# Language-Based Automatic Assessment of Cognitive and Communicative Functions Related to Parkinson's Disease

**Gabriel Murray**
Computer Information Systems
U. of the Fraser Valley
Abbotsford, BC, Canada
gabriel.murray@ufv.ca

**Lesley Jessiman**
Psychology
U. of the Fraser Valley
Abbotsford, BC, Canada
lesley.jessiman@ufv.ca

**McKenzie Braley**
Psychology
U. of the Fraser Valley
Abbotsford, BC, Canada
mckenzie.braley@student.ufv.ca

## Abstract

We explore the use of natural language processing and machine learning for detecting evidence of Parkinson's disease from transcribed speech of subjects who are describing everyday tasks. Experiments reveal the difficulty of treating this as a binary classification task, and a multi-class approach yields superior results. We also show that these models can be used to predict cognitive abilities across all subjects.

## 1 Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disease worldwide, affecting more than one percent of individuals above the age of 60 (deRijk et al., 2000; von Campenhausen et al., 2005). PD is associated with the gradual degeneration of dopaminergic neurons in the substantia nigra pars compacta in the basal ganglia (Bottcher, 1975; Samii et al., 2004). Dopamine depletion originating in the basal ganglia leads to an under-activation of the frontal lobes, where motor functions and executive processing are predominantly housed. Fronto-striate pathway disturbances lead to motor impairments such as resting tremors, muscular rigidity, bradykinesia and postural disturbances (Samii et al., 2004; von Campenhausen et al., 2005). Motor-related speech deficits are also observed. One of the most common speech problems is a marked decrease in the volume of the PD sufferer's voice, known as *aphonia* (Nutt et al., 1992). PD can also impair the individual's use of vocal parameters, preventing them from appropriately stressing and emphasizing particular words (Dubois, 1991). Short bursts of speech coupled with long pauses (Darley et al., 1975), accelerated speech (*tachiphemia*), compulsive repetition of words or phrases (*palilalia*) (Boller et al., 1975), and stuttering (Lebrun, 1996) are also observed in some individuals with PD. All of the aforementioned speech and language impairments stem from PD-related motor decline.

A gradual decline in dopaminergic neurons in the basal ganglia and a subsequent disturbance of the fronto-striate loop also leads to language impairments related to an executive processing dysfunction. The research shows that PD results in deficits in word-finding/verbal fluency (Gurd and Oliveira, 1996; Henry and Crawford, 2004; Matison et al., 1982; Randolph et al., 1993; Zec et al., 1999), syntactical processing (Arnott et al., 2005; Illes, 1989; Grossman et al., 1992; Grossman et al., 1996; Grossman et al., 2000; Hochstadt et al., 2006; Kemmerer, 1999; Kemmerer, 1999; Lieberman et al., 1992; Natsopoulos et al., 1991; Ullman et al., 1997), and speech error monitoring (McNamara et al., 1992). There is also evidence that PD individuals score lower on measures of pragmatic communication abilities such as conversational appropriateness, speech acts, stylistics, gestures and prosodics (McNamara and Durso, 2003).

*Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 63–74
Santa Fe, New Mexico, United States, August 20, 2018.
https://doi.org/10.18653/v1/P17

Many of the language deficits reported have been attributed to impaired working memory, namely executive function of working memory (Grossman et al., 1992; Grossman et al., 2000; Kemmerer, 1999). It is worth noting that one of the most well-documented problems in the PD and cognition literature is of course working memory decline (Dirnberger and Jahanshahi, 2013; Gabrieli et al., 1996; Lee et al., 2010). Other cognitive deficits associated with PD are set-shifting deficits (Gauntlett-Gilbert et al., 1999), poor Theory of Mind (Bora et al., 2015), and visual working memory impairments (Zhao et al., 2018).

Given that PD results in changes in the comprehension and production of language and also the awareness of one's own communicative ability, it would seem reasonable to assume that language could be used as a diagnostic tool and a means of monitoring the progression of PD. The aim of this work is thus to automatically detect evidence of PD by extracting linguistic features from textual transcripts generated by participants with PD. Although there is some research that has looked at the acoustic features of speech to detect PD, an examination of linguistic features from textual transcripts is a more neglected area. We first show that it is difficult to approach this as a binary classification task (i.e., with or without PD), particularly because of linguistic similarities between healthy older adults and older adults with PD. We subsequently show that better prediction performance can be had by treating automated detection as a multi-class classification problem. Specifically, we classify participants into one of three groups: healthy younger adults (HYA), healthy older adults (HOA), and older adults with PD (PD). Finally, we show that the same set of linguistic features can be used to predict cognitive performance scores across all subjects.

The structure of this paper is as follows. In Section 2, we discuss related work on using machine learning and speech and language processing to detect age-related conditions, as well as research on linguistic abilities and cognitive functions. In Section 3 we describe how the data in this study were collected, including the participant cohorts, the description tasks given to them, and the cognitive scores that were measured. In Section 4, we describe the linguistic features, machine learning models, and evaluation metrics used. Section 5 presents a series of experiments and key results. We conclude and discuss future work in Section 6.

## 2 Related Work

In the past few years, there has been an increase in research on the detection of aging pathologies using speech and language processing techniques. For example, using spoken language samples elicited in the clinical setting, Roark and colleagues (2011) were able to discriminate older adults with mild cognitive impairment (MCI) from those who showed no evidence of MCI. Masrani et al. (2017b) used domain adaptation techniques that exploit existing data resources from the source domain of Alzheimer's disease (AD) to improve detection in the target domain of MCI. Fraser, Meltzer and Rudzicz (2015) were able to distinguish individuals with probable AD from individuals without AD using only short samples of their verbal responses on a picture description task. The four features that emerged from the verbal responses were semantic impairment, acoustic abnormality, syntactic impairment, and information impairment. Masrani et al. (2017a) also recently explored the task of automatically detecting evidence of dementia within blog data.

However, the detection of PD using computational linguistics remains a relatively neglected area, particularly when compared to research on the detection of MCI and AD. The automatic detection of PD has tended to look at acoustic features extracted from speech signals (Bocklet et al., 2013; Orozco-Arroyave et al., 2016; Pompili et al., 2017). However, Garcia et al. (2016) note the necessity of extracting linguistic features from text to detect PD. The authors explain that computational linguistics can address many of the limitations currently present in the literature on the PD-associated linguistic impairments. For instance, research on language ability is often conducted in controlled and artificial settings, whereby participants must process arbitrary strings of letters or words (Lieberman et al., 1992; Hochstadt et al., 2006). Moreover, the use of linguistic features is often manually coded by researchers. In manual coding, researchers use their subjective interpretations to rate language use. As an example, Murray (2000) asked PD and Huntingdon's disease individuals to describe a picture. Judges then rated the responses for "informativeness." Garcia and colleagues (2016) explain that computational linguistics can be used to assess naturally produced speech, avoiding the confound of biased human interpretation. Using support vector machines with a leave-one-out cross-validation approach, the authors found that semantic fields and grammatical features detected PD with significant rates in accuracy. Garcia and colleagues (2016) also found that although word repetitions were unable to accurately detect PD diagnoses, repetitions could accurately predict performance on neuropsychological batteries.

Interestingly, findings from the Nun Study reveal that language ability in early adulthood is a reliable predictor of cognitive function in later life. Indeed, Kemper and colleagues (2001) showed that language skills in younger adulthood, as measured by grammatical complexity and idea density in written autobiographies, can predict the likelihood of dementia in older adulthood. Riley et al. (2005) found that low idea density in early life is a significant predictor of later aging pathologies. Specifically, low idea density in young adulthood correlated significantly with older adult cognitive impairment. Post-mortem examinations also revealed an association between early life low idea density and AD-related neuropathology. It thus seems reasonable to assume that linguistic features can also be used to predict general cognitive performance in healthy older adults and older adults with PD. Additionally, it is possible that some typically ageing older adults may have age-related cognitive deficits. We use linguistic features of task descriptions to detect evidence of PD and also to predict general cognitive performance across all three groups of HYA, HOA, and PD.

## 3 Corpus Description

In this section we describe the two tasks that were used to generate data, as well as the cognitive tests that were measured.

### 3.1 Script Generation Task

A total of 10 everyday tasks were used in this experiment. An independent panel of five people generated a list of everyday tasks that would not be biased in terms of gender, age and culture. Out of the everyday tasks generated, the 10 tasks most frequently cited were used. Each participant's responses were transcribed using a recording booklet, each of which displayed the individual task at the top of each page. All of the PD participants were recruited at PD support branches where the researchers gave talks on PD, language and cognition. At the end of the talks, the researchers asked for volunteers for their research. If individuals wished to participate in the research, they later contacted the researcher by phone or email.

All of the PD participants were diagnosed by neurologists from the Tayside and Fife medical trusts as having idiopathic PD. The mean number of years since diagnosis of PD was 9.4 (SD = 3.2). The Hoehn and Yahr's (1967) scale of motor impairment revealed three individuals were in stage II (bilateral involvement) and nine were in stage III (mild to moderate disability with impairment to balance). The HOA participants were drawn from an older adult research participant database and the HYA participants were recruited via convenience sampling.

All of the participants were told the title of each of the tasks (e.g. to write and post a letter). The participants were asked to provide sufficient detail to enable someone who was unfamiliar with the task to complete it successfully using the scripts they provided. None of the participants were given any form of constraint or boundary such as not to provide personal information or to only include principal, high-level actions. All of the participants were provided with an example of a script: drying the dishes (pick up the tea-towel, pick up the wet dish from the draining board, rub the tea-towel all over the dish until it is dry and place the dish in the cupboard in its usual place). When the experimenter was satisfied the participant fully understood the instructions, the experiment began.

None of the participants were corrected or aided by the experimenter once the experiment had started, unless the participant forgot the target item. The experiment took between 30 and 60 minutes to complete and all participants were offered a break after each task. All of the participants were debriefed on completion of the experiment.

### 3.2 Directions Task

In this experiment, the participants were shown a list of 36 destinations. 18 had been rated as being very familiar or familiar to most people and the remaining 18 had been rated as being relatively familiar or unfamiliar to most people. From the list of 36 destinations, the participants were asked to pick five places that they knew exactly how to get to and five places that they knew of but were only relatively familiar with how to get there. The list of destinations was presented to the participants in a random order and not according to their level of familiarity.

The participants were asked to mark the items with an F if they were familiar with them and a U if they were less familiar with them. Once they had marked five of each, the experimenter confirmed their level of familiarity verbally, e.g. "so you are familiar with directions to a vet?" or "so you are not as familiar with the directions to a zoo?"

The participants were then asked to provide directions for each of their choices, with the choices ordered randomly. They were asked to provide as clear and precise directions as possible. All participants were asked to give directions from a point they were most comfortable with, e.g. from their house to the zoo.

Note: Participant recruitment for the directions task was the same as used in the script generation task.

### 3.3 Demographic Information

Here we briefly describe basic demographic information about the participants across the two tasks. In the PD group, the average age was 64.1 and the group was evenly split between males and females. The HOA group had an average age of 69.1 and featured two males and 15 females. The HYA group had an average age of 27.17 and contained four males and five females. All participants were Caucasian and were British nationals.

### 3.4 Cognitive and Depression Scores

Here we briefly describe three scores that we analyze in this study: two cognitive scores, and one depression score.

**Phonological Abilities Test (PAT)** The PAT is made up of a series of phonological abilities tasks. The PAT was thus designed to identify reading difficulties early on in young children (Muter et al., 1997). The six tests within the PAT are 1. rhyme detection, 2. rhyme production, 3, word completion, 4, phoneme deletion, 5. speech rate and 6. letter knowledge. The first four tests measure phonological awareness. The fifth test measures speech rate (repeating the word buttercup 10 times as quickly as possible) and the sixth measures knowledge of letters (supplying the name or the sound of each of the twenty-six letters of the alphabet). Only the first four phonological awareness tasks were used in the research.

**Alternate Uses Test (AUT)** The AUT is a measurement of mental inflexibility. The AUT asks participants to produce as many uses for common objects (e.g. brick, or paper) as they can think of. Providing obvious and conventional uses for objects is thought to reflect convergent thinking. An example is suggesting you can use a brick to build a house or use paper to write a letter. Divergent thinking is, however, reflected in responses such as using a brick to make a sculpture or using paper to make a mask for a ball. The diminished capacity to provide uncommon uses of an object is believed to be symptomatic of the inability to switch from one mental set to another and thus the AUT is often employed as a measure of executive function (Lezak, 2004). In this work, we focus on the AUT uncommon uses score (AUTU).

**Beck Depression Inventory (BDI)** The BDI-short consists of 13 items. It is used within a clinical and research setting to measure levels of depression. The BDI is frequently used because it is easy to administer and score. It has the capacity to determine the presence and the level of depression but is unable to measure the frequency and duration of depressive illness (Lezak, 2004). It measures levels of depression by asking the individual to make self-reports about how they are feeling.

## 4 Experimental Setup

In this section we describe the features, machine learning models, and evaluation metrics used in these experiments.

### 4.1 Features

We use a wide variety of linguistic features derived from the subjects' transcripts. The features are entirely derived from the transcripts, as the original speech recordings were not preserved. The features fall into the following categories, and for key features we provide a short handle that can be referred to in the results section.

**Psycholinguistic** We use several psycholinguistic features. Words are scored for their concreteness (CNC), imageability (IMG), typical age of acquisition (AOA), and familiarity (FAM). We also derive SUBTL scores for words, which indicate how frequently they are used in everyday life (subtl1 and subtl2). Masrani et al. (2017b) found similar features useful for detecting MCI.

**Dependency Parse Features** All sentences are parsed using spaCy's dependency parser[1]. We extract several features, including the branching factor of the root of the dependency tree (maxroot_sc), the maximum branching factor of any node in the dependency tree (maxchild_sc), sparse bag-of-relations features, and the type-token ratio for dependency relations (tt_dep).

**Sentiment** We use the SO-Cal sentiment lexicon (Taboada et al., 2011), which associates positive and negative scores with sentiment-bearing words, indicating how positive or negative their sentiment typically is. These are summed over sentences, and then averaged over each document.

**GloVe Word Vectors** Words are represented using GloVe vectors[2], and the vectors are summed over sentences. We then create a document vector that is the average of the sentence vectors. The first five dimensions of the document vectors are used as features (denoted as vdim1 $\cdots$ vdim5 in later discussion).

**Lexical Cohesion** We measure cohesion using the average cosine similarity of adjacent sentences in a document, using the GloVe vectors.

**Sentence and Document Length** We include the average number of words per sentence (avelen), and average number of sentences per document (num_sens).

**Part-of-Speech Tags** We use spaCy's part-of-speech tagger, and use a sparse bag-of-tags representation for the most frequent tags, as well as the type-token ratio for tags (tt_pos).

**Other Lexical Features** Finally, we use a bag-of-words representation for the most common 200 non-stopwords in the dataset, and also calculate the type-token ratio for words (type/token).

### 4.2 Models and Evaluation

In these experiments we primarily use Random Forest regression and classification models, though in the final set of experiments we compare several machine learning methods, including an ensemble of models. We employ a leave-one-out cross-validation procedure.

In the following section, we report results at two levels. At the *document level*, each data instance is an individual description generated by a subject, and the features are derived from each single description. At the *participant level*, each data instance is a participant (subject) and the features are aggregated over all of that participant's descriptions. When doing prediction at the document level, we ensure that a participant cannot have instances in both the training and testing folds.

For evaluation, we report accuracy scores and compare model accuracy with the baseline accuracy that is achieved when always predicting the majority class. We also report the area under the curve (AUC), where 0.5 indicates random classification performance and 1 is perfect classification performance.

[1] https://spacy.io/

[2] https://nlp.stanford.edu/projects/glove/

| Model | AUC | Acc. |
|---|---|---|
| Random Forest | 0.913 | 0.927 |
| Baseline | 0.5 | 0.78 |

Table 1: Predicting Younger vs. Older

## 5 Experimental Results

In this section we describe the sequence of experiments we carried out, with both positive and negative results.

### 5.1 Binary Classification of Parkinson's Disease

Our first experiment demonstrates the difficulty of treating the automatic detection of PD as a binary classification task. We treat the healthy older adults (HOA) and healthy younger adults (HYA) as a single class (the non-PD class) and subjects with PD as the other class (PD). The goal is to use the extracted linguistic features to detect evidence of PD, at both the document level and participant level.

However, at both the document level and participant level, the classification results are essentially random, with AUC scores of 0.49 and 0.51, respectively. Similarly, accuracy levels are below the baseline performance of a system that simply predicts the majority class. We analyze this result in the next set of experiments.

### 5.2 Binary Classification of Older vs. Younger Cohorts

One interpretation of the negative results from the previous section is that the task is difficult because of linguistic similarities between healthy older adults and older adults with PD, and that the cohort of healthy younger adults is linguistically distinct from both older groups.

To test this, we trained a new binary classification model to predict younger vs. older subjects. One class contains the HYA cohort and the other class contains HOA + PD subjects.

The results support our hypothesis, with extremely high accuracy in discriminating between younger and older subjects. Table 1 shows the participant-level prediction scores, with an AUC score of 0.913 using the random forest regression model. The two older groups are highly similar to one another in many respects, with the younger cohort being distinct.

Figures 1 and 2 show some of the similarities between the two older groups and that the younger group is distinct; specifically, the healthy younger adults show higher sentiment and higher SUBTL scores, and the two older groups are similar to each other in terms of those features. This pattern is reflected in many of the other features as well, e.g. younger adults have higher syntactic complexity and lower type-token ratios than the older group.

Given the positive results on this task, we next move away from treating the healthy older adults and healthy younger adults as a single group, and move towards employing a machine learning model that can separate age-related language differences from language differences relating to PD.

### 5.3 Multi-Class Prediction: Healthy Younger, Healthy Older, and Subjects with Parkinson's

Based on the results of the previous two sets of experiments, we reformulated the problem as a multi-class prediction, with three distinct classes HYA, HOA, and PD. We again use the same set of linguistic features described earlier, and random forest classification models. We report accuracy but not AUC scores since this is no longer a binary classification task.

Table 2 summarizes the accuracy scores for document-level and participant-level prediction. Document-level prediction is only at baseline levels, which is not surprising given that many of the documents are very short (some are 1-2 sentences). However, prediction at the participant-level is substantially better than baseline performance, with an overall accuracy of 0.63.

Summarizing the results so far, the first experiment illustrates the difficulty of treating PD detection as a binary classification task. The second experiment explains why, showing that healthy older adults and subjects with PD have linguistic similarities, while healthy younger adults are distinct. This third

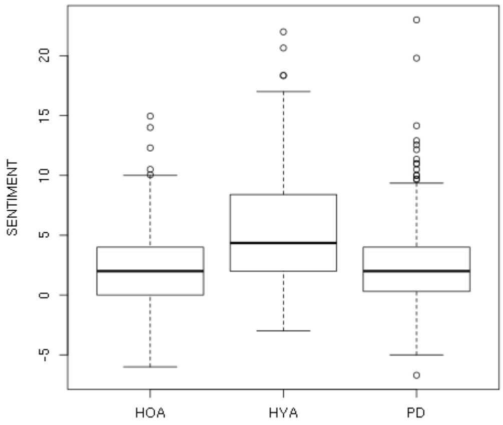

Figure 1: Sentiment by Group

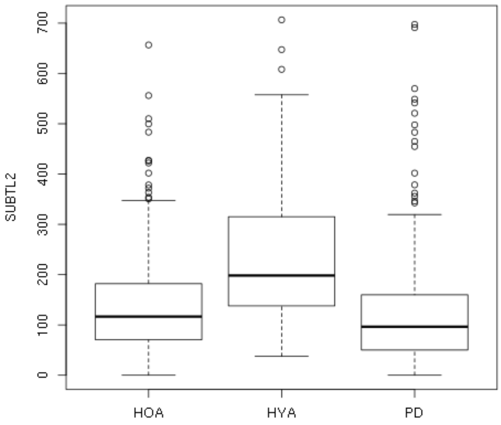

Figure 2: SUBTL2 Scores by Group

| Model | Document-Level | Participant-Level |
|---|---|---|
| Random Forest | 0.52 | 0.63 |
| Baseline | 0.54 | 0.41 |

Table 2: Accuracy for Multi-Class Prediction

experiment shows that performance is substantially better than baseline performance when approaching the task as a multi-class problem.

### 5.4 Prediction of Cognitive and Depression Scores

Our final set of experiments moves beyond the prediction of discrete classes, and we instead try to predict the cognitive abilities of all subjects in all cohorts. This is motivated partly by the above experimental results, and by the hypothesis that some healthy older adults might have mild age-related cognitive impairment, even though they have not been diagnosed with PD or any form of dementia.

As described in Section 3, we recorded a variety of cognitive and depression measures for each subject. In this final experiment, we test whether we can use the same linguistic features as the previous experiments for predicting cognitive and depression scores across all participants.

Table 3 summarizes the results for automatic prediction of three of the test scores, BDI, AUTU, and PAT. For both BDI and PAT, the best machine learning models are able to outperform a baseline that predicts the mean value of the training observations. The ensemble of models yields the lowest MSE on predicting BDI scores, while the Lasso and Random Forest regression methods give the lowest MSE on predicting PAT scores. On predicting AUTU scores, no machine learning model fares better than the baseline. This is owing to the fact that there is relatively little variation in scores amongst subjects. For BDI, the ensemble approach gives results that are significantly better than kNN and Random Forests,

| Model | BDI | AUTU | PAT |
|---|---|---|---|
| Least Squares | 7.46 | 226.26 | 166.81 |
| Lasso | 7.37 | 112.51 | **75.92** |
| kNN | 8.69 | 100.37 | 95.32 |
| Random Forest | 8.89 | 83.53 | 82.21 |
| Ensemble | **6.69** | 98.35 | 76.60 |
| Baseline | 8.25 | **82.54** | 94.21 |

Table 3: MSE for Predicting Cognitive and Depression Scores

| Variable | SS | df | MSE | F | P |
|---|---|---|---|---|---|
| BDI | 113.74 | 2,38 | 56.87 | 10.38** | .00 |
| PA | 1591.10 | 2,38 | 795.55 | 15.44** | .00 |
| AUTU | 1112.08 | 2,38 | 556.04 | 11.91** | .00 |

Note. N=41. *p<.05, **p<.01

Table 4: A One-Way Analysis of Variance of Neuropsychology Test Scores by Group

| | HYA vs. HOA | | | HYA vs. PD | | | PD vs. HOA | | |
|---|---|---|---|---|---|---|---|---|---|
| Variable | Mean Diff. | SE | p | Mean Diff. | SE | p | Mean Diff. | SE | p |
| avelen | 1.27 | 1.31 | .60 | 4.34** | 1.36 | .01 | -3.07* | 1.19 | .04 |
| sentiment | 3.26** | .74 | .00 | 2.59** | .77 | .01 | .67 | .67 | .58 |
| vdim1 | -3.85* | 1.49 | .04 | -.22 | 1.55 | .99 | -3.63* | 1.35 | .03 |
| vdim4 | -1.23 | .61 | .12 | .36 | .63 | .84 | -1.58* | .55 | .02 |

Note. *p<.05, **p<.01

Table 5: Tukey HSD Post Hoc Comparisons of Group for Average Length of Script, Sentiment, Vdim1 & Vdim4

according to paired t-tests. For AUT, the only statistically significant differences are that least squares regression is significantly worse than the Random Forests, Lasso, ensemble, and baseline approaches. For PAT, the ensemble and Lasso approaches are again significantly better than least squares regression.

Figures 3, 4, and 5 show feature importance scores for some of the features that were most useful in predicting AUTU, BDI, and PAT, respectively. An individual feature's importance score is determined by how useful that feature was in reducing MSE, on average, when it was used as a split in the decisions trees used within the Random Forests model. For example, length and sentiment features are very useful for all three prediction tasks.

We also perform statistical analyses to further explore linguistic ability and cognitive functioning. First, a one-way Analysis of Variance (ANOVA) was used to examine an effect of group (3 levels: HYA, HOA & PD) on the cognitive tests, as illustrated in Table 4. Analyses revealed main effects of group on BDI scores ($F$ (2, 38) = 10.38, $p < .01$), PAT scores, ($F$ (2, 38) = 15.44, $p < .01$), and AUTU scores ($F$ (2, 38) = 11.91, $p < .01$). The results indicate that group has a significant effect on all three of the cognitive tests.

A one-way ANOVA was also used to examine an effect of group on the linguistic features. Analyses revealed main effects of group on average length of script ($F$ (2, 38) = 5.81, $p$ = .01, $\eta^2$ = .23), sentiment ($F$ (2, 38) = 10.15, $p < .01$, $\eta^2$ = .35), vdim1 ($F$ (2, 38) = 4.92, $p$ = .01, $\eta^2$ = .21), and vdim4 ($F$ (2, 38) = 4.53, $p$ = .02, $\eta^2$ = .19). Post hoc comparisons were performed using the Tukey HSD test, as illustrated in Table 5. Tukey HSD comparisons revealed significant differences between the groups for the following measures ($p < .05$): average length of scripts was significantly lower in the PD group ($M$ = 13.60, $SD$ = 2.69) compared to the HOA group ($M$ = 16.67, $SD$ = 3.74) and the HYA group ($M$ = 17.93, $SD$ = 3.20). The number of sentiment items was also significantly higher in the HYA group ($M$ = 5.40, $SD$= 2.51) than the HOA group ($M$ = 2.51, $SD$ = 1.02) and the PD group ($M$ = 2.82, $SD$ = 2.09). The HOA group had greater mean vdim1 values ($M$ = 6.91, $SD$ = 4.64) than the HYA group ($M$ = 3.06, $SD$ = 1.20) and the PD group ($M$ = 3.28, $SD$ = 3.69). Finally, mean vdim4 values were significantly lower in the PD group ($M$ = -.34, $SD$ = .81) than the HOA group ($M$ = 1.25, $SD$ = 1.58).

Spearman's rank correlation coefficients were performed to measure correlations between the linguistic features observed in the scripts and the cognitive assessment scores. Correlations were performed within each group. While there were no significant correlations within the HYA and HOA group, significant correlations did emerge in the PD group. The AUTU score formed positive correlations with the features vdim4 ($r_s$ = .79, $p < .01$) and vdim1 ($r_s$ = .74, $p < .01$). Moreover, scores on the BDI were negatively correlated with the feature vdim1 ($r_s$ = -.76, $p < .01$).

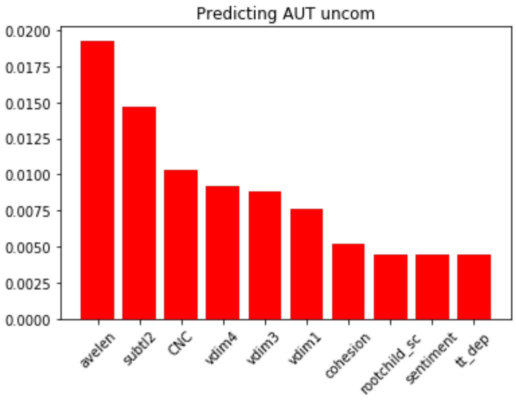

Figure 3: Feature Importance: AUTU

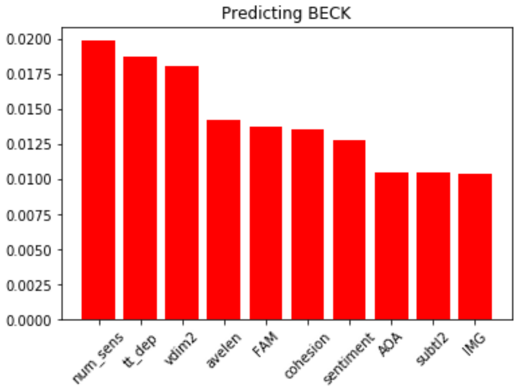

Figure 4: Feature Importance: BDI

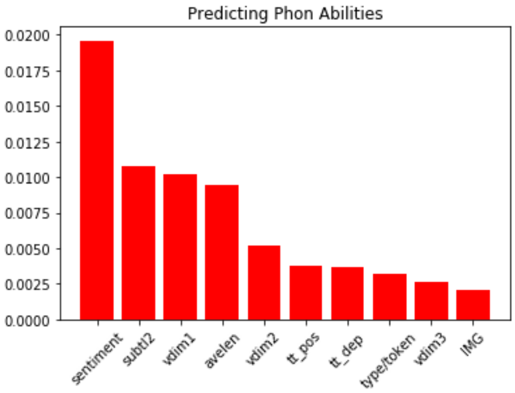

Figure 5: Feature Importance: PAT

## 6 Conclusion

In this set of experiments, we have used natural language processing and machine learning to automatically detect evidence of PD in task transcripts generated by subjects. We first showed that it is difficult to approach this as a binary classification task, particularly because of linguistic similarities between healthy older adults and older adults with PD. We subsequently showed that a multi-class classification approach yields better results. Finally, we used the same set of linguistic features to predict scores of cognitive ability across all subjects.

The vast majority of previous work on automatically detecting Parkinson's disease from speech has focused on using acoustic features. Like Garcia et al. (2016), we demonstrated that linguistic features can be very useful for this task. In future work where we have both speech recordings and transcripts, we will investigate the use of multi-modal features.

Future work will also include further experiments on automatically predicting cognitive ability scores, as we have collected numerous other cognitive measures for the subjects who participated in these tasks.

## References

Wendy L Arnott, Helen J Chenery, Bruce E Murdoch, and Peter A Silburn. 2005. Morphosyntactic and syntactic priming: an investigation of underlying processing mechanisms and the effects of parkinson's disease. *Journal of Neurolinguistics*, 18(1):1–28.

Tobias Bocklet, Stefan Steidl, Elmar Nöth, and Sabine Skodda. 2013. Automatic evaluation of parkinsons speech-acoustic, prosodic and voice related cues. In *Proc. of Interspeech, Lyon, France*.

F Boller, Albert M. L., and F Denes. 1975. Palilalia. *British Journal of Disorders of Communication*, 10:92–97.

E Bora, M Walterfang, and D Velakoulis. 2015. Theory of mind in parkinson's disease: A meta-analysis. *Behavioural Brain Research*, 292:515–520.

J Bottcher. 1975. Morphology of the basal ganglia in parkinson's disease. *Acta Neurologica Scandinavica*, 52:7–87.

F. L Darley, A. E Aronson, and J. R Brown. 1975. Hypokinetic dysarthria. pages 171–197.

M. C deRijk, L. J Launer, K Berger, M. M Breteler, J. F Dartigues, M Baldereschi, and A Hofman. 2000. Prevalence of parkinsons disease in europe: A collaborative study of population-based cohorts. *Neurology*, 54:S21–S23.

Georg Dirnberger and Marjan Jahanshahi. 2013. Executive dysfunction in parkinson's disease: a review. *Journal of neuropsychology*, 7(2):193–224.

Bruno Dubois. 1991. Cognitive deficits in parkinson's disease. *Handbook of neuropsychology*, 5:195–240.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2015. Linguistic features identify alzheimers disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

John DE Gabrieli, Jaswinder Singh, Glenn T Stebbins, and Christopher G Goetz. 1996. Reduced working memory span in parkinson's disease: Evidence for the role of frontostriatal system in working and strategic memory. *Neuropsychology*, 10(3):322.

Adolfo M García, Facundo Carrillo, Juan Rafael Orozco-Arroyave, Natalia Trujillo, Jesús F Vargas Bonilla, Sol Fittipaldi, Federico Adolfi, Elmar Nöth, Mariano Sigman, Diego Fernández Slezak, et al. 2016. How language flows when movements dont: an automated analysis of spontaneous discourse in parkinsons disease. *Brain and language*, 162:19–28.

Jeremy Gauntlett-Gilbert, Richard C Roberts, and Verity J Brown. 1999. Mechanisms underlying attentional set-shifting in parkinsons disease. *Neuropsychologia*, 37(5):605–616.

Murray Grossman, Susan Carvell, Matthew B Stern, Stephen Gollomp, and Howard I Hurtig. 1992. Sentence comprehension in parkinson's disease: The role of attention and memory. *Brain and language*, 42(4):347–384.

Murray Grossman, Jenifer Mickanin, Keith M Robinson, and Mark D'Esposito. 1996. Anomaly judgments of subject–predicate relations in alzheimer's disease. *Brain and Language*, 54(2):216–232.

Murray Grossman, Julia Kalmanson, Nechama Bernhardt, Jennifer Morris, Matthew B Stern, and Howard I Hurtig. 2000. Cognitive resource limitations during sentence comprehension in parkinson's disease. *Brain and Language*, 73(1):1–16.

JM Gurd and RM Oliveira. 1996. Competitive inhibition models of lexical–semantic processing: Experimental evidence. *Brain and Language*, 54(3):414–433.

Julie D Henry and John R Crawford. 2004. Verbal fluency deficits in parkinson's disease: a meta-analysis. *Journal of the International Neuropsychological Society*, 10(4):608–622.

Jesse Hochstadt, Hiroko Nakano, Philip Lieberman, and Joseph Friedman. 2006. The roles of sequencing and verbal working memory in sentence comprehension deficits in parkinsons disease. *Brain and language*, 97(3):243–257.

Margaret M Hoehn, Melvin D Yahr, et al. 1967. Parkinsonism: onset, progression, and mortality. *Neurology*, 50(2):318–318.

Judy Illes. 1989. Neurolinguistic features of spontaneous language production dissociate three forms of neurodegenerative disease: Alzheimer's, huntington's, and parkinson's. *Brain and language*, 37(4):628–642.

David Kemmerer. 1999. Impaired comprehension of raising-to-subject constructions in parkinson's disease. *Brain and Language*, 66(3):311–328.

Susan Kemper, Lydia H Greiner, Janet G Marquis, Katherine Prenovost, and Tracy L Mitzner. 2001. Language decline across the life span: findings from the nun study. *Psychology and aging*, 16(2):227.

Yvan Lebrun. 1996. Cluttering after brain damage. *Journal of Fluency Disorders*, 21(3-4):289–295.

Eun-Young Lee, Nelson Cowan, Edward K Vogel, Terry Rolan, Fernando Valle-Inclan, and Steven A Hackley. 2010. Visual working memory deficits in patients with parkinson's disease are due to both reduced storage capacity and impaired ability to filter out irrelevant information. *Brain*, 133(9):2677–2689.

Muriel Deutsch Lezak. 2004. *Neuropsychological assessment*. Oxford University Press, USA.

Philip Lieberman, Edward Kako, Joseph Friedman, Gary Tajchman, Liane S Feldman, and Elsa B Jiminez. 1992. Speech production, syntax comprehension, and cognitive deficits in parkinson's disease. *Brain and language*, 43(2):169–189.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017a. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP 2017*, pages 232–237.

Vaden Masrani, Gabriel Murray, Thalia Field, and Giuseppe Carenini. 2017b. Domain adaptation for detecting mild cognitive impairment. In *Proc. of Canadian AI, Edmonton, Canada*.

Rena Matison, Richard Mayeux, Jeffrey Rosen, and Stanley Fahn. 1982. tip-of-the-tongue phenomenon in parkinson disease. *Neurology*, 32(5):567–567.

Patrick McNamara and Raymon Durso. 2003. Pragmatic communication skills in patients with parkinsons disease. *Brain and language*, 84(3):414–423.

Patrick McNamara, Loraine K Obler, Rhoda Au, Raymon Durso, and Martin L Albert. 1992. Speech monitoring skills in alzheimer's disease, parkinson's disease, and normal aging. *Brain and Language*, 42(1):38–51.

Laura L Murray. 2000. Spoken language production in huntington's and parkinson's diseases. *Journal of Speech, Language, and Hearing Research*, 43(6):1350–1366.

Valerie Muter, Charles Hulme, and Margaret J Snowling. 1997. *The phonological abilities test*. The Psychological Corporation.

Dimitris Natsopoulos, Z Katsarou, S Bostantzopoulou, George Grouios, G Mentenopoulos, and J Logothetis. 1991. Strategies in comprehension of relative clauses by parkinsonian patients. *Cortex*, 27(2):255–268.

John G Nutt, John P Hammerstad, and Stephen T Gancher. 1992. *Parkinson's disease: 100 maxims*. Mosby Inc.

JR Orozco-Arroyave, F Hönig, JD Arias-Londoño, JF Vargas-Bonilla, K Daqrouq, S Skodda, J Rusz, and E Nöth. 2016. Automatic detection of parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500.

Anna Pompili, Alberto Abad, Paolo Romano, Isabel P Martins, Rita Cardoso, Helena Santos, Joana Carvalho, Isabel Guimarães, and Joaquim J Ferreira. 2017. Automatic detection of parkinsons disease: An experimental analysis of common speech production tasks used for diagnosis. In *International Conference on Text, Speech, and Dialogue*, pages 411–419. Springer.

Christopher Randolph, Allen R Braun, Terry E Goldberg, and Thomas N Chase. 1993. Semantic fluency in alzheimer's, parkinson's, and huntington's disease: Dissociation of storage and retrieval failures. *Neuropsychology*, 7(1):82.

Kathryn P Riley, David A Snowdon, Mark F Desrosiers, and William R Markesbery. 2005. Early life linguistic ability, late life cognitive function, and neuropathology: findings from the nun study. *Neurobiology of aging*, 26(3):341–347.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

A Samii, J Nutt, and B Ransom. 2004. Parkinson's disease. *The Lancet*, 363(9423):1783–1793.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Michael T Ullman, Suzanne Corkin, Marie Coppola, Gregory Hickok, John H Growdon, Walter J Koroshetz, and Steven Pinker. 1997. A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of cognitive neuroscience*, 9(2):266–276.

Sonja von Campenhausen, Bornschein Bernhard, Wick Regina, Bötzel Kai, Sampaio Cristina, Poewe Werner, Oertel Wolfgang, Siebert Uwe, Berger Karin, and Dodel Richard. 2005. Prevalence and incidence of parkinson's disease in europe. *European Neuropsychopharmacology*, 15(4):473–490.

Ronald F Zec, Edward S Landreth, Sally Fritz, Eugenia Grames, Ann Hasara, Wade Fraizer, James Belman, Stacy Wainman, Matthew McCool, Carolyn OConnell, et al. 1999. A comparison of phonemic, semantic, and alternating word fluency in parkinsons disease. *Archives of Clinical Neuropsychology*, 14(3):255–264.

Guohua Zhao, Feiyan Chen, Qiong Zhang, Mowei Shen, and Zaifeng Gao. 2018. Feature-based information filtering in visual working memory is impaired in parkinson's disease. *Neuropsychologia*.