Automatic detection of turn-taking events in continuous EEG data from spontaneous dialogue

Pablo Brusco [1,2], Juan Kamienkowski [1,2,3], Agustín Gravano [1,2]

[1] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[3] Departamento de Física, FCEyN, Universidad de Buenos Aires, Argentina
pbrusco@dc.uba.ar, juank@dc.uba.ar, gravano@dc.uba.ar

Abstract

This paper describes a series of machine learning experiments for automatically classifying the type of turn-taking transition based on features extracted from the EEG signal. These were conducted on a corpus of unrestricted dialogues between pairs of subjects, with simultaneous recordings of speech and EEG from both subjects. Our preliminary results indicate that the listener's EEG signal contains useful information for predicting whether the current speaker will either yield the conversational turn or continue talking. In other words, we show it is possible to detect the listener's perception of turn-yielding cues present in the speaker's speech. Similarly, from the current speaker's EEG signal we can extract useful information for detecting whether the speaker will yield the turn or continue talking. These results may lead to new tools valuable for the development of brain-computer interfaces.

Index Terms: EEG, speech, turn-taking, machine-learning.

1. Introduction

When engaged in dialogue, humans typically alternate conversational turns in a swift, well-coordinated manner, with relatively few silences and overlaps [1]. This is possible in part due to a set of acoustic, prosodic, lexical and syntactic cues produced by speakers at the end of speech segments, signaling upcoming turn transitions. In the literature, these are called TURN-TAKING CUES and include changes in pitch, intensity, speech rate and voice quality, among others, occurring in the final 500 miliseconds of speech segments [2, 3]. Further, these cues can be perceived by listeners, who can thus anticipate turn endings and prepare to start speaking [4, 5].

After extensive research in Linguistics, Speech Processing and related fields, turn-taking has recently become a topic of interest in the Neuroscience community [6, 7]. Here, it is important to distinguish between two mental processes involved in conversation, which are believed to occur at different levels. On the one hand, there exists considerable documentation of the neural networks involved in processing linguistic content [6, 8, 9, 10]. On the other hand, very little is known about the more automatic (i.e., less conscious) system for dynamically monitoring turn-taking activity [6, 7], although there exists recent evidence suggesting that the two systems are indeed separate [11].

The main goal of our research project is to find evidence of neural correlates of turn-taking phenomena. Given the aforementioned evidence of acoustic-prosodic turn-taking cues that occur in the final portion of speech segments, we hypothesize that we should be able to observe specific patterns in the interlocutors' neural activity, related to their perception of such cues and to the preparation for their upcoming response.

For this purpose, we collected a corpus of dialogues between pairs of subjects, with simultaneous recordings of speech and electroencephalography (EEG). In the present paper, we describe the first analyses of these data, consisting of a series of machine learning (ML) experiments for automatically classifying the type of turn-taking transition based on features extracted from the EEG signal. Two notable differences between this and previous studies are that 1) our EEG data were concurrently collected from **two** subjects talking to each other, and 2) the conversations between subjects were **unrestricted**. This differences enables us to explore three vital aspects of communication: the social, pragmatic, and dynamic features.

Visualization of raw EEG data from different conditions offers no clear cues to distinguish between classes. Thus, we think ML is a particularly convenient tool to use in this case, complementing more traditional statistical tests. In addition, several Brain-Computer Interfaces (BCI) applications rely on the use of ML algorithms applied to EEG signals nowadays. For example, Lotte et al. give an extensive review on this topic, and compare different approaches to the problem of estimating the best class for unlabeled data from the EEG signal [12]. These comparisons are made in terms of performance, and the authors provide guidelines for different types of problems.

2. Materials

For the present study, we collected a corpus of ten dyadic conversations with simultaneous recordings of speech and EEG activity from each pair of participants.

2.1. Tasks and sessions

The experimental task consisted in two subjects playing a series of OBJECTS GAMES (first described in [2]). Each subject used a separate laptop computer and could not see the screen of the other subject. Subjects sat facing each other in a booth, with an opaque curtain hanging between them, so that all communication was verbal.

In an Objects Game, each subject's laptop displayed a game board with 5–7 objects. Both subjects saw the same set of objects at the same position on the screen, except for one (the target). One subject (the describer) was instructed to describe the position of an object on her screen to the other (the follower), whose task was to position the same object on her own screen. Subjects could discuss freely about the location of the target

object, and were later awarded 1–100 points based on how well the Follower's target location matched the Describer's. Each session consisted of a minimum of 15 and a maximum of 30 instances of the Objects Game, with subjects alternating in the Describer and Follower roles. At the end of the session, subjects were paid a fixed amount of money for their participation, plus a bonus based on the number of awarded points.

The sessions were recorded at the University of Buenos Aires in April 2014. A total of 20 subjects (10F, 10M) participated in the study. Their ages ranged from 19 to 43 years ($M = 26.4$, $SD = 6.3$), and 18 were right-handed. All subjects were native speakers of Argentine Spanish, lived in the Buenos Aires area at the time of the study, and agreed to join the study by signing a consent form.

2.2. EEG registration and preprocessing

For each participant, EEG activity was recorded on a dedicated PC at 1024Hz, at 128 electrode positions on a standard 10–20 montage, using the Biosemi Active-Two system.[1] Also, four reference electrodes were placed at both mastoids and ear lobes, and four electrodes in the left and right external ocular canthi and under the eye and above the eyebrow. The data for each participant ($128 + 8 = 136$ channels) was recorded with separated references and ground, and also separately amplified. The amplifiers were connected by an optic fiber and the data from both participants was put together with a copy of audio signals before it was digitalized ($136 \times 2 + 2$ channels).

The EEG data were digitally downsampled to 512Hz, and imported into MATLAB using the EEGLAB toolbox [13] with linked mastoids as reference. After an initial visual inspection, we excluded one participant due to a very high electrical noise in most of the electrodes.

The EEG was band-pass filtered between 0.1 and 100 Hz, and a notch filter was applied between 49 and 51 Hz to remove the line signal. The intervals between trials were removed, and we applied Independet Component Analysis (ICA) to the remaining data in order to remove mainly ocular artifacts – but also some noisy channels and muscular artifacts. Artifactual ICs were selected using EyeCatch [14] and ADJUST plugins [15], and supervised by an expert.

2.3. Audio registration and preprocessing

The audio for each subject was recorded on a separate channel of a TASCAM DR-100 digital recorder, at a sampling rate of 44.1 kHz with 16-bit precision, using a Rode HS-1 head-mounted close-talking microphone. Each session was later downsampled to 16 kHz, and saved as two separate mono wav files, one for each subject.

Low-resolution copies of the audio signals were also included in the EEG recordings, for later synchronization with the wav files. This synchronization was performed by finding the time offset that maximizes the cross-correlation between the two audio copies.

2.4. Definition of epochs

First, we define an INTER-PAUSAL UNIT (IPU) as a maximal speech segment from a single speaker that is surrounded by pauses longer than 100 ms. IPUs in our corpus were manually aligned to the audio signal by trained annotators.

[1] Biosemi, Amsterdam, Holland, http://www.biosemi.com

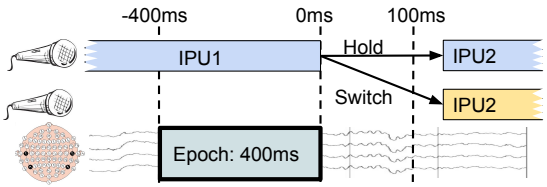

Figure 1: *Epoch selection.*

Next, each transition from one IPU to the next was automatically labeled as illustrated in Figure 1: a transition was labeled HOLD (H) when the current speaker continued talking after a short pause, and SWITCH (S) when the other speaker started talking after a short pause. Transitions containing any overlapping speech were discarded, since IPU2 would trample on any turn-taking cues produced in IPU1. Transitions with IPU2 shorter than 500ms were also discarded, to avoid BACKCHANNELS – short utterances such as *yeah* or *uh-huh* used to display attention and invite the speaker to continue [16]. Given their high frequency and their radical differences with Switches, we chose to leave backchannels for future study. Lastly, only IPUs longer than 400ms were considered, so as to have enough material for our analysis.

The EEG signal from each subject was band-pass filtered between 0.2 and 30 Hz, and later epoched to the final 400ms of IPU1 (see Figure 1). In this way, each IPU transition yielded two epochs: one with 400ms of EEG data from the utterer of IPU1 (we call this the SPEAKING condition), and one with 400ms of EEG data from the other speaker (the LISTENING condition). After completing this pipeline, we ended up with a mean of 271 Hold-type epochs per subject (SD=112), and 72 Switch-type epochs per subject (SD=29).

3. Methods

We conducted three different ML experiments: (T1) predict whether an epoch corresponds to a speaker or a listener; (T2) predict if an epoch in the listening condition is followed by a Hold or a Switch; and (T3) predict if an epoch in the speaking condition is followed by H or S.

T1 is supposed to be the least difficult task, since speaking and listening are known to produce distinct patterns of brain activity [17]. Therefore, we use this task as a sanity check, given the novelty and difficulty of the research topic, as well as for gaining insights into the whole process. T2 is the main goal of the present paper: finding neural evidence of the listener's **perception** of turn-taking cues produced by the other speaker. Finally, T3 consists in finding neural evidence of the current speaker's **production** of turn-taking cues – i.e., detecting whether the speaker will yield the turn or continue talking.

In these tasks, we define an INSTANCE as the 128-dimensional time series of EEG amplitudes for a given epoch. That is, an instance is formed by the amplitudes of the 128 EEG channels sampled over an epoch's 400ms. The CLASSES are Speaker vs. Listener in T1, and Hold vs. Switch in T2 and T3. For these tasks, we only considered data from 10 subjects in our corpus, for which we had completed the manual alignment of IPUs at the time we started the analysis.

From each instance we extracted 5248 features using the mean amplitude of the signal over sliding windows of different widths, as described in Algorithm 1. Once the features had been extracted, we conducted 10-fold cross-validation classification experiments using the Random Forests (RF) learning algorithm

Algorithm 1 Extract features from instance

FeatureVector $\leftarrow \{\}$
for each EEG channel ch **do**
  for each w **in** $\{50, 100, 200, 400\}$ ms **do**
    Slide a window W of width w, time step 25ms.
    At each step:
      1. Compute mean amplitude of ch over W.
      2. Append this value to *FeatureVector*.
return *FeatureVector*

[18] included in Python's *Sklearn* software package.

Regarding the choice of ML algorithm, in preliminary analyses we experimented with Support Vector Machines (SVM), Bagging techniques and Logistic Regression, in addition to RF. SVM had the best performance, closely followed by RF. However, we chose to continue with RF for two reasons – it is considerably faster to train, and provides a straightforward assessment of the relative predictive power of individual features. In the literature, RF has been shown to achieve results on several EEG classification tasks that are comparable to those of classical classifiers such as SVM, Neural Networks and Linear Discriminant Analysis, among others [19]. Also, RF appears to be robust to high-dimensional, sparse data [20], which is the case for our dataset.

Since this is work in progress, we split our dataset into a DEVELOPMENT set containing 80% of the epochs from every subject, and a CONTROL set with the remaining 20%. All the results we present in this work are based on the development set, leaving the control set untouched for future work.

For evaluating the performance of each model, we computed a ROC curve by combining the probability scores from the test instances in all 10 folds (this would be analogous to 'averaging' the ROC curves from the 10 folds). ROC curves are a useful measure of classifier performance, and are preferred over accuracy and similar measures that are known to be sensitive to unbalanced classes. We also compute the Area Under the Curve (AUC), which summarizes ROC curves in one number (AUC=1.0 is a perfect classification; 0.5 is chance).

Lastly, we assess the statistical significance of the results via non-parametric PERMUTATION tests. In a permutation test, we train and test a classifier on 100 labeled-permuted versions of the data. Each permutation consist in randomly shuffling all instance labels, keeping the classes distribution. In this way, the performance is expected to decrease, since the structural dependencies between features and classes are lost [21].

As an example, Figure 2 shows the resulting ROC curves for a particular subject for the task of predicting Hold vs. Switch in the listening condition. The model's ROC curve is shown in green, and those of the permutations, in blue. The significance (p-value) is subsequently calculated as the percentage of permutations' AUC that are greater than the classifier's AUC.

4. Results

Table 1 summarizes the results of our three ML experiments for each individual subject, using RF, 10-fold cross-validation and Permutation tests of size 100, as described in Section 3.

4.1. Model performance

The results for task T1 (our sanity check) show that our method manages to extract information from the EEG data, useful for separating listeners from speakers. For all but one speakers, AUC values are significantly better than chance ($p < 0.01$); for subject 23-1, the difference approaches significance ($p \approx 0.08$).

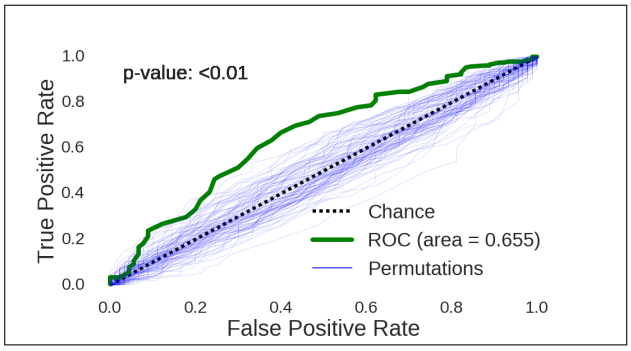

Figure 2: *ROC courve for subject 25-2, task T2. The curve of our RF model (green) lies above all permutation curves (blue).*

The second column of Table 1 corresponds to task T2 (S vs. H in the listening condition).[2] We note that, though most results are not significant, three particular subjects do show significant results (23-1, 24-2, 25-2). That is, for three subjects, our trained models significantly outperform chance in discriminating S and H transitions based on the **listener's** EEG signal. Finally, the third column of Table 1 shows the results for task T3 (S vs. H in the speaking condition). In this case, for four subjects (21-2, 24-2, 25-1, 25-2), our trained model outperforms chance in separating H from S based on the **speaker's** EEG signal.

It is worth noting that both in T2 and T3 our models performed better with subjects who had more instances of Switch transitions (the minority class) in our corpus. Symmetrically, the worst-performing models were trained on subjects who had fewer instances of S. This suggests that the availability of more training data should increase the performance of our models, thus encouraging further research.

4.2. Predictive power of individual features

Next we assess the relevance of individual features, aiming at determining which channels and time intervals are most informative in our three ML tasks. Figure 3 shows the IMPORTANCE of features assigned by RF over time (averaged for all 128 channels), comparing different window widths in which the signal is averaged for all subjects in task T2. Each point represents the time at which the window starts, together with its standard deviation among channels and subjects. We note that 50ms-wide windows have a higher relevance for all points.

Figure 4 shows which areas of the brain are more informative over time on windows of width 50ms, for the subject with the most significant results in task T2. Each topological map uses a cromatic scale, such that darker areas are the more informative ones. This figure has a few things worth noting. First, the dark areas seem to be very specific. This might be related with the way a classifier works; when two or more variables are highly correlated – which is the case for neighbor EEG channels –, tree-based classification algorithms such as RF select the most informative one, after which the remaining ones are given a lower priority, since they become less informative. Thus, activity based topological maps should be smoother. Second, the left posterior cortex attained greater significance in the

[2] The counts of S instances in the listening and speaking conditions differ due to a minor bug in the IPU selection criterion, which included a few of the shorter IPUs by mistake. This may have slightly decreased the performance of our models, and will be corrected in future work.

Subject	Task T1: Spk vs. Lst AUC	p	# instances	Task T2: S vs. H (Lst) AUC	p	# instances	Task T3: S vs. H (Spk) AUC	p	# instances
(21-1)	0.62	$<$ **0.01**	Spk:285 Lst:347	0.54	0.22	S:80 H:267	0.52	0.26	S:60 H:225
(21-2)	0.66	$<$ **0.01**	Spk:335 Lst:295	0.52	0.35	S:70 H:225	0.63	**0.02**	S:68 H:267
(22-1)	0.64	$<$ **0.01**	Spk:182 Lst:180	0.49	0.53	S:32 H:148	0.56	0.18	S:32 H:150
(22-2)	0.79	$<$ **0.01**	Spk:178 Lst:183	0.46	0.73	S:33 H:150	0.51	0.41	S:30 H:148
(23-1)	0.57	0.08	Spk:116 Lst:144	0.58	**0.05**	S:40 H:104	0.40	0.88	S:24 H:92
(23-2)	0.67	$<$ **0.01**	Spk:137 Lst:119	0.40	0.85	S:27 H:92	0.48	0.54	S:33 H:104
(24-1)	0.63	$<$ **0.01**	Spk:372 Lst:424	0.53	0.29	S:76 H:348	0.52	0.28	S:78 H:294
(24-2)	0.76	$<$ **0.01**	Spk:407 Lst:386	0.57	**0.04**	S:92 H:294	0.59	**0.02**	S:59 H:348
(25-1)	0.64	$<$ **0.01**	Spk:315 Lst:391	0.51	0.31	S:80 H:311	0.60	**0.01**	S:80 H:235
(25-2)	0.77	$<$ **0.01**	Spk:383 Lst:325	0.65	$<$ **0.01**	S:90 H:235	0.63	$<$ **0.01**	S:72 H:311

Table 1: *Summary of results for our three classification tasks for individual subjects. Spk: Speaking condition; Lst: Listening condition; S: Switch transition; H: Hold transition*

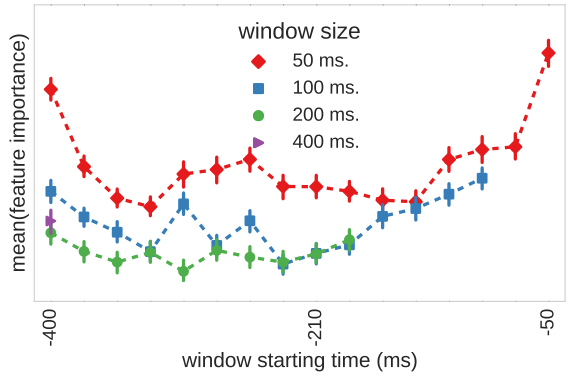

Figure 3: *Predictive power of features in task T2. Features importance averaged over all subjects and channels over time.*

final 150ms, which is consistent with previous reports (see for instance [22]). Further analysis should be made on consistency of the relevant brain regions on different subjects.

5. Discussion and conclusions

In this work we ran three machine learning experiments using Random Forests classifiers. The first task (T1) consisted in predicting whether a portion of EEG signal was produced by the speaker or the listener of an utterance. Our classifiers significantly outperformed chance in all but one subject. In this case several improvements in performance could be achieved, but since we conducted this task mainly as a sanity check, we simply conclude that the signal contains information useful for distinguishing between the two conditions.

The main goal of this study is our second task (T2) – predicting from the EEG signal whether listeners will start speaking or keep listening to the interlocutor after an utterance. In this case, we found promising results for three subjects, for which the classifiers effectively captures information about this turn-taking event and significantly outperform chance. Moreover, the performance of our classifiers seems to improve with the amount of available data, as expected when the method captures the dependencies between the instances and their classes. Similar results were obtained in our third task (T3) – predicting if the speaker will continue or stop talking, based on an utterance-final portion of EEG signal. It is important to note that these experiments focus on capturing the production/perception of turn-taking cues, and not on the more complicated mechanisms involved in deciding when to take the floor or continue talking. These results may become useful for anticipating turn completions in BCI applications.

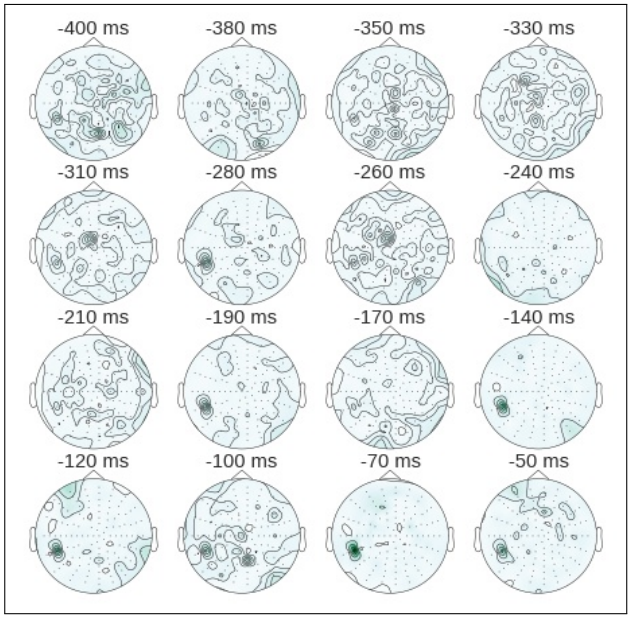

Figure 4: *Predictive power of features by position (T2, 25-2).*

There is much room for improving our classifiers. Our current methods do not exploit crucial characteristics of the data. For example, we do not model the temporal dependency of the EEG samples, since features are extracted as if they were independent of one another. This could be addressed with techniques such as Hidden Markov Models (HMMs), which are also well suited for modeling poorly time-aligned events.

There is currently a source of errors in our data arising from the fact that Switch and Hold transitions were estimated automatically, since no transcriptions were available at the time we started this study. In consequence, our Switch transitions could contain instances of interruptions and backchannels, and some of the IPUs we chose to discard could constitute valid Switch transitions. We will correct this by manually labeling all turn-transitions and repeating the analysis.

Finally, data from nine additional subjects are still available for further experimentation. We plan to use these data not only to extend the experiments we have conducted so far, but also to run new inter-subjects classification tasks.

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[2] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[3] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 9, no. 1, p. 1, 2012.

[4] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, pp. 23–25, 2011.

[5] M. Zellers, "Duration and pitch in perception of turn transition by swedish and english listeners," in *Proc. from FONETIK*, 2014, pp. 41–46.

[6] J. C. Hoeks and H. Brouwer, "Electrophysiological research on conversation and discourse," *Holtgraves, TM (Ed.)*, pp. 365–386, 2014.

[7] S. C. Levinson, "Turn-taking in human communication–origins and implications for language processing," *Trends in cognitive sciences*, vol. 20, no. 1, pp. 6–14, 2016.

[8] S. Bögels, L. Magyari, and S. C. Levinson, "Neural signatures of response planning occur midway through an incoming question in conversation," *Scientific reports*, vol. 5, 2015.

[9] A. D. Friederici, "Event-related brain potential studies in language," *Current neurology and neuroscience reports*, vol. 4, no. 6, pp. 466–470, 2004.

[10] M. Kutas and K. D. Federmeier, "Thirty years and counting: Finding meaning in the n400 component of the event related brain potential (erp)," *Annual review of psychology*, vol. 62, p. 621, 2011.

[11] D. Foti and F. Roberts, "The neural dynamics of speech perception: Dissociable networks for processing linguistic content and monitoring speaker turn-taking," *Brain and language*, vol. 157, pp. 63–71, 2016.

[12] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 4, no. 2, p. R1, 2007.

[13] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[14] N. Bigdely-Shamlo, K. Kreutz-Delgado, C. Kothe, and S. Makeig, "Eyecatch: data-mining over half a million eeg independent components to construct a fully-automated eye-component detector," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 5845–5848.

[15] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.

[16] A. Gravano, J. Hirschberg, and Š. Beňuš, "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, 2012.

[17] A. Hiraiwa, K. Shimohara, and Y. Tokunaga, "Eeg topography recognition by neural networks," *IEEE Engineering in Medicine and Biology Magazine*, vol. 9, no. 3, pp. 39–42, 1990.

[18] L. Breiman, "Manual on setting up, using, and understanding random forests v3. 1," *Statistics Department University of California Berkeley, CA, USA*, vol. 1, 2002.

[19] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. E. John, L.-O. Wahlund, Y. Dodge, and T. Dierks, "Application and comparison of classification algorithms for recognition of alzheimer's disease in electrical brain activity (eeg)," *Journal of neuroscience methods*, vol. 161, no. 2, pp. 342–350, 2007.

[20] N. Gunduz and E. Fokoue, "Robust classification of high dimension low sample size data," *arXiv preprint arXiv:1501.00592*, 2015.

[21] M. Ojala and G. C. Garriga, "Permutation tests for studying classifier performance," *Journal of Machine Learning Research*, vol. 11, no. Jun, pp. 1833–1863, 2010.

[22] R. S. Gisladottir, D. J. Chwilla, and S. C. Levinson, "Conversation electrified: Erp correlates of speech act recognition in underspecified utterances," *PloS one*, vol. 10, no. 3, p. e0120068, 2015.