Evaluating the Performance of Transformer-based Language Models for Neuroatypical Language

Duanchen Liu[1,2], **Zoey Liu**[1], **Qingyun Yang**[1,3], **Yujing Huang**[1], **Emily Prud'hommeaux**[1]
[1]Department of Computer Science, Boston College, Chestnut Hill MA, USA
[2]MIT, Cambridge MA, USA [3]Cornell Tech, New York NY, USA
liudc@mit.edu, {liuaal,huangac,prudhome}@bc.edu, qny2@cornell.edu

Abstract

Difficulties with social aspects of language are among the hallmarks of autism spectrum disorder (ASD). These communication differences are thought to contribute to the challenges that adults with ASD experience when seeking employment, underscoring the need for interventions that focus on improving areas of weakness in pragmatic and social language. In this paper, we describe a transformer-based framework for identifying linguistic features associated with social aspects of communication using a corpus of conversations between adults with and without ASD and neurotypical conversational partners produced while engaging in collaborative tasks. While our framework yields strong accuracy overall, performance is significantly worse for the language of participants with ASD, suggesting that they use a more diverse set of strategies for some social linguistic functions. These results, while showing promise for the development of automated language analysis tools to support targeted language interventions for ASD, also reveal weaknesses in the ability of large contextualized language models to model neuroatypical language.

1 Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder with an estimated global prevalence of 1 in 100 worldwide (Zeidan et al., 2022). The majority of people diagnosed with ASD today are verbal (Rose et al., 2016) and have average or above average intellectual ability (Christensen et al., 2016). Nevertheless, young adults with ASD are employed at significantly lower rates than their peers with other neurodevelopmental conditions, including learning disabilities and intellectual disability (Shattuck et al., 2012).

Difficulty with social communication is one of the diagnostic criteria for ASD (American Psychiatric Association, 2013) and is reported to be one of the strongest contributors to negative professional outcomes (Hurlbutt and Chalmers, 2004; Baldwin et al., 2014). For this reason, language skills are commonly targeted for intervention in individuals with ASD (Parsons et al., 2017), but it can be difficult to identify specific areas in need of remediation. Most prior work on quantifying pragmatic deficits in ASD has relied on manual analysis of speech transcripts (Loukusa et al., 2007; Paul et al., 2009; Conlon et al., 2019), a process that is time consuming and requires expertise.

In this paper, we use a spoken language corpus collected specifically for training automatic systems to explore social communication and pragmatics in ASD. The corpus consists of transcribed conversations between adults with and without ASD as they engage with neurotypical interlocutors in collaborative tasks designed to resemble workplace activities. We review the careful manual process of assigning pragmatic feature values and dialog act labels to each utterance. Following recent prior work on these features in other contexts, we propose a BERT-based framework (Devlin et al., 2019) for automatically assigning these values and labels.

Although our models achieve higher accuracy than previous neural and non-neural approaches to these labeling tasks on this dataset, we observe that models for some features show significantly weaker performance on utterances produced by individuals with ASD. An error analysis with logistic mixed-effects regression reveals that these models fail to recognize unusual or idiosyncratic strategies for conveying certain social and pragmatic meanings. Our results, while showing promise for the automated analysis of social communication in ASD, point to an unsurprising but potentially problematic bias in models trained primarily on news and web data toward neurotypical language.

2 Background

Much of the prior work on extracting social communication and discourse features from conversa-

Proceedings of the 29th International Conference on Computational Linguistics, pages 3412–3419
October 12–17, 2022.

tions has focused on dialogue acts, yielding both a large number of corpora and nearly as many distinct annotation schemes (Stolcke et al., 2000; McCowan et al., 2005; Zhang et al., 2017; Bunt et al., 2019). Conversational corpora, mostly written, have also been manually annotated at the utterance level for specific pragmatic features, including politeness (Danescu-Niculescu-Mizil et al., 2013); uncertainty (Vincze, 2014; Farkas et al., 2010); and informativeness, formality, and implicature (Lahiri, 2015). Early work relied primarily on bag-of-words models with statistical classifiers, but recently transformer-based models have been used with success for many of these tasks (Aljanaideh et al., 2020; Hayati et al., 2021; Wu et al., 2020; Żelasko et al., 2021; Wu et al., 2021). We follow this prior work in our use of BERT (Devlin et al., 2019) for predicting feature labels.

While many recent studies have explored automated analysis of language in ASD, particularly in children (Parish-Morris et al., 2016; Adams et al., 2021; Salem et al., 2021), the most relevant to ours is Yang et al. (2021), which introduced a corpus of conversations between adults with and without ASD and neurotypical conversational partners, partially annotated for three pragmatic features. We go beyond this work in three ways. First, we complete the annotation and introduce a new feature, dialog act (see Section 3.2), which has not been studied previously in ASD language. Second, we use BERT directly rather than using BERT embeddings for prediction. Lastly, we give a statistical analysis of performance across diagnostic groups.

3 Data

3.1 Participants and tasks

The corpus used in the present study includes data previously described in Yang et al. (2021), which consists of conversations between neurotypical **conversational partners (CPs)** ($n = 11$) and adult **experimental participants (EPs)** 18-30 years of age with ASD ($n = 16$) and with typical development (TD, $n = 9$). ASD EPs met the diagnostic criteria for ASD on the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002). All EPs and CPs were monolingual speakers of American English, with no history of intellectual disability, language impairment, or hearing difficulties.

Each EP engaged with a CP in two collaborative discussion tasks. In the map task (Anderson et al., 1991), the EP and the CP were each given a map of the same area with slight differences in labels and obstacles. The EP was tasked with giving verbal directions to their CP to lead them to their marked position on the map. In the island task (Klippel et al., 1984), the EP and CP jointly viewed a set of labeled pictures of various items and discussed which items they would choose to bring to a deserted island. The conversations were transcribed by a team of three undergraduate research assistants yielding a total of 9433 utterances: 3091 produced by EPs with ASD; 1846 by EPs with TD; 2842 by CPs of ASD EPs; and 1654 by CPs of TD EPs.

The number of participants, while small, is quite typical for work that involves manual analysis of language in ASD, particularly for adults (Mawhood et al., 2000; Young et al., 2005; Parsons et al., 2017). We additionally note that our methods are applied to individual utterances rather than individual participants. The overall number of utterances in our dataset is on par with that of many widely used NLP datasets hand-labeled for similar features (Danescu-Niculescu-Mizil et al., 2013; Braley and Murray, 2018; Wang et al., 2019), and the statistical analyses we employ are appropriate for the size and distribution of our corpus.

3.2 Pragmatic feature annotation

The transcripts had previously been partially annotated by trained undergraduates with dual majors in Computer Science and either Linguistics or Psychology for three ordinal features: **politeness, uncertainty,** and **informativeness** (Yang et al., 2021). In our work, we completed this annotation process and introduced a new categorical feature, **dialog act**. For the ordinal features, two annotators, from a pool of four trained undergraduates double majoring in Computer Science and either Linguistics or Psychology, assigned to each utterance a rating on a three-point scale, achieving inter-annotator agreement as measured by Krippendorf's α (Artstein and Poesio, 2008) of 0.7, 0.76, and 0.83 for politeness, uncertainty, and informativeness, respectively. The final score for each utterance was the mean of the annotations for that feature. Example utterances with their respective scores for each of these features are shown in Table 1, and a brief overview of the annotation guidelines is provided in Appendix A. We refer the reader to Yang et al. (2021) for further details.

For the new categorical feature, dialog act, the annotator assigned to each utterance one from a

Task	Utterance	Politeness	Uncertainty	Informativeness	dialog act
Map	Do you have a pond on your map?	2	2	2	request for information
Map	Not at my area, no.	2	1	1	polar answer
Island	I don't care.	1	1	1	providing opinion
Island	You're on a deserted island.	2	1	2	providing information

Table 1: Examples of manual annotations for each task.

dialog act	Description
Request for Information	A request for factual information: *do you have a pond on your map?*
Providing Information	Answering a request for information or providing factual information unprompted: *I don't see that on my map*
Request for Opinion	A request for an opinion or suggestion: *what do you think?*
Providing Opinion	Answering a request for opinion or providing an opinion unprompted: *I think we should have the pot*
Polar Answer	Answering a polar question: *yeah*, *no*, *mm-hm*
Command	An utterance giving instruction or direction including indirect instruction: *and then could you go left?*
Filler	Filler words or phrases used to fill pauses in the conversation: *hm*, *anyways*, *okay so*
Backchannel	An utterance that indicates the participant is listening and understanding: *okay*, *mm-hm*, *gotcha*, *sounds good*
Nicety	Utterances which primarily serve to express apology, gratitude, or to otherwise maintain a pleasant conversation: *sorry, I didn't mean to cut you off*, *no you're good*
Comment	An utterance that contains extraneous commentary on the task, such as narrating or explaining the participant's actions: *So this is like Easy Street haha*, *How the heck do I say this?*
Interjection	Short exclamations or interjections such as *ah*, *oops*, *yay*, *wow*
Fragment	Short abandoned or interrupted utterances that are too incomplete to classify

Table 2: Descriptions of the dialog acts used in the annotations.

set of 12 possible dialog acts chosen specifically for the two tasks, which are illustrated in Table 2. Two annotators first independently annotated 60% of all utterances ($\alpha = 0.83$). Disagreements were then resolved via discussion. Finally, the remaining utterances were annotated by one of the two annotators and later reviewed by the other annotator. More details are found in Appendix A.

4 Method

We remind the reader that the goal of this work is not to identify features that distinguish typical development from ASD, as in prior work on applying NLP to language in autism (see Section 2). Instead, we aim to exploit known effective approaches to develop robust models for predicting linguistic features tied to social and pragmatic aspects of communication known to be impacted in ASD in order to support targeted communication interventions. Crucially, the models we develop must perform similarly for individuals with and without autism.

We begin with three baseline classification models: majority class, where every sample is assigned the most frequent label; stratified, where labels are assigned randomly according to their distribution in the training data; and random, where labels are assigned uniformly randomly from the set of possible labels. Previous work (Yang et al., 2021) on a subset of this dataset using a different cross-validation strategy has shown that these baselines yield competitive performance to bag-of-words models and existing statistical models (Meyers et al., 2019) trained on separate corpora of written texts for these features.

We compare these baselines to neural models trained with BERT (Devlin et al., 2019) using MaChamp (van der Goot et al., 2021) and its default parameters for classification tasks.[1] We did not explore statistical models here since neural models were shown to be substantially better in Yang et al. (2021). To learn how models might perform differently for participant groups whose linguistic features are potentially atypical, we measure model performance separately for each speaker group: ASD EPs, TD EPs, CPs when interacting with ASD EPs, and CPs when interacting with TD EPs. Model performance was indexed with raw accuracy and weighted F1 scores.

[1] Here we trained separate models for each feature; training models on the combined features yielded very weak results.

Features		ASD		TD		CP (with ASD)		CP (with TD)	
	Model	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**
Politeness	*Majority*	0.75	0.83	0.75	0.83	0.80	0.86	0.77	0.84
	MaChamp	**0.85**	**0.86**	**0.89**	**0.89**	**0.89**	**0.90**	**0.90**	**0.90**
+ context	*MaChamp*	**0.85**	**0.86**	**0.89**	**0.89**	**0.89**	**0.90**	**0.90**	**0.90**
Uncertainty	*Majority*	0.48	0.63	0.48	0.63	0.49	0.63	0.57	0.70
	MaChamp	**0.74**	0.76	**0.75**	0.76	0.76	**0.77**	**0.78**	**0.79**
+ context	*MaChamp*	0.73	**0.77**	0.75	**0.77**	**0.77**	**0.77**	**0.78**	**0.79**
Information	*Majority*	0.37	0.53	0.32	0.48	0.37	0.53	0.42	0.58
	MaChamp	**0.80**	**0.81**	**0.81**	0.81	**0.80**	0.80	0.82	0.82
+ context	*MaChamp*	**0.80**	0.80	**0.81**	**0.82**	**0.80**	**0.82**	**0.83**	**0.83**
Dialog Act	*Stratified*	0.13	0.13	0.13	0.13	0.15	0.15	0.15	0.14
	MaChamp	0.73	**0.74**	0.77	0.77	0.77	**0.78**	0.78	0.79
+ context	*MaChamp*	**0.75**	0.73	**0.79**	**0.78**	**0.80**	**0.78**	**0.80**	**0.80**

Table 3: Comparison of F1 scores and accuracy across different groups of speakers given each feature, using **held-out transcript** cross-validation; + *context* indicates that training input included the preceding utterance; "CP" stands for conversational partner; we present only the result from the best baseline among the three.

Prior work has demonstrated that incorporating contextual information (in this case, previous utterances) is useful for predicting dialog act labels for both human-human (Liu et al., 2017) and human-chatbot (Khatri et al., 2018) interactions. To see whether a similar approach will be effective in our setting, we applied an evaluation scheme of **held-out transcript** [2], where we iteratively held out the data from one full transcript (i.e., the full conversation between one EP and one CP) as the test set and used the data from the remaining transcripts as the training set. This allowed us to incorporate contextual information by embedding the preceding utterance for feature prediction without enabling the models to learn individual speaker characteristics.

5 Results

Table 3 shows, unsurprisingly, that neural models, trained both with and without using the previous utterance as context, consistently outperform baselines for all four features, particularly the more challenging task of dialog act labeling.

To investigate whether adding contextual information was helpful in feature prediction, we compared for each feature the performance of the models with context to those without using logistic mixed-effects regression. The dependent variable was whether the feature value predicted by the model matched the manually assigned value; the fixed effect was the model (with or without the context of the previous utterance); and participant identity was included as a random intercept to control for repeated measurements of the same speaker. Though the results revealed no significant difference for the three ordinal features, including context appears to help improve model performance for dialog act ($\beta = 0.14$, $p < 0.001$). All further results presented will pertain to the models trained with prior contextual information.

We now turn to the question of whether there are differences in neural model performance on the utterances of ASD vs. TD experimental participants (EPs), as well as the utterances of conversational partners (CPs) of EPs with ASD vs. CPs of EPs with TD. Again, we applied logistic mixed-effects regression. The regression structure was similar to that described above, except that the group to which the speaker of the utterance belongs was used as the fixed effect (ASD EP vs. TD EP; or CP of ASD vs. CP of TD). Without using speaker identity as a random intercept, we found a significant effect of speaker group for EPs ($\beta = -0.31$, $z = -3.423$, $p < 0.001$) for politeness, which indicates that model predictions for politeness are more accurate for the TD group; we observed no such effect for CP groups. For dialog acts, there was also significant difference in accuracy between groups for EPs ($\beta = -0.38$, $z = -5.46$, $p < 0.001$), indicating that the models were more accurate for TD than for ASD utterances; a similar but weaker pattern was observed for CPs ($\beta = -0.22$, $z = -2.93$, $p < 0.01$). No significant differences were observed for uncertainty or informativeness. Running the same analysis using speaker identity as a random intercept, the significant differences for politeness and dialog acts are weaker but maintained for EPs (politeness: $\beta = -0.32$, $z = -2.00$,

[2] Held-out-speaker yielded comparable results.

Task	Utterance	Feature	Manual annotation	Model prediction
Map	Uh I don't mean to have you turn the map around it's just kind of how I think	Politeness	3	2
Island	Oh yeah that's right okay yeah that's that's smart	Politeness	3	2
Map	We're going to go past the duck	Dialog act	Command	Providing Information

Table 4: Examples of incorrectly classified ASD utterances.

$p < 0.05$; dialog acts: $\beta = -0.32$, $z = -2.15$, $p < 0.05$), suggesting that the observed differences in model accuracy may be driven by certain individuals, a finding that aligns with prior observations of heterogeneity in ASD language.

To qualitatively understand why our models might be more accurate for TD utterances, we inspected some of the incorrect predictions for ASD utterances. At times we observe similar unusual communication features in multiple ASD subjects, while other strategies appear to be idiosyncratic. Table 4 shows a few examples of misclassified ASD utterances. We observed many utterances like the first in this table, in which the ASD EP struggles to explain his reasoning, and many like the second, in which ASD EP evaluates the quality of his CP's prior statement. These strategies tend to be rare among TD EPs. In the third example, the EP uses "we" to politely give commands, a choice that, while easily recognized as a command-giving strategy by our annotators, was unique to that EP and was consistently misclassified.

6 Conclusions

Using a corpus of collaborative conversations between adults with and without ASD and their neurotypical conversational partners, we outline a framework for automatic identification of linguistic features associated with social communication. Although transformer-based models were able to achieve strong performance overall, when comparing results between diagnostic groups, we found that models fall short on the language of participants with ASD, especially in cases of politeness rating and dialog act labeling. This suggests that as powerful as transformer models are in capturing certain linguistic aspects of (written) data produced by (presumably) neurotypical speakers (see Linzen and Baroni (2021) for a review), they do not suffice in characterizing language in ASD, a finding that has broad implications for work applying these models to any potentially atypical language.

7 Ethical Considerations

All work described here was carried out with the approval of the Institutional Review Boards of all of the participating institutions. In accordance with our IRB protocol, we plan to release the full set of annotations of the corpus to interested researchers who can demonstrate completion of their institution's human subjects protection training curriculum.

References

Joel R. Adams, Alexandra C. Salem, Heather MacFarlane, Rosemary Ingham, Steven D. Bedrick, Eric Fombonne, Jill K. Dolata, Alison Presmanes Hill, and Jan van Santen. 2021. A pseudo-value approach to analyze the semantic similarity of the speech of children with and without autism spectrum disorder. *Frontiers in Psychology*, 12:3089.

Ahmad Aljanaideh, Eric Fosler-Lussier, and Marie-Catherine de Marneffe. 2020. Contextualized embeddings for enriching linguistic analyses on politeness. In *Proceedings of the 28th International Conference on Computational Linguistics*.

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

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

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Susanna Baldwin, Debra Costley, and Anthony Warren. 2014. Employment activities and experiences of adults with high-functioning autism and asperger's disorder. *Journal of autism and developmental disorders*, 44(10):2440–2449.

McKenzie Braley and Gabriel Murray. 2018. The group affect and performance (gap) corpus. In *Proceedings of the ICMI 2018 Workshop on Group Interaction Frontiers in Technology (GIFT)*.

Harry Bunt, Volha Petukhova, Andrei Malchanau, Alex Fang, and Kars Wijnhoven. 2019. The DialogBank: Dialogues with interoperable annotations. *Language Resources and Evaluation*, 53(2):213–249.

Deborah L Christensen, Jon Baio, Kim Van Naarden Braun, Deborah Bilder, Jane Charles, John N Constantino, Julie Daniels, Maureen S Durkin, Robert T Fitzgerald, Margaret Kurzius-Spencer, et al. 2016. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network. *MMWR Surveillance Summaries*, 65(3):1.

Olivia Conlon, Joanne Volden, Isabel M Smith, Eric Duku, Lonnie Zwaigenbaum, Charlotte Waddell, Peter Szatmari, Pat Mirenda, Tracy Vaillancourt, Teresa Bennett, et al. 2019. Gender differences in pragmatic communication in school-aged children with autism spectrum disorder (asd). *Journal of Autism and Developmental Disorders*, 49(5):1937–1948.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12.

Shirley Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331.

Karen Hurlbutt and Lynne Chalmers. 2004. Employment and adults with asperger syndrome. *Focus on autism and other developmental disabilities*, 19(4):215–222.

Chandra Khatri, Rahul Goel, Behnam Hedayatnia, Angeliki Metanillou, Anushree Venkatesh, Raefer Gabriel, and Arindam Mandal. 2018. Contextual topic modeling for dialog systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 892–899.

Friederike Klippel, Penny Ur, and John H Klippel. 1984. *Keep talking: Communicative fluency activities for language teaching*. Cambridge University Press.

Shibamouli Lahiri. 2015. Squinky! a corpus of sentence-level formality, informativeness, and implicature. *arXiv preprint arXiv:1506.02306*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

Yang Liu, Kun Han, Zhao Tan, and Yun Lei. 2017. Using context information for dialog act classification in DNN framework. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2170–2178, Copenhagen, Denmark. Association for Computational Linguistics.

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

Soile Loukusa, Eeva Leinonen, Katja Jussila, Marja-Leena Mattila, Nuala Ryder, Hanna Ebeling, and Irma Moilanen. 2007. Answering contextually demanding questions: Pragmatic errors produced by children with asperger syndrome or high-functioning autism. *Journal of Communication Disorders*, 40(5):357–381.

Lynn Mawhood, Patricia Howlin, and Michael Rutter. 2000. Autism and developmental receptive language disorder—a comparative follow-up in early adult life. i: Cognitive and language outcomes. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41(5):547–559.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Wilifried Post, Dennis Reidsma, and Pierre Wellner. 2005. The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, pages 137–140.

Benjamin S Meyers, Nuthan Munaiah, Andrew Meneely, and Emily Prud'hommeaux. 2019. Pragmatic characteristics of security conversations: an exploratory linguistic analysis. In *2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 79–82. IEEE.

Julia Parish-Morris, Mark Liberman, Neville Ryant, Christopher Cieri, Leila Bateman, Emily Ferguson, and Robert Schultz. 2016. Exploring autism spectrum disorders using HLT. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 74–84.

Lauren Parsons, Reinie Cordier, Natalie Munro, Annette Joosten, and Renee Speyer. 2017. A systematic review of pragmatic language interventions for children with autism spectrum disorder. *PloS one*, 12(4):e0172242.

Rhea Paul, Stephanie Miles Orlovski, Hillary Chuba Marcinko, and Fred Volkmar. 2009. Conversational behaviors in youth with high-functioning asd and asperger syndrome. *Journal of autism and developmental disorders*, 39(1):115–125.

V Rose, David Trembath, Deb Keen, and Jessica Paynter. 2016. The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research*, 60(5):464–477.

Alexandra C Salem, Heather MacFarlane, Joel R Adams, Grace O Lawley, Jill K Dolata, Steven Bedrick, and Eric Fombonne. 2021. Evaluating atypical language in Autism using automated language measures. *Scientific Reports*, 11(1):10968.

Paul T Shattuck, Sarah Carter Narendorf, Benjamin Cooper, Paul R Sterzing, Mary Wagner, and Julie Lounds Taylor. 2012. Postsecondary education and employment among youth with an autism spectrum disorder. *Pediatrics*, 129(6):1042–1049.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197.

Veronika Vincze. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1844–1853, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021. A context-aware hierarchical bert fusion network for multi-turn dialog act detection. *arXiv preprint arXiv:2109.01267*.

Christine Yang, Dora Liu, Qingyun Yang, Zoey Liu, and Emily Prud'hommeaux. 2021. Predicting pragmatic discourse features in the language of adults with autism spectrum disorder. In *Proceedings of the Association for Computational Linguistics Student Research Workshop (ACL-IJCNLP SRW)*, pages 284–291.

Edna Carter Young, Joshua J Diehl, Danielle Morris, Susan L Hyman, and Loisa Bennetto. 2005. The use of two language tests to identify pragmatic language problems in children with autism spectrum disorders. *Language, Speech, and Hearing Services in Schools*, 36:62–72.

Jinan Zeidan, Eric Fombonne, Julie Scorah, Alaa Ibrahim, Maureen S Durkin, Shekhar Saxena, Afiqah Yusuf, Andy Shih, and Mayada Elsabbagh. 2022. Global prevalence of autism: A systematic review update. *Autism Research*.

Piotr Żelasko, Raghavendra Pappagari, and Najim Dehak. 2021. What helps transformers recognize conversational structure? importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, 9:1163–1179.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *Proceedings of ICWSM*.

A Appendix: Annotation guidelines

The **politeness** feature measures how well an utterance contributes to a polite and non-demanding dialogue, marked by agreeableness, positivity, and willingness to compromise. An utterance with a low politeness rating of 1 is given to utterances expressing negative comments or frustration (*you're wrong*, *ugh I don't know*) and utterances which use a more blunt way of phrasing commands (*go back*). A high politeness rating of 3 is given to utterances with niceties (e.g., *thanks*, *sorry*) or affirmative words (*wonderful*, *awesome*) and indirect phrasing of commands (*if you could make a turn*).

The **uncertainty** feature measures the amount of uncertainty expressed about the correctness or legitimacy of the utterance. An utterance with a low uncertainty rating of 1 shows no uncertainty at all, or contains only a few filler words. A rating of 2 indicates some hesitation. It is given to polar questions, either-or questions, short abandoned utterances, and utterances containing many filler words (*um*, *uh*) or hedge phrases (*I guess*). An utterance with high uncertainty (rating of 3) has open questions (*what do you see?*) or expresses explicit uncertainty or confusion (*I have no idea*).

The **informativeness** feature is defined as a measure for the overall information content and specificity of an utterance. Utterances provide no information, contain only polar answers (*yes*, *no*) or vague words with low specificity (*that thing*, *over there*) are given a low informativeness rating of 1. For the map task, a rating of 2 is given to utterances that contain words for general objects and do not specify a specific location on the map (*another path*), while a high informativeness rating of 3 is given to utterances which contain proper nouns or labels or descriptions that point specific location on the map (*near the red pandas*). In the island task, a rating of 2 is given to utterances which contain only a short phrase indicating the item (*I want the fishing pole*), and a rating of 3 is given to utterances which contain multiple item words or a longer explanation of the items (*the fishing pole is good for catching fish*).

Rather than using one of the many existing (and conflicting) sets of **dialog acts**, we devised a small set specific to this dataset and commonly observed characteristics of language in ASD. When assigning dialog act labels, the annotators were instructed to consider the surrounding utterances, in order to fully capture the function of the utterance in the larger conversation. A complete list of each dialog act with a description and examples for each one can be found in Table 2.