EXPLORING THE PRESIDENTS CORPUS

A PREPRINT

Jomar Alcantara
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
alcantaj@aston.ac.uk

Peter Sawyer
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
p.sawyer@aston.ac.uk

George Vogiatzis
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
g.vogiatzis@aston.ac.uk

Felipe Campelo Franca Pinto
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
f.campelo@aston.ac.uk

October 17, 2019

Keywords Mild Cognitive Impairment · Alzheimer's Disease · Machine Learning · Diagnosis · Natural Language Processing

ABSTRACT

This is the paper's abstract ...

1 Introduction

Alzheimer's Disease (AD) and other forms of dementia affect a significant proportion of the global older adult population and the impact on morbidity and mortality rates is considerable. Dementia and AD are currently the leading cause of death in England and Wales and the sixth leading cause of death in the United States (US). A recent report commissioned by the Alzheimer's Society in 2015 estimated that in the UK by 2015 there would be approximately 855,000 people rising to 1 million by 2021. This represents 1 in 79 of the general population rising to 1 in 14 of those aged 65 or over [**?**]. Worldwide, there are 46 million people with a diagnosis of dementia globally and that number is expected to hit 131.5 million by 2050 [**?**]. From a financial perspective, they estimated annual spend on dementia healthcare is 4.3 billion of which approximately 85 million is spent on diagnosis. The total financial burden of dementia (excluding the costs associated with early onset dementia) is 26.3 billion annually. Globally, this picture is a lot bleaker. Another report by Alzheimer's Disease International suggests that in 2015 there were The report also states that the worldwide cost of AD in 2018 is estimated to be in the region of one trillon US dollars.

AD is a neurodegenerative disease in which a definitive diagnosis can only be produced at post-mortem. However, there are a number of psychological and physiological indicators that can indicate that dementia is present. From a physiological perspective, researchers have identified two proteins called beta-amyloid and tau. In a typical case, tau accumulates and eventually forms tangles inside neurons and beta-amyloid clumps into plaques which slowly builds up between neurons. At a certain point, the levels of beta-amyloid rise and trigger a more rapid spread of tau throughout the brain. Eventually, due to this and other changes, neurons lose their ability to communicate and the brain starts to shrink. This leads to some of the more psychological symptoms, those who have dementia demonstrate cognitive deficits such as problems with episodic and semantic memory, organizing and planning, difficulties with language, problems with executive function and visuospatial deficits [**?**]. In addition, these symptoms are often accompanied by emotional problems such as depression and behavioural difficulties. As more neurons die throughout the brain, a person with Alzheimer's gradually loses the ability to think, remember, make decisions and function independently.

Despite this growing problem and an increasing understanding about how AD affects the brain there are no medications that improve the prognosis of those with AD. All the medications that are currently on the market are designed to manage symptoms. Whilst there are numerous investigational drugs in development for the treatment of AD, a larger than normal percentage (99.6%) of these drugs fail in clinical trials (in contrast to anti-cancer drugs which have a 80% failure rate) [**?**]. Researchers have proposed that a possible reason for the lack of success is that the drugs treatments are initiated too far along in the progression of the disease and thus much of the degeneration of the brain has already taken place [**?**]. Research has started to be more focused on AD at it's earliest stages which some literature describes as 'Mild Cognitive Impairment (MCI) due to AD'.

One of the challenges of this approach is differentiating natural cognitive decline due to aging with decline due to a form of cognitive impairment or dementia. This challenge is often complicated further due to the large variation in the cognitive abilities and educational background of individuals. Albert and his team have worked to define clinical criteria which professionals can use to diagnose MCI due to AD and differentiate this from age-associated memory impairment and age-associated cognitive decline. One of the most important observations from this piece of work is that a diagnosis of MCI requires evidence of intra-individual change and optimally requires evaluation at two or more points [**?**], and this is essentially to place more importance on the trajectory of a person's cognitive abilities rather than a person's cognitive ability in general.

There has been a significant research in the area of language deterioration as a means of detecting Alzheimer's Disease. This usually takes the form analysis of speech recorded as part of a cognitive assessment such as the Picture Description Task [**?**, **?**]. Given that language samples are relatively easy to collect, research has moved towards analysis of spontaneous speech. An good example of this type of research is the study conducted by Berisha and Liss which looked at the differences in language use between two US presidents, Ronald Reagan (RR), who would go on to receive a diagnosis of Dementia and George H. W. Bush (GHWB) who acted as a matched control based on Age [**?**]. They found several differences in language use over time which they felt acted as indicators of RR's difficulties with language due to dementia. These significant differences were in the number of unique words used per speech, the use of non-specific nouns and fillers and low-imageability verbs [**?**].

This study replicates work done by Berisha and Liss and extends this by adding Donald Trump (DJT) as an alternative, more appropriate comparison to RR as he is much closer in age than GHWB. This work contributes to our understanding of the pre-diagnosis phase of Alzheimer's Disease and it's effect on language. We explore a wider range of features than previously looked at with this dataset and our hypothesis is that there will be features of RR's language which change more noticably over time when compared with the two closest appropriate controls. In addition, our assumption is that any language deterioration found will be non-linear and so we will be looking at non linear and linear correlations within the data to see if decline is constant over time

2 Methodology

We took 46 transcripts of Ronald Reagans (RR) press conferences from 1981 to 1988 and compared them with 134 press conferences by George H. W. Bush (GHWB) and 29 press conferences conducted by Donald J. Trump (DJT). We analyzed transcripts for lexical features shown to change longitudinally with dementia. These features encompassed, word level, sentence level and document level features and included a number of features contained in the study by Berisha and Liss with the aim of replicating and extending on their findings. We compared the trends described in the transcripts of RR and GHWB, but also included DJT. Berisha and Liss originally made the comparison as it GHWB (GHWB - age at the start of presidency - 64 years, 222 days) was the closest match to RR in terms of age (RR - age at the start of presidency, 69 yeaars and 349 days). However, with the inauguration of Trump, he now is the closest comparible president in terms of age (DJT - age at the start of presidency - 70 years, 220 days). It would be interesting to look at a comparison of RR and DJT to see whether the comparisons made by Berisha and Liss hold true with this more appropriate match (in terms of age). DJT as with GHWB has no known diagnosis of AD. We used the press conference transcripts in the American Presidency Project (APP) archive as a data source for this project. The APP is a comprehensive and organized searchable database of presidential documents, including transcripts of speeches, transcripts of news conferences, and other public documents.

2.1 Pre-processing

To generate the files necessary for analysis, we downloaded each transcript and performed the following changes. We omitted the prepared statement by the president and any speech by other individuals and started each transcript at the beginning of the first answer to a question by a member of the press. We filtered any annotations that were added to the transcript, including any references or clarifications, and any laughter. It's worth noting that there appears to be a difference in how 'hesitations' were marked down between each president, for RR & DJT hesitations were marked by

a single hyphen whereas for GHWB hesitations are marked by a double hyphen. In order to maintain consistency when parsing through the documents, We have changed both types of hesitation to be marked by a single hyphen. We also omitted one word sentences as this data would, from a theoretical perspective, not be relevant for language analysis. We did not control for the length of the document, but generated features which would normalise by the length of the document. We therefore was able to include all press conferences by all presidents analysed where there was a question and answer session conducted at least in part by the president in question (2 press conferences of GHWB were omitted due to a lack of a question and answer session).

2.2 Feature Generation

We calculated the following features for each transcript in turn using the Natural Language ToolKit (NLTK) package [**?**], Linguistic Inquiry and Word Count (LIWC) [**?**] and Python. Words were stemmed to their root form using the Porter2 (Snowball) stemmer algorithm.

2.2.1 Measures of lexical diversity

Lexical diversity is a measure of the variety of language used within a given document. A document is said to have high lexical diversity if the number of unique words is large. We constructed four features of lexical variation. Firstly we looked at the number of unique words. To do this we were able to split each transcript into individual words and changed them to lowercase using NLTK and were then count the number of unique words that appeared in each transcript. We also used the TTR formula, Brunet's Index and Honore's Statistic as other measures of lexical diversity [**?**].

Type token ratio (TTR) is the ratio obtained by dividing the types (The total number of different words) by the tokens (the total number of words in an utterance).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{1}$$

Brunet's Index (W) differentiates itself from TTR, as it is not impacted by the length of the text itself. Brunet's Index is defined by the following equation:

$$W = N^{V(-0.165)} \tag{2}$$

where N is the total length of the utterance being measured and V is equal to the total vocabulary being used by the subject. Brunet's Index usually has a score of between 10 and 20, with high numbers indicating a more rich vocabulary compared to low numbers.

Honore's Statistic is based on the idea that vocabulary richness is implied when a speaker uses a greater amount of unique words. This is indicated by the following equation:

$$R = (100 \log N)/(1 - V1/V) \tag{3}$$

where v1 is equal to the number of unique words, V is the total vocabulary used and N is the total number of words in the utterance being measured.

2.2.2 Fillers, Non-Specific Nouns and Low Imageability Verbs

Fillers, Non-Specific Nouns and Low Imageability Verbs were features used by Berisha and Liss in their research [**?**] and were taken from work done by Bird et al [**?**]. Fillers can be described as a potentially meaningless word that marks a pause or hesitation in speech. In those with MCI and AD, these words can be used to temporarily disguise problems in thought processes or word finding difficulties. Non-Specific Nouns refer to a category or an unspecified member of a given category, once again this can be characterised as a compensatory strategy for word finding difficulties. Imageability is characterized, according to Berisha and Liss, as the ease with which a word provokes a mental image of what the word describes.

Category	Words
Fillers	*"um", "uh", "er", "ah", "like", "okay", "right", "you know", "well", "so", "basically", "actually", "literally"*
Non Specific Nouns	*"something", "anything", "thing", "everything"*
LI Verbs	*"be", "come", "do", "get", "give", go", "know", "look", "make", "see", "tell", "think", "want"*

Table 1: Examples of words belonging to the categories Fillers, Non-Specific Nouns and Low Imageability Verbs

2.2.3 Usage of parts of speech

Using a Part of Speech tagger (PoS) on each transcript analyses each sentence within the transcript and assigns a 'tag' to each word based on the function the word has in a sentence. At a basic level this can be divided into the eight defined parts of speech: 'nouns', 'pronouns', verbs', 'adjectives', 'adverbs', 'conjunctions', 'prepositions' and 'interjections' but can be further subcategorised. We used the PoS tagger built into NLTK to tag each transcript in turn and used these the counts from each of these eight categories in our analysis. In addition to frequency counts we also normalised these features by dividing the frequency count by the number of words in the document to take into account transcript length.

2.2.4 Linguistic Inquiry and Word Count(LIWC)

The LIWC is a text processor which analysis words within a given a document and compares the words in the LIWC dictionary file. The LIWC dictionary file contains over 6000 words which are categorised into approximately 90 different types. For example the word 'cried' is part of five word categories: sadness, negative emotion, overall affect, verbs and past focus [**?**]. Each word in the document or set of documents is searched for in the dictionary file and if found, the appropriate type is incremented by 1. What is output is an analysis of word usage in reference to each category. This is particularly useful in analysing both structural language changes but also the content or themes of language for a given transcript or set of transcripts.

As with Berisha and Liss, we analysed the transcripts over time in order to determine if there was an underlying temporal trend that may indicate that different parts of language increase or decrease over time. In AD, cognitive abilities are said to deteriorate over time and therefore it may be useful to analyse temporal trends that indicate a potential deterioration of cognition. In order to get a accurate measure of deterioration over time, we time stamped each transcript in relation to number of days from the first transcript to the date of the transcript being analysed as the press conferences were not evenly spaced out during the presidencies.

3 Results

One of the most important thing to note is the wide variety of samples between the three presidents and also the varying timescales. RR participated in 46 press conferences over eight years (an average of 5.75 a year) which is the fewest number of press conferences given by an American president during their term of office. GHWB participated in 136 press conferences over four years (an average of 34 a year) and DJT participated in 29 press conferences to date (an average of 19.3 per year). Equally, there are differences in the average number of words. RR produced an average of 3424 words per conference compared to 2608 by GHWB (unpaired $t = 4.434$, $p<0.001$) and DJT at 1849 words (unpaired $t = 6.524$, $p<0.001$).

	RR	GHWB	DJT
Total Words	3423.91 (416.42)	2607.72 (1210.38)	1848.65 (1549.38)
Unique Words	894.13 (85.15)	667.76 (218.67)	481.82 (221.29)
Mean Length of Utterance	23.17 (1.402)	18.71 (2.067)	13.84 (1.619)

Table 2: Means and Standard Deviations of general features for each set of transcripts

In terms of more specific language differences between the presidents, we found that RR used significantly more unique words, non-specific nouns and low imageability verbs than GHWB and DJT (see Table 3.3). The mean length of utterance for RR was significantly greater than that of GHWB and DJT. Some of these differences are due to the length of the sample, particularly in the case of DJT where his average sample is almost half the sample of RR. It could also be said that this could be down to differences in linguistic abilities or speaking style [**?**, **?**]. However, we can certainly see that as controls GWHB and DJT are comparative in relation to non-specific nouns and LI verbs.

	RR v GHWB	RR v DJT	GHWB v DJT
Total Words	**4.434*** **	**6.524*** **	**2.899** **
Unique Stems	**10.878*** **	**10.111*** **	**4.148*** **
Mean Length of Utterance	**16.175*** **	**25.084*** **	**13.484*** **
Non Specific Nouns	**7.877*** **	**3.426** **	-0.574
LI Verbs	**2.656** **	**3.420*** **	1.628

* denotes p<0.05
** denotes p<0.01
*** denotes p<0.001

Table 3: RR T-tests vs GWB and DJT

3.1 Longitudinal Analysis

We then looked at the data from a longitudinal perspective as we are interested seeing whether we can track various language variables and their progress over time. We ran a number of Pearsons correlations with transcript index number as a time reference and the dependant variables (Table 3.4). For our controls, we found them to be stable for the most part with the main highlights being a decrease in Adverb usage for DJT (R=-0.36, p=0.049) and a steady but not severe decline in a number of variables for GWHB, namely total word count, unique words, low imageability words and verb usage.

For RR, his decline is more marked and more widespread through his language use. We noticed an significant increase in adverb (R=0.41, p=0.004) and pronoun usage (R=0.65, p<0.001), as well as a slight usage increase in Non-specific nouns(R=0.30, p=0.03). There was a highly significant decrease in number of unique words (R=-0.56, p<0.001) and noun usage (R=-0.70, p<0.001). Also very significant decrease in adjective usage (R=-0.40, p=0.005) and a significant decrease in total word count (R=-0.31, p=0.03).

	RR	GHWB	DJT
Word Count	**-0.31***	**-0.21***	0.08
Unique Words	**-0.56*** **	**-0.25** **	0.16
Non Specific Nouns	**0.30***	-0.08	-0.03
LI Verbs	-0.19	**-0.20** **	0.02
Nouns Normalised	**-0.70*** **	-0.03	0.14
Verbs Normalised	**0.36** **	**0.24*** **	-0.03
Adjectives Normalised	**-0.40** **	0.08	-0.34
Adverbs Normalised	**0.41*** **	0.02	**-0.36***
Pronouns Normalised	**0.65*** **	0.13	0.07

* denotes p<0.05
** denotes p<0.01
*** denotes p<0.001

Table 4: Pearson Correlations for Features

Given the number of hypotheses being tested, it is necessary to apply some correction for family wise error rate and therefore a Hochberg step-up procedure was applied to the results of the analysis. We also used the Benjamini-Yekutieli procedure for controlling for False Discovery rate and came up with a final list of important features for each president. For RR, this was 22 features, for GHWB this was 14 features and there were no significant features for DJT after applying these methods. The features were generated from multiple sources and therefore there was some significant overlap between the features. In these cases, the normalised features were prefered to raw count features. Where there was doubt as to the most appropriate feature, all features have been included.

	RR R-Squared	GWB R-Squared	DJT R-Squared
ppron	**0.700*** **	0.2	0.35
social	**0.698*** **	0.2	0.35
NounsNormalised	**-0.689*** **	0.2	0.35
function	**0.670*** **	0.2	0.35
conj	**0.644*** **	0.2	0.35
PronounsNormalised	**0.631*** **	0.2	0.35
Analytic	**-0.626*** **	0.2	0.35
NN	**-0.601*** **	0.2	0.35
male	**0.585*** **	0.2	0.35
UniqueWords	**-0.578*** **	0.2	0.35
WDT	**-0.577*** **	0.2	0.35
shehe	**0.529***	0.2	0.35
VBZ	**-0.521***	0.2	0.35
JJ	**-0.518***	0.2	0.35
Fillers	**-0.518***	**-0.518***	0.35
EX	**-0.518***	**-0.518***	0.35
achieve	**-0.518***	**-0.518***	0.35

* denotes p<0.05
** denotes p<0.01
*** denotes p<0.001

Table 5: Pearson Correlations for Features

	RR R-Squared	GWB R-Squared	DJT R-Squared
ppron	**0.700*** **	0.2	0.35
social	**0.698*** **	0.2	0.35
NounsNormalised	**-0.689*** **	0.2	0.35
function	**0.670*** **	0.2	0.35
conj	**0.644*** **	0.2	0.35
PronounsNormalised	**0.631*** **	0.2	0.35
Analytic	**-0.626*** **	0.2	0.35
NN	**-0.601*** **	0.2	0.35
male	**0.585*** **	0.2	0.35
UniqueWords	**-0.578*** **	0.2	0.35
WDT	**-0.577*** **	0.2	0.35
shehe	**0.529***	0.2	0.35
VBZ	**-0.521***	0.2	0.35
JJ	**-0.518***	0.2	0.35
Fillers	**-0.518***	**-0.518***	0.35
EX	**-0.518***	**-0.518***	0.35
achieve	**-0.518***	**-0.518***	0.35

* denotes p<0.05
** denotes p<0.01
*** denotes p<0.001

Table 6: Pearson Correlations for Features

4 Discussion

President Reagan received his diagnosis of AD in August 1994 but using transcripts of speeches he made in his two terms as President (January 1981 - January 1989) we have be able to identify certain changes in his use of language that we might ascribe to the onset of MCI and early AD. Despite differences in our methodology, our research supports some of the findings of Berisha and Liss in that we find an increase in non-specifc noun usage. Compared to our controls (GWHB and DJT), we find some slight trends with GWHB but no such trends with DJT in his speech albeit his samples of speech span a shorter amount of time.

A criticism of Berisha and Liss's work is the problems they had with normalising the transcripts in terms of length. This was also a problem in the work of Garrard et al [**?**, **?**]. Whilst it is important to control for outliers, there are other ways in which we can control for length of sample. In this paper, we controlled for transcript length by dividing any features that were raw counts by the total length of the transcript. Given the increased amounts of hypothesis testing, it became necessary to control for false discoveries.

Interestingly, when we normalised the various types of words used by the presidents we found some interesting patterns that further differentiated RR from the controls. Whilst Non-specific nouns increased over time, we found that noun usage in general significantly decreased and pronouns increased similarly significantly. The increase in pronoun for those with early AD has been identified in literature, although there are only a few studies that explore this [**?**, **?**]. Wendlestein et al propose that the increased used of pronouns is an expression of an impaired ability to adapt language to the listener's needs [**?**]. Almor et al attributed this reliance on pronouns due to a impaired working memory [**?**].

The decrease in overall noun usage has also been identified as a feature. Jarrold et al found that AD patients would use more pronouns, verbs and fewer nouns than controls [**?**]. Wendlestein in their investigations into noun usage found that decreased later on in AD progression and was unaffected in the pre-clinical stages of AD [**?**]. Our results are supported by existing literature and this potentially means that language analysis in the way we have structured it may have diagnostic or prognostic properties. One of the most striking differences between this study and the original is when looking at Unique Words. Whilst Unique Words as a feature is significant, which was a finding in the original study, when we controlled via normalisation, we found that this was not significant.

Another observation is that when conducting fitting a linear model to this data does not tell the whole story. For example, consider the following two plots. One shows a model fit to a linear model, and the other shows a model fit using local polynomial regression model which aims to model the data more closely.

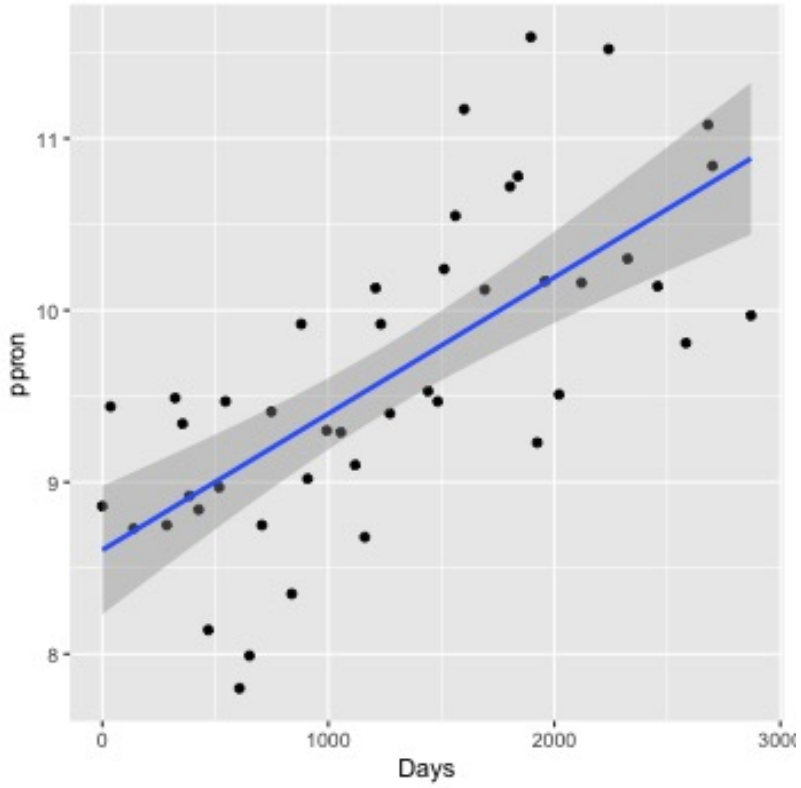

Figure 1: RR Personal Pronouns - Linear Model

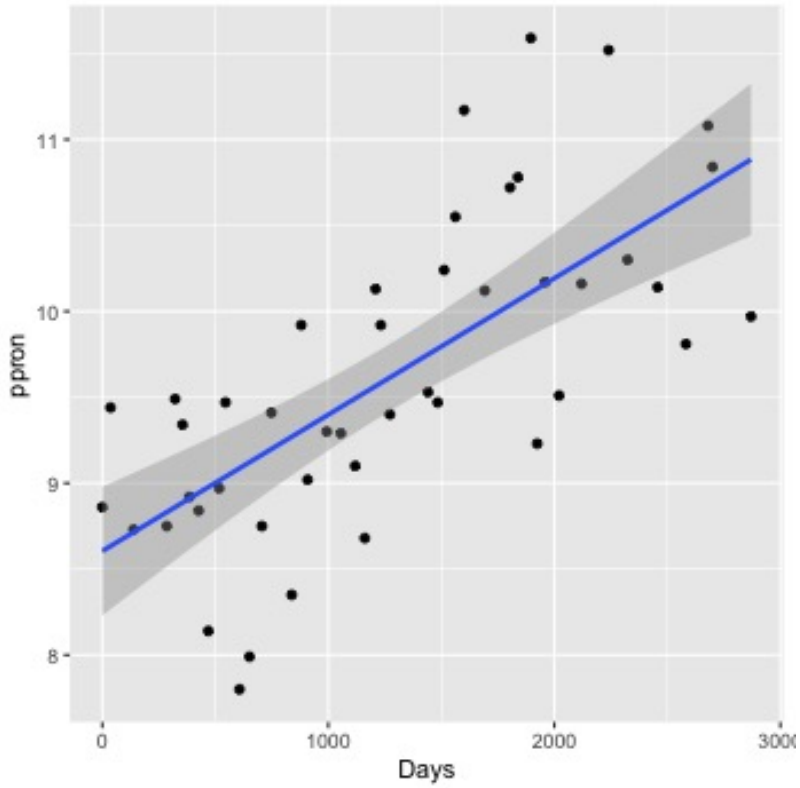

Figure 2: RR Personal Pronouns - Local Regression Model

There are limitations of this research. Whilst in terms of age, DJT is certainly more suitable as a control to match with RR, in some ways they held very different styles of press conferences in that RR preferred to do solo press conferences and DJT has shown a preference for doing joint press conferences which have an impact on the amount of language produced. This artifact of the data is in itself notable as it illustrates the problems we may have with smaller amounts of speech. Also, the problem of finding an appropriate control is a common one in this domain. Given that those with MCI and early dementia have such variable presentations, it might prove of limited value in matched pairs design.

With further work, it is not feasible to the vast array of samples over a timeframe, as we have had with the president corpus and so it would be worth exploring how the quality of these predictions might lessen when faced with considerably fewer samples and over a smaller time period. It would also be worth extending this research further to encompass more of the linguistic features Fraser used in her work [**?**] to see if there are any further insights to be gained. In addition, this replication and extension has demonstrated the potential utility of using longitudinal data as a means of comparing language use of a person at two or more time periods and using this information as a diagnostic aid for MCI and therefore more work would be helpful from a longitudinal perspective to see if this approach may be valid in moving towards a solution for this particular problem.

5 Conclusions

The results of this work show that we can track a person's use of language through time in a number of ways, and it is possible for an individual to be his or her own control. This is important as it means the heterogenous nature of the MCI population does not impact results as much as if we were comparing those with MCI to controls. Equally, it would be helpful to have controls to ascertain what would be usual to expect in the decline of language in a healthy older adult.

References