Published in final edited form as:

Proc Conf Assoc Comput Linguist Meet. 2021 August ; 2021: 284–291. doi:10.18653/v1/2021.acl-srw.29.

Predicting pragmatic discourse features in the language of adults with autism spectrum disorder

Christine Yang,

Duanchen Liu,

Qingyun Yang,

Zoey Liu,

Emily Prud'hommeaux

Department of Computer Science, Boston College

Abstract

Individuals with autism spectrum disorder (ASD) experience difficulties in social aspects of communication, but the linguistic characteristics associated with deficits in discourse and pragmatic expression are often difficult to precisely identify and quantify. We are currently collecting a corpus of transcribed natural conversations produced in an experimental setting in which participants with and without ASD complete a number of collaborative tasks with their neurotypical peers. Using this dyadic conversational data, we investigate three pragmatic features – politeness, uncertainty, and informativeness – and present a dataset of utterances annotated for each of these features on a three-point scale. We then introduce ongoing work in developing and training neural models to automatically predict these features, with the goal of identifying the same between-groups differences that are observed using manual annotations. We find the best performing model for all three features is a feedforward neural network trained with BERT embeddings. Our models yield higher accuracy than ones used in previous approaches for deriving these features, with F1 exceeding 0.82 for all three pragmatic features.

1 Introduction

Autism spectrum disorder (ASD) is a neurological disorder associated with impairments in communication that can have a life-long impact on relationships, professional success, and personal independence (Ketelaars et al., 2010; Whitehouse et al., 2009; Hendricks, 2010). Although some percentage of individuals with ASD are not verbal from a young age, most go on to acquire spoken language but experience challenges in social aspects of communication related to discourse and pragmatic expression (Eales, 1993; Young et al., 2005). This atypicality in language has been recognized since the disorder was first named nearly eighty years ago (Kanner, 1943), and unusual language usage is one of the criteria used in the primary diagnostic instruments for ASD (Lord et al., 2002; Rutter et al., 2003).

yangael@bc.edu .

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 284–291 August 5–6, 2021.

One challenge for clinicians, however, is that there are no existing assessment tools for quantifying atypicality in discourse or pragmatics that can highlight communication deficits associated specifically with ASD while ruling out those associated with unrelated language disorders.

Most previous work on identifying pragmatic features that index atypicality in expressive language relies on careful manual annotations of transcripts of spontaneous spoken language (Volden and Lord, 1991; Bishop et al., 2000; Adams, 2002; Gorman et al., 2016; Canfield et al., 2016). Deploying complex annotation schemes like these, however, is time consuming and requires training and expertise, rendering this sort of detailed linguistic analysis impractical in the clinical intervention settings in which it would be most useful. Work on computational approaches for automatically identifying these features in the expressive language of individuals with ASD has focused exclusively on the language of children. In addition, this prior research has generally been applied to expressive language produced in a semi-structured context with an examiner or parent rather than spontaneous conversational speech with a peer (Prud'hommeaux et al., 2014; Losh and Gordon, 2014; Parish-Morris et al., 2016; Goodkind et al., 2018).

Our work addresses these aforementioned shortcomings in the previous work on pragmatic expression in ASD. In this paper, we describe an annotated corpus of conversations between adults with and without ASD and their neurotypical interlocutors as they engage in several collaborative tasks. Using this corpus, we investigate the degree of politeness, uncertainty, and informativeness in these conversations with the goal of identifying distinctive pragmatic features of ASD. We focus on these three features in particular because they are specific, remediable, and relevant in the collaborative discourse domain.

When data collection is complete, we will release the transcribed and annotated dataset to researchers who have completed their institution's human subjects training. The dataset will be unique in that it is produced by adults, a subgroup of the ASD population that is both understudied and underserved. In addition, the dataset will consist entirely of spontaneous conversations with a peer, a rarity in ASD datasets. To our knowledge there is no single corpus manually annotated with all three features of politeness, uncertainty, and informativeness. Moreover, our corpus is already larger than any existing *spoken* language (as opposed to *textual*) corpus available for these features.

With our annotated corpus, we propose several neural models for classifying utterances according to these features, and we explore whether our automated methods of generating these pragmatic features can be used to distinguish adults with ASD from their neurotypical peers as effectively as features derived via manual annotation. Our models outperform prior approaches to all three classification tasks, often by very wide margins. Although our predicted annotations do not capture all of the between-group differences observed using the manual annotations, we see promise in our approach.

2 Data Collection

2.1 Participants and tasks

We have collected spoken language data in a collaborative dyadic setting from adults 18 to 30 years of age with high-functioning ASD (n = 14) and with typical development (TD, n = 8). The ASD participants met the criteria for a diagnosis of ASD on the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2002). All participants met the following eligibility criteria: (1) performance IQ (PIQ) ≥ 80; (2) verbal IQ (VIQ) ≥ 80; (3) monolingual speaker of American English; and (4) no history of language impairment, auditory processing disorder, or hearing difficulty. This data collection is ongoing and is being conducted with the approval of the Institutional Review Boards of the two participating universities.

Each ASD or TD participant is paired with a neurotypical conversational partner (CP, n = 11), and together they engage in collaborative tasks involving verbal communication and deliberation. The two tasks we focus on in this paper include a map task and a deserted island task. In the map task, styled after Anderson et al. (1991), each participant is given a map of the same area, but with slight differences in the place names and locations of obstacles. Each map is marked with an X to show where that participant is located on the map. The experimental participant must give verbal directions to the conversational partner to lead them to their position on the map. In the deserted island task, a widely used method of eliciting natural conversation in second language instruction, the two participants are given a selection of labeled pictures of various items. They must agree on which of these items they would like to have with them on a deserted island. They are also given some specific categories of items to decide upon, such as items meant for entertainment or items that would be used to escape.

The conversations are recorded and then manually transcribed using Praat (Boersma and Weenink, 2001). Thus far, we have collected and transcribed conversations from 22 pairs of participants, with 14 experimental participants in the ASD group, 8 experimental participants in the TD group, and 11 neurotypical conversational partners, resulting in a corpus of 9,267 total utterances produced by experimental participants, with 5,742 utterances produced by experimental participants in the ASD group and 3,525 utterances produced by experimental participants the TD group. In the transcriptions, an utterance is defined as a C-unit, “an independent clause with its modifiers” which cannot be further split up without losing the primary meaning of the utterance (Loban, 1976). Each utterance is marked with a punctuation to denote the utterance type as an exclamation, question, abandoned utterance, interrupted utterance, or regular utterance. Additionally, we transcribe discourse markers, filler words, unfilled pauses, partial or interrupted words, sound effects or onomatopoeia, and verbal expressions of affirmation, negation, or exclamation.

2.2 Pragmatic feature annotation

After transcription, the transcripts are then annotated for politeness, uncertainty, and informativeness (Meyers et al., 2019), with each utterance receiving two annotations from a set of three trained human annotators. Each feature is given a rating on a scale from 1

to 3, with 1 representing the smallest degree of politeness, uncertainty, or informativeness, and 3 representing the highest degree of that feature. To measure the degree of agreement between the annotators, we calculate Krippendorf's alpha (Artstein and Poesio, 2008) for each feature, the results of which can be seen in Table 1. The final annotation of each feature for every utterance is then taken to be the average of the two annotators. We note that, although certain words are often helpful for determining the score of an utterance for a given feature, we do not rely on a list of specific lexical items or keywords. Example utterances and their corresponding scores are shown in Table 2.

These three features were chosen for a number of reasons. First, they are specific and interpretable, and as such, they are ideal features for targeted remediation. Secondly, they are especially relevant for and important in collaborative conversation; interviews, narratives, or monologues might be better analyzed using other features. Third, there are existing corpora labelled for these features and available toolkits for extracting these features, which allows us to compare our work against prior baselines and will enable us to leverage external corpora in our future work. Finally, we note that politeness, in particular, has been cited as an area of deficit in ASD (Frith, 1994; Sirota, 2004).

Politeness—The *politeness* feature is a measure of how well an utterance contributes to a polite and collaborative dialogue, marked by agreeableness, positive attitudes, and willingness to compromise. A low politeness rating of 1 is given to utterances expressing frustration or criticism ("no you're wrong", "ugh how do I do this?") and utterances which use a more blunt way of phrasing commands ("go left"). A high politeness rating of 3 is given to utterances containing niceties (e.g., "thanks", "sorry") or highly positive words ("perfect", "awesome") and utterances that use a polite or indirect way of phrasing commands ("if you could make a left", "you want to make a left").

Uncertainty—The *uncertainty* feature is defined to be a measure of the amount of uncertainty expressed about the correctness, validity, or permissibility of the utterance. A low uncertainty rating of 1 is given to utterances which express no uncertainty at all, or contain only a few filler words. A medium uncertainty rating of 2 is given to polar questions, either-or questions, short abandoned utterances, and utterances containing many filler words ("um", "uh") or hedge phrases ("I guess", "I'm assuming"). A high uncertainty rating of 3 is given to open questions ("where are you?") and utterances expressing explicit uncertainty or confusion ("I have no idea").

Informativeness—The *informativeness* feature is defined as a measure for the overall information content and specificity of an utterance. A low informativeness rating of 1 is given to utterances which contain only polar answers ("yes", "no") or vague words with low specificity ("thing", "over there"). In the map task, a medium informativeness rating of 2 is given to utterances which contain words for general objects and do not specify a specific location on the map, and a high informativeness rating of 3 is given to utterances which contain proper nouns or labels or descriptions that can only point to one specific location on the map. In the island task, a rating of 2 is given to utterances which contain only an item word or a short phrase explaining the item, and a rating of 3 is given to utterances which contain multiple item words or a longer explanation of the items.

3 Models

After the transcripts are annotated for the pragmatic features described above, we train a number of machine learning models on the annotated data, with the goal of eventually being able to bypass the manual annotations and automate the annotation process using these predictive models. The models are given the transcribed and tokenized utterance converted to all lowercase and are tasked with predicting the categorical label for politeness, uncertainty, and informativeness based on the manual transcriptions.

3.1 Baselines

We start with several different baseline models, shown in Table 4. The majority baseline always predicts the most frequent class; the stratified baseline makes random predictions proportional to the distribution of classes in the training set, and the random baseline predicts a random class every time.

We also evaluate against existing pre-trained models for rating politeness, uncertainty, and informativeness (Meyers et al., 2018). The results of this baseline can be seen in the "Existing Models" row in Table 4. The pre-trained *politeness* classifier is an SVM and is trained on the Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013), which includes 4,353 sentences of text conversations from public forums on Wikipedia and Stack Exchange. The pre-trained *uncertainty* classifier is a logistic regression model trained on the Szeged Uncertainty Corpus (Vincze, 2014), which includes more than 9,000 annotated sentences from corpora from different genres. The pre-trained *informativeness* classifier is a logistic regression model trained on the `SQUINKY!` corpus (Lahiri, 2015), which includes 7,000 utterances annotated for informativeness, implicature, and formality.

Additionally, because the scales used in the pre-trained classifiers for politeness and informativeness are continuous and differ from our own categorical annotation scale, we use thresholding to convert the predictions to our scale. For example, to convert a continuous scale from 0 to 1 into a categorical scale from 1 to 3, we map any scores less than 0.33 to be 1, scores between 0.33 and 0.67 to be 2, and scores greater than 0.67 to be 3. Since the pre-trained uncertainty classifier only predicts a binary result of either 0 or 1 corresponding to certain or uncertain, we map their 0 rating to our 1 rating and their 1 rating to our 3 rating.

3.2 Neural model architecture

We apply several methods for extracting sentence embeddings from the utterances in our dataset. First we use a basic *sequences* embedding in which each unique word appearing in the training data is assigned a unique identification number, and each utterance is then converted to a vector composed of the identification numbers for the words in the utterance, with padding for dimension consistency. With the sequence embeddings, we use a bidirectional LSTM model trained for 20 epochs with a batch size of 128.

Additionally, we also use word embeddings from pre-trained word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models, representing each utterance summing all of the vectors for the component words. Each utterance is represented with these pretrained embeddings in the embedding layers of our models, which are implemented in Keras[1]. For

the word2vec model, we use the Google News model which includes about 100 billion word vectors with a dimension of 300[2]. For the GloVe model, we use the pre-trained Stanford GloVe model trained on data from Wikipedia and Gigaword which includes around 6 billion word vectors with a dimension of 100 (Pennington et al., 2014). With the word2vec and GloVe embeddings, we use a convolutional neural network (CNN) model with global max pooling, trained for 20 epochs with a batch size of 128.

The last type of embeddings that we employ are the contextualized word representations of BERT (Devlin et al., 2019). Rather than integrating classification within the BERT architecture, we extract the 768-dimensional embeddings from the BERT-base model, and use them within a feedforward neural network with two hidden layers (Schuster et al., 2020) to predict the three points on each of the three annotation scales. The complete information for the parameterizations of our baseline and neural models is provided in Table 3.

3.3 Model evaluation

All our models are trained and evaluated with 5-fold cross validation. For each fold, the accuracy, precision, recall and F1 of the predictions are calculated. Then the averages of these metrics across the 5 folds are computed as the indexes to evaluate model performance.

4 Results

4.1 Manual annotations

Given the manual annotations, we examine whether there are significant differences between the ASD and the TD participant groups in terms of the three pragmatic features, using t-tests for significance testing. As shown in Table 5, the manual annotations reveal significant differences between the ASD and TD participants for politeness and informativeness in the map task, and uncertainty and informativeness in the island task. ASD participants are more polite, less uncertain, and less informative compared to TD participants in the map task. However, the results are reversed in the island task, where ASD participants are less polite, more uncertain, and more informative than TD participants.

The difference in politeness between the two tasks could be partially due to the nature of the two tasks, as the map task requires the experimental participant to give instructions and commands to their conversational partner and thus presents greater opportunity and need for phrasing their statements in a more polite way. In contrast, in the island task, the two participants have equal roles, and there may be less need for phrasing statements more politely. These results suggest ASD participants tend to be more polite than their TD peers in tasks in which they have a leading or authority role. Furthermore, the structure of the task could also contribute to the difference in uncertainty in the two tasks. In the map task, the participant giving instructions has a clear, factual set of information to convey to their partner, while the island task is more subjective and requires more discussion between the two participants to agree on a set of items. This would suggest that ASD participants exhibit

[1] https://keras.io/
[2] https://code.google.com/archive/p/word2vec/

more uncertainty than their TD peers in open-ended tasks which require more discussion and exchange of opinion.

4.2 Model predictions

The prediction results for all our models are presented in Table 4. Overall, the majority classifier performed the best among the baselines tested and had a fairly high accuracy already. This was especially true for politeness, where the majority baseline had an F1 measure of 0.77. This is likely due to the distribution of the politeness ratings, since most statements fell into the neutral category of 2 for politeness, being neither particularly polite or impolite. Despite the high performance of the majority baseline however, all four models trained on our own data generally performed substantially better than all the baseline classifiers, especially for uncertainty and informativeness. The BERT model seemed to perform the best overall across all three features, while the sequences model also performed well for politeness and informativeness. In terms of the F1 measure, the feedforward model trained with BERT embedding outperforms the majority baseline by 0.1 for politeness, 0.33 for uncertainty, and 0.42 for informativeness.

Since our goal is to investigate the differences in pragmatic expression between the two participant groups, we want our model to be able to capture the same group differences seen in the manual annotations. To this end, we take the output for each group predicted from the best-performing model, the feedforward model using BERT embedding, and perform a t-test between the two groups as well. The results of significance testing based on model predictions are then compared to those given manual annotations. As presented in Table 5, the BERT model fails to capture the group tendencies for uncertainty and informativeness in the map task and politeness and uncertainty in the island task, showing the opposite results as the manual annotations. However, it does seem to show the same group tendencies for politeness in the map task and informativeness in the island task, but it does not reveal statistically significant differences for any of the features.

5 Conclusions and Future Work

From the results of our study, we can see that there exist significant and quantifiable differences in pragmatic expressions between adults with ASD and their neurotypical peers. Moreover these differences are not fixed or consistent across all situations, but rather they may vary depending on the open-ended nature of the task, the roles involved, and the general context of the discourse. Relying on manual annotations of this sort, however, would not be practical or feasible in a clinical setting or for monitoring the efficacy of an intervention.

To determine whether these annotations can be carried out automatically, we introduced several potential models trained on the annotated data. Although all of our models outperformed one or more of the baselines, the BERT model generally is superior for all three features. None of the models, however, were able to capture the statistically significant differences we observe in the manual annotations. There is still more work to be done in fine-tuning the model to capture between-group differences which are vital to our study of the pragmatic expression of adults with ASD.

In our future work, we plan to extend the current study in at least three directions. First, we would like to employ different model architectures, leveraging external labeled corpora, with more systematic comparisons to see whether the differences between ASD and TD groups seen in manual annotations can be fully automatically derived. Second, after a long hiatus, we have recently resumed collecting data, with the goal of including 20 participants with ASD and 20 with typical development. Third, we aim to include annotations of other pragmatic features such as coherence and dialog acts in order to examine the differences of these features between ASD and neurotypical groups more comprehensively.

References

Adams Catherine. 2002. Practitioner review: The assessment of language pragmatics. Journal of child psychology and psychiatry, 43(8):973–987. [PubMed: 12455920]

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

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

Bishop Dorothy VM, Chan Janet, Adams Catherine, Hartley Joanne, and Weir Fiona. 2000. Conversational responsiveness in specific language impairment: Evidence of disproportionate pragmatic difficulties in a subset of children. Development and psychopathology, 12(2):177–199. [PubMed: 10847623]

Boersma Paul and Weenink David. 2001. Praat, a system for doing phonetics by computer. Glot international, 5:341–345.

Canfield Allison R, Eigsti Inge-Marie, de Marchena Ashley, and Fein Deborah. 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.

Danescu-Niculescu-Mizil Cristian, Sudhof Moritz, Jurafsky Dan, Leskovec Jure, and Potts Christopher. 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, Sofia, Bulgaria. Association for Computational Linguistics.

Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Eales Martin 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. [PubMed: 8106302]

Frith Uta. 1994. Autism and theory of mind in everyday life. Social development, 3(2):108–124.

Goodkind Adam, Lee Michelle, Martin Gary E, Losh Molly, and Bicknell Klinton. 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 Kyle, Olson Lindsay, Hill Alison Presmanes, Lunsford Rebecca, Heeman Peter A, and van Santen Jan PH. 2016. Uh and um in children with autism spectrum disorders or language impairment. Autism Research, 9(8):854–865. [PubMed: 26800246]

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

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

Ketelaars Mieke P, Cuperus Juliane, Jansonius Kino, and Verhoeven Ludo. 2010. Pragmatic language impairment and associated behavioural problems. International Journal of Language & Communication Disorders, 45(2):204–214. [PubMed: 22748032]

Lahiri Shibamouli. 2015. Squinky! a corpus of sentence-level formality, informativeness, and implicature.

Loban Walter. 1976. Language Development: Kindergarten through Grade Twelve. NCTE Committee on Research Report No. 18. ERIC.

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

Losh Molly and Gordon Peter 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. [PubMed: 24915929]

Meyers Benjamin S., Munaiah Nuthan, Meneely Andrew, and Prud'hommeaux Emily. 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.

Meyers Benjamin S, Munaiah Nuthan, Prud'hommeaux Emily, Meneely Andrew, Wolff Josephine, Alm Cecilia Ovesdotter, and Murukannaiah Pradeep. 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.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Gregory S., and Dean Jeffrey. 2013. Distributed representations of words and phrases and their compositionality. In Conference on Neural Information Processing Systems, pages 3111–3119.

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

Pennington Jeffrey, Socher Richard, and Manning Christopher D.. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Prud'hommeaux Emily, Morley Eric, Rouhizadeh Masoud, Silverman Laura, van Santeny Jan, Roarkz Brian, Sproatz Richard, Kauper Sarah, and DeLaHunta Rachel. 2014. Computational analysis of trajectories of linguistic development in autism. In 2014 IEEE Spoken Language Technology Workshop (SLT), pages 266–271. IEEE.

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

Schuster Sebastian, Chen Yuxing, and Degen Judith. 2020. Harnessing the linguistic signal to predict scalar inferences. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5387–5403, Online. Association for Computational Linguistics.

Sirota Karen Gainer. 2004. Positive politeness as discourse process: Politeness practices of high-functioning children with autism and asperger syndrome. Discourse Studies, 6(2):229–251.

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

Volden Joanne and Lord Catherine. 1991. Neologisms and idiosyncratic language in autistic speakers. Journal of Autism and Developmental Disorders, 21:109–130. [PubMed: 1864825]

Whitehouse Andrew JO, Watt Helen J, Line EA, and Bishop Dorothy VM. 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. [PubMed: 19340628]

Young EC, Diehl JJ, Morris D, Hyman SL, 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. [PubMed: 15801508]

Table 1:

Percent agreement and interrater reliability (Krippendorf's α) for pragmatic feature annotation.

Feature	Agreement	α
Politeness	91.58%	0.57
Uncertainty	85.62%	0.75
Informativeness	91.62%	0.90

Table 2:

Samples manual annotations for each task.

Task	Utterance	Politeness	Uncertainty	Informativeness
Map	How the heck am I supposed to say this?	1	1	1
Map	It's near the Irrigation Pond.	2	1	3
Map	Okay so we're going to have to go down one block.	3	1	2
Map	Can you describe where you're at?	2	3	1
Map	Yeah it is by some trees.	2	1	2
Map	Yeah it is by some trees.	2	1	2
Island	I don't care.	1	1	1
Island	I would say the matches first, because you can set off a signal, like a signal fire.	2	1	3
Island	Or you could keep the matches?	2	2	2
Island	Fishing pole, definitely, um.	2	1	2
Island	We'll put them off to the side.	3	1	1
Island	Do we want to go with these four?	3	2	1
Island	You want to do the dog?	3	2	2
Island	If we wanna trying get off the island we probably want some rope or something.	3	2	3
Island	We could use logs and stuff to tie up and make some kind of raft trying to get back to civilization.	3	1	3

Table 3:

Summary of model parameters.

Parameters	Sequence (LSTM)	GloVe (CNN)	word2Vec (CNN)	BERT (Feedforward NN)
CV Folds	5	5	5	5
Epochs	20	20	20	20
Batch size	128	128	128	8
Embedding dimension	300	100	300	768
Layers	1 bidirectional hidden layer, 1 dense layer	3 convoluted layers, 2 max pooling layers, 1 global max pooling layer, 1 dense layer	3 convoluted layers, 2 max pooling layers, 1 global max pooling layer, 1 dense layer	2 hidden linear layers
Dropout	0.5	0.5	0.5	0.5
Loss function	categorical cross entropy	categorical cross entropy	categorical cross entropy	categorical cross entropy
Optimizer	RMSprop	RMSprop	RMSprop	Adam

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2022 November 03.

Table 4:

Comparison of accuracy, precision, and recall for the baselines and models tested. The best baseline in each column and the best proposed model in each column are rendered in boldface.

	Politeness				Uncertainty				Informativeness			
Baselines	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**
Majority	**.84**	**.71**	**.84**	**.77**	**.62**	.38	**.62**	**.47**	**.56**	.31	**.56**	**.40**
Stratified	.71	**.71**	.71	.71	.46	**.46**	.46	.46	.38	.38	.38	.38
Random	.20	**.71**	.20	.31	.20	.45	.20	.28	.20	.40	.20	.27
Existing Models	.73	.70	.73	.72	.55	.34	.55	.42	.55	**.49**	.55	.52

Model	**Embeddings**												
LSTM	Sequences	**.87**	.86	**.87**	.86	.72	.70	.72	.71	**.82**	.81	.82	.81
CNN	GloVe	.86	.82	.86	.84	.67	.64	.67	.65	.74	.72	.74	.73
CNN	word2vec	.84	.80	.84	.82	.69	.63	.69	.66	.76	.74	.76	.75
Feedforward NN	BERT	.85	**.88**	.85	**.87**	**.84**	**.82**	**.84**	**.83**	**.82**	**.82**	**.83**	**.82**

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2022 November 03.

Table 5:

Speaker averages for pragmatic features, comparing the manually annotated values and values predicted by the BERT model which has the highest F1 measures.

	Manual Annotations		BERT Model Predictions	
Map Task	**ASD**	**TD**	**ASD**	**TD**
Politeness	**2.0005** **	1.9645	**2.0626**	2.0444
Uncertainty	1.4124	**1.4334**	**1.399**	1.3805
Informativeness	1.6044	**1.7145** ****	**2.0631**	2.0444
Island Task	**ASD**	**TD**	**ASD**	**TD**
Politeness	2.0332	**2.0743**	**2.1798**	2.1597
Uncertainty	**1.3894** ****	1.223	1.367	**1.4021**
Informativeness	**1.7395** ***	1.5169	**2.1798**	2.1597

Asterisks indicates a significant difference between the two groups (** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$).