Early Detection of MCI that progresses to dementia - An Interdisciplinary Approach

Jomar Alcantara

Department of Computer Science

School of Engineering and Applied Sciences

Aston University

Birmingham, United Kingdom

Early Detection of MCI that progresses to dementia - An Interdisciplinary Approach

Early Detection of MCI that progresses to dementia - An Interdisciplinary Approach

Jomar Alcantara
Department of Computer Science
School of Engineering and Applied Sciences
Aston University
Birmingham, United Kingdom

March 31, 2020

Abstract

My original contribution to knowledge is....

Acknowledgements

Please note: An editor has not been used in the construction of this thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

Alzheimer's Disease (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 [1]. 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 [2]. 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 [3]. 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) [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 [4].

We can characterise a person's progression through AD in three stages. Whilst these stages are called different things in the literature, we will stick to the following naming convention throughout this thesis

a) Stage 1: Preclinical AD

b) Stage 2: Mild Cognitive Impairment due to AD

c) Stage 3: Dementia due to AD

Research has started to be more focused on stage 2 however a problem with focusing here is there are no clear defined boundaries between these three stages.

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 [5], 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. The criteria for MCI is detailed below [5].

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 a 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 abilities - 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 dementia.

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.

1.2 Statement of the Problem

Studies that explore ways in which we can diagnose MCI / AD generally follow one of two main approaches, the analysis of biomarkers such as concentrations of amyloid-β 1-42 (Aβ42) in cerebrospinal fluid (CSF) and analysis of 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, that of analysing the cognitive abilities of patients in memory clinics, 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 comparatively 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. This has a number of benefits for countries with less developed healthcare systems, or where the burden of healthcare is more extreme.

One of the most common ways in which clinicians traditionally use the analysis of cognitive abilities to make an early diagnosis of dementia 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 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 has been 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) [7] or the Montreal Cognitive Assessment (MoCA) [8].

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. However, the utility of diagnosing dementia at the point where clinical intervention is warranted is limited because at this stage both psychological and pharmacological interventions have been shown to not be effective [2, 4]. In order to further our understanding of the progression of dementia it is important to detect the signs

of dementia before they are clinically apparent.

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 [9]. The case of Iris Murdoch, who had a confirmed diagnosis of dementia, illustrates this theory well. Le et al [10] 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 when, it is theorized, the symptoms of AD manifested themselves more significantly. 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.

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 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 for differences over time may be confounded by improved performance at tasks via practice effects. 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.

McKhann and his team were tasked with developing diagnostic guidelines for dementia in such as way that the criteria were flexible enough to be used by general healthcare providers without access to specialist medical equipment.the DSM 5 [3]. They state those suffering with mild dementia generally encounter impaired language functions (speaking, reading, writing)—symptoms include: difficulty thinking of common words while speaking, hesitations ;speech, spelling, and writing errors. More specifically, 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 who examined speeches and public interviews of former US president Ronald Reagan [11]. They found that Reagan's speeches towards the end of his presidency suffered from difficulties in word-finding, in-

appropriate 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 looked at whether linguistic ability in early life was associated with cognitive function and AD in later life [12] . 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 [11] 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 [13] 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 asserted that the language forms we learn last are the first to deteriorate as we generally learn language in small simple units initially and build syntax and complexity as we are more comfortable with language. However it is important to note

that different variants of dementia show different deficits in terms of language productions. Regardless, dividing language in this way is useful as it allows us to detect deteriorations in different parts of language usage and therefore may provide a way of discriminating between different forms of dementia.

It is clear from both the clinical diagnostic criteria and supporting research that language is impacted in those with AD. However, whilst there is a move towards research aimed at looking more specifically at MCI we currently lack the measures that are sensitive enough to detect MCI. Given that we know language is affected before a clinical diagnosis of dementia is usually made [11, 12, 10], it makes sense to explore whether language on it's own can provide markers that may indicate a cognitive impairment that could progress to dementia. 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 [14]. I theorize that we may be able to enable an earlier diagnosis of those with MCI and AD using samples of spontaneous speech, natural language processing (NLP) and machine learning (ML).

There is a large body of research that looks at the decline in language in those with MCI and AD [15, 14]. However there is conflicting evidence in these studies about which declining language factors are associated of MCI and AD [15, 14]. 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. Another area of focus for research of this nature is the process of collecting appropriate language samples. Whilst collecting samples of language is comparatively unintrusive, researchers recognise that these samples require a rich sample of language that potentially cannot be generated by tasks such as the picture description task.

Therefore, it would be useful to explore 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. As is the case with other quantitative measures of cognitive ability, contrasting individuals with MCI and Early AD is challenging due to the variation in an individual's baseline speech capacity. It is therefore prudent to measure cognitive decline over time as suggested by Albert et al [5]. There is some evidence to support this approach, Berisha et al [11] demonstrated 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.

1.3 Research Questions

The overall objective of the research described in this thesis is to investigate the deterioration of language in people diagnosed with MCI and the early stages of AD. In order to do this I aim to answer the following research questions

1. What research has been conducted in this field so far?

2. How can we use technology to best enable the collection of language samples?

3. Can we develop a complete data processing pipeline that processes data efficiently

4. What can machine learning tell us about how language affects us?

This thesis' original contributions are an increased understanding of how language deteriorates over time in those with MCI and early AD, the exploration

via a delphi method of ideas for the collection of language features via smartphone or other internet of things devices and the development and execution of a pilot study to test the feasibility of collecting high quality language samples via in-home technology. I also contribute to the research into which language features are important in the diagnosis of MCI and early AD over time. Finally, I explore the potential of one class classification as a new approach to classifying those with MCI and Early AD vs controls. The potential impact of this research in this area is immense.

1.4 Structure of thesis

There is a lot research in this domain from a psychological perspective. In chapter 2, I look at the background and related work in the domain of language deterioration in those with MCI and AD from the psychological perspective. This includes looking at the current ways in which psychologists collect and analyse language samples and exploration of the features of language that psychologists feel are important in this area. In addition, I take a systematic look at the work of the experts in the fields of machine learning and natural language processing in the growing area of research for NLP and Machine Learning and it's application to this problem.

The start of my research aims to answer questions around how we can use technology to aid us in collecting language samples that may facilitate the process of diagnosis. Currently, language samples are collected in memory clinics and physician's offices and is quite a time consuming process. More recently, work has been done to automate this process via a web based app called Talk2Me [16], which seeks to use different psychological tests to assess cognitive decline through analysis of language. There are some potential flaws in this method of data collection. In other areas, the use of smartphone technology has allowed

the collection of language for the diagnosis of disorders such as Parkinson's disease and so it appears plausible to be able to use this technology and potentially other technologies such as Amazon Echo to facilitate the collection of language samples. The work described in chapter 3 therefore describes an exploration of experts opinions' on the following questions.

1. How should we collect these language samples - Smartphones? IoT Devices?

2. What tasks should we use to elicit these language samples

3. At what frequency should we collect these language samples at

An adapted Delphi survey methodology is used to develop consensus on the answers to these questions and the results of these are documented in this chapter. The resulting pilot study is described in Chapter 6.

Once we have a way to collect language samples, the question to consider is how we may best analyse this data. There are a number of factors that one must consider when answering the question. We look at the effectiveness of automatic speech recognition at translating speech into language, as well as the minimum quality of data that needs to be passed to any pipepline we develop. The Talk2Me app collects and then generates approximately 2000 lexico-syntactic, acoustic and semantic features [**?**]omeili2019 as part of their data collection framework. Chapter 4 discusses the challenges connected to constructing an effective data pipeline and contains two experiments.

Whilst Chapter 4 is more theoretical in it's content, Chapter 5 documents the first experiments that use this pipeline on three existing datasets. These datasets are, the Presidents Press Conferences, The Novels of three authors and the DementiaBank dataset.

In chapter 6 Pilot study of the methodology developed

Finally, in chapter 7. I discuss conduct a general discussion of the results of my research. I also think about the strengths and weaknesses of my work and suggest ways in which the research have been improved. Lastly, I look at a number of areas for future work which can build on this research.

2. BACKGROUND AND RELATED WORK, SYSTEMATIC REVIEW OF NLP AND MACHINE LEARNING RESEARCH

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

2.1.1 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, specifically 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
    (a) List Learning
    (b) Story Memory
2. Visuospatial / Constructional
    (a) Figure Copy
    (b) Line Orientation
3. Language
    (a) Picture Naming
    (b) Semantic Fluency
4. Attention
    (a) Digit Span
    (b) Digit Coding
5. Delayed Memory
    (a) List Recall
    (b) List Recognitions
    (c) Story Memory
    (d) Figure Recall

However, whilst the authors claim that the RBANS is adequately sensitive in person's with MCI [**?**] and there is some research to support this assertion [**?**], 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. However, Duff et al warned that caution should be exercised when using the RBANS in a MCI population as it has lower sensitivity in this population [**?**].

2.1.2 Digit Span Test

The Digit Span Test (DST) 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.

Research has consistently shown that MCI patients have a significantly lower digit span score in both normal and reverse order versions of this test and a study by Muangpaisan has shown that the reverse order version of this test can, to some degree, predict a diagnosis of MCI [**?**]. Muangpaisan also found that Age, Gender and Education have an impact on the performance of the tests. Emrani et al found that the DST revealed 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. [**?**] He also identified 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 [**?**].

2.1.3 Rey Auditory-Verbal learning test

Rey's Auditory Verbal Learning Test (RAVLT+) looks a wide range of neuropsychological processes including short-term auditory-verbal memory, learning and 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 repeat it 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 accuracy comparible to MRI-based biomarkers [**?**]. Another study by Schoenberg found that RAVLT to best distinguish patients suspected of Alzheimer's disease from the psychiatric comparison group [**?**].

2.1.4 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 psychomotor 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 participants achieved significantly lower scores than other groups in the DSST task [**?**].

2.1.5 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.

2.1.6 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 [**?**].

Willers et al [**?**] studied Twenty aMCI patients, twenty AD and 21 controls matched by age, sex and education level. They found 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.

Vadikolias investigated the impact of education on naming tests [**?**] and found that 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 lower 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 [12]. 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.

2.1.7 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. In order to make this task repeatable, Taylor et al designed a comparable set of figures which have proven to be of equivalent difficulty [**?**, **?**].

Salvadori found that 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 preempt the delayed recall portion of the test and spend more time attempting to memorise the details) [**?**].

2.1.8 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 retrieval 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 Parkinson's disease in the initiation but not the inhibition condition.

2.1.9 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.

Bezdicek found that the GPT predicted performance on the MoCA and concluded that in addition to being a measure of motor skills, there were results that showed that the GPT could also provide information about a participants cognitive skills [**?**]. Whilst this was specifically using a population of patients with Parkinson's disease, given the MoCA and GPT have shown degree of correlation, there remains an opportunity for research that explores the use of the GPT with MCI and Early AD populations.

Darweesh et al, used the Purdue Pegboard Test (PPT) to assess manual dexterity in a healthy older adult population and followed their participants for between eight and twelve years until an indication of the onset of a neurodegenerative disease [**?**]. In this time 227 (4%) of their participants went on to develop a diagnosis of dementia. They noted that higher PPT scores were associated with lower risk of incident neurodegnerative disease and noted significant associations of PPT scores with all forms of dementia and this potentially highlights a deterioration of motor function in the pre-clinical phase of dementia.

2.1.10 Visual Search

The visual search task requires participants to be watching a computerised display in which a target appears either by itself or surrounded by other elements

which are used as a distraction. These other elements can be vary in similarity to the target. This is inherently a measure of attention shifting and processing of various similarity and as one would expect, the visual search time is increased in the presence of distractors. This effect is magnified in participants who have a cognitive impairment.

Research has shown there to be deficits in both AD and MCI populations in this task [**?**], and those with MCI exhibited less severe deficits compabut that these deficits were not as apparent in MCI populations, there remained a significant enough difference that could be used to differentiate healthy controls from those with MCI. In a comparison of visual search performance between those with aMCI and healthy controls, Tales noted significantly poorer performance in the aMCI group. However, she also noted a good deal of heterogeneity in the aMCI group which illustrates that whilst the aMCI group have essentially the same condition, presentations within this population can differ markedly. The results from this study also illustrates the use of a non-memory task as a means of diagnosing dementia as patients who went on to be diagnosed with dementia whilst this study was in progress had significantly poorer visual search performance.

2.1.11 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 cognitive decline due to a cognitive impairment. The FCSRT starts with a learning phase in which participants are asked to look at a card containing four pictures (e.g., grapes) for an item that belongs to a named category(e.g., fruit). After these four items are identified, immediate

cued recall of these four items is tested and this is repeated for a total of 16 items. Following this learning phase is the recall phase in which participants are asked to recall all the items identified without cues and any items which are not recalled are then cued. There are three outcome measures for this test, free recall (the number of items recalled without cues), total recall (the number of items recalled with or without cues) and cue efficiency which is identified as follows [7].

$$CueEfficiency = (totalRecall - freeRecall)/48 - freeRecall, range 0.0 - 1.0) \tag{2.1}$$

Grober found that patients with impaired free recall with four times more likely to develop dementia than those with intact free recall [7]. This test has also been 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 researchers to used the test to categorise prodromal or the amnestic syndrome of the medial temporal lobe by FSCRT score [**?**].

2.1.12 Conclusions

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 a clinical setting. In regard to the tests, a common theme is a lack of studies and therefore evidence into the utility of these tests with an MCI 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.

Another 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.

Finally, it can be said that in testing for MCI and comparing these results with a cohort of those with a similar diagnosis has limited utility. This is because there is huge variability in the presentation of MCI within participants that even if controlled for with a matched pairs design for age, gender and education will produce inconsistent results. There is an argument therefore for the use of longitudinal studies, where the comparisons on the performance of these tasks are with the participants themselves. There are not many studies which use a longitudinal approach, although those that do show promising results.

2.2 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.

2.2.1 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 [**?**]. The Cookie Theft picture (pictured below) depicts a scene of a home typical of the period of time when it was created and would generally not require participants to use any complicated vocabulary to describe. In this task participants are asked to describe the picture presented to them in as much detail as possible. This task 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 [**?**]

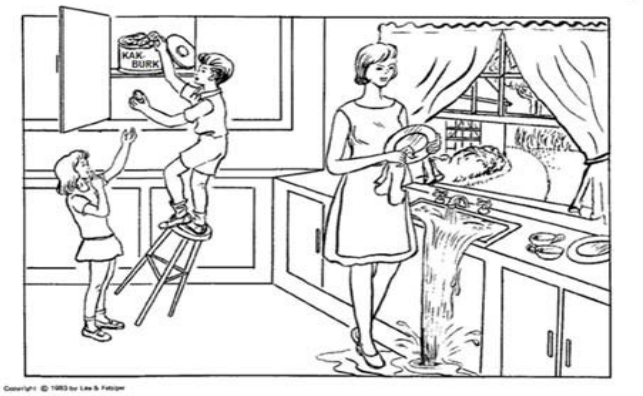

Fig. 2.1: Cookie Theft Picture - From Kaplan and Goodglass (1983)

The picture description task does a fine job of eliciting descriptive language however because of the specific content the language produce could be considered quite limited. There is some disagreement as to the benefits of this using this methodology. This task is reported as being useful to lexico-semantic disorders [14, **?**] as the language being generated is primarly nouns and deixis (words to identify items and words to put those items into context). However, Ash [**?**]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.

2.2.2 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 with the Cinderella being the one most typically used [**?**]. 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 actions, motivations and internal reactions to given events. This task is particularly useful as the procedure reduces the demands on memory, due to the participant being able to access the picture book during the description and is therefore able rule out memory as a confounding variable for any results observed. As noted above, Ash [**?**] felt that this task was interchangable with the Picture Description task. However, other research felt that this was a studier test of lexical and semantic abilities as well as syntactic complexity because this task requires interpretation and elaboration in additional to a simple description [**?**].

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 [**?**]. 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.

2.2.3 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 [14]. 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, although it can be argued that content should not affect the type of language being generated unless it is narrow topic or the participant is constrained in how they answer a given question. 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 [**?**]. There have been some attempts to use interviews to assess language production in AD with promising results [**?**, **?**].

2.2.4 Conclusions

One can view the different types of tasks above as a continuum where picture tasks represent a much more controllable task with a lot of supporting research but which generates a much more constrained set of language that is atypical of normal speech in terms of the cognitive functions used.

2.3 How do we analyse language, issues and debates

There are a wide variety of approaches that we can take in the analysis of language and there is no real consensus on the best approach to this particular problem. This section looks at the different ways in which researchers have looked at the problem as well as discussing some of the areas of contention.

2.3.1 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 word production where this is defined as the ability to form and express words in accordance with certain criteria (see above for a full description of verbal fluency as task). 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 other cognitive processes that are used in lengthier speech tasks. More specifically, single word language tasks do not look at the interaction between language, executive functions and reasoning abilities. They also do not require the logical and efficient organisation of ideas.

Connected language tasks, such as the picture description task and inter-

views described above are much more reflective of the 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. But regardless of how they are framed, they are potentially much more useful in terms looking at all the processes involved in language generation with the drawback that you will not be able to isolate specific parts of the language generation process. A review by Boschi et al concludes that analysis of connected speech is potentially useful in guiding clinicians to identify language disorders [14] and also highlights the role of NLP and Machine learning in assisting in this endeavour.

2.3.2 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 this is an area that should not be ignored. Whilst the study of language from a pragmatic perspective is much more complex, it is perhaps one of the most vital areas to study because of the number of different cognitive processes involved.

2.3.3 Semantic Content

Another approach to linguistic analysis in this field is the idea of measuring semantic content and complexity. According to Emery (2000) [13] 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 [**?**] and this was used to successfully differentiate between people who would later go on to develop AD [12]. An example of semantic content measures is Type to Token Ratio (described below) which is 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 [**?**] and this has carried through in research involving machine learning [**?**, **?**].

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).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{2.2}$$

2.3.4 Thematic and Content elements in relation to the Picture description task

A number of studies looked at the accuracy of picture descriptions using counts of thematic and content elements within a picture. Whilst called diffferent names such as 'pictorial themes', 'relevant observations and 'semantic units', they all represented the same idea. The only difference between the studies was the number of thematic elements that 'scored' correctly. 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 Lukatela et al [**?**].

Smith, Chenery and Murdoch 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.

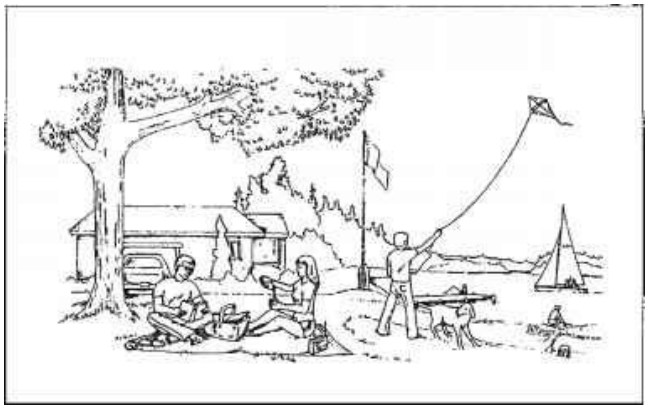

Fig. 2.2: Picnic Scene taken from the Western Aphasia Battery (WAB).

Sajjadi et al 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 found that whilst each AD group differed significantly from the others and also from controls, the measures were not able to 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.

2.3.5 General Information Units or Content Information Units

Some studies used a more general concept of content, using terms such as "general information units" or "content information units" and this could be defined as "the smallest non redundant meaningful fact or inference," and was counted 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 [**?**].

2.3.6 Conciseness of information

1‘ Conciseness has been defined as the number of words a speaker uses to express ideas. The theory is that people with AD would need more words to convey ideas because of difficulties with word-finding and compensatory behaviours such as circumlocutions and repetitions. Conciseness has previously been 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 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 [12]. 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 and specifically looked at empty phrases (defined as common utterances which contribute 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 than did controls.

2.3.7 Efficiency

Efficiency is the rate at which meaningful information is conveyed over time, and can be 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 AD produced fewer content units over time on average than controls, he attributed these differences to increased circumlocutions and repetitions in the AD group [**?**]. Murray used an similar measure in which fillers, irrelevant words, revisions or false stars, vague or non-specific vocabulary and inaccurate output were group together as 'performance deviations" and were divided by the total number of minutes in the 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) [**?**, **?**] found that 'Go-ahead utterances' - instances in dialogue in which a speaker provides responses do not add anything in a conversation beyond a minimal response, were significantly more frequent in those with AD than healthy controls.

2.3.8 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 et al found that controls produced more total words than patients with AD but found no difference between mild and moderate groups [**?**].

2.4 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.

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.

Asgari, Kaye and Dodge (2017) [**?**] 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 [**?**]. They were able to successfully used machine learning algorithms to distinguish between these two groups with an accuracy of 84%.

2.4.1 Formulaic Language

Fraser, Meltzer and Rudzicz (2015) [**?**] looked at connected speech using the DementiaBank corpus. They found that there were four factors which they identified as important in 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) [**?**] 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.

2.5 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.

2.5.1 Coherence

Coherence, in lay terms, can be defined as the ability to maintain awareness of the topic at hand. It can be separated into local coherence which is related to themes of the immediately preceding utterance and global coherence which looks at how closely an utterance is related to the topic currently being discussed. Chapman et al used picture descriptions of Norman Rockwell prints within a frame analysis, with frames being defined as the context in which the picture is viewed [**?**]. The authors identified aspects of content, including whether the frames of interpretation being offered were typical, atypical, incorrect or had no frame. They also looked propositions supporting frames and propositions disrupting frames as measures of coherence. They examined these variables with early stage AD, old-elderly and normal controls. Healthy older adults and normal controls produced significantly more typical frames and more frame supporting information than 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.

2.5.2 Perseveration

Perseveration can be broadly defined as the repetition of a response regardless of the absence or cessation of stimulus which could have generated an appropriate response in the first place. A typical example could be the idea of conversation moving from introductions to a more general conversation, someone who has difficulties with perseveration will struggle with the move between social contexts and repeat language that would be more used in an introduction context.

One study examined verbal preservation in the description of Norman Rockwell prints [**?**]. The presented participants with a number of similar pictures that had significantly different contexts and asked participants to describe these prints. They divided 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. In all tasks, the AD group produced significantly more perseverations than controls but there were no significant differences between the two groups in the picture task in isolation. The authors felt that this was because picture description was an easier task, as it was a visual task in contrast to the other tasks which tapped into other cognitive processes.

2.5.3 Empty Speech

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 in the 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.

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, Bayles and Tomoeda 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[**?**].

2.5.4 Conclusions

This section has looked at a number of different attributes which researchers have cited as being important in the detection of MCI and AD. Whilst in some cases there is a broad consensus about a given feature in the vast majority of cases there is some uncertainty. Some of the difficulty lies in different approaches used particularly around the criteria for experimental groups which means that it is difficult to be able to compare studies, like for like, as the populations of the participants being studied vary in subtle or obvious ways. It is important for any studies moving forward to use comparable standards when it comes to their experimental groups, such that the picture can be made clearer.

2.6 Argument for one class classification

2.7 State of literature into Machine Learning and Natural Language processing techniques

Diagnosing dementia through language analysis has a large background in terms of psychological research. Increasingly current research has called for the use of machine learning as a way of assisting in the process of diagnosis [**?**, 14]. This section will look at what areas of natural language processing (NLP) and machine learning (ML) could potentially be used as tools to help in this domain as well as any research that has applied machine learning to this problem.

2.7.1 Natural Language Processing

Natural Language Processing as a topic can be defined as the intersection between Machine Learning and Linguistics and looks at a number of language tasks, particularly important in this domain is how to enable a computer to process and analyse large amounts of language data. Some other tasks include speech recognition, natural language understanding and natural language generation. 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 (see Figure 2.3).

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

2. Syntactic: for grouping and sequencing elements 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.

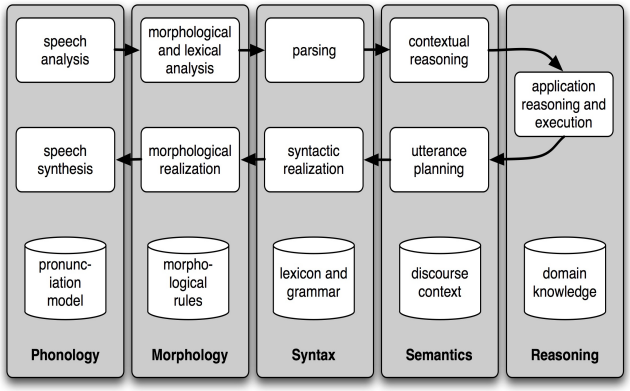

Fig. 2.3: How Natural Language Processing tasks are subdivided.

In order to facilitate these and other tasks, a number of frameworks have been developed that automate some of the more common processing tasks. A brief review of these frameworks follows.

Natural Language Toolkit

The Natural Language Toolkit (NLTK) was originally designed as part of a computational linguistics course at the University of Pennsylvania but has evolved to become an open-source framework which includes methods and modules for a wide array of Natural Language tasks (see Table 2.2 for a full list) [**?**]. The NLTK has been successfully used in both teaching and research purposes and as part of this provides access to large array of databases and corpora which have potential uses in common tasks.

Language processing task	**Functionality**
Accessing corpora	standardized interfaces to corpora and lexicons
String processing	tokenizers, sentence tokenizers, stemmers
Collocation discovery	t-test, chi-squared, point-wise mutual information
Part-of-Speech tagging	n-gram, backoff, Brill, HMM, TnT
Machine Learning	decision tree, maximum entropy, naive Bayes, EM, k-means
Chunking	regular expression, n-gram, named-entity
Parsing	chart, feature-based, unification, probabilistic, dependency
Semantic interpretation	lambda calculus, first-order logic, model checking
Evaluation metrics	precision, recall, agreement coefficients
Probability and estimation	frequency distributions, smoothed probability distributions
Applications	Graphical concordance, parsers, WordNet browser, chatbots
Linguistic Framework	manipulate data in SIL Toolbox format

Tab. 2.1: NLTK tasks and functionality

CoreNLP

Stanford CoreNLP provides a set of human language technology tools. This comes as an installable package that requires JAVA and can be interacted with through the command line or by an accompanying API. Whilst NLTK is much more oriented to parsing text and generating features, CoreNLP does this and adds more functionality such as a Named Entity Recogniser and sentiment analysis functions. At it's centre is the Stanford parser, which parses text and can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc.

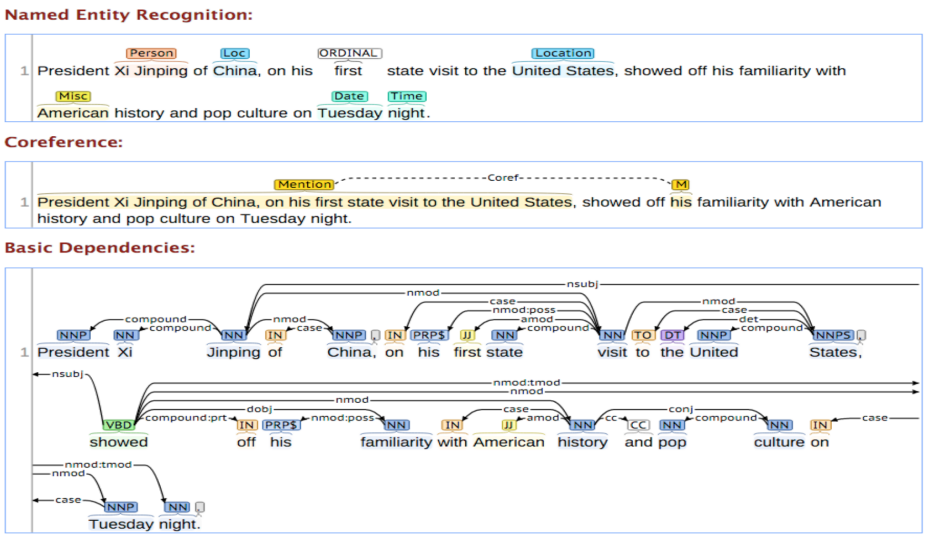

Fig. 2.4: Depiction of how Core NLP marks up a sentence

It's intended for CoreNLP to provide tools and frameworks that will enable

higher level language analysis whilst remaining domain agnostic.

Linguistic Inquiry and Word Count

The Linguistic Inquiry and Word Count (LIWC) is a piece of software that counts words and assigns these counts to categories [**?**]. The rationale behind this is that the words we speak inform an observer of the state of mind of an individual. A typical example of this might be someone with depression who may typically be more inwardly focused and therefore talking in the first person is common[**?**]. The 2015 version of the LIWC has over 90 different categories [**?**] and some supporting research suggests that it is effective at identifying themes and associating these themes with mental health difficulties such as depression and anxiety [**?**].

Whilst not necessarily a tool that can parse data and look at trends from a linguistic perspective, a framework like this could be used as another tool to explore themes within a persons speech from a more qualitative angle. To the authors knowledge, there is no research that looks at speech changes from this perspective.

2.7.2 Traditional methods of Machine Learning

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

One of the first attempts to use machine learning and natural language techniques to look was conducted by Thomas [**?**] 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 most frequent class. They detail several lexical approaches to the

problem of detecting and rating AD. The approaches they looked at relied primarily on character n-gram techniques but also explore correlation of usage of frequency of different types of speech. Their results act as a proof in concept of the utility of using a pure computational approach to the diagnosis of dementia using spontaneous speech. The were able to obtain 95% accuracy dementia vs controls, 70% accuracy classifying dementia into two categories and 50% accuracy classifying dementia into four categories. They suggest further exploration of characteristics to classify [**?**].

Bucks et al [**?**] took a more traditional approach using frequency rates per 100 words of various word types such as nouns, verbs, adjectives and pronouns and some measures of lexical richness to attempt to discriminate between healthy older participants and participants with probable AD. They found that AD patients had higher mean pronoun rate, adjective rate and verb rate but lower noun rate vs normal older controls [**?**]. It's important to note that the clinical group averaged a score on the MMSE of 15 (sd=6.8) which puts the clinical group in the 'moderately impaired range'. At this stage, the two groups should be reasonably easy to differentiate in terms of language and other cognitive abilities.

Fraser et al [**?**] built upon the work by Bucks by adding an acoustic analysis element to her analysis using the DementiaBank corpus, as well as significantly expanding the list of linguistic features they used. They used features generated from the Stanford parser (Now known as Core NLP) and computed the frequency of the occurrence of different parts of speech which they then normalised by the total number of words in each utterance. They also computed ratios between these parts of speech such as pronouns to nouns. They also included measures of syntactic complexity such as mean length of sentences and measures of lexical richness such as TTR. Their analysis showed was able

to achieve 81% accuracy in distinguishing individuals with AD from controls based on short samples of speech. They found that there are four factors which they felt were important in discriminating between these two groups: semantic impairment, acoustic abnormality, syntactic impairment and information impairment.

Zimmerer et al [**?**] took a slightly different approach in that they developed the Frequency in Language Analysis Tool (FLAT) which looked at the degree of formulaicity in an individual sample. The rationale behind this is that people with impaired cognitive function are able to depend on formulaic sentences for language generation to ease the burden of language generation on cognitive symptoms. They were able find that those with probable AD produced more formulaic language than controls.

Orimaye et al (2017) [**?**] 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.

2.7.3 The case for Deep Learning

One of the criticisms of traditional learning models is it's reliance on features that are generated for them. As we can see from the brief look at some of the research in this field, researchers have taken a number of different approaches with success. However, there does not seem to be a consensus. As machine learning in general moves away from traditional machine learning models to deep learning techniques, it poses the question can deep learning assist with this problem?

One of the main benefits of deep learning is that it does not rely on pre defined features being fed to the model. Instead, deep learning models take raw data in with some amount of preprocessing and generates it's own features. This move away from relying on features solves one of the difficulties we have with the current psychological literature, namely that there are some disagreement on the which

One of the drawbacks of attempting to use natural language processing and machine learning in this context is the lack of data. This is the case with all machine learning techniques, but more so with deep learning. There have been ways to 'create' more data such as data augmentation.

2.7.4 Conclusions

We can see that both Natural Language Processing and Machine Learning techniques have a lot to offer, Indeed there has been a lot of research which have used pre-existing datasets to explore this area with promising results. One of the difficulties with the current research is the approach of trying to discriminate between those AD and healthy controls. This is not necessarily a problem in the real world as it is trivially easy to do for trained clinicians. A more interesting, but potentially harder to problem to solve is to discriminate between those with MCI and healthy controls, and more importantly to track their decline over time. Further, to date the authors are unaware of any research in which these techniques are applied to newly created samples of language.

2.8 Discussion

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 speciality clinics and thus

have limited accessibility as frontline screening and diagnostic tools for AD. More importantly, nonspecialists are often inaccurate at identifying early AD and MCI. Thus, there is an increasing need for additional noninvasive and/or cost-effective tools, allowing effective frontline identification of subjects in the preclinical or early clinical stages of AD who could be suitable for monitoring in speciality 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.

However the literature has identified a number of challenges when approaching this problem. Firstly, the clinical features that combine to meet the diagnostic criteria for Dementia or it's variants are continuous in nature and heterogenous between patients and are also impacted by other variables. For example as shown above, cognitive performance is affected in part by a patients educational attainment and a patients ability to live independently is impacted by a their physical health as much as their cognitive health. The challenge therefore is to find features that are minimally impacted by other factors, or that can be controlled for by a experimental design such as a matched pairs design to control for educational attainment.

Another challenge lies with the recruitment of suitable individuals who may notice a decline in cognition to the point where we might classify them as having MCI, but these individuals deduce that there is little to no value to admitting there is a problem and seeking help whilst their symptoms are 'manageable'.

Finally there is a large amount of variability in the presentations of those with MCI and early dementia, and this is compounded by an similar amount of variability in the criteria researchers have used for experimental groups and

the approaches researchers have used to tackle this problem. This had led to a confused literature. Recently a call for research that has consistent inclusion / exclusion criteria has been made along with some proposed definitions of MCI and it's subgroups [**?**]. Researchers have identified the analysis of language impairment as an area of promise to explore in the diagnosis of MCI and early AD and recent developments in natural language processing and machine learning techniques have the potential to assist in this research. Indeed, given the increased burden on the diagnosis of MCI and AD on professionals there has been a call to use technology to potentially ease this burden [14]. A small but growing 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.

2.8.1 Future Work

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) [**?**] describes connectivity in such as a way that correlates directly with what Mueller (2018) [**?**] 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 organise and explore what is truly going on in those with MCI and early AD.. 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.

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, in conjunction with the devel-

opment of a model of language, sensitive enough to detect subtle deterioration in language use to act as an additional cognitive marker to aid diagnosis could potentially move some way to providing this.

Future research should also be directed towards developing non-intrusive ways of detecting subtle changes in natural language such that any perceived deterioration that could indicate the presence of MCI or AD could be flagged up early. Machine learning approaches seem to be the most logical approach for achieving this aim as language could be collected in non-intrusive ways and passed to a machine learning algorithm for preliminary classification. Despite the excellent quality of datasets, for example the DementiaBank dataset, 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 some researchers could claim are limited in scope. Finally, the recent resurgence in the use of neural networks and deep learning could provide the answer to the confused literature in terms of features. A key benefit of deep learning is it's ability to automate the process of feature engineering. So there is an opportunity to explore the use of deep learning, to not only develop new features but also validate existing features independently.

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.

2.9 Introduction

Dementia has been identified as one of the 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 [2]. The report also states that the worldwide cost of dementia in 2018 is estimated to be in the region of one trillion US dollars.

A lot of work has gone into trying to find ways of improving the early diagnosis of Alzheimer's Disease (AD) and Mild Cognitive Impairment (MCI) with research focused on two areas - identifying biological markers and analyzing the cognitive decline of those who are suspected to have the disease [15]. As described above [2], 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 [14]. I theorize that we may be able to enable an earlier diagnosis of those with MCI and AD using samples of spontaneous speech, natural language processing (NLP) and machine learning (ML).

There is a large body of research that looks at language deterioration in those suspected to have MCI or AD [15, 14]. However there is conflicting evidence in these studies about which declining language factors are associated with MCI and AD [15, 14]. 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. Another area of focus for research of this nature is the process of collecting appropriate language samples. Whilst collecting samples of language is comparatively unintrusive, researchers recognise that these samples require a rich sample of language that potentially cannot be generated by tasks such as the picture description task. Therefore, it would be useful to explore 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 [11], 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 question to be addressed in this systematic review is how has the field of machine learning and natural language processing addresses language deterioration in the diagnosis of Mild Cognitive Impairment and Early Alzheimer's Disease. The potential impact of this research in this area 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 however the absence of a single test and the complexity of AD can create significant delays in diagnosis. 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.

The remainder of this article is organized as follows. Section 5.6 gives an account of the process of this Systematic Review. Our results are described in Section 5.7. We discuss the results and implications in Section 5.8. Finally, Section 5.9 gives the conclusions.

2.10 Methodology

A systematic literature review (SLR) describes a process which aims to identify, evaluate and interpret the research and literature in a given area. They are designed to provide a complete and exhaustive summary of the current evidence relevant to an identified research question. SLR's conduct a thorough search of all literature following a pre-defined protocol that specifies focused research questions, identifies criteria for the selection of studies and assessment of their

quality, and forms to execute the data extraction and synthesis of results.

Common motivations for conducting an SLR are:

1. to summary all the evidence about a topic.
2. find gaps in the research.
3. to provide a ground for a fundament to new research.
4. and to examine how the current research supports a hypothesis.

Performing an SLR comprises the following steps:

1. identify the need for performing the SLR.
2. formulate research questions.
3. execute a comprehensive search and selection of primary studies.
4. assess the quality and extract data from the studies.
5. interpret the results.
6. report the SLR.

2.10.1 Search strategy

The main research question this SLR aims to address is: "How has the field of machine learning and natural language processing addressed language deterioration in the diagnosis of Mild Cognitive Impairment and Early Alzheimer's Disease?".

This SLR builds upon these questions and additionally presents the results of the other two additional research questions. Further, the key terms related to MCI and the other names by which it is known were included in the search string to ensure that relevant studies about machine learning and the diagnosis

of a disease were also retrieved, even if not specifically mentioned in the paper's title or abstract.

To address the research questions, a search string was defined using the PICO approach, which decomposes the main research question into four parts:

1. Population - Studies that present research on mild cognitive impairment and dementia. Mild Cognitive Impairment and Dementia keywords were selected from the Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT).

2. Intervention - Intervention: ML or Statistical Modelling techniques that focus on classification. The ML keywords were selected from the branch "Machine Learning Approaches" of the "2012 ACM Computing Classification System". The MS keywords were selected by A2.

3. Comparison

4. Outcome - Outcome: Prognosis on dementia and comorbidies. The prognosis keywords were provided by A4.

Using this process, the main research question was decomposed into four research questions:

- RQ1: What features / data characteristics of text (variables, determinants and indicators) that are considered when applying the ML techniques (n-grams, PoS Tagging etc)?

- RQ2: Which NLP and ML or Statistical Learning techniques are being used in dementia research?

- RQ3: What are the goals of the studies that employ NLP / ML techniques for prognosis of dementia?

- RQ4: Do the studies focus on time as factor?

The automated searches were performed in the Pubmed, Web of Science, Scopus and IEEE databases. Table 1 shows the search string used for the Pubmed automated search, but note that this search string was adapted to each of the other databases' search context.

(dementia OR MCI OR Mild Cognitive Impairment OR Alzheimer's OR Mild Neurocognitive Disorder OR AD) AND TOPIC: (machine learning OR Data Mining OR Decision Support System OR NLP OR Natural Language Processing) AND TOPIC: (prognosis OR prognostic estimate OR predictor OR prediction OR model OR patterns OR diagnosis OR diagnostic OR forecasting OR projection OR Deep Language Model OR Deep Neural Network) AND TOPIC: (classification OR regression OR kernel OR support vector machines OR Gaussian Process OR Bayesian Network OR Factor Analysis OR Deep Learning OR Neural Networks OR Maximum Likelihood OR Principal Component Analysis OR Markov OR Linear Model OR Mixture Model OR Perceptron Algorithm OR Logical Learning OR relational learning OR Supervised Learning OR Unsupervised Learning OR clustering OR Decision Tree) AND TOPIC: (Language OR Cognitive OR Speech OR Conversation OR Connected Speech OR Picture Description OR Discourse Analysis OR Verbal Fluency)
Searched - 4th April 2019 - Generated 1257 Articles

Tab. 2.2: Example of Search Terms for Web of Science database

Scopus	Web of Science	PubMed	IEEE Xplore
1002	991	376	230

Tab. 2.3: Number of articles found from each database - complete search terms appear in Appendix B

2.10.2 Study selection

A total of 1490 unique papers were identified through the searches conducted above. Each paper was initially reviewed by just the title and the abstract. This yielded a total of 25 potential papers. Papers were then evaluated by the JA based on the inclusion and exclusion criteria (see Table 3). Where JA could not reach a decision, PS and GV were consulted and majority vote on inclusion was conducted.

Inclusion Criteria	Exclusion Criteria
Be a primary study in English; AND address research on dementia and comorbidities; AND address at least one ML or MS technique; AND address a prognosis related to dementia and comorbidities; AND use cognition or language decline as a factor for analysis.	Be a secondary or tertiary study; OR be written in another language other than English; OR do not address a research on dementia and comorbidities; OR do not address at least one ML or MS technique; OR do not address a prognosis related to dementia and comorbidities.

Tab. 2.4: Inclusion and Exclusion Criteria

After the identification of these papers, a one-iteration backward snowballing process was carried out using the reference lists of the original set of papers looking for studies that were missed in the original searches. This resulted in 41 additionally identified papers. Throughout the whole selection process, PS and GV acted as additional assessors in the case where there was uncertainty about whether a paper should be included.

In total, 66 papers were selected to be fully read. A quality assessment questionnaire (see Table 4) was developed based on Kitchenham's guidelines and was used to minimize the chance of bias in the selection process. Studies were graded on a 12 point scale and any studies which scored less than 8 points

were excluded for quality reasons. The ones that successfully passed the filtering criteria described earlier had their relevant data extracted.

Variable	Definition
Conditions Studied	For which dementia disorder is the study deriving a prognosis.
Database used in the study	Name and origin of the data source used to derive the prognosis of the studied dementia.
Dataset Categories	Classes in which the data units were divided into.
Follow-up period	Period of time, which the data units were followed.
Techniques used	Natural Language techniques AND/OR ML techniques that were used to build the diagnostic models.
Features generated	NLP generated features used in building the diagnostic models.
Aim of the Study	The goal of the built diagnostic models.

Tab. 2.5: Quality Assessment Questionnaire

In this phase, a paper could also be rejected due to inclusion and exclusion criteria because the selection process adopted an inclusive approach. This means that during the reading of the titles and abstracts, in the case where the information provided was incomplete or too general it was selected to be fully read in the posterior phase. A common example is the case when the data analysis technique specified in the abstract was merely "classification", so it was not possible to know if any machine learning occurred.

In total, 37 studies composed the final set of included primary studies and had their relevant data extracted, 7 papers were rejected due quality reasons, and 34 papers were rejected due to failing the inclusion and exclusion criteria. One reason for the high number of the latter was the decision to exclude the papers that used solely statistical methods as data analysis techniques to build the prognostic models. The selected studies were also assessed for the risk of cumulative evidence bias. This was done by checking, in the case of the same research group with different studies in the final set of included primary studies, if it was justified having both studies (i.e different samples).

2.10.3 Data collection

For the data collection, a base extraction form was defined in the protocol, but later in the study it was evolved based on the research group discussions. Table 4 lists and defines the collected variables.

In addition to these variables other basic data about the studies was collected, these were: title, authors, journal/source, year and type of publication. No summary measures were used. Summary tables were used for the synthesis of results and no additional analyses were carried out

2.11 Results

2.11.1 Features of Language

One of the first steps in building an accurate diagnostic model using language is to analyse the effectiveness of the various language features that are used.

Asgari, Kaye and Dodge (2017) [**?**] 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 [**?**]. They were able to successfully used machine learning algorithms to distinguish between these two groups with an accuracy of 84%.

Measures of Semantic Complexity

Intro!

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).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{2.3}$$

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:

$$W = N^{V(-0.165)} \tag{2.4}$$

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:

$$R = (100 \log N)/(1 - V1/V) \tag{2.5}$$

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.

Guinn, Singer and Habash [**?**], they found in their corpus that there was no significant difference between interviewers and those with dementia when applying these measures. However, when they compared these results with a control dataset, they did find a significant difference in Honore's statistic. They explained this interesting results by suggesting that the interviewer used was trying to match (intentionally or unintentionally) the lexical richness of the person they were interviewing. In comparing healthy older adults with those suffering from dementia, they found that there was an increase in lexical diversity which is contrary to other research. They did note that conversations involving those with dementia contained roughtly 50% less speech than those with controls, and that TTR in particular is not a suitable measure because it is more sensitive to length. They also note that Brunet's Index and Honore's Statistic are better statistics that control for the total length of the conversation and controls had statistically greater lexical diversity on those measures.

2.11.2 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.

2.11.3 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.

2.11.4 N-grams and skip-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 [**?**] 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) [**?**] 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.

Skip-grams are a variant of n-grams in which word tokens are skipped intermittently while creating n-grams. For example, take the sentence 'I am going to London', there are four conventional bigrams: 'I am', 'am going', 'going to', and 'to London' - using skip-grams, we might skip a word to create additional bigrams such as: 'I going' and 'going London'. Orimaye defined k-skip-n-grams as a set of n-gram tokens with the following equation, where n is the specified n-gram (e.g. 2 for a bigram and 3 for a trigram), m is the number of tokens in a given sentence, k is the number of word skip between n-grams given that k ¡ m and a = {1,..., m-n}

One problem with this approach is existing research currently uses language generated from picture description tasks. Given the nature of these tasks, the language generated is relatively constrained in comparison to language generated spontaneously.

$$T_{n-gram} = W_a, ..., W_{a+n-k}, ..., W_{a+n}, ..., W_{m-n}, ..., W_{(m-n)+n-k}, ..., W_m \tag{2.6}$$

Thus for the sentence 'I am going to London', 1-skip-2-grams will give {I going, am to, going London} and 1-skip-3-grams will give {I going to, I am to, am to London, am going London}

2.11.5 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 [**?**]

2.11.6 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.

2.11.7 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.

Lu's Syntactic Complexity Analyser Ygnve measure

2.11.8 Semantic features

. We compute semantic similarity using the average and minimum cosine distance between each pair of one-hot embeddings of utterances, and the cosine

cutoff (i.e., the number of pairs of utterances whose the cosine distance is below a certain threshold). We compute word specificity and ambiguity based on tree depth and the number of senses in WordNet [54]. We also extract multiple WordNet measures of similarity: Resnik [68], Jiang-Coranth [69], Lin [47], Leacock-Chodorow [70], and Wu-Palmer [71].

2.11.9 Syntactic features

Guinn, Singer and Habash [**?**] found that syntactic features such as Noun rate, verb rate, adjective rate and pronoun rate were not significantly different between interviewers and those with dementia in their dataset, although they did note that there was a slightly higher but non significant use of pronouns.

2.11.10 Pragmatic features.

We train a general 100-topic latent Dirichlet allocation (LDA) model [72] on the Wikipedia corpus for generalizability. LDA is a generative statistical model used to determine unlabeled topics in a document. For each transcript, we extract the probabilities of each LDA topic. Next, we extract features related to rhetorical structure theory (RST), which is a classic framework for discourse parsing in which partitions of text are arranged in a tree structure by pragmatic relations such as Elaboration or Contrast [73].

Go-ahead Utterances

Go-ahead utterances are defined as short one or two syllable responses that do not contribute to the conversation beyond a minimal response. The function of these go-ahead utterances can be to validate what the other person is saying, or to agree / disagree with what is being said. Another function can be that they wish speaker is indicating that they have nothing futher to add to the conversation and a signal to the other speaker to continue with what they are

saying. In Curry, Singer and Habash's research, they found that in comparing interviewers and those with dementia, that interviewers used significantly fewer go-ahead utterances. In comparing controls and those with dementia, they found that there was a relative lack of go-ahead utterances in the controls which implies that controls had a lot more contributions to make in their conversations than those with dementia.

2.11.11 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) [**?**] 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.

2.11.12 Number of syllables and Characters

2.11.13 Number of fillers

2.11.14 Readability

Flesch reading score, Flesch-kincaid grade level

2.11.15 Polarity

2.11.16 Frequency

Mean values of frequency, age of acquisition, imageability, familiarity, arousal, dominance and valence based on lexical norms

2.11.17 Dysfluencies

Curry, Singer and Habash noted that in comparing controls and those with Dementia, that those with Dementia had a significantly higher number of pauses per word and a much higher incidence of words that were truncated in mid-speech. In comparing interviewers with those with dementia, they also showed other signs of difficulties with fluency with higher rates of incomplete words, filler words and repeated words.

2.12 Machine Learning methods

The next stage of building statistical or Machine Learning models is to take some features and use these feature to build models able to classify participants into various categories. There are various types of classification models that have been used to tackle this problem, this section looks which models have been used in the current literature.

Machine Learning technique used	Paper Number
Logistic Regression	2
Support Vector Machines	4
Naive Bayes	1
Decision Trees	1, 2
Multi-layered perceptrons	2

2.12.1 Traditional Machine Learning methods

Logistic Regression

Logistic Regressions is a probabilistic approach to classification and is dependant on the assumption that your input space can be separated into two distinct regions, one for each class, by a linear boundary. In 2D space, this boundary is a line and in 3D space, this boundary is defined as a plane.

In the results of this survey.

Support Vector Machines

Support Vector Machine (SVM) was originally proposed as an algorithm for classification problems; it is a relatively new technique compared to the other ML approaches. The classification process consists of mapping the data points (usually the study subjects) into a feature space composed of the variables that characterize these data points, except for the outcome variable. Then, the algorithm finds patterns in this feature space by defining the maximum separation between two or more classes, depending on the problem to be solved. Contrary to some regression techniques, SVMs are not dependent on a pre-determined model for data fitting, although there are still algorithm specifications to be considered (e.g. choice of a kernel function); instead, it is a data-driven algorithm that can work relatively well in a scenario where sample sizes are small compared to the number of variables, reason why it has been widely employed by diagnostic studies in tasks related to the automated classification of diseases.

Regarding the SLR results, SVMs were present in ??/41 selected studies, in 38 proposed models, and being by far the most used machine learning technique in the dementia diagnosis research. These numbers account for the traditional SVM and variations. In all of the ?? selected studies the SVMs focused at binary classifications where the task was to discriminate mild cognitively impaired

(MCI) patients that will or will not develop Alzheimer's Disease (AD). In the general case, the problem is posed as either MCI converters versus MCI non-converters, or progressive MCI versus stable MCI classification. This outlines a situation in which a regression problem (when will the MCI patients convert to AD?) is formulated as a classification problem (which MCI patients will convert to AD in X months?) to be solved. Reasons for this could be due to limitations in the data used, i.e. the limited follow-up periods of the subjects included in the studies.

Linear Discriminant Analysis

Decision Trees

A Decision Tree (DT) is a classification algorithm in which the learned knowledge is represented in a tree structure that can be translated to if-then rules. DT's learning process is recursive and starts by testing each input variable as to how well each of them, alone, can classify the labeled examples. The best one is selected as a root node for the tree and its descendant nodes are defined as the possible values (or relevant ratios) of the selected input variable. The training set is then classified between the descendant nodes according to the values of the selected input variable. This process is repeated recursively until no more splits in the tree are possible. Like SVMs, DTs do not depend on a pre-defined model and are mostly used to find important interactions between variables. Being intuitive and easy to interpret, DTs have been used in prognostic studies as a tool for determining prognostic subgroups.

In this SLR, DTs were the second most frequently used ML technique, present in 6/37 selected studies and proposed in 7 models. It was employed for the same reason as SVM; except for one study that investigated the evolution of patients diagnosed with cognitive impairment no dementia (CIND) to

AD.

Curry, Singer and Habash - Alz (66.7 - Pres, 66.7 - Recall) and Control (67.9 - Pres, 67.9 Recall)

Naive Bayes

Curry, Singer and Habash - Alz(80.8 - Pres, 0.75 - Recall) and Control (79.3 - Pres, 82.1 Recall) - In the naive bayes classifier, they identified that pauses, go-ahead, fillers and incomplete words as the most significant features.

Artificial Neural Networks

An Artificial Neural Network (ANN) is a methodology that performs multifactorial analyses, which is desirable in the health area as medical decision-making problems are usually dependent of many factors. An ANN is composed of nodes connected by weighted edges in a multi-layer architecture that comprises: an input layer, one or more hidden layers and an output layer. In the training process, inputs and outputs values are known to the network, while the weights are incrementally adjusted so that the outputs of the network are approximate to the known outputs. Despite being a powerful predictor, ANNs are 'black boxes', which means that they are not able to explain their predictions in an intuitive way, contrary to DTs or BNs. Also, they require the specification of the architecture to be used beforehand (i.e. the number of hidden layers).

K-nearest neighbours

K Nearest Neighbors (KNN) is a classification algorithm that takes a data point from an unknown class and assigns it as an input vector in the feature space. Then, the classification process follows by assigning the unknown class data point to the class in which the majority of the K nearest data points belong to. The distance between data points is usually measured by Euclidean distance,

but it is possible to employ other measures. KNN is one of the simplest ML classification algorithms and have been used in a wide range of applications; however, it can be computationally expensive in a highly dimensional scenario. Further, it considers all features to be equally weighted, which can be a problem if the data has superfluous attributes.

Best Results

Machine Learning technique used	Paper Number
Logistic Regression	2
Support Vector Machines	4, 5
Naive Bayes	1
Decision Trees and Random Forest Classifiers	1, 2, 5
Multi-layered perceptrons	2

2.12.2 Deep Learning methods

Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia - Orimaye, Wong and Wong (2018)

In this paper, Orimaye et al use deep-deep neural networks language models (D2NNLM) to learn linguistic changes that distinguish the language of patients with MCI and AD-type dementia from the healthy controls using higher order n-grams. An ordinary DNNLM uses lower order n-gram N-dimensional sparse vectors as descrete feature representations to train the neural network with multiple hidden layers.

2.12.3 What type of data is used by the studies?

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.

2.12.4 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 [**?**]. The Cookie Theft picture (pictured below) depicts a scene of a home typical of the period of time when it was created and would generally not require participants to use any complicated vocabulary to describe. In this task participants are asked to describe the picture presented to them in as much detail as possible. This task 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 [**?**]

Fig. 2.5: Cookie Theft Picture - From Kaplan and Goodglass (1983)

The picture description task does a fine job of eliciting descriptive language however because of the specific content the language produce could be considered quite limited. There is some disagreement as to the benefits of this using this methodology. This task is reported as being useful to lexico-semantic disorders [14, **?**] as the language being generated is primarly nouns and deixis (words to identify items and words to put those items into context). However, Ash [**?**]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.

There are a number of existing corpus which use this tasks as the foundation of their data. The most well known is the Pitt Corpus of the Dementia Bank database. The participants of the corpus are mainly healthy controls, those with MCI and those with probable and/or possible AD.

2.12.5 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 with the Cinderella being the one most typically used [**?**]. 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 actions, motivations and internal reactions to given events. This task is particularly useful as the procedure reduces the demands on memory, due to the participant being able to access the picture book during the description and is therefore able rule out memory as a confounding variable for any results observed. As noted above, Ash [**?**] felt that this task was interchangable with the Picture Description task. However, other research felt that this was a studier test of lexical and semantic abilities as well as syn-

tactic complexity because this task requires interpretation and elaboration in additional to a simple description [**?**].

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 [**?**]. 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.

2.12.6 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 [14]. 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, although it can be argued that content should not affect the type of language being generated unless it is narrow topic or the participant is constrained in how they answer a given question. 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 [**?**]. There have been some attempts to use interviews to assess language production in AD with promising results [**?**, **?**].

2.12.7 Conclusions

One can view the different types of tasks above as a continuum where picture tasks represent a much more controllable task with a lot of supporting research but which generates a much more constrained set of language that is atypical of normal speech in terms of the cognitive functions used.

2.12.8 What are the goals of the studies that employ ML or Statistical Learning techniques for diagnosis of MCI or AD?

2.12.9 Do the studies focus on a one point in time or looking at cognitive deterioration over time?

2.13 Discussion and conclusions

2.13.1 Discussion of the current evidence

One of the criticisms of traditional learning models is it's reliance on features that are generated for them. As we can see from the brief look at some of the research in this field, researchers have taken a number of different approaches with success. However, there does not seem to be a consensus. As machine learning in general moves away from traditional machine learning models to deep learning techniques, it poses the question can deep learning assist with this problem?

One of the main benefits of deep learning is that it does not rely on pre

defined features being fed to the model. Instead, deep learning models take raw data in with some amount of preprocessing and generates it's own features. This move away from relying on features solves one of the difficulties we have with the current psychological literature, namely that there are some disagreement on the which

One of the drawbacks of attempting to use natural language processing and machine learning in this context is the lack of data. This is the case with all machine learning techniques, but more so with deep learning. There have been ways to 'create' more data such as data augmentation.

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 speciality clinics and thus have limited accessibility as frontline screening and diagnostic tools for AD. More importantly, nonspecialists are often inaccurate at identifying early AD and MCI. Thus, there is an increasing need for additional noninvasive and/or cost-effective tools, allowing effective frontline identification of subjects in the preclinical or early clinical stages of AD who could be suitable for monitoring in speciality 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.

However the literature has identified a number of challenges when approaching this problem. Firstly, the clinical features that combine to meet the diagnostic criteria for Dementia or it's variants are continuous in nature and heterogenous between patients and are also impacted by other variables. For example as shown above, cognitive performance is affected in part by a patients

educational attainment and a patients ability to live independently is impacted by a their physical health as much as their cognitive health. The challenge therefore is to find features that are minimally impacted by other factors, or that can be controlled for by a experimental design such as a matched pairs design to control for educational attainment.

Another challenge lies with the recruitment of suitable individuals who may notice a decline in cognition to the point where we might classify them as having MCI, but these individuals deduce that there is little to no value to admitting there is a problem and seeking help whilst their symptoms are 'manageable'.

Finally there is a large amount of variability in the presentations of those with MCI and early dementia, and this is compounded by an similar amount of variability in the criteria researchers have used for experimental groups and the approaches researchers have used to tackle this problem. This had led to a confused literature. Recently a call for research that has consistent inclusion / exclusion criteria has been made along with some proposed definitions of MCI and it's subgroups [**?**]. Researchers have identified the analysis of language impairment as an area of promise to explore in the diagnosis of MCI and early AD and recent developments in natural language processing and machine learning techniques have the potential to assist in this research. Indeed, given the increased burden on the diagnosis of MCI and AD on professionals there has been a call to use technology to potentially ease this burden [14]. A small but growing 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.

2.13.2 Methodological Issues

2.13.3 Limitations

2.13.4 The future of the field

One of the main patterns

One-class classification is a traditional machine learning model in which the object is to recognize instances of a concept by only using examples of the same concept. In the training of this model, instances of only a single object class are available during training. In this context,all other classes (any instances that deviate from the single object class) are referred to as 'alien' classes. During the testing of the model, it may encounter alien classes and the job of the model is to distinguish between objects of the known class from objects belonging to any 'alien' classes. This approach to classification has broad applications and is mainly used when there is a high cost of obtaining samples of alien classes.

In this context, we can see from

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) [**?**] describes connectivity in such as a way that correlates directly with what Mueller (2018) [**?**] 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 organise and explore what is truly going on in those with MCI and early AD.. 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.

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, in conjunction with the devel-

opment of a model of language, sensitive enough to detect subtle deterioration in language use to act as an additional cognitive marker to aid diagnosis could potentially move some way to providing this.

Future research should also be directed towards developing non-intrusive ways of detecting subtle changes in natural language such that any perceived deterioration that could indicate the presence of MCI or AD could be flagged up early. Machine learning approaches seem to be the most logical approach for achieving this aim as language could be collected in non-intrusive ways and passed to a machine learning algorithm for preliminary classification. Despite the excellent quality of datasets, for example the DementiaBank dataset, 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 some researchers could claim are limited in scope. Finally, the recent resurgence in the use of neural networks and deep learning could provide the answer to the confused literature in terms of features. A key benefit of deep learning is it's ability to automate the process of feature engineering. So there is an opportunity to explore the use of deep learning, to not only develop new features but also validate existing features independently.

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.

2.14 Conclusions

We can see that both Natural Language Processing and Machine Learning techniques have a lot to offer, Indeed there has been a lot of research which have used pre-existing datasets to explore this area with promising results. One of

the difficulties with the current research is the approach of trying to discriminate between those AD and healthy controls. This is not necessarily a problem in the real world as it is trivially easy to do for trained clinicians. A more interesting, but potentially harder to problem to solve is to discriminate between those with MCI and healthy controls, and more importantly to track their decline over time. Further, to date the authors are unaware of any research in which these techniques are applied to newly created samples of language.

3. DELPHI METHODOLOGY AND DEVELOPING CONSENSUS ON HOW BEST TO COLLECT LANGUAGE SAMPLES USING TECHNOLOGY

3.1 Introduction

Here is the text of your introduction.

$$\alpha = \sqrt{\beta} \tag{3.1}$$

3.1.1 Subsection Heading Here

Write your subsection text here.

3.2 Conclusion

Write your conclusion here.

4. DEVELOPMENT OF A PIPELINE THAT PROCESSES LANGUAGE DATA ACCURATELY

4.1 Background

As Komeili in her paper points out, machine learning is a tool may allow earlier detection and management of change in language [16]. But as with all machine learning models, the benefits of accuracy only come when large amounts of data is used to train the model. With our target population, this problem is further exacerbated due to the relatively low incidence of people with MCI or early AD and the relatively high cost of collecting this data in this particular population.

One potential way forward is to look not at the ill but at the healthy. This has inverts the problem of scarcity and cost of data collection, and so we now have an abundance of data that we can now use. This, in theory, would allow us to look specifically at any deviations from the healthy. This process is known as one-class classification.

4.2 One-class classification

In traditional machine learning, there are numerous classification algorithms that are widely used. Their utility is classifying instances of data into one (Binary classification) or more (multiclass classification) categories. This ia a fine way to look at classification in the vast majority of cases, but some difficulties do arise under certain circumstances.

5. ANALYSIS OF THE PRESIDENTS CORPUS, THREE AUTHORS AND DEMENTIABANK DATASETS

5.1 Background

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 (who would go on to receive a diagnosis of Dementia) and George H.W. Bush who acted as a matched control based on Age [11]. They found several differences in language use which they felt acted as indicators of Reagan'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 [11].

This study replicates work done by Berisha and Liss and extends this by adding Donald Trump as an alternative, more appropriate comparison to Ronald Reagan as he is much closer in age than George H.W. Bush. This experiment will look at the features originally identified by Berisha and Liss, as well as any others that have potential as discussed in the literature review above.

5.2 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 29 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 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. 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.

5.2.1 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 GHWB 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 GHWB were ommited due to a lack of a question and answer session).

5.2.2 Feature Selection

We calculated the following features for each transcript in turn using the NLTK (see section 2 for a description) [**?**] and Python.

Measures of lexical variation

We constructed two 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 (see section 2 for a description) for a feature that measures lexical diversity independent of sample size [**?**].

Fillers, Non-Specific Nouns and LI Verbs

For these features, we counted the number of occurrences for different categories (see table for list of categories tracked and the words counted). The features were used by Berisha and Liss in their research [11] and were taken from work done by Bird et al [**?**].

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

Tab. 5.1: Categories and Words Counted

Usage of parts of speech

This section involves using a Part of Speech tagger (PoS) which analyses a sentence 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.

5.3 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 variances 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)
Non Specific Nouns	12.72 (4.63)	6.78 (4.32)	7.41 (8.75)
LI Verbs	124.22 (17.89)	103.45 (51.87)	84.48 (75.78)

Tab. 5.2: Means and Standard Deviations of important features

In terms of unique words, we found that RR used significantly more unique words, non-specific nouns and low imageability verbs than GHWB and DJT (see Table 3.3). 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 [11, 10]. 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 Words	**6.832*** **	**11.403*** **	**4.137*** **
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

Tab. 5.3: RR T-tests vs GWB and DJT

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

Tab. 5.4: Pearson Correlations for Features

5.4 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 the findings of Berisha and Liss in that we both find a significant decrease in unique words over time and an increase in non-specifc noun usage. Compared to our controls (GWHB andDJT), 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.

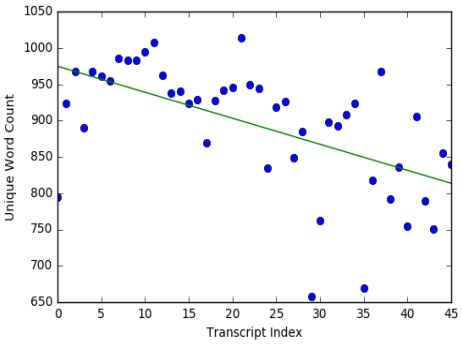

Fig. 5.1: Ronald Reagan - Unique Words over time

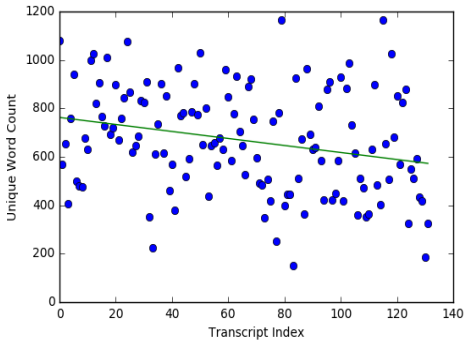

Fig. 5.2: George H.W. Bush - Unique Words over time

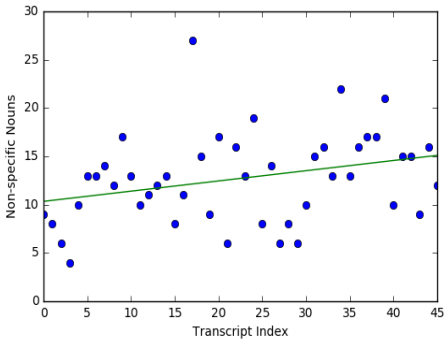

Fig. 5.3: Ronald Reagan - Non-specifc Nouns over time

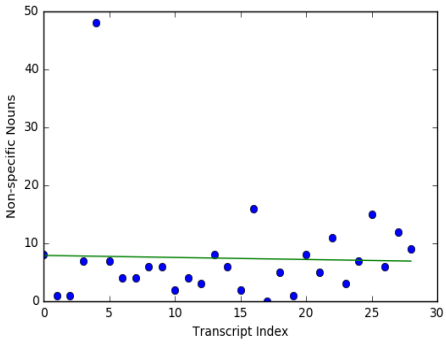

Fig. 5.4: Donald J. Trump - Non-specific Nouns over time

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 [**?**, 10]. Whilst it is important to control for outliers, there are other ways in which we can control for length of sample.

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 preclinical 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.

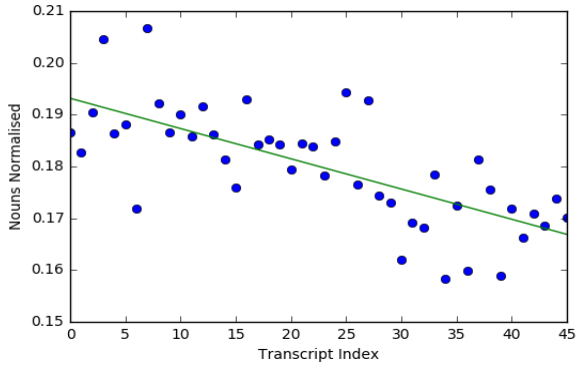

Fig. 5.5: Ronald Reagan - Nouns Normalised over time

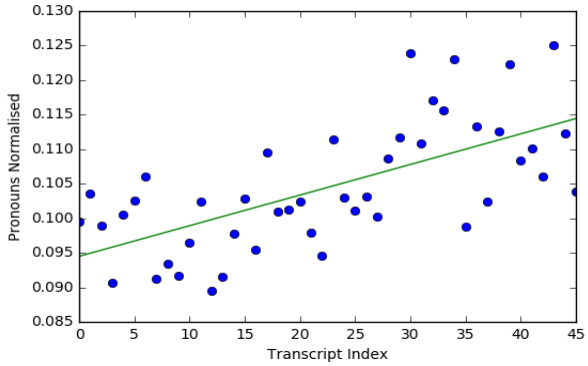

Fig. 5.6: Ronald Reagan - Pronouns Normalised over time

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.

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.

5.5 Introduction

A diagnosis of dementia is generally made when there is a decline in brain function due to physical changes in the brain [5]. It affects a significant proportion of the global older adult population and the impact on morbidity and mortal-

ity rates is considerable. Dementia is currently the leading cause of death in England and Wales and the sixth leading cause of death in the United States (US). A 2014 report commissioned by the Alzheimer's Society estimated that in the UK by 2015 there would be approximately 855,000 people rising living with dementia increasing to 1 million by 2021 [1]. This represents 1 in 79 of the total UK population rising to 1 in 14 of those aged 65 or over [1]. Worldwide, there are 46 million people with a diagnosis of dementia globally and that number is expected to hit 131.5 million by 2050 [2]. From a financial perspective, the cost burden is also significant. The estimated annual spend on dementia healthcare in the UK 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. The worldwide cost of dementia in 2018 was estimated to be in the region of one trillon US dollars [2].

There are different types of dementia including Alzheimer's Disease (AD), vascular dementia, dementia with Lewy bodies and fronto-temporal dementia, all of which are currently incurable. AD is a progressive neurodegenerative disease and is the most common type of dementia, responsible for approximately 60% to 80% of all cases [**?**]. Currently, a definitive diagnosis for AD 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 have been identified as two of the key proteins that have a role in initiation and progression of AD. In a typical case beta-amyloid clumps into plaques which slowly build up between neurons with tau proteins accumulating and eventually forming tangles inside neurons. At a certain point, the levels of beta-amyloid rise triggering a more rapid spread of tau throughout the

brain. Eventually, due to these and other changes, neurons lose their ability to communicate and the brain starts to shrink[**?**]. This leads to the development of some of the psychological symptoms associated with AD, primarily cognitive deficits such as problems with episodic and semantic memory, organizing and planning, difficulties with language, problems with executive function and visuospatial deficits [3]. These cognitive symptoms are often accompanied by emotional problems such as depression and behavioural difficulties. As more neurons die throughout the brain, a person with AD gradually loses the ability to think, remember, make decisions and function independently.

Despite the increasing prevalence of AD and an improved understanding about how it affects the brain there are no medications that improve prognosis. 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 of these drugs fail in clinical trial stage of the drug discovery process ((99.6% failure rate vs 80% for systemic anti-cancer drugs) [4]. Cummings et al 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 occurred [4]. Research focus has now started to shift to the earlier stages of AD (i.e. symptomatic pre-dementia phase of AD) which some literature describes as 'Mild Cognitive Impairment (MCI) due to AD'.

One of the challenges associated with the early detection of AD is differentiating natural age associated memory impairment and cognitive decline due to aging from with decline due to AD. This challenge is often complicated further due to the large variation in the cognitive abilities and educational background of individuals. The work of Albert et al helps to address this by the development of clinical criteria which professionals can use to diagnose MCI due to

AD. 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 [5], 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 method of detecting AD at an earlier stage. This usually takes the form of recording speech whilst patients' undertake 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. The work of Berisha and Liss is a good example of this, this study compared 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 [11]. They found several differences in language use over time which they felt acted as indicators of RR's difficulties with language due to AD. Differences in features of language identified to be statistically significant included the number of unique words used per speech, the use of non-specific nouns and fillers and low imageability verbs [11].

Berisha and Liss have developed some very interesting ideas about how we might track changes over time in various lexical features which have been associated with the development of Alzheimer's Disease. However we feel that cognitive decline in pre-clinical AD is not accurately modeled as a linear process and therefore we explored the application of generalised additive models (GAMs) to this data which have no assumed understanding of the distribution of the data. We refine these ideas with the aim of exploring the potential for a protocol that can be used for analyzing language deterioration.

This study extends the work of Berisha and Liss in a number of ways. Our

hypotheses are:

1. Language use of RR will be significantly different to our controls (GHWB, DJT).

2. Modelling the dynamics of lexical features over time can be improved using non-linear models compared to linear models.

5.6 Methodology

We took 46 transcripts of press conferences given by RR (from 1981 to 1988) and compared them with 134 press conferences (from 1989 to 1993) given by GHWB and 29 press conferences (from 2016 to 2019) conducted by DJT. We analysed transcripts for language features (described below) shown to change longitudinally with AD. These language features are analysed at the word level, sentence level and document level.

In the original study, GHWB was selected as the comparator president as he was the closest match in terms of age to RR (GHWB - age at the start of presidency: 64 years and 222 days vs RR: 69 years and 349 days). However, since his inauguration, DJT is now the closest comparible president in terms of age (DJT - age at the start of presidency: 70 years and 220 days). We included DJT who like GHWB has no known diagnosis of AD to determine whether the comparisons made by Berisha and Liss hold true with this closer presidential match in terms of age.

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. These documents are open access and can be downloaded at any time from the APP archive.

5.6.1 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 answers to questions as this data would, from a theoretical perspective, not be relevant for language analysis. We did not make any alterations to the length of the document, but instead 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).

5.6.2 Feature Generation

Features were generated by running each transcript through a number of natural language processing libraries. We used the Natural Language ToolKit (NLTK) package [**?**] in python to stem each transcript using the Snowball Stemmer and also completed a part of speech (POS) tagging process. We also ran each transcript through Linguistic Inquiry and Word Count (LIWC) [**?**] software. We completed all NLP tasks in Python and completed all statistical analysis in R..

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).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{5.1}$$

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:

$$W = N^{V(-0.165)} \tag{5.2}$$

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:

$$R = (100 \log N)/(1 - V1/V) \tag{5.3}$$

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.

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 [11] 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[11], 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"*

Tab. 5.5: Examples of words belonging to the categories Fillers, Non-Specific Nouns and Low Imageability Verbs

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.

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.

Longitudinal Analysis

As with Berisha and Liss[11], 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.

We looked also compared the results of linear models and a non linear model in terms of best fit. The non linear model we used was the generalised additive model [**?**]. The selection of this non linear model was due to the authors not wanting to assume an underlying distribution of the data and therefore this was the most appropriate model to use.

5.7 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)

Tab. 5.6: 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). 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 [11, 10]. 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

Tab. 5.7: RR T-tests vs GWB and DJT

5.7.1 Longitudinal Analysis

We then looked at the data from a longitudinal perspective as we were 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 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 sig-

nificant 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

Tab. 5.8: Pearson Correlations for Features

To get a better idea of decline over time, we also ran Pearson's correlations using time elapsed in days from the first transcript instead a transcript index. This was to control for variation in the number of days between transcripts. Given the number of hypotheses being tested, we felt it was 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 but we

have included all features in Table 6.

	RR R-Squared	GWB R-Squared	DJT R-Squared
ppron	**0.700*** **	0.007	-0.196
social	**0.698*** **	0.204	0.032
NounsNormalised	**-0.689*** **	-0.023	0.194
function	**0.670*** **	-0.169	-0.243
conj	**0.644*** **	-0.452	-0.187
PronounsNormalised	**0.631*** **	0.131	0.002
Analytic	**-0.626*** **	-0.013	0.366
NN	**-0.601*** **	-0.197	0.116
male	**0.585*** **	0.024	0.042
UniqueWords	**-0.578*** **	-0.257	0.204
WDT	**-0.577*** **	-0.173	-0.105
shehe	**0.529***	-0.005	0.064
VBZ	**-0.521***	-0.157	0.136
JJ	**-0.518***	-0.206	0.062
Fillers	-0.076	**-0.359****	-0.032
EX	-0.193*	**-0.337****	0.339
achieve	-0.248	**0.334****	0.169

* denotes p<0.05

** denotes p<0.01

*** denotes p<0.001

Tab. 5.9: Pearson Correlations for Features

In terms of comparing the general additive model and linear models. We calculated the predicted residual error sum of squares (PRESS) statistic for each feature is a form of cross-validation used in regression analysis to provide

a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. There were no significant differences (paired t-test: t=1.8875, df = 21, p-value = 0.07299).

Feature	GAM Press Statistic	LM Press Statistic
ppron	18.591104613	20.001055542
social	21.681732112	24.906646115
NounsNormalised	0.002818941	0.002797798
function	46.546071705	46.031993602
conj	8.812330188	8.354974826
PronounsNormalised	0.002284437	0.002356129
Analytic	2272.200519272	2239.224863610
Pronoun	44.557282749	46.894576916
NN	101685.870974277	96485.575598505
male	5.191815443	5.278559152
UniqueWords	255114.636660673	238042.427437454
WDT	1987.149134694	1958.781236655
nouns	276723.222762842	262192.120091097
nouns/100	27.670762078	26.219212009
UniqueStems	170450.645283172	166302.725530118
shehe	5.188136314	5.265385185
VBZ	13889.871831536	13721.175422577
JJ	24950.440222526	24692.043386181
article	13.500676675	13.468083961
Adjectives	30210.006270270	29865.240707991
Adjectives.100	3.020898615	2.986524071
Dic	40.099289973	38.501060715

Tab. 5.10: Comparison of GAM and Linear Model using the PRESS statistic

5.8 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. 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 preclinical 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.

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 [**?**, 10]. 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. When we did this, we found that there was a significant decrease in the number of unique words (R=-0.56, p<0.001) used however when we controlled via normalisation, we found that this was not a significant feature (R=-0.172, p=0.25).

Overall, we are able to show differences in the a number of language features between RR, GHWB and DJT over time and the psychological literature confirms our findings in reference to these changes.

Our next hypothesis involved exploring whether linear models were the most appropriate way to track this longitudinal data. In our analysis, we found there was no significant difference in terms of mean sum of squares errors between the linear model and the generalised additive model however this does not mean that these are equivalent. For example, considering the plot below. We compare a model fit to a linear model with a generalised additive model which aims to model the data more closely. It is clear to see that both models can track the movements of language use over time however, the linear model is quite a rigid model which does not allow for different inflection points. We can see that RR's decline is not linear, and that there are periods of relative stability and some periods where the decline is more severe. We argue that these more nuanced changes over time can only be modelled by a model such as a GAM.

In her paper on developing diagnositic criteria for Mild Cognitive Impairment, Albert et all [5] states 'it is important to obtain longitudinal assessments of cognition, whenever possible' and 'obtaining objective evidence of progressive declines in cognition over time is important for establishing the accuracy of the diagnosis, as well as for assessing any potential treatment response.' As we have

shown, it is not the case that language declines in a linear way and whilst we have a number of data points that we can use to model RR's decline, in a clinical setting it would be impractical to have upwards of 40 data collection points. We feel that it is important that we have more than just two. In the figures below, we can see that if we measured language at Reagan's first transcript to a point after 700 days the vast majority of points would point to a decline in language. However, the next figure shows what would happen if we measured language from Reagan's second transcript, just 34 days later. In this case, we can see that in some cases we would mark Reagan as improving when it is clear that he is not.

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 artefact of the data is in itself notable as it illustrates the problems we may have with smaller amounts of speech and the problem of finding an appropriate control is a common one in this domain, given the considerable variation in factors such as age, language ability and education. As mentioned above, we are lucky enough to have numerous samples of data collected over a long period of time and this is not something that can be easily translated to a clinical protocol. Finally, the variability in the number and quality of the transcriptions raises some doubts as to the results of DJT. Is it the lack of instances of data, or that the number of words per transcript is significantly less than with RR or GHWB. This does have further implications for using a protocol such as this in a more general way as it will impact how data is collected.

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 may 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.

5.9 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 that 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 group of MCI patients with a group of 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.

The results of this work also identify some clues as to what could work in a clinical setting, or how we might be able to collect data in a way that accurately tracks language decline without making incorrect assumptions.

From a clinical perspective, we can see that using samples (albeit for this dataset) is able to track languages changes over a given time frame. This potentially means that we can use a similar methodology to collect regular language samples in settings such as memory clinics and GP's surgeries and even poten-

tially in the people's homes and that these language samples, have the potential to act as an early warning sign for Mild Cognitive Impairment that potentially will identify patients at risk of developing Alzheimer's Disease.

6. PILOT STUDY OF THE METHODOLOGY DEVELOPED

6.1 Introduction

Here is the text of your introduction.

$$\alpha = \sqrt{\beta} \tag{6.1}$$

6.1.1 Subsection Heading Here

Write your subsection text here.

6.2 Conclusion

Write your conclusion here.

7. GENERAL DISCUSSION, CONCLUSIONS AND FUTURE WORK

7.1 Introduction

Here is the text of your introduction.

$$\alpha = \sqrt{\beta} \tag{7.1}$$

7.1.1 Subsection Heading Here

Write your subsection text here.

7.2 Conclusion

Write your conclusion here.

BIBLIOGRAPHY

[1] Alzheimer's Society. Alzheimer's Society - Dementia UK: Second Edition. Technical report, 2014.

[2] Martin Prince, Anders Wimo, Guerchet M, Ali GC, Wu YT, and Prina M. World Alzheimer Report 2015 The Global Impact of Dementia An analysis of prevalence, incidence, cost and trends. *Alzheimer's Disease International*, 2015.

[3] Guy M. McKhann, David S. Knopman, Howard Chertkow, Bradley T. Hyman, Clifford R. Jack, Claudia H. Kawas, William E. Klunk, Walter J. Koroshetz, Jennifer J. Manly, Richard Mayeux, Richard C. Mohs, John C. Morris, Martin N. Rossor, Philip Scheltens, Maria C. Carrillo, Bill Thies, Sandra Weintraub, and Creighton H. Phelps. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):263–269, may 2011.

[4] Jeffrey L Cummings, Travis Morstorf, and Kate Zhong. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's Research & Therapy*, 6(4):37, 2014.

[5] Marilyn S. Albert, Steven T. DeKosky, Dennis Dickson, Bruno Dubois, Howard H. Feldman, Nick C. Fox, Anthony Gamst, David M. Holtzman, William J. Jagust, Ronald C. Petersen, Peter J. Snyder, Maria C. Car-

rillo, Bill Thies, and Creighton H. Phelps. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279, may 2011.

[6] McHugh PR Folstein MF, Folstein SE. Mini-mental state. A grading the cognitive state of patiens for the clinician. *J Psychiatr Res*, 12:189–198, 1975.

[7] Ellen Grober, Amy E Sanders, Charles Hall, and Richard B Lipton. Free and Cued Selective Reminding Identifies Very Mild Dementia in Primary Care. 24(3), 2010.

[8] Daniel Davis, Sam Creavin, Jennifer Yip, Anna Noel-Storr, Carol Brayne, and Sarah Cullum. Montreal Cognitive Assessment for the diagnosis of Alzheimer's disease and other dementias. *Cochrane Database of Systematic Reviews*, (10), 2015.

[9] Peter J Nestor, Tim D Fryer, and John R Hodges. Declarative memory impairments in Alzheimer ' s disease and semantic dementia. 30:1010–1020, 2006.

[10] Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461, 2011.

[11] 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.

[12] 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.

[13] Olga B. Emery. Language Impairment in Dementia of the Alzheimer's Type: A Hierarchical Decline? *International Journal of Psychiatry in Medicine*, 30(2):145–164, jul 2000.

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

[15] Vanessa Taler and Natalie A. Phillips. Language performance in Alzheimer's disease and mild cognitive impairment: A comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008.

[16] Majid Komeili, Chloé Pou-Prom, Daniyal Liaqat, Kathleen C Fraser, Maria Yancheva, and Frank Rudzicz. Talk2Me: Automated linguistic data collection for personal assessment. *PLOS ONE*, 14(3):e0212342, mar 2019.