Thesis

Jomar Alcantara

2018-09-17

0.1 Abstract

0.2 Acknowledgements

Contents

List of Figures

List of Tables

Chapter 1

Introduction

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

Introduction to the problem of dementia in the context of the wider world including quality of life and financial implications. Exploration of dementia as a syndrome rather than a disease, and a look at the different variants of dementia. A look at the rationale behind research into the early diagnosis of dementia as well as a brief look at what has been done in the area (wide context, so pharmacological and psychological).

Chapter 2

Background

2.1 Introduction

Alzheimer's disease 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 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 [**?**]. 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 and that number is expected to hit 131.5 million by 2050 [**?**]. The report also states that the worldwide cost of AD in 2018 is estimated to be in the region of one trillon US dollars.

At present there are no drugs that improve the prognosis of those suffering with AD, all the drugs that are on the market are designed to manage symptoms. Whilst there are numerous investigational drugs 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) [**?**]. A possible reason for the lack of success is that the drugs treatments are initiated too far along in the progression of the disease and thus much of the degeneration of the brain has already taken place. Research has therefore been aimed at earlier diagnosis.

Alzheimer's Disease (AD) is a neurodegenerative disease in which, from a physiological perspective, the brain develops neurofibrillary tangles and neuritic plaques along with the deterioration and loss of cortical neurons and synapses. However, a definitive diagnosis of Dementia can only produced at post-mortem. From a clinical psychology perspective, those who have dementia have cognitive deficits such as problems with episodic and semantic memory, organising and planning, difficulties with language and visuospatial deficits[**?**]. In addition,

these symptoms are often accompanied by emotional difficulties such as depression and irritability and behavioural difficulties.

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. [**?**]. The case of Iris Murdoch illustrates this theory well. Le et al [**?**]found, in their analysis of three writers and the novels they wrote, that Iris Murdoch's work declined subtly over time, but there was a steep drop off in the use of language in her last novel. If this theory holds true more generally, it should be possible to detect subtle cognitive changes in language and memory before a clinical diagnosis can be formed.

One of the most common ways in which clinicians traditionally make an early diagnosis of cognitive impairment is through the use of the Mini Mental State Examination (MMSE) [**?**]. 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 [**?**]. 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 [**?**]. Whilst the MMSE is useful as a brief screening tool it has it's limitations. The MMSE was not specifically created to screen for dementias and therefore does not interrogate key aspects of cognitive impairment known to be affected in dementia. It also has limited value in assessing under educated subjects and a meta-analysis on the effectiveness of the MMSE as a diagnostic tool for dementia to showed that it's accuracy was low (sensitivity between 78.4% and 85.1% and specificity between 81.3% and 87.8%)

As the MMSE is shown to have low accuracy specifically in the diagnosis of dementia, it becomes necessary for professionals to employ the use of other tools or measures such as the Free Cued Selective Reminding Test (FCSRT) or the Montreal Cognitive Assessment. These tests have the benefits of being much more accurate at diagnosing cognitive impairment and discriminating between dementia and other types of cognitive impairment at the cost time and training of psychological professionals such as clinical psychologists in administering these tests. Given the burden on these professionals is likely to increase due to a growing elderly population or in the case of developing countries where the clinician / patient ratio is already unsustainable, it seems useful to find less burdensome ways do aid professionals in diagnosing dementia.

The two main ways in which diagnosis is performed is through assessment of memory and language. Tests of memory are classically among the most accurate ways of diagnosing dementia, however they 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 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 consensus is that this decline is typically accelerated by the presence of dementia [**?**]. Given this statement, it seems logical to conclude that one of the key variables that distinguishes language decline in healthy individuals vs those with dementia is the rate of change in which the decline occurs.

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

One of the most famous pieces of research on the topic of language decline in dementia was by Berisha and Liss (2015) [**?**] who examined speeches and public interviews of former US president Ronald Reagan. They found that Reagan's speeches towards the end of his presidency suffered from difficulties in word-finding, inappropriate phrases and uncorrected sentences which are hallmarks of language deterioration associated with Alzheimer's Disease. It turned out later to be the case that he had Alzheimer's Disease. Another classical study by Snowdon et al (1996) [**?**] looked at whether linguistic ability in early life was associated with cognitive function and AD in later life. They found that idea density (defined as the number of expressed propositions divided by the number of words) was a key predictor in predicting whether nuns would go on to develop AD in later life. They found that those who would go on to develop AD all had low idea density in early life and they found no AD present in those with high idea density in early life. As we can see, just with these two pieces of research the range of language deficits in those who suffer with AD are wide and varied and differ as the disease progresses.

Emery [**?**] completed a literature review looking at all the potential language deficits that could exist in those with AD and / or MCI. She divided these deficits into four levels of language: Phonology, Morphology, Syntax and Semantics. She proposed that language and the processes involved in language are hierarchical in nature and that language moves from simple units of construction (Phonology and Morphology), and build layers of complexity and sophistication (Syntax and Semantics). She found that people with AD generally had intact Phonology and Morphology but more impaired Syntax and Semantics. She stated that the language forms we learn last are the first to deteriorate. We generally learn

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

Table 2.1: Search Terms and Number of Results.

language in small simple units initially and build syntax and complexity as we are more comfortable with language.

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

2.1.1 Aims and Methodology

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

A search of the literature was conducted using ProQuest (PsychArticles), SCOPUS, Web of Science. The following results were found (Table 1). All papers were then reviewed for relevance by reading the abstract and full text where appropriate and a shortlist was compiled. An additional search through references of shortlisted papers was also conducted and any papers who upon further review appeared relevant were added to the shortlist. Papers were included where researchers used machine learning to classify participants as MCI, AD or Healthy using language. We excluded any papers that focused on other forms of dementia or cognitive impairment, as well as any papers in which the language being analysed was not English. This resulted in 17 journal articles and/or conference papers which form this review.

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 [**?**]. In this task participants are asked to describe a picture presented to them in as much detail as possible. The picture itself depicts a familiar domestic scene and would not require participants to use any vocabulary beyond that learned in childhood. It was originally designed to assess Aphasia, but has shown itself to be useful in the assessment of language for the purposes of diagnosis of MCI and AD as well [**?**]

The picture description task does a fine job of eliciting descriptive language, however because of the limited content has limited use. The task in itself just a descriptive task, and therefore elicits a certain type of language. There is some disagreement as to the benefits of this using this methodology. This task is reported as being useful to lexico-semantic disorders [**?**, **?**] 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.

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.

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 such as Cinderella. [**?**] This task requires ordering the story in a structured and coherent framework. It also requires comprehension and understanding of the stories characters and the events depicted, as well as an awareness of a character's goals and internal responses to given events. This task is particularly useful, as the procedure reduces the demands on memory and is therefore able rule out memory as a confounding variable for any results observed. As noted above, Ash [**?**] felt that this task was interchangable with the Picture Description task and because this task requires elaboration rather than simple description, is a

sturdier test of lexical and semantic abilities as well as syntactic complexity. [?]

Given the relative strengths of the Narrative description task vs Picture description task, there are few pieces of research research that have used Machine Learning to analyse features from Narrative picture tasks [?]. This could be down 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, the idea of employing 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 [?]. 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 other forms of interview as a middle ground. The semi structured nature of these interviews means that there is some replicability but does not constrain the participant in answering questions.

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

2.3 How do we analyse language, issues and debates

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.

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.

2.4 Semantic Content

Another approach to linguistic analysis in this field is the idea of measuring semantic content and complexity. According to Emery (2000) [**?**] in which she states that Semantic and Syntactic skills deteriorate first in people with MCI and AD. If this is true, then psychological measures of semantic and syntactic skills should be able to pick up signs of deterioration and act as markers for possible MCI and AD. An example of these measures 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 [**?**]. Type to Token Ratio, Brunet's Index and Honore's Statistic are two more examples of validated measures which are used to measure the lexical diversity of a given piece of text. 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 [**?**, **?**].

2.4.1 Picture-related content thematic elements

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

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

Smith, Chenery and Murdoch (1989) applied Hier's methodology for constructing pictorial 'themes' with the Picnic Scene from the Western Aphasia

Battery (WAB) with a control and patients with moderate to moderately severe AD. The authors found no difference in the number of semantic elements produced but did not that the group with moderate to moderately severe AD took more time and more syllables to communicate these elements.

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

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

2.4.2 General Information Units

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

2.4.3 Conciseness of information

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

Ahmed, de Jager et al examined idea density with patients who had confirmed AD post mortem. They found that those with AD produced fewer total semantic units than controls but there was no significant difference between the groups with regards to idea density. The study of "empty speech" by Nicholas et al examined conciseness with measures thought to contribute to the "non-specifity" of discourse in AD, such as empty phrases (defined as common idioms

contributing no relevant information), deictic terms (e.g. "this", "that" without referents), indefinite terms (e.g. "thing" or "stuff"), pronouns without proper noun antecedants, and repetitions. In their study they found that AD patients produced more of these behaviours than did controls.

2.4.4 Efficiency

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

2.4.5 Lexical richness and diversity

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).

$$TTR = numberOfUniqueWords/totalNumberOfWords. \tag{2.1}$$

Brunet's Index(W)

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

$$W = N^{V(-0.165)} \tag{2.2}$$

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 (R)

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

$$R = (100 \log N)/(1 - V1/V) \tag{2.3}$$

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

2.4.6 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.

2.5 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.

2.5.1 N-grams

One of the first features discussed as a potential predictor of MCI or AD is the n-gram. An n-gram is a contiguous sequence of n items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. For example, given the sequence of words "to be or not to be", this extract is said to contain six 1-gram sequences (to, be, or, not, to, be), five 2-gram sequences (to be, be or, or not, not to, to be), four 3-gram sequences(to be or, be or not, or not to, not to be) and so on. This is useful as, given a large portion of text or speech, we can predict the probability of a word being close by to a given word. A number of researchers have used n-grams as features. One of the first attempts to use machine learning

and natural language techniques to look was conducted by Thomas [?] 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.

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%.

2.5.2 Mean length of utterance (MLU)

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

2.5.3 Proportion of verbs to nouns plus verbs

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

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

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

2.5.5 Formulaic Language

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

the classification of participants as either healthy or AD. These four factors were semantic impairment, acoustic abnormality, syntactic impairment and information impairment and were based on existing measures of semantic and syntactic complexity. Zimmerer (2016) [**?**] 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.

2.6 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.

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

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

2.6.2 Coherence

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

representation, or failure to access frame knowledge.

2.6.3 Perseveration

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

2.6.4 Speech intonation / prosody

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

2.6.5 Discourse Fluency

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

The study of "empty speech" by Nicholas et al was one of the first to examine aspects of fluency in the connected speech of persons with AD. They found that adults with AD had significantly more repetitions than controls. Similarly, Tomoeda et al found more aborted phrases, revisions and ideational repetitions

in the AD group than in controls. Several other studies support the idea that in AD population there are a greater number of repetitions and revisions than in healthy controls. However, there are some studies which contradict these findings (Ahmed et al).

2.6.6 Speech Monitoring

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

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

2.7 State of literature into Natural Language processing techniques

As we can see, diagnosing dementia through analysis has a large background in terms of psychological research. However, a lot of more current research has called for the use of machine learning as a way of assisting in the process of diagnosis [**?**].

2.7.1 Traditional methods of Machine Learning and Natural Language Processing

2.7.2 Deep Learning

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

It is challenging to characterise Dementia as a disease as it's more syndrome. The clinical features that combine to meet the diagnostic criteria for Dementia

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

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

2.9 Conclusions and Future Work

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

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

One of the areas for research to study is a careful examination of the features that are being used to measure language deterioration. For example, Zimmerer (2016) [**?**] 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 conceptualise as part of the larger problem. Some work needs to be done in producing a consistent list of measures that are validated using existing datasets and can be used for future research moving forward.

2.9.1 Future Work

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

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

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

Chapter 3

Research Questions and Hypotheses

The broad question is can we use Machine Learning to assist in the diagnosis of dementia through natural language. There have been a number of studies from the psychological perspective that look at the changes in language in those with Dementia. These

Chapter 4

Experiment: Presidential Debates revisited

4.1 Background

Aim: To replicate and extend the work by Berisha and Liss. To use traditional machine learning techniques to test whether they can differentiate between Reagan, Bush and Trump using similar features. To test to see whether Deep Learning can also differentiate between the three.

4.2 Data

Dataset: the presidential press conferences given by Presidents Ronald Reagan, George H.W. Bush and Donald Trump.
Paper: This experiment replicates work done by Berisha and Liss and extends this by adding Donald Trump as an alternative comparison to Ronald Reagan. This experiment will look at the features originally recommended by Berisha and Liss, as well as any others that have potential as discussed in the literature review above.

Chapter 5

Experiment: Three Authors - Murdoch, Christie and James

5.1 Background

5.2 Data

Aim: To replicate the work by Le et al. To use traditional machine learning techniques to see if it can differentiate between the three authors using similar features. To test to see whether deep learning can also differentiate between the three. To add 'pre-trained' layers from Presidential Debates experiment to see if this improves accuracy.
Dataset is the literary novels of Iris Murdoch, Agatha Christie and P.D. James.
Paper:

Chapter 6

Experiment: Using the DementiaBank Corpus

Aim: To replicate and extend the work of Orimaye et al. To use traditional machine learning techniques to see if it can categorise people into dementia or healthy categories using the samples given. To see whether can use pre-trained layer from presidential debates to improve accuracy. It doesn't seem relevant to use pretrained layer from Experiment four, but it might be worth looking at the results of this.
Dataset: The DementiaBank Corpus.
Paper: To replicate and extend the work of Orimaye et al. To compare traditional machine learning techniques (already completed by Orimaye) and deep learning.

Chapter 7

Experiment: Longitudinal Data - Current Cohort

Aim:
Dataset: To be created.

7.0.1 Dataset creation

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

Patient Health Questionnaire (PHQ-9)

Free Cued Selective Reminding Test (FCSRT)

Mini Mental State Examination(MMSE)

Written Description (Cookie Theft Picture)

Semi-Structured Interview

Paper:

7.0.2 Ethics

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

7.0.3 What if someone reveals suicidal ideation or other risks?

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

7.0.4 What happens at the end of the data collection phase if there is a suspected decline in a participant's cognitive function

Chapter 8

Overall results and Discussion

Chapter 9

Conclusions and Future work