Qualifying Report

Jomar Alcantara

2019-01-07

Abstract

Contents

List of Tables

List of Figures

1 Overview of Problem

Dementia has been identified as one of those fast growing difficulties facing the world. A recent report suggests that in 2015 there were 46 million people with a diagnosis of dementia and that number is expected to hit 131.5 million by 2050 [3]. The report also states that the worldwide cost of dementia in 2018 is estimated to be in the region of one trillion US dollars.

In 2009, the Department of Health designed it's National Dementia strategy and as part of this made early diagnosis and support one of it's key priorities. A lot of work has gone into trying to find ways of improving the early diagnosis of Alzheimer's Disease (AD) and Mild Cogntive Impairment (MCI) with research focused two areas - identifying biological markers and analyzing the cognitive decline of those who are suspected to have the disease. As described above, the numbers of those suffering from AD and MCI are going to increase as the population ages and thus it is important that we utilize technology wherever possible to aid clinicians in the detection of MCI and AD. At the present time diagnosis is typically conducted at memory clinics by trained clinicians. I theorize that we may be able to enable an earlier diagnosis of those with MCI and AD using natural language.

There is a large body of research that looks at the decline in language in those with MCI and AD. However there is conflicting evidence in these studies about which declining language factors are associated of MCI and AD. Research therefore should look at these features in more detail and a clarification of this currently disorganised picture should go some way to helping researchers further understand the disease and it's progression. One of the difficulties in research of this nature is the process of collecting appropriate language samples. Part of this work therefore is to see whether spontaneous discourse, such a semi-structured interview, has the ability to put pressure on both the cognitive and linguistic systems in the same way as traditional cognitive tests such that it might be able to distinguish between healthy controls, those with MCI and those with AD. There is some evidence to support this. Berisha et al [1], has shown through a longitudinal language analysis of spontaneous speech that there are marked differences in this process between those who would go on to have a diagnosis of AD and a healthy control.

The potential impact of this research is immense. Research has shown that early diagnosis of people with AD or MCI improves sufferers quality of life and can, in some cases, slow the progress of the disease. Early diagnosis can increase the number of research opportunities for understanding the early stages of dementia and how the disease progresses so that more research can be conducted which may, in the future, lead to new treatments and other interventions.

1.1 What is this subsection called?

Full Provisional Title of the Research: Title here
Research Questions and Hypotheses The broad question is can we use Machine Learning to assist in the early diagnosis of dementia and/or mild cognitive

impairment through language produced naturally rather than language that is produced as part of a formal cognitive assessment. There have been a number of studies from the psychological perspective that look at the changes in language in those with Dementia. These studies have identified a number of speech and language differences between healthy controls and those with dementia. Some of these features have been validated in work using machine learning to classify healthy controls from those with dementia. However, these have been done in lab settings, with formal cognitive assessment tasks such as the 'picture description task'. While this is help to learn that machine learning can be used to differentiate between healthy controls and those with dementia, this doesn't take the burden off healthcare professionals who have to conduct the assessments and other professionals who may be used to transcribe these assessments.

Machine Learning itself is undergoing a shift in focus at the present time. I will describe this shift using the context of my proposed research. In traditional machine learning, the transcripts from cognitive assessments are used to generate features such as idea density. These features are then used to classify patients. A company in Toronto, Canada has taken this approach and currently uses over 400 acoustic and language features to make this classification and to generate a psuedo MMSE score.

Deep Learning takes a different approach by bypassing the feature generating stage. Deep learning uses neural networks as a fundamental building block.

1.2 Aims and Rationale

The aim of my research is to find less burdensome ways of detecting dementia without the use of invasive procedures (taking bloods, or using medical equipment such as MRI's and EEG's) and without resorting to to time-consuming and expensive psychological tests. There is a lot of research into the analysis of language as a bio-marker for MCI and Early Dementia. Given that sampling a person's language is relatively effortless, my research looks at whether we can find bio-markers of MCI and Early Dementia in natural language.
Concerns: Language and Memory are quite naturally intertwined and it would be difficult to test one without some reliance on the other. I'm not going to control for memory problems as a potential confound, but does this weaken the research? How do I defend this?

- Hypothesis. I predict that, given a number of samples of language over a given time from, is it possible to detect language decline in humans such that you can differentiate between normal language decline and decline in those with MCI.
- Hypothesis. I predict that my new neuropsychological battery of tests will be more sensitive at detecting mild cognitive impairment than existing measures.

- My first hypothesis is whether monitoring language samples and other measures of cognitive decline over the course of the year is enough to detect cognitive decline in a population of those with Subjective Cognitive Decline and Mild Cognitive Impairment.
- My second hypothesis is that those with MCI show a different language profile to controls, if matched similarly for education and other significant factors.

1.3 Research Challenges

2 Literature Review

2.1 Introduction

Alzheimer's Disease is a neurodegenerative disease in which, from a physiological perspective, the brain develops neurofibrillary tangles and neuritic plaques along with the deterioration and loss of cortical neurons and synapses. Whilst a definitive diagnosis of Dementia can only produced at post-mortem, there are a number of clinical indicators from psychological perspective that can indicate that dementia is present. From a clinical psychology perspective, those who have dementia have cognitive deficits such as problems with episodic and semantic memory, organising and planning, difficulties with language and visuospatial deficits[**?**]. In addition, these symptoms are often accompanied by emotional difficulties such as depression and irritability and behavioural difficulties.

AD and other forms of dementia affect a significant proportion of the geriatric population in the world today and is currently the sixth leading cause of death in the US and was named the leading killer of women in the UK. According to a recent report commissioned by the Alzheimer's Society in 2015, they estimate the prevalence of AD in the UK at approximately 815,000 people. This represents 1 in 14 of those aged 65 or over and 1 in 79 of the general population [2]. From a financial perspective, they estimate an annual spend of £4.3 billion of which approximately £85 million is spent solely on diagnosis and that the total impact of AD (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 46 million people with a diagnosis of dementia globally and that number is expected to hit 131.5 million by 2050 [3]. The report also states that the worldwide cost of AD in 2018 is estimated to be in the region of one trillon US dollars.

Despite this growing problem, at present there are no drugs that improve the prognosis of those suffering with AD. All the drugs that are 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) [4]. 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. It is therefore important to focus AD at it's earliest stages which some literature describes as 'Mild Cognitive Impairment (MCI) due to AD'.

Current thinking suggests that the cognitive deficits associated with AD often begin before the clinical symptoms of the disease become apparent. Researchers propose that neurofibrillary tangles and other associated physiological effects of AD develop over time and alter cognitive function until a threshold is reached and clinical symptoms become more obvious [5]. The case of Iris Murdoch illustrates this theory well. Le et al [**?**]found, in their analysis of three writers and the novels they wrote, that Iris Murdoch's work declined subtly over time, but there was a steep drop off in the use of language in her last novel. If this theory holds true more generally, it should be possible to detect subtle cognitive changes in language and memory before a clinical diagnosis can be formed. However, one of the challenges of this approach is differentiating natural cognitive decline due to aging with decline due to a form of dementia. Albert et al 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. They note that the diagnosis of MCI requires evidence of intraindividual change and optimally requires evaluation at two or more points.

1. Concern regarding a change in cognition - A person or an informant should express concern that there is a change in cognitive ability in comparison to previous level of performance.

2. Impairment in one or more cognitive domains - There should be evidence of lower performance in one or more cognitive domains beyond what would be expected of a person given their age and education.

3. Preservation of independence in functional - Whilst persons with MCI are expected to be able to maintain independence, it is common to experience mild problems in complex functional tasks which they may have been able to perform previously. This might mean that they take more time or be less efficient at completing these tasks, or it may be that may make more mistakes.

4. Not demented - The deterioration should be mild to the point that there is no significant loss of functioning in social or occupational contexts.

In addition to meeting the above criteria, a clinician must rule out other conditions or factors that could account for the decline in cognition with the goal to increase the likelihood that the underlying cause of this decline is AD.

Many researchers have studied the early detection of MCI AD. These studies usually follow two main approaches. The analysis of biomarkers and the examination of patients who have demonstrated decreasing cognitive abilities. The first approach yields reliable results in the detection of AD in its moderate and advanced states but does not perform well during the early stages of the

disease. The second approach has gained more attention in recent years, due to the fact that in clinical practice it has shown promise in the early detection of AD. In addition, the analysis of the decline of cognitive abilities is comparitively inexpensive and less invasive that the first approach which commonly requires the collection of a sample of cerebro-spinal fluid which is painful for the patient involved.

One of the most common ways in which clinicians traditionally make an early diagnosis of cognitive impairment is through the use of the Mini Mental State Examination (MMSE) [6]. The MMSE is a brief questionnaire consisting of eleven questions which tests cognitive aspects of mental function and requires only 5-10 minutes to administer [6]. The MMSE is chosen due to it's effectiveness at assessing a person's cognitive mental state at a specific point in time, as well as being as sensitive to changes as a more detailed and complex assessment such as the Wechler Adult Intelligence Scale [6]. Whilst the MMSE is useful as a brief screening tool it has it's limitations. The MMSE was not specifically created to screen for dementias and therefore does not interrogate key aspects of cognitive impairment known to be affected in dementia. It also has limited value in assessing under-educated subjects and a meta-analysis on the effectiveness of the MMSE as a diagnostic tool for dementia to showed that it's accuracy was low (sensitivity between 78.4% and 85.1% and specificity between 81.3% and 87.8%). As the MMSE is shown to have low accuracy specifically in the diagnosis of dementia, it becomes necessary for professionals to employ the use of other tools or measures such as the Free Cued Selective Reminding Test (FCSRT) or the Montreal Cognitive Assessment (MoCA). These tests have the benefits of being much more accurate at diagnosing cognitive impairment and discriminating between dementia and other types of cognitive impairment at the cost time and training of psychological professionals such as clinical psychologists in administering these tests. Given the burden on these professionals is likely to increase due to a growing elderly population or in the case of developing countries where the clinician / patient ratio is already unsustainable, it seems useful to find less burdensome ways to aid professionals in identifying those at risk of developing dementia.

The two main ways in which diagnosis is performed is through assessment of memory and language. Tests of memory are classically among the most accurate ways of diagnosing dementia, however these tests suffer from the same reliance on trained clinicians to administer these tests in a clinical setting. Language however is a lot easier to collect and can be done in more naturalistic settings. As with memory, these tests can be done over time and would be able to chart a patients language degeneration over time. Given that language is less intrusive to test and requires a lot of the cognitive processes that may be impacted by AD, a lot of research has focused on measure decline in the use of language in those with AD. There are a number of difficulties to watch out for with this approach. There are a wide number of factors that are involved in language degeneration in the elderly, and consequently there will be an expected amount of variability between subjects. The administration of such tests may induce nervousness and discomfort which may impact performance, and also repeatedly administering

the same language tests to tests for differences over time by be confounded by improved performance at tasks via practice effects. Also, generally speaking the language tests that are currently being used do not necessarily describe patients 'real performance' in language production. However, there is enough promise in this approach such that it could help further our understanding of the disease, it's progression and the parts of the brain affected in the early stages.

According to the DSM 5 [**?**], those with mild dementia suffer from noticeable word finding difficulty. They may substitute general terms for more specific terms and may avoid the use of specific names of acquaintances. There may be grammatical errors involving subtle omission or incorrect use of articles, prepositions, auxiliary verbs, etc. Those who have progressed from Mild to Major depression also have difficulties with expressive or receptive language. They will often use general-use phrases such as "that thing" and "you know what I mean" and prefers general pronouns rather than names. With severe impairment, sufferers may not even recall names of closer friends and family. Idiosyncratic word usage, grammatical errors, and spontaneity of output and economy of utterances occur. Echolalia (meaningless repetition of another person's spoken words) and automatic speech typically precede mutism. With the wide range of deficits someone with AD can suffer, it makes sense to try to categorise these deficits in some way.

One of the most famous pieces of research on the topic of language decline in dementia was by Berisha and Liss (2015) [1] who examined speeches and public interviews of former US president Ronald Reagan. They found that Reagan's speeches towards the end of his presidency suffered from difficulties in word-finding, inappropriate phrases and uncorrected sentences which are hallmarks of language deterioration associated with Alzheimer's Disease. It turned out later to be the case that he had Alzheimer's Disease. Another classical study by Snowdon et al (1996) [7] looked at whether linguistic ability in early life was associated with cognitive function and AD in later life. They found that idea density (defined as the number of expressed propositions divided by the number of words) was a key predictor in predicting whether nuns would go on to develop AD in later life. They found that those who would go on to develop AD all had low idea density in early life and they found no AD present in those with high idea density in early life. As we can see, just with these two pieces of research the range of language deficits in those who suffer with AD are extremely variable and can differ from patient to patient as the disease progresses. The consensus among researchers that this language degeneration is typically accelerated by the presence of dementia [1] and that a potential indicator of dementia is the rate of change in which the decline occurs relative to a fixed point in time rather than a comparison across a cohort of individuals.

Emery [**?**] completed a literature review looking at all the potential language deficits that could exist in those with AD and / or MCI. She divided these deficits into four levels of language: Phonology, Morphology, Syntax and Semantics. She proposed that language and the processes involved in language are hierarchical in nature and that language moves from simple units of construction (Phonology and Morphology), and build layers of complexity and sophistication (Syntax and

Semantics). She found that people with AD generally had intact Phonology and Morphology but more impaired Syntax and Semantics. She stated that the language forms we learn last are the first to deteriorate. We generally learn language in small simple units initially and build syntax and complexity as we are more comfortable with language.

It is clear from both the clinical diagnostic criteria and supporting research that language is impacted in those with AD. However, one of the costs of analysing language is that there is a huge burden on trained practitioners, be it clinical psychologists, audio transcribers and text encoders to facilitate the process of collecting data and analysis. The field of machine learning and natural language processing has been suggested as a way to improve the accuracy and lessen the human cost of this research as well as provide new insights into the difficulties that AD suffer in terms of language decline [8].

2.1.1 Aims and Methodology

The purpose of this review is to look at the current standards of cognitive tests available that can be used as a diagnostic tool for MCI. I then go on to explore what techniques for assessing language have been used within the field of cognitive and clinical psychology. I then go on to look at what techniques have been developed in the field of machine learning (ML), deep learning (DL) and natural language processing (NLP) that might enable the automated analysis of language easier as well as an obstacles and/or limitations of current technology. Finally, I look at some studies which have already looked at the intersection of these two domains.

A search of the literature was conducted using ProQuest (PsychArticles), SCOPUS, Web of Science. The following results were found (Table 1). All papers were then reviewed for relevance by reading the abstract and full text where appropriate and a shortlist was compiled. An additional search through references of shortlisted papers was also conducted and any papers who upon further review appeared relevant were added to the shortlist. Papers were included where researchers used machine learning to classify participants as MCI, AD or Healthy using language. We excluded any papers that focused on other forms of dementia or cognitive impairment, as well as any papers in which the language being analysed was not English.

2.2 Existing Neuropsychological Measures of Cognitive Impairment, and Repeatable Battery of Neuropsychological Tests

In terms of standardised cognitive tests, there are two main aims. Does the test distinguish accurately between normal aging, MCI and AD (diagnostic utility) and, does the test distinguish between those individuals with MCI who will then go on to develop AD and those individuals with MCI who don't then go on to develop AD (prognostic utility). This section of the literature review

Database	Number of Results	Search Terms
ProQuest(PsychArticles)	1484 Results	Language AND Decline AND Dementia
ProQuest(PsychArticles)	486 Results	Language AND Decline AND Dementia AND Speech
ProQuest(PsychArticles)	159 Results	Machine Learning AND Dementia AND Language
Web of Science	1207 Results	Language AND Decline AND Dementia
Web of Science	151 Results	Language AND Decline AND Dementia AND Speech
Web of Science	34 Results	Machine Learning AND Dementia AND Language
Scopus	791 Results	Language AND Decline AND Dementia
Scopus	91 Results	Language AND Decline AND Dementia AND Speech
Scopus	29 Results	Machine Learning AND Dementia AND Langauge
Scopus	1292 Results	"Language Deficits" AND Dementia

Table 1: Search Terms and Number of Results.

outlines a number of different cognitive tests that have been used to measure cognitive impairments as well as their perfomance in terms of both diagnostic and prognostic utility.

2.2.1 Repeatable Battery for the Assessment of Neuropsychological Status

The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) was originally developed as an assessment tool for dementia, specfically looking at detecting a characterizing very mild dementia. The authors felt that there was a shortfall of appropriate measures that were sensitive enough to milder impairments as well as a number of other shortcomings of existing tests. They met a number of design goals for this new battery of tests that addressed these shortcomings. The RBANS consists of a number of sub-tests across five distinct domains.

1. Immediate Memory
2. Visuospatial / Constructional
3. Language
4. Attention
5. Delayed Memory

However, whilst the authors claim that the RBANS is adequately sensitive in person's with MCI, other research points out that the RBANS has poor sensitivity in detecting MCI. Another drawback of this battery is the lack of executive function measures and object naming tasks. Research has shown that the RBANS has good test-retest reliability and convergent validity, pa

2.2.2 Digit Span Test

The Digit Span Test is predominantly used to measure a person's working memory capacity, specifically the capacity used to store and recall numbers. A participant is presented with a series of numbers of fixed length and is asked to recall those numbers in normal or reverse order. The series of numbers gets progressively longer until such time as a participant fails three or more times out of eight presentations.

MCI patients had significantly lower digit span score, in both normal and reverse order versions of this test. Research has shown that Digits Backwards can, to some degree, predict a diagnosis of MCI. They found that Age, Gender and Education have an impact on the performance of the tests.

Found no differences between specifically amnestic MCI and controls, but could differentiate between Mixed MCI and the other groups with mixed MCI recalling fewer correct responses than other groups. Notably, there was an attenuated recency effect in those with mixed MCI. Kessels identifies that that there are working memory deficits in MCI patients and these worsen with AD patients. Kessels identifies that both MCI and AD have impaired performance on all three conditions of the digit span test. No differences were found between forward and backward conditions in any of the groups. However, available tests may not detect subtle impairments.

2.2.3 Rey Auditory-Verbal learning test

Rey's Auditory Verbal Learning Test (RAVLT+) looks a wide range of neuropsychological processess including short-term auditory-verbal memory, retention of information. Participants are given a list of 15 unrelated words, repeated over five different trials and are asked to repeat. Another list of 15 unrelated words are given and the client must again repeat the original list of 15 words and then again after 30 minutes. Several studies have shown that an impairment in RAVLT score reflect well the underlying pathology caused by AD. Thus making the RAVLT an effective early marker to detect AD in persons with memory complaints. Moradi investigated to what extent the RAVLT scores are predictable based on MRI data using machine learning approaches, as well as to find out what the most important brain regions are for the estimation of RAVLT scores. They found a highly significant cross validated correlation between the estimated and observed RAVLT immediate and RAVLT Percent Forgetting. Further, they found that the conversion of MCI subjects to AD in 3-years could be predicted based on either observed or estimated RAVLT scores with an accuract comparible to MRI-based biomarkers. Study found that RAVLT to best distguish patients suspected of Alzheimer's disease from the psychiatric comparison group.

2.2.4 Digit Symbol Substitution Test

The Digit Symbol Substitution Test (DSST) involves a key consisting of the numbers 1-9 and a corresponding unique symbol. Below this key is a series

of numbers from 1-9 in a randomized order and repeated multiple times. The participant is asked to fill in the corresponding symbol for each number. The task requires that the participant move between the key and the randomized sequence such that they may retrieve the correct answer from the key, hold this in short-term memory and transcribe the key in the appropriate place. Among those with no disorder in cognition, mobility and mood, being in the lowest DSST quartile compared to the highest was associated with nearly twice the odds of developing one or more clinical or subclinical disorders. Associations were stronger for incident clinical disorders in cognition. Slower psychomotor speed may serve as a biomarker of risk of clinical disorders, mobility and mood. While in part attributable to vascular brain disease, other potentially modifiable contributors may be present. Further studying the causes of psychomoto slowing with ageing might provide novel insights into age-related brain disorders. Pascoe compared patients with PD with Normal Cognition (PD-N) with those with PD and MCI (PD-MCI) and healthy participants. PD-MCI pariticipants achieved significantly lower scores than other groups in the DSST task.

2.2.5 Corsi Block Tapping Task

The original Corsi apparatus consists of a series of nine blocks arranged irregularly on a 9 x 11in board. The blocks are tapped by the adminstrator in a randomized sequence of increasing length. Immediately after each sequence is demonstrated, the task is to mirror the sequence back to the adminstrator with the sequences increasing in length until a mistake is made. The CBT performances were significantly less satisfactory in the multi-domain MCI category than in other groups (Healthy adults, Amnesic MCI and Non-amnesic MCI with executive function impairment).

2.2.6 Verbal Fluency

There are a number of tests that characterise this category of verbal fluency, but generally fall into two categories. Letter (Phonemic) fluency involves the generation of as many words as possible which begin with a specified letter. Category fluency involves the generation of as many words as possible that fall into a specfied category. Both these tasks impose demands on a number of different cognitive processes namely, executive function, verbal retrieval and recall, giving appropriate answers while monitoring previous answers and inhibit inappropriate responses. However, these two tasks require different strategies when attempting them. Letter fluency relies on search strategies based on lexical representations whereas category fluency requires a search for semantic extensions of a superordinate term, meaning that semantic associations within the lexicon must be intact in order for the task to be carried out successfully.

There have been numerous studies which have documented the impact that AD has on verbal fluency tests in both categories. A review carried out by Henry et al, found that performance in both letter fluency and category fluency was impaired in those with AD vs controls, but found a larger effect for tests of

semantic fluency.

2.2.7 Naming Tests

Word-finding difficulty is a common symptom of AD and these deficits usually occur during the early stages of the disease progression. As such , a test of a patients ability to find words (known as confrontation naming) is a common way to measure cognitive decline. One common way to do this is the Boston Naming Test (BNT) which comprises 60 items on a spectrum of very frequent to very infrequent. This has been reduced subsequently to two thirty item versions and four fifteen item versions, which correlate significantly with the original sixty item version and the benefit of the shorter versions of the test is that it facilitates testing of individuals with AD who may suffer from fatigue or limited attention span. Twenty aMCI patients, twenty AD and 21 controls matched by age, sex and education level were evaluated. AD patients obtained significantly lower total scores on the BNT than aMCI patients and controls. aMCI patients did not obtain significant differences in total scores but showed significantly higher semantic errors compared to controls. Semantic processing is impaired during confrontation naming in aMCI. Higher educational attainment in aMCI subjects were correlated with better performance in verbal and non-verbal tasks during repeated examinations over 1-year. Subjects with a lowere level of education performed worse than patients with a high level of education who presented a more stable clinical score. The explanation for this is the idea of a 'cognitive reserve' in participants with a higher education such that this provides a buffer that, while not preventing the physiological symptoms of AD, can potentially delay the clinical onset of cognitive symptoms that characterise AD. This theory is supported by Snowdon who found a relationship between early life linguistic ability and the density of neurofibrillary tangles in his nun study [7].

Whilst these studies and tests provide evidence that there are word finding difficulties in those with MCI, on their own they do not provide sufficient ability to diagnose MCI or provide an prognosis of disease progression.

2.2.8 Rey-Osterrieth Complex Figure Task

The Rey-Osterrieth Complex Figure (ROCF) is a task widely as a test of visuo-spatial skill and visual memory. The task, which was originally designed by Rey (1944) and standardised by Osterrieth (1944), asks a participant to copy a complex geometrical figure (known as the immediate copy condition and to recall and reproduce the figure from memory without warning (known as the delayed recall condition). In the immediate copy condition, the complexity of the figure requires an integrative cognitive ability. The reproduction of such a complex structure involves processes such as planning and organizational strategies that are related to executive functions. Patients with vascular MCI had a worse performance in the immediate copy of the ROCF compared to individuals with degenerative MCI, despite their significant impairment in terms of general cognitive status and visual memory. Evidence shows that patients with disorders

that possibly involve attention and executive functions are characterized by a more disorganised approach when copying the ROCF compared to controls.

One of the difficulties with the presentation of this task in a repeated battery will be the fact that it turns from an incidental memory task (it is a memory task but the participant is not forewarned that they will be required to memorise the picture) into an intentional memory task (given the previous exposure to the task, it would be expected that a participant would pre-empt the delayed recall portion of the test and spend more time attempting to memorise the details).

2.2.9 Hayling Sentence Completion Test

Inhibitory deficits are a common in all stages of dementia. This is usually tested using Stroop test, however this has the drawback of lacking ecological validity. Therefore researchers have moved towards using the Hayling Sentence Completion Test (HCST) which uses skills such as word retreval as well as the ability to inhibit ones responses where appropriate and is therefore a much more ecologically valid task. There are two parts to the HCST. In the first part, participants have to complete a sentence by providing a word that best fits the given sentence (this is known as the initiation condition). In this second part, participants have to complete sentences by inhibiting an impulse to give the word that best first the sentence as in the first part, and producing a semantically unconnected word. Performance in both conditions is measured by the time taken for the participant to initiate a response, and in the inhibition condition also by the correctness of the word.

Martyr et al compared healthy controls with patients with dementia and patients with parkinson's disease. They found that a high proportion of Category A errors (producing a word that fits the sentence when instructed otherwise) was a factor in performance loss for participants with dementia. Findings suggest that the HSCT may be sensitive to verbal suppression deficits and may provide insight into inhibitory control in participants with dementia. Patients with Dementia were significantly slower than controls in the initiation and inhibition conditions vs healthy controls, and slower than patients with parkinsons disease in the initiation but not the inhibition condition.

2.2.10 Serial Non-word Repetition

2.2.11 Grooved Pegboard Test

The Grooved Pegboard Test (GPT) was originally designed to cover a variety of different psycho-motor functions including hand-eye coordination and motor speed. However, some studies have shown a correlation of performance in the GPT and measures of cognitive performance such as the Montreal Cognitive Assessment.

2.2.12 Visual Search

In research, there have been shown to be deficits in both AD and MCI populations. Whilst these deficits were not as apparent in MCI populations, there remained a difference that could be used to differentiate healthy controls from those with MCI.

2.2.13 Continuous Picture Naming

2.2.14 Free Cued Selective Reminding Test

The Free Cued Selective Reminding Test (FCSRT) was borne out of the premise that by controlling the conditions of learning, a measure of memory is possible that is not confounded by normal age related changes in cognition. Theoretically speaking, any controlled learning test should be able to discriminate between cognitive decline due to age and cogntive decline due to a cognitive impairment. /newline /par FCSRT-IR (The instant recall version of the Free Cued Selective Reminding Test) performance has beeen shown to distinguish patients with MCI who then went on to develop AD, from those with MCI that did not then go on to develop AD and this led the authors to define prodromal or the amnestic syndrome of the medial temporal lobe by FSCRT-IR performance (Sarazin et al, 2007 /newline /par The sensitivity and specificity of the different test forms in correctly classifying patients were compared to each other using a previously established FCSRT-IR cut off score. /newline

2.2.15 Conclusion

I have looked at a number of different cognitive tests and a battery of tests that aim to have high diagnostic and prognostic capabilities in the MCI population. However, particularly with the RBANS, there lacks sufficient sensitivity in differentiating those with MCI from healthy elderly individuals such that this tool could be used with a level of confidence in the results. In regard to the tests, a common theme is a lack of studies and therefore evidence into the utility of these tests with this particular population. However if, as researcher, we aim to investigate this population further then a benchmark battery of tests which is sensitive enough in both diagnostic and/or prognostic utility should be a goal.

A criticism of the current literature is the lack of consistency with regard to the experimental groups. For example, some studies focus on differentiating between an MCI group, a AD group and a healthy controls group whereas other studies may further subdivide the MCI category according to a number of factors such as Amnesic or Non-Amnesic MCI, Mixed MCI (sometimes called Multi-domain MCI in the literature). Given the inconsistency in defining the experimental groups the confusing and often conflicting results that researchers produce is to be expected. Future research should look at the standardization of the operational definition of cognitive impairment in MCI may result in more consistent predictions of progression to AD.

2.3 Types of Language Assessment

One of the key debates when looking at how to analyse language is the type of task provided to elicit language production in participants. In the literature researchers have primarily focused on Picture Description tasks but have also suggested other ways in which we might collect data.

2.3.1 Picture Description Tasks

One of the most commonly used tasks to measure language is the Picture Description task. An example of this is part of the Boston Diagnostic Aphasia Examination (BDAE), called the Boston Cookie Theft picture description task [**?**]. In this task participants are asked to describe a picture presented to them in as much detail as possible. The picture itself depicts a familiar domestic scene and would not require participants to use any vocabulary beyond that learned in childhood. It was originally designed to assess Aphasia, but has shown itself to be useful in the assessment of language for the purposes of diagnosis of MCI and AD as well [9]

Figure 1: Cookie Theft Picture.

The picture description task does a fine job of eliciting descriptive language, however because of the limited content has limited use. The task in itself just a descriptive task, and therefore elicits a certain type of language. There is some disagreement as to the benefits of this using this methodology. This task is reported as being useful to lexico-semantic disorders [8, 10] as the language being generated is primarly nouns and deixis (words to identify items and words to put those items into context). However, Ash [11]felt that there was no difference in using this task vs Story Narration (described below). In explaining the differences, it is worth noting that these researchers were using differing variables

and this could explain their different perspectives.

In terms of Machine Learning research in this area, a number of researchers have used transcripts based on picture description tasks [12, 13, 14, **?**] and have successfully extracted linguistic features that could differentiate between AD and controls.

2.3.2 Narrative description task

The story narration task is designed to study a participant's ability to describe and elaborate on a story which is depicted using a series of pictures. The stories depicted are usually based on children's books or famous stories such as Cinderella. [15] This task requires ordering the story in a structured and coherent framework. It also requires comprehension and understanding of the stories characters and the events depicted, as well as an awareness of a character's goals and internal responses to given events. This task is particularly useful, as the procedure reduces the demands on memory and is therefore able rule out memory as a confounding variable for any results observed. As noted above, Ash [11] felt that this task was interchangable with the Picture Description task and because this task requires elaboration rather than simple description, is a sturdier test of lexical and semantic abilities as well as syntactic complexity. [16]

Given the relative strengths of the Narrative description task vs Picture description task, there are few pieces of research that have used Machine Learning to analyse features from Narrative picture tasks [15]. This could be due to the availability of data and the absence of any meaningful sets of transcripts of participants performing this task. However, this could be an interesting direction to take research in the future to see if features generated from this task could be used to predict MCI or AD.

2.3.3 Interviews

Interviews can also be used to elicit language in a more natural way by asking questions to guide a conversation between speakers. There are three types of interviews: unstructured, structured and semi-structured. Structured interviews tend to produce very limited speech and therefore has never been used in this area [8]. Unstructured interviews are open ended and generally do not conform to any particular pattern. They use generic themes such as family or hobbies to guide the conversation. Whilst this is the most ecologically valid form of conversation and therefore language generation, it's unstructured nature means that the protocol cannot be consistent and therefore reproduced. Semi-structured interviews are therefore preferred over other forms of interview as a middle ground. The semi structured nature of these interviews means that there is some replicability but does not constrain the participant in answering

questions.

The analysis of interviews can be difficult to analyse as both the content can vary even between participants. It is also difficult to measure as there are no pre-defined task goals in comparison to the other two methods. Nevertheless, this is the most naturalistic setting for looking at language production and can be used to look at the syntactic and semantic parts of language generation [10]. There have been some attempts to use interviews to assess language production in AD with promising results [17, 18] .

2.4 How do we analyse language, issues and debates

2.4.1 Single Word Language tasks vs Connected Language tasks

Part of the reason we need to pay attention to how we ask participants to generate data is understanding how we wish to analyse the data afterwards. As discussed above, the different methodologies to collect data generate different types of language. There are two main approaches which we have looked at to analyse language, using frequencies of words and combinations of words and measures of syntax and semantics. There are other less common methods of analysing language but these are beyond the scope of this review.

Single word tasks such as the Boston Naming Test and other such standardized language tests generally target a participants verbal fluency where this is defined as the ability to form and express words in accordance with certain criteria. For example, a common test of verbal fluency would be to ask a participant to name as many words in a given category in a limited time frame. These categories can be objects such as animals or sports, or can be words beginning with a specific letter. There are a number of benefits to using single word tasks. From a research methodology perspective, using a standardized test allows researchers to target a very specific process in language generation and isolate factors that impact performance in language well. However, this approach does not take into account the symbiotic nature of language processes.

Connected language tasks, such as the picture description task and interviews described above are a much more reflective of processes involved in natural language generation. They are able to be relatively constrained in the language available to be used, for example in the picture description task, or they can be unconstrained in the form of interviews.

2.4.2 Semantics vs Pragmatics

When navigating the English language it is necessary to distinguish between what a sentence says in both semantic and pragmatic terms. Semantic meaning

refers to the meaning of the words in a sentence local only to the given sentence. Another way to put this is, semantics considers the meaning of words without taking into account the context in which these words are spoken. Pragmatic meaning refers looks at the same sentence in terms of words and grammar but takes into account the situation or context in which these words are spoken. Emery in her literature review looks at all levels of language tasks except for pragmatics, however

2.5 Semantic Content

Another approach to linguistic analysis in this field is the idea of measuring semantic content and complexity. According to Emery (2000) [**?**] in which she states that Semantic and Syntactic skills deteriorate first in people with MCI and AD. If this is true, then psychological measures of semantic and syntactic skills should be able to pick up signs of deterioration and act as markers for possible MCI and AD. An example of a semantic complexity measure is the concept of idea density. Formally, idea density is defined as the average number of propositions per sentence [19] and this was used to successfully differentiate between people who would later go on to develop AD [7]. Some examples of semantic content measures are Type to Token Ratio, Brunet's Index and Honore's Statistic which are used to measure the lexical diversity of a given piece of text and/or utterance. This has also shown to be effective in differentiating between MCI, AD and Controls, with those with language impairments [20] and this has carried through in research involving machine learning [21, 22].

2.5.1 Picture-related content thematic elements

Several studies examined the amount of thematic elements expressed that were directly relevant to picture stimulus in picture description tasks. The studies used a variety of phrases to denote these thematic elements, these included 'pictorial themes', 'relevant observations and 'semantic units' but are notionally similar. The only difference was the number of thematic elements that 'scored' for the studies in question. Nicholas et al identified eight thematic elements of the Cookie Theft picture and used the number of elements as an outcome measure in different groups. He found that patients with AD expressed significantly fewer content elements than controls.

Hier, Hagenlocker and Shindler assessed content using a similar list of thematic elements. They divided their participants into early-stage and late-stage AD, as well as including a control group. The late-stage AD group produced significantly fewer relevant observations than the early stage group, and the AD group combined produced fewer relevant observations than controls. This study was replicated by Vuorinen et al (2001).

Smith, Chenery and Murdoch (1989) applied Hier's methodology for constructing pictorial 'themes' with the Picnic Scene from the Western Aphasia Battery (WAB) with a control and patients with moderate to moderately severe AD. The authors found no difference in the number of semantic elements produced but did not that the group with moderate to moderately severe AD took more time and more syllables to communicate these elements.

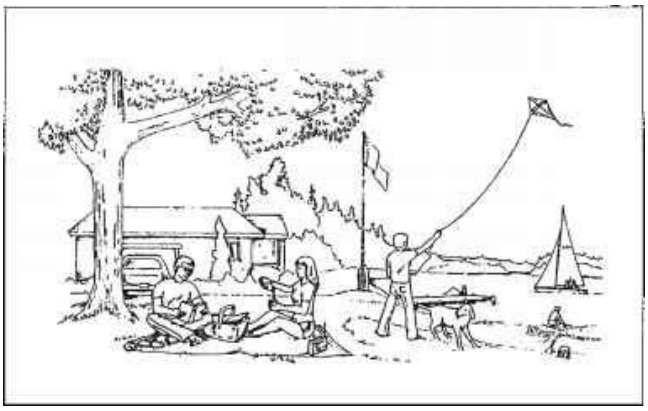

Figure 2: Picnic Scene taken from the Western Aphasia Battery (WAB).

Sajjadi et al (2012) examined 10 pictorial themes in picture description the Comprehensive Aphasia Test and found that the group with mild AD produced similar themes than controls. Bschor et al. (2001) examined Cookie Theft picture descriptions at four stages of AD. They identified 12 elements of the cookie theft picture and found that whilst each AD group differed significantly from the others and from controls, the measures did not distinguish between MCI and normal controls.

Finally, a number of studies used composite measures which contained thematic elements and other unspecified information units resulting in a list of 23 possible information units of the Cookie Theft picture. The authors felt that this provided a wider, more liberal range of relevant content and thus subtler differences could be noted. Studies using these features found some differences between AD and controls, and some could differentiate between different stages of AD.

2.5.2 General Information Units

Some studies used a more general concept of content, defining "information units" as "the smallest non redundant meaningful fact or inference," whether or not the information conveyed was specific to the context in which the conversation happened. Giles et al, for example, studied adults with minimal, mild or

moderate AD vs controls and found that adults with AD produced fewer overall information units than controls.

2.5.3 Conciseness of information

Conciseness has been defined as the number of words a speaker uses to express ideas. The theory is that people with AD will need more words to convey ideas because of word-finding related behaviours such as circumlocutions and repetitions. Conciseness is calculated by dividing the number of ideas expressed by the total number of words in a measure commonly referred to as idea density but also known as lexical index, information content and information unit conciseness index. Snowdon et al examined written discourse from the Nun study and found that low idea density in early life was associated with reduced cognitive performance in later life. Riley et al extended these findings by concluding that early-life idea density was associated with lower brain weight, higher degree of cerebral atrophy and increased neurofibrillary pathology in later life.

Ahmed, de Jager et al examined idea density with patients who had confirmed AD post mortem. They found that those with AD produced fewer total semantic units than controls but there was no significant difference between the groups with regards to idea density. The study of "empty speech" by Nicholas et al examined conciseness with measures thought to contribute to the "non-specifity" of discourse in AD, such as empty phrases (defined as common idioms contributing no relevant information), deictic terms (e.g. "this", "that" without referents), indefinite terms (e.g. "thing" or "stuff"), pronouns without proper noun antecedants, and repetitions. In their study they found that AD patients produced more of these behaviours than did controls.

2.5.4 Efficiency

Efficiency is the rate at which meaningful information is conveyed over time, calculated by dividing the total number of information units by the duration in seconds of the speech sample. Smith et al, 1989 found that 18 adults with moderately severe AD produced fewer content units per minute on average than controls, and attributed this different to increased circumlocutions and repetitions. Murray 2010, used an analogous measure which he called "performance deviations per minute", in which fillers, irrelevant words, revisions or false stars, vague or non-specific vocabulary and inaccurate output (e.g. paraphasias) were divided by the total number of minutes in the speech sample; this measure was lower for those with AD than those with depression, and also healthy controls. The authors suggested that discourse information measures may help disentangle the similarities in symptoms of early AD versus depression in older adults. Guinn (2012, 2015) [23, 18] found that 'Go-ahead utterances' - instances in dialogue in which a speaker provides responsees do not add anything in a conversation beyond a minimal response, were significantly more frequent in those

with AD than healthy controls..

2.5.5 Lexical richness and diversity

2.5.6 Type token ratio(TTR)

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).

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

2.5.7 Brunet's Index(W)

Brunet's Index (W) differentiates itself for 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.

2.5.8 Honore's Statistic (R)

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.

2.5.9 Quantity - Total number of words

Several studies report that adults with moderate AD produce fewer words than controls on picture description, however other studies found no differences in total words among groups of controls and patients with MCI or AD. Murray and Nicholas et al investigated normal controls, patients with AD and older adults with depression and found no group differences in total words. In contrast, Lira 2014 found that controls produced more total words than patients with AD but found no difference between mild and moderate groups.

2.6 Syntax and Morphology (Language Form)

Syntax can be defined as the rules that govern how words can be combined to form sentences, whilst Morphology is the system that governs the structure of words and the construction of word forms. Multiple studies of language decline in dementia included at least one measure of syntax or syntactic complexity. Common constructs included words per clause, grammatical form (measures of an appropriate use of syntactic conjections, tenses, conditionals, subordinate clauses and passive constructions), measures of phrase length and proportions of words in sentences. Some researchers have explored the use of formulaic language in those with dementia, the theory being that well practiced phrases are less effortful and therefore place low load on the cognitive abilities of those with AD. The general hypothesis motivating these studies is that either working memory limitations or semantic memory limitations in AD affect one's ability to use complex constructions.

2.6.1 N-grams

One of the first features discussed as a potential predictor of MCI or AD is the n-gram. An n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. For example, given the sequence of words "to be or not to be", this extract is said to contain six 1-gram sequences (to, be, or, not, to, be), five 2-gram sequences (to be, be or, or not, not to, to be), four 3-gram sequences(to be or, be or not, or not to, not to be) and so on. This is useful as, given a large portion of text or speech, we can predict the probability of a word being close by to a given word. A number of researchers have used n-grams as features. One of the first attempts to use machine learning and natural language techniques to look was conducted by Thomas [22] who was able to successfully demonstrate the ability of machine learning algorithms to analyse n-grams as well as other features to outperform a naive rule-based classifier which always selects the modal class. Orimaye et al (2017) [13] investigated the use of machine learning algorithms to detect differences primarily in n-gram use to distinguish between those with a diagnosis of AD and healthy controls. Their main finding supported n-grams as the most significant predictor. One of the criticisms is the use of picture description tasks and n-grams. Because the language generated by this task is content specific the n-grams generated are only specific to the task given and cannot be generalised.

Asgari, Kaye and Dodge (2017) [17] used another form of word frequency measurement. Using recordings of unstructured conversations (with standardized preselected topics across subjects) between interviewers and interviewees they grouped spoken words using Linguistic Inquiry and Word Count (LIWC) which is a technique used to categorize words into features such as negative and positive words [24]. They were able to successfully used machine learning

algorithms to distinguish between these two groups with an accuracy of 84%.

2.6.2 Mean length of utterance (MLU)

Murray found that MLU was not a distinguishing factor among health adults, adults with depression and adults with AD. Ripich et al found a decrease in MLU in adults with severe AD over time, and this was supported by findings of Le et al in their studies of authors [**?**]

2.6.3 Proportion of verbs to nouns plus verbs

Kave and Levy used a verb index to capture syntactic complexity and found that adults with AD expressed the same amount of verbs to nouns plus verbs as adult controls.

2.6.4 Syntactic Complexity - Composite measures of MLU, syntactic errors and verbs

Ahmed et al, and Ahmed, Haigh et al found differences in syntactic complexity between adults with MCI and controls, and between MCI and moderate AD stages. The differences in syntactic complexity were not significant when individual measures were tested, but were apparent using a composite score consisting of MLU, words in sentences, syntactic errors, nouns with determiners, and verbs with inflections.

2.6.5 Formulaic Language

Fraser, Meltzer and Rudzicz (2015) [**?**] looked at connected speech using the DementiaBank corpus. They found that there were four factors which informed the classification of participants as either healthy or AD. These four factors were semantic impairment, acoustic abnormality, syntactic impairment and information impairment and were based on existing measures of semantic and syntactic complexity. Zimmerer (2016) [12] looked at whether language was more formulaic in those suffering from AD. He proposed that those who suffer from AD rely on formulatic sentences, for example 'Noun-Verb-Noun', and this is done to reduce language complexity. He noticed a significant difference in the use of formulaic sentences between AD and Healthy Controls.

2.7 Pragmatic Language

The pragmatic language domain refers to the social rules for language for the purposes of communication including, using language to achieve goals, using

information from the context to achieve these goals and using the interaction between people to initiate, maintain and terminate conversations.

2.7.1 Questions, turn-taking, unsure statements, egocentric comments

One study, Ripich et al, examined several pragmatic abilities among patients at different stages of AD. The severe AD group asked more questions over time than the other group. The authors argued that question-asking was a compensatory strategy, and as a result, adults in late-stage AD may have had insight into their communication problems. Potentially, there were a number of flaws of the methodology if their research. Firstly, a caregiver was asked to administer the picture description task in order to mirror a more typical communicative interaction. Whilst this improved the ecological validity of the study, the caregivers delivery would be varied and inconsistent. Secondly, due to the constrained nature of the picture description task, it is unlikely that it would be able to capture the pragmatic skills that are typioical of conversations in everyday life.

2.7.2 Coherence

Coherence refers to the appropriate maintenance of topic in discourse and is thematically related to the immediately preceding utterance (local coherence) and by how closely an utterance relates to the general topic at hand (global coherence). One study looked at coherence, Chapman et al used picture descriptions of Norman Rockwell prints within a frame analysis, with frames being defined as internalized knowledge structures. The authors identified aspects of content, including typical frames of interpretation, atypical, incorrect or no frames, propositions supporting frames and propositions disrupting frames as measures of coherence. They examined these variables with early stage AD, old-elderly and normal controls. Old-elderly and normal controls produced significantly more typical frames and more frame supporting information that the AD group. The authors attributed AD patients' difficulties to memory deficits, attentional deficits, visual perceptual problems, disruption of internalized frame representation, or failure to access frame knowledge.

2.7.3 Perseveration

One study examined verbal preservation in the description of Norman Rockwell prints, dividing the total number of words within perseverations by total number of words in the speech sample. The authors also calculated rate of perseveration on two other language tasks: confrontation naming and generative naming. Across all tasks, the AD group produced significantly more perseverations than controls; however, on the picture description task alone, there were no significant differences between the two groups. The authors theorised that

this was because picture description was an easier taks because there was the visual stimulus, similar to the argument made by Bschor et al.

2.7.4 Speech intonation / prosody

There are a number of studies, that have examined "melodic line", which is a subjective measure of speech prosody defined as "the appropriate use of intonational contour, including alterations in pitch, volume and duration". For instance, Forbes-McKay compared melocdic line using a "simple" picuture such as the Cookie Theft picture vs a complex picture. The number of pictorial themes differentiated the simple tasks from the complex. Results showed no group differences on simple picture tasks (Cookie Theft and Tripping Woman) but there were differences in melodic line for the complex picture tasks. Fraser et al examined several acoustic features of speech in patients with AD ussing machine learning methods that captureed both rate and phonation patterns, and found acoustic abnormalities to be a significant factor. Konig et al, also used an automated feature analysis examining vocal and temporal aspects of speech among controls, and patients with MCI or AD, and reported a classification of 81%.

2.7.5 Discourse Fluency

Verbal fluency is a term used in neuropsychological contexts generally referring to timed, word-generation tasks, while in speech-language pathology contexts, "fluency disorders" are defined as interruptions int he flow of speaking characterized by atypical rate, rhythm and repetitions in sounds, syllables, words and phrases. "Fluency", in the literature of discourse of adults with AD, typically refers to the smoothness or flow of spoken language. Abnormalities of fluency in this population are typically characterised by filled and unfilled pauses, word repetitions, circumlocutions, and revisions. In contrast with fluency disorders (i.e. stuttering), abnormailities in the fluency of adults with MCI or AD are rarely manifested at the phonological level.

The study of "empty speech" by Nicholas et al was one of the first to examine aspects of fluency in the connected speech of persons with AD. They found that adults with AD had significantly more repetitions than controls. Similarly, Tomoeda et al found more aborted phrases, revisions and ideational repetitions in the AD group than in controls. Several other studies support the idea that in AD population there are a greater number of repetitions and revisions than in healthy controls. However, there are some studies which contradict these findings (Ahmed et al).

2.7.6 Speech Monitoring

Speech monitoring is related not only to word retrieval deficits and to pragmatic language skill but also to "anosognosia" define as the awareness of one's own deficits. McNamara et al investigated word error monitoring skills by comparing uncorrected and repaid errors in adults with AD vs health controls. The AD group was equally impaired in error monitoring as the controls. Severity of naming deficits correlated negatively to the amount of uncorrected errors. The authors suggested that this impairment was related to the executive function dificulties in the clinical groups. The authors did not report correlations between error monitoring and executive function test scores, however, which could have strengthened that hypothesis.

Another measure of speech monitoring used in the picture description literature is "response to word-finding delays," defined as "whether patients appear unaware of their problem, produce an approximation of the target word or actively search and produce the target word". Response to word-finding delays differed significantly between minimal AD and normal controls (forbes-McKay and Venneri). The measure was based on Goodglass and Kaplan's scale for rating discourse on the Boston Diagnostic Aphasia Examination (BDAE) and is composed of clinical judgement of behaviour that is rated on a likert-type scale ranging fdrom 1 to 7 (7 being no abnormality).

2.8 State of literature into Natural Language processing techniques

As we can see, diagnosing dementia through analysis has a large background in terms of psychological research. However, a lot of more current research has called for the use of machine learning as a way of assisting in the process of diagnosis [8]. Generally speaking, Natural Language Processing (NLP) systems mimic the semiotic perspective on language. That is to say that we can subdivide NLP processes into different components. This broadly fits with the literature review conducted by Emery.

1. Morphological/Lexical: providing the basic language elements or vocabulary, such as words, their roots and inflections.

2. Syntactic: for grouping and sequencing elementds within samples of language (usually sentences).

3. Semantic: for knowing the meaning of an utterance, usually defined in terms of the 'truth value' of the logical propositions that are thought to be expressed by sentences.

4. Pragmatic: for understanding the context and purpose of an utterance.

2.8.1 Traditional methods of Machine Learning and Natural Language Processing

2.8.2 Deep Learning

2.8.3 How can Machine Learning be applied to this problem

One of the drawbacks of attempting to use natural language processing and machine learning in this context is the lack of data.

2.9 The problems with attempting early diagnosis of dementia and the ethics of actively screening for it

It is challenging to characterise Dementia as a disease as it's more syndrome. The clinical features that combine to meet the diagnostic criteria for Dementia or it's variants are continuous in nature and also impacted by other variables. For example, cognitive performance is affected by previous education and ability to live independently is impacted by a person's physical health. Also, the individual who may develop symptoms that indicate the onset of dementia needs the insight to recognise them, or family members will need to do so and also deduce that there is value to admitting there is a problem and seeking help before the deterioration is too great.

To detect dementia early requires an understanding of how this presents early in it's progression, but as described above there are a number of potential ways in which individuals with dementia could be impacted but no consistent list of criteria has been produced. Even for AD, the most prevalent form of dementia, clinical diagnosis can only be confirmed after death through autopsy. In addition, population studies have shown that people over the age of 80 show similar changes in the brain in those that do not have dementia. Therefore, it's difficult to say with certainty that any symptoms can definitely cause dementia despite the fact that these symptoms are commonly present in those that have dementia.

2.10 Conclusions and Future Work

It appears that there are a number of approaches that have been shown to be effective in various experimental settings. The fact that there are a numerous array of different perspectives to tackling this question, and even within these perspectives there are a range of various methodologies leaves the field with a sense of confusion as to the most effective way to solving the problem.

Current state-of-the-art diagnostic measures of AD are invasive (CSF analysis), expensive (neuroimaging), and time-consuming (neuropsychological assessment). Furthermore, these measures are limited to specialty clinics and thus

have limited accessibility as high-throughput, or frontline, screening and diagnostic tools for AD. More importantly, nonspecialists are often inaccurate at identifying early AD and MCI [5]. Thus, there is an increasing need for additional noninvasive and/or cost-effective tools, allowing frontline identification of subjects in the preclinical or early clinical stages of AD who could be suitable for monitoring in specialty clinics and for early treatment. Implementation of effective screening instruments will allow diagnosis earlier in the course of dementia, even at the point when memory function is still essentially within the normal range. This strategy would enable an earlier, and potentially more effective, prevention and treatment of AD with a special focus to preserve cognitive functions

Identification of language impairment is key idea in the diagnosis of AD and early diagnosis can alter the prognosis and change the management of this degenerative disorder as well as provide new opportunities to study the progression of the disease. Given the burden on the diagnosis of AD on professionals there has been a call to use technology to potentially ease this burden [8]. It has already been shown that analysis of speech and language has shown markers that pre-date the official diagnosis of dementia [7, 1]. A significant amount of research has gone into the use of machine learning techniques to potentially look at the automated classification of participants with MCI and/or AD, however this is a new area of research and there are some gaps in our knowledge.

One of the areas for research to study is a careful examination of the features that are being used to measure language deterioration. For example, Zimmerer (2016) [12] describes connectivity in such as a way that correlates directly with what Mueller (2018) [14] calls Fluency. Whilst these are very nuanced measures which differ slightly in the form they take, the sheer range of measures and features being produced make it difficult to conceptualise as part of the larger problem. Some work needs to be done in producing a consistent list of measures that are validated using existing datasets and can be used for future research moving forward.

2.10.1 Future Work

Future research should also be directed towards developing non-intrusive ways of detecting subtle changes in language based in the home, such that any deterioration that could indicate the presence of MCI or AD and be flagged up early in the progression of any potential cogntive impairment for further review via referral. Machine learning approaches seem to be the most logical approach for achieving this aim. Further, despite the quality of datasets being used to 'backtest' these algorithms, further research should look at generating additional datasets to increase the validity of the results found so far as well as using other methods to generate data other than Picture Description tasks which we could claim are limited in scope.

The recent resurgence in the use of neural networks and deep learning is because of it's significantly improved performance on many problems and it's ability to scale from small to large datasets. Another key benefit of deep learning is it's ability to automate the process of feature engineering. Feature engineering, with this particular problem is an interesting debate as there are currently numerous ways in which to try to generate features as I have described above.

This area of research is extremely promising in its early results and the impact of successful research would be life changing for both individuals and the health of the worlds aging population in general.

As described above, there is a large gap in the knowledge of how, and it what circumstances does MCI progress to AD as well as how best to assess MCI over time and at this time, to the authors knowledge, there is no consensus in how to approach these two problems. Teng et al suggests that work should focus on the MCI population and concentrate on developing a consensus neuropsychological battery that could yield predictable rates of progression to AD. This, and with the development of a model of language, sensitive enough to detect subtle deterioration in language use to act as an additional cognitive marker to aid diganosis could potentially move some way to providing this.

3 Project Proposal

3.1 Experiment 1 - Pilot Study of Neuropsychological Tests

3.1.1 Methods

I recruited a cohort of 9 volunteers which which to complete the battery of Neuropsychological tests. We compiled a series of tests (see literature review for a comprehensive list, explanation and rationale for each test) and completed each version of the test on three consequetive days. The results of which were compiled and analysed. The hypothesis is that there would be no difference in performance between one battery of tests and another. An additional hypothesis is that the language tasks included on each task would be enough to collect an appropriate sample for linguistic analysis, such that we may be able to detect cognitive decline in a clinical population.

3.1.2 Participants

3.1.3 Results

3.1.4 Discussion

3.1.5 Conclusion

3.2 Experiment 2 - Initial Study of the Presidents corpus

3.2.1 Background

3.2.2 Methods

I took 46 transcripts of Ronald Reagan's (RR) press conferences from 1981 to 1988 and compared them with 134 press conferences by George H. W. Bush (GHWB) and 25 press conferences conducted by Donald J. Trump (DJT). I analyzed transcripts for lexical features shown to change longitudinally with dementia (for a comprehensive review of these, see the literature review above). For this collection of documents, I generated a number of features which looked at a number of different aspects of each document. 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. These findings were. number of unique words, non-specific nouns and fillers and low imageability (LI) verbs. Imageability is characterized, according to Berisha and Liss, as the ease with which a term gives rise to a sensory .mental image. I compared the trends described in the transcripts of RR and GHWB, but also included DT. 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. I 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.

3.2.3 Pre-processing

To generate the files necessary for analysis, I downloaded each transcript and performed the following changes. I omitted the prepared statement by the president and any speech by other individuals. I started each transcript at the beginning of the first answer to a question. I 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 hesitations were marked by a single hyphen whereas for GWB hesitations are marked by a double hyphen. In order to maintain consistency when parsing through the documents, I have changed both types of hesitation to be marked by a single hyphen. I also omitted one word sentences as this data would, from a theoretical perspective, not be relevant for language analysis. I did not control for the length of the document, but generated features which would normalise by the length of the document. I therefore was able to include all press conferences by both Ronald Reagan and George H.W. Bush where there was a question and answer session conducted at least in part by the sitting president (2 press conferences of GWB were ommited due to a lack of a question and answer session).

3.2.4 Results

3.2.5 Discussion

3.2.6 Conclusion

3.3 Chapter 3: Experiments with existing datasets: Presidential Press Conferences, Three Authors - Murdoch, Christie and James, DementiaBank dataset.

Aim: To replicate and extend the work by Berisha and Liss. To use traditional machine learning techniques to test whether they can differentiate between Reagan, Bush and Trump using similar features. To test to see whether Deep Learning can also differentiate between the three.
Dataset: the presidential press conferences given by Presidents Ronald Reagan, George H.W. Bush and Donald Trump.
Paper: This experiment replicates work done by Berisha and Liss and extends this by adding Donald Trump as an alternative comparison to Ronald Reagan. This experiment will look at the features originally recommended by Berisha and Liss, as well as any others that have potential as discussed in the literature review above.
Aim: To replicate the work by Le et al. To use traditional machine learning techniques to see if it can differentiate between the three authors using similar features. To test to see whether deep learning can also differentiate between the three. To add 'pre-trained' layers from Presidential Debates experiment to see if this improves accuracy.
Dataset is the literary novels of Iris Murdoch, Agatha Christie and P.D. James.
Paper: Aim: To replicate and extend the work of Orimaye et al. To use traditional machine learning techniques to see if it can categorise people into dementia or healthy categories using the samples given. To see whether can use pre-trained layer from presidential debates to improve accuracy. It doesn't seem relevant to use pretrained layer from Experiment four, but it might be worth looking at the results of this.
Dataset: The DementiaBank Corpus.
Paper: To replicate and extend the work of Orimaye et al. To compare tradi-

tional machine learning techniques (already completed by Orimaye) and deep learning.

3.4 Chapter 4: Validation of new repeatable neuropsychological battery for detecting MCI.

Aim:
Dataset: To be created.

3.5 Chapter 5: Longditunal Study

3.5.1 Dataset creation

The aim is to recruit a set of 50 participants with MCI / Early Dementia (diagnosed or undiagnosed) and a set of 50 controls with a similar age and broadly similar educational background (a known potential confound in assessing cognitive decline). The rationale for recruiting this number is that there will be an expected drop out rate and so Paper:

3.5.2 Ethics

Need to discuss confidentiality right at the start of each conversation and to reiterate at every meeting. I will also provide an information sheet which details the confidentiality statement. The spoken confidentiality statement will be as follows. "Anything you say to me today will be confidential unless I feel that there is a risk to you, or someone else. If this is the case, I may need to speak to or write to third parties such as your GP. Is this ok?"

3.5.3 What if someone reveals suicidal ideation or other risks?

As part of the battery of tests for this experiment, there is an opportunity for a participant to reveal that he or she has had suicidal ideation in the past two weeks (Patient Health Questionnaire, Question 9). This is something to explore further. As I have been trained in psychological risk assessments, it seems appropriate to carry out a further risk assessment. Optionally with the tape recorder turned on, or off as requested. Any text recorded during the risk assessment will not be included in the language analysis. Following this, some information about local services such as Improving Access to Psychological Therapies (IAPT), Samaritans and other appropriate services will be provided. A discussion will then be made about whether it is appropriate to notify the GP of the disclosure. In the event that there is imminent risk to self or others, a discussion will be made about who I need to contact. (Crisis Team and GP). In the event that the participant scores 0 on Q9 of the PHQ, no risk assessment will be undertaken.

3.6 Timeline of Future Work, Publication plan and Provisional Table of Contents for final Thesis

Year One
Year Two
Year Three

4 Future Work

4.1 Introduction

4.2 Project Plan

4.3 Risk Assessment

4.4 Ethical Considerations

4.5 Thesis - Provisional Table of Contents

In this section I provide a provisional table of contents based on the work I have either completed or proposed above.

1. Introduction
    (a) Overview
    (b) Background and Context
    (c) Aims and Objectives
    (d) Challenges and Achievements
2. Literature Review
    (a) Introduction
    (b) Assessing Mild Cognitive Impairment using cognitive tests.
    (c) Assessing Mild Cognitive Impairment using Language.
    (d) Natural Language Processing and Machine Learning
    (e) Conclusion
3. Further Exploration of the Presidents Corpus
    (a) Introduction
    (b) Methodology
    (c) Results
    (d) Discussion
    (e) Conclusion

4. Validation of a Repeatable Neuropsychological Battery of Tests
    (a) Introduction
    (b) Methodology
    (c) Pilot Study Results
    (d) Longitundinal Study results
    (e) Discussion
    (f) Conclusion
5. Longitudnal Study
    (a) Introduction
    (b) Methodology
    (c) Experiment 1 Results
    (d) Experiment 2 Results
    (e) Discussion
    (f) Conclusion
6. General Discussion
    (a) Discussion
    (b) Future Work
    (c) Conclusion

5 Conclusion

References

[1] Visar Berisha, Shuai Wang, Amy LaCross, and Julie Liss. Tracking Discourse Complexity Preceding Alzheimer's Disease Diagnosis: A Case Study Comparing the Press Conferences of Presidents Ronald Reagan and George Herbert Walker Bush. *Journal of Alzheimer's Disease*, 45(3):959–963, 2015.

[2] John Sturrock. Dementia UK - Second Edition. 2016.

[3] Martin Prince, Anders Wimo, M Guerchet, Gemma-Claire Ali, Yu-Tzu Wu, and Matthew Prina. World Alzheimer Report 2015 The Global Impact of Dementia. Technical report, 2015.

[4] Alois Alzheimer and Alzheimer Aware. The Need for Early Detection and Treatment in Alzheimer's Disease. *EBioMedicine*, 9:1–2, 2016.

[5] Peter J Nestor, Philip Scheltens, and John R Hodges. Advances in the early detection of Alzheimer's disease. *Nature Medicine*, 12(8):961–966, 2006.

[6] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[7] David A. Snowdon, Susan J. Kemper, James A. Mortimer, Lydia H. Greiner, David R. Wekstein, and William R. Markesbery. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *Journal of the American Medical Association*, 275(7):528–532, 1996.

[8] Veronica Boschi, Eleonora Catricalà, Monica Consonni, Cristiano Chesi, Andrea Moro, and Stefano F. Cappa. Connected speech in neurodegenerative language disorders: A review, 2017.

[9] Elaine Giles, Karalyn Patterson, and John R. Hodges. Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408, 1996.

[10] Seyed Ahmad Sajjadi, Karalyn Patterson, Michal Tomek, and Peter J. Nestor. Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, 26(6):847–866, 2012.

[11] S Ash, E Evans, J O'Shea, K Chandrasekaran, A Boller, L Burkholder, E Camp, D Weinberg, J Haley, K Kitain, and C McMillan. Differentiating primary progressive aphasias in connected speech production. *Dementia and Geriatric Cognitive Disorders*, 34:246, 2012.

[12] Vitor C. Zimmerer, Mark Wibrow, and Rosemary A. Varley. Formulaic Language in People with Probable Alzheimer's Disease: A Frequency-Based Approach. *Journal of Alzheimer's Disease*, 53(3):1145–1160, 2016.

[13] Sylvester O. Orimaye, Jojo S.M. Wong, Karen J. Golden, Chee P. Wong, and Ireneous N. Soyiri. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18(1):1–13, 2017.

[14] Kimberly D. Mueller, Rebecca L. Koscik, Bruce P. Hermann, Sterling C. Johnson, and Lyn S. Turkstra. Declines in connected language are associated with very early mild cognitive impairment: Results from the Wisconsin Registry for Alzheimer's Prevention. *Frontiers in Aging Neuroscience*, 9(JAN):1–14, 2018.

[15] Kathleen C. Fraser, Jed A. Meltzer, Naida L. Graham, Carol Leonard, Graeme Hirst, Sandra E. Black, and Elizabeth Rochon. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55(1):43–60, 2014.

[16] Juliana Onofre De Lira, Karin Zazo Ortiz, Aline Carvalho Campanha, Paulo Henrique Ferreira Bertolucci, and Thaís Soares Cianciarullo Minett.

Microlinguistic aspects of the oral narrative in patients with Alzheimer's disease. *International Psychogeriatrics*, 23(3):404–412, 2011.

[17] Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, 3(2):219–228, 2017.

[18] Curry Guinn, Ben Singer, and Anthony Habash. A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CICARE 2014: 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-Health, Proceedings*, pages 98–103, 2015.

[19] Walter Kintsch and Janice Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3):257–274, 1973.

[20] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, 14(1):71–91, 2000.

[21] Shuai Wang. *Automatic Tracking of Linguistic Changes for Monitoring Cognitive-Linguistic Health*. PhD thesis, 2016.

[22] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, 3(February 2005):1569–1574, 2005.

[23] Curry Guinn and Anthony Habash. Language Analysis of Speakers with Dementia of the Alzheimer's Type. Technical Report AAAI Technical Report FS-12-01, 2012.

[24] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The Development and Psychometric Properties of LIWC2015. *UT Faculty/Researcher Works*, 2015.

A Title of Appendix A

Text of Appendix A is Here

B Title of Appendix B

Text of Appendix B is Here

C Title of Appendix C

Text of Appendix C is Here.