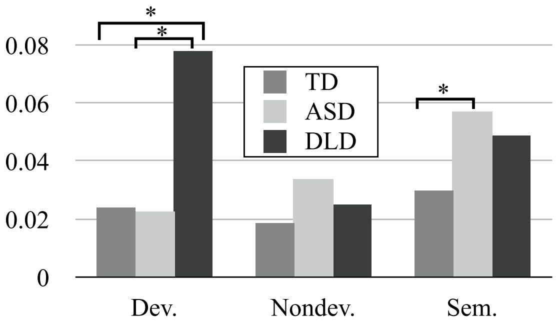

Figure 1: Error rates by diagnostic group (*p <0.05).

The process of manually identifying sentences with atypical or unusual language was relatively painless, but determining the specific error types is subjective and time-consuming, and requires a great deal of expertise. In addition, although we do observe significant differences between groups, it is not clear whether the differences are sufficient for diagnostic classification or discrimination.

We now propose automatically extracting from the transcripts various measures of linguistic likelihood, complexity, and surprisal that have the potential to objectively capture qualities that differentiate 1) the three types of errors described above, and 2) the three diagnostic groups discussed above. In the next three sections, we will discuss the various linguistic features we extract; methods for using these features to classify each sentence according to its error type for the purpose of automatic error-detection; and methods for using these features, calculated for each subject, for diagnostic classification.

4 Features

N-gram cross entropy. Following previous work in both error detection (Gamon et al., 2008; Leacock and Chodorow, 2003) and neurodevelopmental diagnostic classification (Gabani et al., 2009), we begin with simple bigram language model features. A bigram language model provides information about the likelihood of a given item (e.g., a word or part of speech) in a sentence given the previous item in that sentence. We suspect that some of the types of unusual language investigated here, in particular those seen in the syntactic errors shown in Table 2, are characterized by unlikely words (*drinked*) and word or part-of-speech sequences (*a games*, *all of out*) and hence might be distinguished by language model-based scores.

We build a word-level bigram language model and a part-of-speech level bigram language model from the Switchboard (Godfrey et al., 1992) corpus. We then automatically generate part-of-speech tags for each sentence (where the tags were derived from the best scoring output of the full syntactic parser mentioned below), and then apply the two models to each sentence. For each sentence, we calculate its cross entropy and perplexity. For a word string $w_1 \ldots w_n$ of length n, the cross entropy H is

$$H(w_1 \ldots w_n) \quad = \quad -\frac{1}{n} \log \mathrm{P}(w_1 \ldots w_n) \quad (1)$$

where $\mathrm{P}(w_1 \ldots w_n)$ is calculated as the product of the n-gram probabilities of each word in the string. The corresponding measure can be calculated for the POS-tag sequence, based on an n-gram model of tags. Perplexity is simply 2^H.

While we would prefer to use a corpus that is closer to the child language that we are attempting to model, we found the conversational style of the Switchboard corpus to be the most effective large corpus that we had at our disposal for this study. As the size of our small corpus grows, we intend to make use of the text to assist with model building, but for this study, we used all out-of-domain data for n-gram language models and parsing models. Using Switchboard also allowed us to use the same corpus to train both n-gram and parsing models.

Surprisal-based features. Surprisal, or the unexpectedness of a word or syntactic category in a given context, is often used as a psycholinguistic measure of sentence-processing difficulty (Hale, 2001; Boston et al., 2008). Although surprisal is usually discussed in the context of cognitive load for language processing, we hoped that it might also capture some of the language characteristics of the semantic errors like those in Table 2, which often contain common words used in surprising ways, and the nondevelopmental syntax errors, which often include strings of function words presented in an order that would be difficult to anticipate.

To derive surprisal-based features, each sentence is parsed using the Roark (2001) incremental top-down parser relying on a model built again on the Switchboard corpus. The incremental output of the parser shows the surprisal for each word, as well as other scores, as presented in Roark et al. (2009). For each sentence, we collected the mean surprisal (equivalent to the cross entropy given the model); the mean syntactic surprisal; and the mean lexical surprisal. The lexical and syntactic surprisal are a decomposition of the total surprisal into that portion due to probability mass associated with building non-terminal structure (syntactic surprisal) and that portion due to probability mass associated with building terminal lexical items in the tree (lexical surprisal). We refer the reader to that paper for further details.

Other linguistic complexity measures The non-developmental syntax errors in Table 2 are characterized by their ill-formed syntactic structure. Following Roark et al. (in press), in which the authors explored the relationship between linguistic structural complexity and cognitive decline, and Sagae (2005), in which the authors used automatic syntactic annotation to assess syntactic development, we also investigated the following measures of linguistic complexity: words per clause, tree nodes per word, dependency length per word, and Ygnve and Frazier scores per word. Each of these scores can be calculated from a provided syntactic parse tree, and to generate these we made use of the Charniak parser (Charniak, 2000), also trained on the Switchboard treebank.

Briefly, words per clause is the total number of words divided by the total number of clauses; and tree nodes per word is the total number of nodes in the parse tree divided by the number of words. The dependency length for a word is the distance (in word tokens) between that word and its governor, as determined through standard head-percolation methods from the output of the Charniak parser. We calculate the mean of this length over all words in the utterance. The Yngve score of a word is the size of the stack of a shift-reduce parser after that word; and the Frazier score essentially counts how many intermediate nodes exist in the tree between the word and its lowest ancestor that is either the root or has a left sibling in the tree. We calculate the mean of both of these scores over the utterance. We refer the reader to the above cited paper for more details on these measures.

As noted in Roark et al. (in press), some of these measures are influenced by particular characteristics of the Penn Treebank style trees – e.g., flat noun phrases, etc. – and measures vary in the degree to which they capture divergence from typical structures. Some (including Yngve) are sensitive to the breadth of trees (e.g., flat productions with many children); others (including Frazier) are sensitive to depth of trees. This variability is a key reason for including multiple, complementary features, such as both Frazier and Yngve scores, to capture more subtle syntactic characteristics than would be available from any of these measures alone.

Although we were not able to measure parsing accuracy on our data set and how it might affect the reliability of these features, Roark et al. (in press) did investigate this very issue. They found that all of the above described syntactic measures, when they were derived from automatically generated parse trees, correlated very highly (greater than 0.9) with those measures when they were derived from manually generated parse trees. For the moment, we assume that the same principle holds true for our data set, though we do intend both to verify this assumption and to supplement our parsing models with data from child speech. Based on manual inspection of parser output, the current parsing model does seem to be recovering largely valid structures.

5 Error Classification

The values for 8 of the 12 features were significantly different over the three error classes, as measured by one-way ANOVA: words per clause, Yngve, dependency, word cross-entropy all significant at $p <$ 0.001; Frazier, nodes per word at $p <$ 0.01; overall surprisal and lexical surprisal at $p <$ 0.05. We built classification and regression trees (CART) using the Weka data mining software (Hall et al., 2009) using all of the 12 features described above to predict which error each sentence contained, and we report the accuracy, weighted F measure, and area under the receiver operating characteristic curve (AUC).

Including all 12 features in the CART using 10-fold cross validation resulted in an AUC of 0.68, while using only those features with significant between-group differences yielded an AUC of 0.65.

Classifier	Acc.	F1	AUC
Baseline 1	41%	0.24	0.5
Baseline 2	33%	0.32	0.5
All features	53%	0.53	0.68
Feature subset	49%	0.49	0.65

Table 3: Error-type classification results.

These are both substantial improvements over a baseline with an unbalanced corpus in which the most frequent class is chosen for all input items (Baseline 1) or a baseline with a balanced corpus in which class is chosen at random (Baseline 2), which both have an AUC of 0.5. The results for each of these classifiers, provided in Table 3, show potential for automating the identification of error type.

6 Diagnostic Classification

In Section 3, we found a number of significant differences in error type production rates across our three diagnostic groups. Individual rates of error production, however, provide almost no classification power within a CART (AUC = 0.51). Perhaps the phenomena being observed in ASD and DLD language are related to subtle language features that are less easily identified than simply the membership of a sentence in one of these three error categories.

Given the ability of our language features to discriminate error types moderately well, as shown in Section 5, we decided to extract these same 12 features from every sentence longer than 4 words from the entire transcript for each of the subjects. We then took the mean of each feature over all of the sentences for each speaker. These per-speaker feature vectors were used for diagnostic classification within a CART.

We first performed classification over the three diagnostic groups using the full set of 12 features described in Section 4. This results in only modest gains in performance over the baseline that uses error rates as the only features. We then used ANOVA to determine which of the 12 features differed significantly across the three groups. Only four features were found to be significantly different across the three groups (words per clause, Yngve, dependency, word cross entropy), and none of them different significantly between the ASD group and the DLD group. As expected, classification did not improve with this feature subset, as reported in Table 4.

Features	Acc.	F1	AUC
Error rates	33%	0.32	0.51
All features	42%	0.38	0.59
Feature subset	40%	0.37	0.6

Table 4: All subjects: Diagnostic classification results.

Recall that 15 of the 20 ASD subjects also met at least one criterion for a developmental language disorder. Perhaps the language peculiarities we observe in our subjects with ASD are related in part to language characteristics of DLD rather than ASD. We now attempt to tease apart these two sources of unusual language by investigating three separate classification tasks: TD vs. ASD, TD vs. DLD, and ASD vs. DLD.

6.1 TD vs. ASD

We perform classification of the TD and ASD subjects with three feature sets: 1) per-subject error rates; 2) all 12 features described in Section 4; and 3) the subset of significantly different features. We found that 7 of the 12 features explored in Section 4 differed significantly between the TD group and the ASD group: words per clause, Yngve, dependency, word cross-entropy, overall surprisal, syntactic surprisal, and lexical surprisal. Classification results are shown in Table 5. We see that using the automatically derived linguistic features improves classification substantially over the baseline using per-subject error rates, particularly when we use the feature subset. Note that the best classification accuracy results are comparable to those reported in related work on language impairment and mild cognitive impairment described in Section 2.

6.2 TD vs. DLD

We perform classification of TD and DLD subjects with the same three feature sets used for the TD vs. ASD classification. We found that 6 of the 12

Features	Acc.	F1	AUC
Error rates	62%	0.62	0.56
All features	62%	0.62	0.65
Feature subset	68%	0.67	0.72

Table 5: TD vs. ASD: Diagnostic classification results.

Features	Acc.	F1	AUC
Error rates	67%	0.67	0.72
All features	80%	0.79	0.75
Feature subset	77%	0.75	0.66

Table 6: TD vs. DLD: Diagnostic classification results.

features explored in Section 4 different significantly between the TD group and the ASD group: words per clause, Yngve, dependency, word cross-entropy, overall surprisal, and lexical surprisal. Note that this is a subset of the features that differed between the TD group and ASD group. Classification results are shown in Table 6. Interestingly, using per-subject error rates for classification of TD and DLD subjects was quite robust. Using all of the features improved classification somewhat, while using only a subset resulted in degraded performance. We see that the discriminative power of these features is superior to that reported in earlier work using LM-based features for classification of specific language impairment (Gabani et al., 2009).

6.3 ASD vs. DLD

Finally, we perform classification of the ASD and DLD subjects using only the first two features sets, since there were no features found to be even marginally significantly different between these two groups. Classification results, which are dismal for both feature sets, are shown in Table 7.

6.4 Discussion

It seems quite clear that the error rates, feature values, and classification performance are all being influenced by the fact that a majority of the ASD subjects also meet at least one criterion for a developmental language disorder. Neither error rates nor feature values could discriminate between the ASD and DLD group. Nevertheless we see that our ASD group and DLD group do not follow the same patterns in their error production or language feature scores. Clearly there are differences in the language patterns of the two groups that are not being captured with any of the methods discussed here.

Features	Acc.	F1	AUC
Error rates	55%	0.52	0.48
All features	58%	0.44	0.40

Table 7: ASD vs. DLD: Diagnostic classification results.

We also observe that the error rates themselves, while sometimes significantly different across groups as originally observed in Volden and Lord, do not perform well as diagnostic features for ASD in our framework. Volden and Lord did not attempt classification in their study, so it is not known whether the authors would have encountered the same problem. There are, however, a number of possible explanations for a discrepancy between our results and theirs. First, our data was gathered from pre-school and young school-aged children, while the Volden and Lord subjects were generally teenagers and young adults. The way in which their spoken language samples were elicited allowed Volden and Lord to use raw error counts rather than error rates. There may also have been important differences in the way we carried out the manual error identification process, despite our best efforts to replicate their procedure. Further development of our classification methods and additional data collection are needed to determine the utility of error type identification for diagnostic purposes.

7 Future Work

Although our classifiers using automatically extracted features were generally robust, we expect that including additional classification techniques, subjects (especially ASD subjects without DLD), and features will further improve our results. In particular, we would like to explore semantic and lexical features that are less dependent on linear order and syntactic structure, such as Resnik similarity and features derived using latent semantic analysis.

We also plan to expand the training input for the language model and parser to include children's speech. The Switchboard corpus is conversational speech, but it may fail to adequately model many linguistic features characteristic of small children. The CHILDES database of children's speech, although it is not large enough to be used on its own for our analysis and would require significant manual syntactic annotation, might provide enough data for us to adapt our models to the child language domain.

Finally, we would like to investigate how informative the error types are and whether they can be reliably coded by multiple judges. When we examined the output of our error-type classifier, we noticed that many of the misclassified examples could be construed, upon closer inspection, as belonging to multiple error classes. The sentence *He's flying in a lily-pond*, for instance, could contain a developmental error (i.e., the child has not yet acquired the correct meaning of *in*) or a semantic error (i.e., the child is using the word *flying* instead of *swimming*). Without knowing the context in which the sentence was uttered, it is not possible to determine the type of error through any manual or automatic means. The seemingly large number of misclassifications of sentences like this indicates the need for further investigation of the existing coding procedure and in-depth classification error analysis.

8 Conclusions

Our method of automatically identifying error type shows promise as a supplement to, or substitute for, the time-consuming and subjective manual coding process described in Volden and Lord (Volden and Lord, 1991). However, the superior performance of our automatically extracted language features suggests that perhaps it may not be the errors themselves that characterize the speech of children with ASD and DLD but rather a preference for certain structures and word sequences that sometimes manifest themselves as clear language errors. Such variations in complexity and likelihood might be too subtle for humans to reliably observe.

In summary, the methods explored in this paper show potential for improving diagnostic discrimination between typically developing children and those with these neurodevelopmental disorders. Further research is required, however, in finding the most reliable markers that can be derived from such spoken language samples.

Acknowledgments

This work was supported in part by NSF Grant #BCS-0826654; an Innovative Technology for Autism grant from Autism Speaks; and NIH NIDCD grant #1R01DC007129-01. 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, Autism Speaks, or the NIH. Thanks to Meg Mitchell and Cheryl Greene for their assistance with this project.

References

American Psychiatric Association. 2000. *DSM-IV-TR: Diagnostic and Statistical Manual of Mental Disorders.* American Psychiatric Publishing, Washington, DC, 4th edition.

Laurence Bartak, Michael Rutter, and Anthony Cox. 1975. A comparative study of infantile autism and specific developmental receptive language disorder. I. The children. *British Journal of Psychiatry*, 126:27145.

Mariss Ferrara Boston, John Hale, Reinhold Kliegl, and Shravan Vasishth. 2008. Surprising parser actions and reading difficulty. In *Proceedings of ACL-08:HLT, Short Papers*.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the ACL*, pages 132–139.

Keyur Gabani, Melissa Sherman, Thamar Solorio, and Yang Liu. 2009. A corpus-based approach for the prediction of language impairment in monolingual english and spanish-english bilingual children. In *Proceedings of NAACL-HLT*, pages 46–55.

Michael Gamon, Jianfeng Gao, Chris Brockett, and Re Klementiev. 2008. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of IJCNLP*.

John J. Godfrey, Edward Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of ICASSP*, volume 1, pages 517–520.

John T. Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd meeting of NAACL*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1).

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

Leo Kanner. 1946. Irrelevant and metaphorical language. *American Journal of Psychiatry*, 103:242–246.

Claudia Leacock and Martin Chodorow. 2003. Automated grammatical error detection. In M.D. Shermis and J. Burstein, editors, *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

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 disorders. *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.

Rebecca McCauley. 2001. *Assessment of language disorders in children*. Lawrence Erlbaum Associates, Mahwah, NJ.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of EMNLP*, pages 324–333.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristina Hollingshead, and Jeffrey Kaye. in press. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

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

Michael Rutter. 1965. Speech disorders in a series of autistic children. In A. Franklin, editor, *Children with communication problems*, pages 39–47. Pitman.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the ACL*.

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