Toward Characterizing the Language of Adults with Autism in Collaborative Discourse

Christine Yang[†], Emily Prud'hommeaux[†], Laura B. Silverman[‡], Allison Canfield[*]
[†]Boston College, Chestnut Hill MA, USA,
[‡]University of Rochester, Rochester NY, USA
[*]Upstate Medical Center, Syracuse NY, USA
{yangael,prudhome}@bc.edu, laura_silverman@urmc.rochester.edu

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental condition associated with life-long deficits in communication that can impact both personal and professional well-being. Although the linguistic features associated with these deficits are routinely observed in clinical settings, they are difficult to quantify. In this paper, we present a growing dataset of conversations between high-functioning adults with ASD and their neurotypical conversational partners as they complete a collaborative task. We compare the linguistic characteristics of the two groups using both manual annotations and computational linguistic features extracted from these conversations. Our results indicate that there are quantifiable differences in the language use of adults with ASD in collaborative discourse scenarios, demonstrating the promise of our methods and dataset.

Keywords: autism spectrum disorder, discourse, pragmatics, spoken language analysis, corpus analysis, speech acts

1. Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition associated with life-long deficits in communication and social engagement. Among these deficits is impaired pragmatic expression, or the inappropriate use of language in a given context (Kanner, 1943; Lord and Paul, 1997; Young et al., 2005; Simmons et al., 2014). Because of the pragmatic difficulties they experience, individuals with ASD face challenges in establishing interpersonal relationships, maintaining satisfactory employment, and achieving independence (Mok et al., 2014; Whitehouse et al., 2009; Hendricks, 2010). Researchers do not agree, however, on precisely what functions are impaired, particularly in high-functioning adults. Analyzing spontaneous spoken language is an effective way to reveal these impairments, but there has been relatively little work on either manually annotating or computationally analyzing spontaneous language data from adults with ASD. As a result, there are no publicly available conversational spoken language datasets produced by adults with ASD.

In this paper, we describe a growing dataset of transcribed conversations between high-functioning adults with ASD or typical development (TD) and their neurotypical conversational partners as they work together to navigate from one location to another on a shared map. Although the number of study participants whose collaborative conversations has been transcribed thus far is modest, the data collection project that is the source of these conversations is ongoing and will include, within the next 18 months, 60 to 75 participants. The dataset includes conversations from two additional collaborative tasks and as well as spoken responses to a variety of widely used clinical instruments. We will be making the transcripts, as well as the manual annotations we have created, available to researchers who can demonstrate that they have completed their institution's human subjects training in the hopes that the data will reveal new and useful information about the strengths and weaknesses in pragmatic expression associated with ASD.

Here we present the results of both manual and automated computational analyses of the data collected so far. Our findings suggests that there are observable and quantifiable differences between adults with ASD and those with typical development on several discourse-level pragmatic dimensions. These results underscore the importance of examining spontaneous conversational speech in adults with ASD and point to the promise of automated computational approaches for clinical language analysis.

2. Related Work

Atypical language has been observed in verbal individuals with autism since the disorder was first described (Kanner, 1943) and continues to serve as a diagnostic criterion in many widely used instruments for diagnosing autism (Lord et al., 2002; Rutter et al., 2003). Atypical use of language for a given context, known as pragmatic expression, seems to be universally affected in autism, even in the absence of structural language impairments in syntax, morphology, and phonology (Eales, 1993; Landa, 2000; Young et al., 2005; Simmons et al., 2014). In high-functioning individuals with autism, impaired pragmatic expression is associated with challenging behaviors (Ketelaars et al., 2010), difficulty developing relationships (Whitehouse et al., 2009), and struggles in maintaining employment (Hendricks, 2010).

The most promising methods for pinpointing the pragmatic features that characterize autism rely on careful manual annotation of transcripts of spontaneous spoken language (Volden and Lord, 1991; Bishop et al., 2000; Adams, 2002; Gorman et al., 2016; Canfield et al., 2016). Carrying out

Dx	PIQ	Age
TD (n=5)	103 (10.5)	22y5m (4y3m)
ASD (n=9)	106 (8.4)	19y11m (1y9m)

Table 1: Summary demographic statistics of our current set of participants: mean (s.d.).

complex annotations schema, however, requires training and expertise, making these methods are impractical to deploy. There has been relatively little work in applying computational methods for identifying these sorts of linguistics features in the language of individuals with ASD, and this work has focused exclusively on the language of children and language produced in a semi-structured context (Prud'hommeaux et al., 2014; Losh and Gordon, 2014; Parish-Morris et al., 2016; Goodkind et al., 2018).

The language resource and accompanying analysis presented here makes several novel contributions. First, this language data is produced by adults, a subgroup of the ASD population that is both understudied and underserved. Second, the dataset consists entirely of spontaneous conversations in a restricted semantic domain. Third, the dataset has been manually annotated to indicate the category of speech act for each turn and a numeric rating on several scales, including politeness, uncertainty, and informativeness.

3. Data Collection

3.1. Spoken Language Data

As part of a project investigating differences in pragmatic expression in adults with ASD, we are collecting spoken language data from high-functioning adult participants with ASD and with typical development (TD). Participants with ASD must meet criteria for a diagnosis of ASD on the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002), as well as the following basic eligibility criteria: (1) full-scale IQ (PIQ) $>$ 80; (2) verbal IQ (VIQ) $>$ 80; (3) monolingual American English speaking; and (4) no history of language impairment, auditory processing disorder, or hearing difficulty. Neurotypical participants are selected in order to match the experimental participants on age, VIQ, PIQ, gender, and ethnicity. Because our data collection is in progress, the participants analyzed here may not yet be matched on all dimensions. Table 1 presents mean values for age, full-scale IQ for our 9 participants with ASD and 5 participants with TD.

Each participant, whether with ASD or TD, is paired with a neurotypical conversational partner, with whom they perform several collaborative tasks that require verbal communication and deliberation. In one task, each person is given a map of the same place, but with slight differences in the place names and the location of obstacles referred to as "road blocks" (Anderson et al., 1991). Each map is also labeled with a mark to show where the conversational partner is located. The participant's task is to give directions to the conversational partner to get them to their position on the map. Currently, we have collected recorded conversations from 14 pairs of participants, with 5 participants in the TD group and 9 participants in the ASD group.

After the spoken data is collected, the recordings are transcribed using Praat (Boersma, 2001). All filler words, discourse markers, and words or sounds of affirmation, negation, or exclamation are included in the transcripts, as these serve as important tools for expression and may be informative for pragmatic analysis. Transcriptions also included annotations for sounds effects or onomatopoeia, partial or interrupted words, and unfilled pauses within an utterance. Utterances are segmented using the concept of the *C-unit*, which is formally defined as "an independent clause with its modifiers" (Loban, 1976). It includes the main clause and all subordinate clauses, and cannot be further segmented without losing its essential meaning. It does not have to be a complete syntactical sentence, and may comprise of a single coordinate clause (using coordinating conjunctions "and", "so", "then", etc.), but not a single subordinate clause (using subordinating conjunctions "because", "if", "when", etc.). Each utterance is punctuated in the transcript with one of several punctuation marks for exclamations, questions, regular statements, abandoned utterances, and interrupted utterances. Our corpus of 14 transcripts currently consists of about 4,463 utterances in total, with 3,019 utterances in the TD group and 1,444 in the ASD group. Of these, we have 2,240 utterances from the participants which we use for the analyses that follow.

3.2. Linguistic Features

For each utterance, we gathered several pragmatic features which we believed could be significant in illuminating the differences between adults with ASD and adults with TD. In our selection of potentially significant linguistic features, we explored both manually annotated features as well as computationally derived features generated by existing models and toolkits, described in the subsections below. The manual annotations investigated potential pragmatic differences identifiable by human observers using a set annotation guidelines, under the assumption that a pragmatic feature perceptible to an annotator would also be perceptible to a conversational partner and thus have an effect in real-world communication and pragmatics. We also use existing computational tools to predict further pragmatic features for each utterance. The ability of the automated features to capture meaningful conclusions is, of course, dependent on the model and corpus used to generate the feature ratings, but it is still worth investigating these features to see whether they may point to some linguistic differences between the two groups that are not captured by the manual annotations. The predicted features are also much easier and less time-consuming to acquire and may thus help us determine which additional features might be worthwhile for future exploration and annotation.

3.2.1. Manually Annotated Features

Two human annotators were assigned to annotate each utterance with a numerical score for politeness, uncertainty, and information content. Each category was originally rated on a discrete scale of 1 to 5, but it was later collapsed to a scale of 1 to 3 as the smaller scale helped improve inter-annotator agreement. Each utterance was treated independently and rated without consideration for the context surrounding it, as the eventual goal is to potentially train a model that can assign these ratings automatically on an utterance by utterance level. Therefore, identical utterances in different contexts were given the same feature ratings. The annotated feature categories are defined as follows:

Politeness: The *politeness* rating is a measure of how positive, agreeable, and non-demanding an utterance is. A most polite utterance shows high positivity and willingness to compromise or admit wrongdoing. An utterance with a politeness score of 2 was given to neutral statements, which included direct questions ("where are you?"), objective observations ("the house is red"), and commands phrased indirectly ("then you wanna go left", "then you're gonna go left"). A high politeness rating of 3 was given to utterances which included positive or affirmative words ("great", "I agree", "true"), acknowledged the speaker's own mistakes, or contained distinct politeness markers ("please", "thank you"). Commands phrased as conditionals ("if you wanna make a left"), suggestions ("how about you go left"), or directed to both of them using the first person plural ("then we need to go left") had a score of 3. Requests using a modal ("could you tell me") were also given a rating of 3. A low politeness score of 1 was given to utterances which contained negative comments or expressed frustration ("how the heck am I supposed to say this?"), criticized the other person, or directly accused the other person of being wrong. Commands phrased as imperatives ("go left") or as necessity for the other person ("you have to go left", "I need you to go left") were also given a score of 1.

Uncertainty: The *uncertainty* rating is a measure of how uncertain the speaker is about a fact or about the accuracy of their utterance. An utterance with an uncertainty rating of 1 showed no clear signs of uncertainty, while a rating of 2 indicated some hesitation (filler words, pauses), hedging ("maybe", "might be"), or qualification to the statement ("if I'm reading this correctly"). Polar questions ("Is it red?") and words or phrases intended as questions for confirmation ("The one by the tree?") were also assigned an uncertainty rating of 2. Questions expecting a one word or phrase answer ("What color is the building?") and questions expecting longer explanations ("How do I get there from here?") were given the highest rating of 3. Directly stating "I don't know" or "I'm confused" was also given a high uncertainty rating of 3.

Information Content: The *information content* rating is a measure of the quantity and specificity of the information words contained in the utterance. Utterances containing no information at all and utterances containing

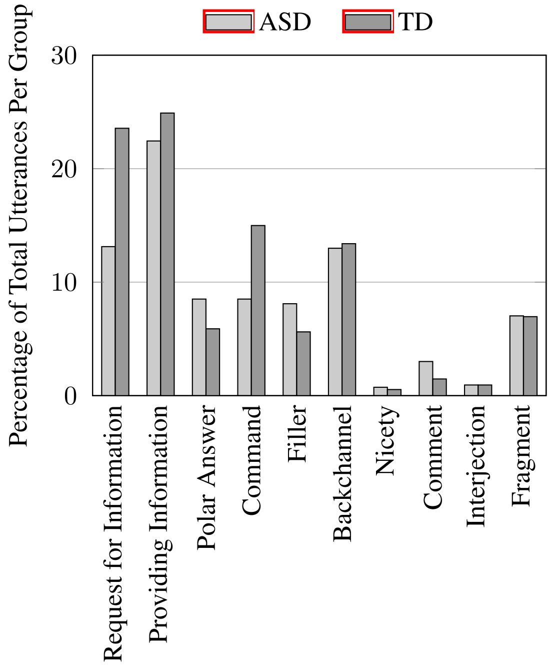

Figure 1: Speech act distribution per group.

some vague pronouns or polar answers ("yes", "I don't have it", "that one") were given a rating of 1. A score of 3 was given to utterances containing directional words ("left", "north", "down") or general object words that could refer to multiple items on the map ("bird", "red roof", "road block"). Utterances with the highest rating of 3 contained proper nouns and specific place names ("Hawk_Meadow", "compost site") or elaborate descriptions of objects or places that could only refer to a specific location on the map ("a big red house with two windows on the left side and four windows in the middle").

Speech Acts: Along with these features, the annotators were also asked to assign a speech act to each utterance. The set of speech acts used to annotate this specific dataset is defined in Table 2.

Between the two annotators, all 14 transcripts were annotated for the features and speech acts detailed above, with 8 transcripts having annotations from both annotators. To determine the inter-annotator agreement, we calculated the percentage of agreement and Cohen's *kappa* (Cohen, 1968) for the utterances that had been annotated by both annotators. Results are shown in Table 3.

The agreement for each category was above 80%, and the kappa scores for uncertainty, information content, and speech act were all in the substantial range of 0.61 to 0.80, as defined by Cohen (Cohen, 1968). The politeness feature had a high agreement but a lower kappa score of 0.54, likely because most of the utterances had a neutral politeness rating, which was true under both the original 5-point scale and the collapsed 3-point scale. This was likely due to the fact that the majority of utterances did not attempt to use any identifiable politeness strategies.

Speech Act	Description
Request for Information	Any request for information, clarification, or confirmation. May take the form of a question, or of a statement or phrase intended to be a question ("You want me to keep going?", "the one by the tree?")
Providing Information	Answering a request for information, or providing information unprompted.
Polar Answer	Answering "yes" or "no" to a polar question.
Command	An utterance that gives instruction or direction to the other person whether in the form of an imperative ("go left."), a suggestion ("how about you go left"), a hypothetical ("if you wanna make a left"), or a question ("do you wanna go left?"). It may be instructing the person on where to go on the map, or on how to act or strategize ("hold on", "how about you tell me where you are first").
Filler	Filler words or phrases used to fill pauses in the conversation ("hm", "anyways", "okay so").
Backchannel	An utterance that indicates the speaker is listening and understanding what the other person is saying ("okay", "mm-hm", "gotcha", "sounds good").
Nicety	Utterances that serve to maintain a polite and collaborative conversation, such as apologizing, expressing gratitude, or reassuring the conversational partner.
Comment	An utterance that contains commentary or an opinion on the task, such as explaining the speaker's own actions or discussing the best strategy to take.
Interjection	Short exclamations or interjections such as "ah", "oops", "yay", "wow", "ew", "awesome", etc.
Fragment	Short abandoned or interrupted utterances that are too incomplete to classify as any other speech act.

Table 2: The set of speech acts used in annotation and their descriptions with some examples.

Feature	Agreement	Kappa (κ) Score
Politeness	89.83%	0.544
Uncertainty	84.18%	0.691
Information Content	85.82%	0.758
Speech Act	81.19%	0.777

Table 3: Inter-annotator Agreement

3.2.2. Automated Computational Features

In addition to these manually annotated features, we also explored several computationally derived features extracted using existing toolkits and models used previously to characterize conversations in collaborative software development (Meyers et al., 2018). In particular, we extracted scores for politeness, uncertainty, formality, informativeness, and implicature. The *politeness* classifier uses an SVM model trained on over 10,000 annotated requests from online forums. It uses the Stanford CoreNLP software to generate dependency parses for preprocessing and assigns each utterance a politeness rating on a continuous scale from 0 to 1, with 1 being the most polite. The *uncertainty* classifier uses a logistic regression model trained on the Szeged Uncertainty Corpus (Vincze, 2014) and assigns each utterance a binary classification of either certain or uncertain. In this package, an uncertain utterance is defined as one for which the "truth value or reliability cannot be determined due to lack of information" (Vincze, 2014). The `squinky` package (Meyers et al., 2018) uses a logistic regression model trained on a corpus of over 7,000 annotated sentences and rates each utterance on formality, informative, and implicature on a scale of 0 to 1, with 1 being the most formal, informative, and implicative respectively. Formality is a linguistic strategy employed to effectively convey as much information as possible while adhering to Grice's maxims. The informative scale is related to the concept of *term informativeness*, and corresponds to how clearly and how directly the intended meaning is communicated. Implicature is a measure of the amount of missing or implied information in an utterance. We refer the reader to Meyers et al. (2018) and Vincze (2014) for further details.

4. Data Analysis

The frequency of speech acts in each group, shown in Figure 1, generally shows a similar distribution across diagnostic groups, with an increased usage in the TD group of requests for information, commands, and providing information. The percentage of speech acts in the TD group that are requests for information is over 10% higher than that of the ASD group. The fact that the TD group

Feature	ASD Average Rating	TD Average Rating
Manually Annotated Features (scale of 1 to 3)		
Politeness***	**2.02**	1.93
Uncertainty*	1.43	**1.48**
Information Content***	1.54	**1.77**
Automated Computational Features (scale of 0 to 1)		
Politeness	**0.448**	0.445
Uncertainty	0.954	**0.958**
Formality	**0.016**	0.014
Informativeness	**0.105**	0.089
Implicature***	0.35	**0.414**

Table 4: Average manual ratings and automated scores for pragmatic features. One asterisk indicates a significant difference between the two groups ($p < 0.05$). Three asterisks (***) indicates a highly significant difference between the two groups ($p < 0.0001$).

made greater use of speech acts which often include more information content might also contribute to the fact that the TD utterances were rated more highly for information content on average.

The two groups of participants generally had similar mean values for the annotated features, as shown in Table 4. For the manually annotated features, the ASD group was slightly more polite while the TD group contained more information content overall. The automated computational features also appeared to have very close averages between the two groups, with the exception of the implicature feature, for which the ASD participants' utterances scored noticeably higher.

To determine if the differences between the two groups were significant, we performed a two-tailed t-test over all the utterances of each participant group, with the results displayed in Table 4. This significance testing revealed that all three manually annotated features were significantly different across the two groups, with the ASD group having more polite utterances and the TD group showing more uncertainty and information content. The t-test also showed that there were generally more significant differences in the manually annotated features than in the automated features, with the exception of the automated implicature score which had a highly significant difference between the two diagnostic groups.

We also note that the two groups might employ different pragmatic strategies for different speech acts. To examine the linguistic features of the utterances for each individual speech act, we calculated average linguistic feature score for each speech act. The results for some speech acts of interest are displayed in Table 5. We again performed a t-test on this data, and from the results we can see that while there are no significant differences between the two groups when requesting information, there are significant differences in levels of politeness and information content when providing information, and in politeness and uncertainty when issuing a command.

Feature	ASD Average Rating	TD Average Rating
Request for Information (196 utterances)		
Politeness	**2.05**	2.04
Uncertainty	**2.12**	2.1
Information Content	2.01	**2.03**
Providing Information (335 utterances)		
Politeness*	**2.01**	1.99
Uncertainty	**1.37**	1.33
Information Content***	2.0	**2.26**
Command (127 utterances)		
Politeness***	**1.73**	1.40
Uncertainty**	**1.46**	1.29
Information Content	2.02	**2.1**

Table 5: Results of significance testing for manually annotated linguistics features in individual speech acts. One asterisk indicates a significant difference between the two groups ($p < 0.05$). Two asterisks (**) indicates a high significant difference between the two groups ($p < 0.01$).

5. Conclusion

Our results indicate that both manual and automated analysis of conversational data in a collaborative environment can reveal interesting and telling differences between the language use of high-functioning adults with autism spectrum disorder and their matched neurotypical peers. These findings provide the beginnings of quantitative support for the qualitative observations that are routinely made in clinical settings. By being able to identify atypical linguistic characteristics of specific utterances in a collaborative work scenario, our methods can contribute to the development of tools for remediating weaknesses in communication for adults with ASD, a historically underserved population. The data that we will release, including both the transcripts and the manual annotations, will serve as a resource for other researchers working to better understand the communicative challenges facing adults with ASD as they seek to find employment and live independently.

6. Acknowledgements

We thank Jill Aldrich for her work recruiting and running participants; and Erin Corcoran, Emily Fabius, and Hannah Jiang for their transcription and annotation of the audio recordings. The project is supported in part by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under award number R21DC017000. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

7. References

Adams, C. (2002). Practitioner review: The assessment of language pragmatics. *Journal of child psychology and psychiatry*, 43(8):973–987.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The hcrc map task corpus. *Language and Speech*, 34(4):351–366.

Bishop, D. V., Chan, J., Adams, C., Hartley, J., and Weir, F. (2000). Conversational responsiveness in specific language impairment: Evidence of disproportionate pragmatic difficulties in a subset of children. *Development and psychopathology*, 12(2):177–199.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345.

Canfield, A. R., Eigsti, I.-M., de Marchena, A., and Fein, D. (2016). Story goodness in adolescents with autism spectrum disorder (asd) and in optimal outcomes from asd. *Journal of Speech, Language, and Hearing Research*, 59(3):533–545.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Eales, M. J. (1993). Pragmatic impairments in adults with childhood diagnoses of autism or developmental receptive language disorder. *Journal of autism and developmental disorders*, 23(4):593–617.

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

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

Hendricks, D. (2010). Employment and adults with autism spectrum disorders: Challenges and strategies for success. *Journal of Vocational Rehabilitation*, 32(2):125–134.

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

Ketelaars, M. P., Cuperus, J., Jansonius, K., and Verhoeven, L. (2010). Pragmatic language impairment and associated behavioural problems. *International Journal of Language & Communication Disorders*, 45(2):204–214.

Landa, R. (2000). Social language use in asperger syndrome and high-functioning autism. *Asperger syndrome*, pages 125–155.

Lord, C. and Paul, R. (1997). Language and communication in autism. In D. Cohen et al., editors, *Handbook of autism and pervasive developmental disorders*, pages 195–225. Wiley.

Lord, C., Rutter, M., DiLavore, P., and Risi, S. (2002). *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services.

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

Meyers, B. S., Munaiah, N., Prud'hommeaux, E., Meneely, A., Wolff, J., Alm, C. O., and Murukannaiah, P. (2018). A dataset for identifying actionable feedback in collaborative software development. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 126–131.

Mok, P. L., Pickles, A., Durkin, K., and Conti-Ramsden, G. (2014). Longitudinal trajectories of peer relations in children with specific language impairment. *Journal of Child Psychology and Psychiatry*, 55(5):516–527.

Parish-Morris, J., Liberman, M., Ryant, N., Cieri, C., Bateman, L., Ferguson, E., and Schultz, R. T. (2016). Exploring autism spectrum disorders using hlt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 74–84.

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

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

Simmons, E. S., Paul, R., and Volkmar, F. (2014). Assessing pragmatic language in autism spectrum disorder: The yale in vivo pragmatic protocol. *Journal of Speech, Language, and Hearing Research*, 57(6):2162–2173.

Vincze, V. (2014). Uncertainty detection in hungarian texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1844–1853.

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

Whitehouse, A. J., Watt, H. J., Line, E., and Bishop, D. V. (2009). Adult psychosocial outcomes of children with specific language impairment, pragmatic language impairment and autism. *International Journal of Language & Communication Disorders*, 44(4):511–528.

Young, E., Diehl, J., Morris, D., Hyman, S., and Bennetto, L. (2005). Pragmatic language disorders in children with autism: The use of two formal tests to distinguish affected children from controls. *Language, Speech, and Hearing Services in Schools*, 36:62–72.