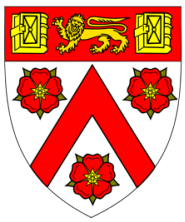
Automatic Detection of Early Signs of Alzheimer's Disease in Speech and Language



Ulla Meeri Petti

Supervisor: **Prof. Anna Korhonen**

Co-supervisor: **Dr. Simon Baker**

Theoretical and Applied Linguistics

University of Cambridge

This dissertation is submitted for a degree of

Doctor of Philosophy

Trinity College March 2024

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Automatic Detection of Early Signs of Alzheimer's Disease in Speech and Language

Ulla Meeri Petti

Abstract

The relevance of Alzheimer's disease (AD) is growing due to ageing population, and an increasing amount of research is being conducted in both treatment development and early detection of the disease. Previous research has shown that changes in language use could be one of the earliest signs of cognitive decline in AD, and these changes could be automatically detected using natural language processing (NLP) and artificial intelligence (AI). While NLP- and AI-based tools could contribute to detecting AD early in a non-invasive, fast, cheap, and accessible way, there is still a major gap between the scientific knowledge and its applicability to clinical practice.

In the current thesis, I first conducted a systematic literature review of the studies looking at automatic speech-based AD detection and identified the key challenges in the state-of-the-art: (1) the lack of longitudinal language data; (2) the lack of replicability, generalisability, and standardisation; and (3) the lack of ethical guidelines. To tackle these issues, I first present a novel corpus of longitudinal transcripts of interview recordings with public figures, and demonstrate the usefulness of this kind of data in understanding longitudinal language changes in AD. Second, I replicate a previous case study on a larger group of individuals, explore the generalisability of the language change, and identify the most informative language features that change consistently across individual speakers. Third, I focus on the standardisation of data collection methods and investigate the role of sample length in analysing AD-related language change. Fourth, I outline the ethical considerations in AI- and NLP-based AD detection from speech and language and provide a list of suggestions that could be incorporated in the development of ethical guidelines and best practices.

This work aims to address some of the main challenges in automatic speech-based AD detection and fill the gaps in the existing literature to contribute to developing robust, fair, and ethical methods for the automatic detection of early signs of AD in speech and language.

Acknowledgements

I wish to thank my supervisor Professor Anna Korhonen for guiding me throughout the PhD and supporting me in following my interests.

I dedicate this thesis to Urmas. I wish you could see where your encouragement at the early stages of developing an interest in linguistics over a decade ago has taken me. There have been countless times where I wish I could share the joys and obstacles of the journey with you, but our early conversations have stayed with me throughout my studies. I am humbled and honoured to join the publishing world sharing your author initials!

On a more upbeat note, I have been incredibly lucky to share the PhD journey with many amazing people! I wish to thank Simon for all the help, goodwill, and patience. I thank Jessica and Winterlight Labs for an unforgettable internship opportunity and time in Toronto. I thank Rune and Jeffrey for inspiring collaboration and discussions that have directed my research and interests. I wish to thank my fellow PhDs and colleagues for endless research and life chats – Alex, Nat, Steph, Lina, Genie, Andromachi, Marju, Language Technology Lab, and Centre for Human-Inspired AI Early Career Community.

I am sincerely thankful to my family and friends. Johan, thank you for being next to me from high school to PhD, baring the ups and downs of the close-up of the grind, and believing in me - I would never be writing the acknowledgements section of my PhD thesis without you.

I thank my mum Katrin for endless love and support, and my friends who have got me through every PhD-breakdown and reminded me that there is “not a cloud in the sky”. Special thanks to Martha, Ten, Jim, Pippa, Matt, Lui, Sasha, Jonners, Meredith, Andy – your kindness, warmth and encouragement at key moments has meant the world to me!

And of course, thank you to Cambridge University Real Tennis Club - my second, third, or often first home. Special thanks to Kees and Peter!

Table of Contents

List of figures

List of tables

1. Introduction

This thesis explores the automatic detection of early signs of Alzheimer's disease (AD) in speech and language using natural language processing (NLP) and artificial intelligence (AI).
In this Introduction section, I will (1) briefly summarise what is AD, (2) discuss the challenges and benefits of the early detection of the condition, (3) summarise how language changes in AD, (4) recap the methods and tools that can be used to detect these changes, and (5) outline the ethical considerations that arise in the process. I will also provide (6) the motivation behind writing this thesis, (7) the contributions this work brings to the field, and (8) outline the chapters and the structure of the thesis.

1.1. What is Alzheimer's disease?

Alzheimer's disease (AD) is a progressive neurodegenerative disease, contributing to 60-70% of dementia cases (WHO, 2023), making it the main cause for dementia (Scheltens et al., 2021). Dementia affects around 50 million people worldwide, and due to the ageing population, the number of individuals suffering from dementia is expected to triple by 2050 (WHO, 2023). AD has been considered one of the greatest healthcare challenges and the most burdening, expensive, and lethal disease of the current century (Scheltens et al., 2016; Scheltens et al., 2021).
Clinical manifestation of AD appears in the loss of synapses and synaptic plasticity, and these changes are parallel to the cognitive decline (Lopez et al., 2019). Some of the cognitive and functional abilities that are affected in AD include memory (Kramer et al., 2004), language and speech (de Lira et al., 2018), executive and visuospatial functioning and attention (Scheltens et al., 2017) and intellectual function (Casserly & Topol, 2009). In the first stages of the disease, the individuals can experience difficulties in encoding and storing new memories; in the later stages, progressive changes in cognition and behaviour appear (Lopez et al., 2019). A more detailed description of AD will be provided in the Background chapter (section 2.1).

1.2. Early detection of Alzheimer's disease

Early detection of dementia and AD can be challenging for several reasons: the symptoms often overlap with normal ageing, clear manifestations do not often appear until several years after onset, diagnosing can be costly and time-consuming and require access to a qualified clinician (Casserly & Toppol., 2009). This contributes to a large number of undiagnosed dementia cases, for example, Prince and colleagues (2013) estimate that 55% of dementia cases in the US remain undiagnosed, and this number is believed to be even bigger in less developed countries (Sadeghian et al., 2021), illustrating the need for cheap, accessible, robust and scalable screening tools.

Although there is currently no cure for AD, recent research has reported cautious positive effects of disease-modifying treatment (Scheltens et al., 2016). For example, a recent study on lecanemab shows some slowing of cognitive decline (van Dyck et al., 2023). Similarly, pharmacological interventions, such as donepezil, rivastigmine patch, or galantamine have shown benefits in the early and moderate stages of the disease, with less clear impact in the later stages (Lopez et al., 2019), advocating for the importance of early detection. Early detection could also allow introducing preventative lifestyle changes, accessing services that could help with adjusting to the condition, lessen the economic and emotional burden of the patients and their families, and empower the affected individuals to make decisions about the future as self-determining agents (Calza et al., 2021; Schictanz et al., 2014; Alzheimer's Association, 2008; Tanaka et al., 2017; Brock, 1993; Porteri et al., 2017; Mattson et al., 2010; Markesbery et al., 2006).

Typical diagnosis tools that are used for AD include blood tests, magnetic resonance imaging (MRI) and positron emission tomography (PET) scans. Blood tests are cost-effective and can be conducted at primary care clinics, however, they are invasive. MRI is non-invasive, but both MRI and PET are costly cannot be conducted at a primary care clinic. Numerous studies suggest that changes in language are among the first manifestations of cognitive decline in AD (Calza et al., 2021; Fang et al., 2017; Yancheva et al., 2015) and could therefore act as early biomarkers of the condition (Luz et al., 2018; Fraser et al., 2016; Satt et al., 2013). In comparison to blood tests, MRI and PET scans, automating language processing could provide a non-invasive and

cost-effective approach to detecting signs of AD, which could benefit clinicians during in-hospital screenings. While these technologies would be useful, they are still in the development stage and are not yet in clinical use. A more detailed overview of the state-of-the-art is provided in the Background section (section 2.2).

1.3. How does language change in Alzheimer's disease?

Language changes in AD interact with changes in other cognitive domains such as attention and memory. Memory impairment typical in AD contributes to many of the changes in language use, for example, individuals with AD often experience difficulties in retrieving words (Appell et al., 1982; Bayles, 1982; Obler, 1983), which can manifest in challenges in verbal naming (Silagi et al., 2015; Lopez et al., 2019), affecting accurate meaning communication (Mirheidari et al., 2017), pausation, speech tempo, fluency and acoustics (Meilan et al., 2012; Pistono et al., 2016; Gosztolya et al., 2016; Hoffmann et al., 2010), lexical diversity (Critchley, 1964; de Ajuriaguerra & Tissot, 1975; Stengel, 1964) and speech content density and quantity (de Lira et al., 2018). Memory and attention deficit also contributes to the tendency to repeat words and concepts (Lopez et al., 2019) which can result in communication errors, lower coherence and information density (Gosztolya et al., 2019). In addition to memory deficit, semantic dysfunction is another major factor contributing to the changes in language in AD (Forbes-McKay & Venneri, 2005; Venneri et al., 2018; Rodriguez-Aranda et al., 2016; Garrard & Carroll, 2006; Hamberger et al., 1995). It has been proposed that the impaired storage of information, together with naming difficulties and semantic dysfunction could help discriminate AD from other overlapping conditions (Weintraub et al., 2012; Rascovsky et al., 2007; Lopez et al., 2019). AD-related changes have also been reported in syntactic complexity (Ahmed et al., 2013; Garrard et al., 2005a) and discourse (Hodges et al., 1992; Lima et al., 2014).

1.4. How can language change in Alzheimer's disease be detected?

While until recently language data was analysed manually, the development of technology has enabled automating the analysis. Automation promotes the inclusion of more data and more detailed analysis revealing patterns that may go unrecognised in manual analysis. Promising

results in AD detection have been achieved using natural language processing (NLP), signal processing (SP), and machine learning (ML). NLP is concerned with understanding, learning, and producing human language using computational tools (Hirschberg & Manning, 2015). SP explores signals and the information they convey and is concerned with how they can be transformed, manipulated, and represented (Oppenheim et al., 1998). ML focuses on the questions concerned with constructing computer programs that can improve automatically based on experience (Mitchell, 1997).

Typical steps in automated language analysis include recording an individual's speech (normally triggered by casual conversation, picture description or recall tasks, verbal fluency tasks or other language tests), transcription (for linguistic analysis) and feature extraction. The choice and extraction of language features has mostly relied on clinical research, using the knowledge of language changes in AD described in section 1.3, and the features have been manually engineered based on domain expertise. However, fully automated analysis could also base the features on data and performance and use all the evidence in language. Figure 1 summarises some of the language and speech features that have been used in previous studies, and links them to the linguistic domain and the underlying AD-related deficit that makes the changes in these features informative. The extracted features can then be analysed depending on the research question. Most studies have focussed on classification tasks, which have achieved over 90% accuracy (Petti et al., 2020), however, the current research direction is moving towards detecting the earliest signs, tracking longitudinal changes and response to treatment. Although challenged by lack of data, automated analysis shows promising results (Luz et al., 2018; Fraser et al., 2016; Satt et al., 2013).

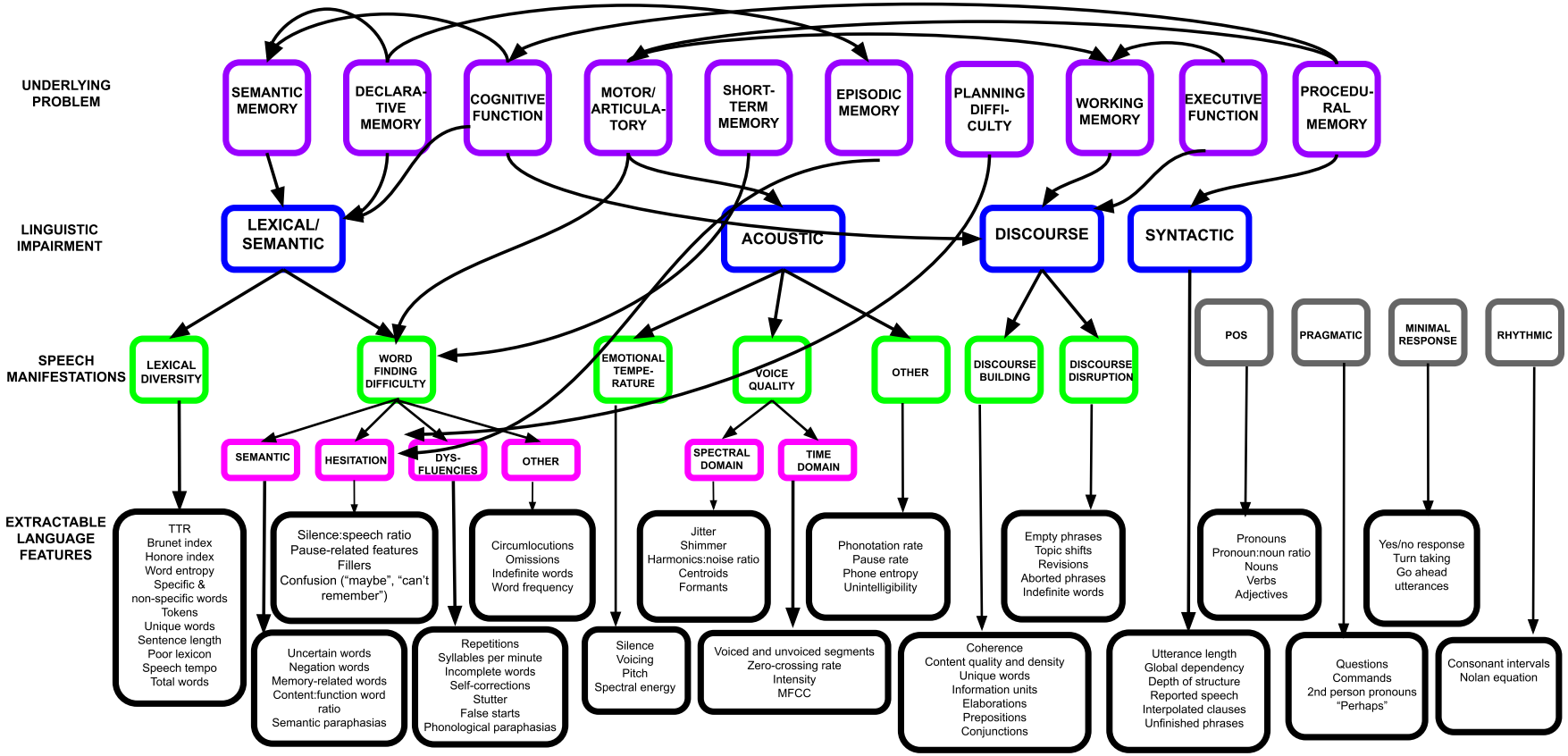


Figure 1: Examples of the language and speech features extracted from previous studies, and their connection to the underlying AD-related problem *(POS: part-of-speech; TTR: type:token ratio; MFCC: mel-frequency cepstral coefficient)*

1.5. Ethical considerations

While the use of speech data, NLP- and AI-based tools has a great potential to contribute to detecting the early signs of AD, numerous ethical concerns that arise in the process must be addressed before practical application of this technology. The concerns can arise on different levels and at different stages of the process. For the start, any research involving human subjects must be based on informed consent, but how will AD affect an individual's ability to understand the research and give consent? Additionally, cognitive testing and producing language can be burdensome to an individual experiencing difficulties in the proposed tasks. It is also important to recognise that the capacity of language tests can be limited, potentially insensitive to cultural differences, and blind to the real challenges the individuals with AD experience in everyday conversational situations. This can paint a skewed picture of the speaker's abilities, which can be both depersonalising and compromise the accuracy of the research outcome. Accuracy can also be affected by small sample size, inconsistent and sparse data, which are common issues in AD-related speech research. Accurate results, however, are extremely important when working in a clinical domain, as false outcomes can lead to otherwise preventable harm, such as wrong or denied treatment, missed interventions, or refused access to services, and cause emotional burden that can have devastating impact on the affected individual. Even when the outcome of the test is accurate, ethical issues related to communicating the results must be considered. When working with speech data, it must also be acknowledged that speech can carry personal content and act as an identifier, which is especially relevant when working with sensitive medical information such as cognitive decline. It is also necessary to consider the potential misuse of the technology being developed – private use of automatic speech-based AD-detection tools may lead to risk of discrimination in employment opportunities, health insurance or legal status, which is why this technology should be aimed for clinical use. Clinician input should also be included in model development to ensure interpretability of the models and language features. In addition, it is also important to consider the equal distribution of research benefits and acknowledge potential biases. AD disproportionately impacts minority populations, highlighting the need to include these

populations in the research on early detection. Ethical considerations are addressed in more detail in Chapter 6.

1.6. Motivations

There are several motivations for writing this thesis. Firstly, while there are many studies focussing on speech-based classification between the individuals with and without AD, very few studies focus on long-term longitudinal decline and exploring the earliest markers of the condition, and the studies that do tend to be based on individual cases. This is largely due to limited availability of longitudinal recordings. I aim to tackle this issue by creating a novel corpus consisting of decades of interview recordings with 20 famous individuals, half of whom will eventually be diagnosed with AD, and explore the potential earliest signs of AD-related change, as well as their generalisability across a larger group of speakers. Detecting these changes early could contribute to timely intervention and help start the treatment early, which, considering the ageing population, could be beneficial to many. In addition, there is currently little standardisation in the data collection and analysis methods in the studies focussing on speech and language changes in AD. I aim to tackle this by exploring the optimal length of the recordings needed for informative and reliable analysis, considering the potentially burdensome process of producing spontaneous speech as well as the computation and analysis time. Another motivation for part of the thesis is the lack of ethical guidelines and suggestions in this fast-developing and promising, but under-regulated area of research. While the research into speech-based early detection of AD using AI and automated analysis could promote the screening process and positively impact the affected individuals, there are numerous ethical considerations that arise and cannot go overlooked. As part of the thesis, I provide a list of suggestions that could be incorporated into ethical guidelines for researchers and clinicians working in this area.

1.7. Contributions

The novel contribution of this thesis include:

- a systematic literature review summarising the state-of-the-art of automatic language-based AD detection;
- a longitudinal corpus of spontaneous speech transcripts from famous individuals, half of whom will eventually be diagnosed with AD;
- a demonstration that this kind of data can provide useful information in tracking AD-related changes in language;
- an identification of language features that change similarly in many individuals who will eventually develop AD;
- an examination of the impact of sample length and the robustness of the language features, informing the standardisation of data collection methods;
- a list of recommendations that could aid the development of ethical guidelines for research in speech-based and AI-aided AD detection.

1.8. Chapters

The Introduction will be followed by a Background chapter, four main chapters and Conclusions. In the Background chapter, I will discuss AD in more detail and provide a systematic literature review on the state-of-the-art of automatic AD detection from speech and language published in JAMIA, as well as give an overview of the impact it has had on the direction of the current thesis and the developments in the area since the publication of the article. In Chapter 3, I will present the corpus I collected and show how this corpus can be useful in examining signs of cognitive decline. In Chapter 4, I will focus on the generalisability of the AD-related longitudinal changes in language and identify the language features that change the most consistently across participants with AD. In Chapter 5, I will address the lack of standardisation in data collection methods, and investigate the impact of the length of the speech sample on the language features and the results, with an aim to establish how much speech data is needed for informative analysis of language changes in AD. In Chapter 6, I will focus on the ethical issues that arise in detecting early signs of AD using speech and AI, and

provide a list of suggestions for ethical research and tool development. These chapters will be followed by the Conclusions chapter.

The chapters are based on and use materials from the following articles published during the PhD:

- Petti, U., Baker, S., & Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, *27*(11), 1784-1797. (Chapter 2)
- Petti, U. & Korhonen, A. (2024). LoSST-AD: A Longitudinal Corpus for Tracking Alzheimer's Disease Related Changes in Spontaneous Speech. (Accepted to LREC-Coling 2024) (Chapter 3)
- Petti, U., Baker, S., Korhonen, A., & Robin, J. (2023a). The Generalizability of Longitudinal Changes in Speech Before Alzheimer's Disease Diagnosis. *Journal of Alzheimer's Disease*, 1-18. (Chapter 4)
- Petti, U., Baker, S., Korhonen, A., & Robin, J. (2023b). How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples. *Digital Biomarkers*, *7*(1), 157. (Chapter 5)
- Petti, U., Nyrup, R., Skopek, J. M., & Korhonen, A. (2023c). Ethical considerations in the early detection of Alzheimer's disease using speech and AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1062-1075). (Chapter 6)

I have carried out and led the work in the co-authored papers included in the thesis, with other co-authors providing only advice, with the following exceptions. Simon Baker acted as a second reviewer in the selection of studies in the systematic literature review in section 2.2, and Winterlight Labs employees performed the cutting of the audio and feature extraction in Petti et al. (2023b) that the analysis in Chapter 5 is based on. All contributions have also been stated in the text. I have run all the experiments and conducted all the analysis reported in this thesis.

2. Background and systematic literature review

In this chapter, I will introduce the previous literature on automatic AD detection from speech and language. This chapter includes three sub-sections: in the first subsection, I will give a more detailed overview of AD; the second subsection is based on a published systematic literature review conducted over the first year of the PhD (Petti et al., 2020); in the third subsection, I will discuss the impacts of the literature review: how it has informed the directions taken in the current thesis as well as the impact it has had on further research. A summary of the chapter will follow the three sub-sections. My own research reported in this chapter has been published in the Journal of American Medical Informatics (JAMIA).

2.1. Alzheimer's disease

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that impacts memory, cognitive, mental, and functional abilities, and eventually results in death (Srivastava et al., 2021; Alzheimer's Association, 2019; DeTure et al., 2019). It was named after Alois Alzheimer who first described the condition in 1906 (Lopez et al., 2019; Breijyeh & Karaman, 2020). AD is the most common cause of neurodegenerative dementia, accounting for approximately 60-70% of all dementia cases and it has been recognised as a global public health priority by WHO (Srivastava et al., 2021; Lopez et al., 2019; DeTure et al., 2019). Other common causes of dementia include conditions like Parkinson's disease with dementia, Lewy body dementia, frontotemporal dementia, and vascular dementia, which all make up around 5-10% of dementia cases (DeTure et al., 2019).

AD is characterised by progressive cognitive decline caused by the degeneration of brain cells and the loss of synaptic plasticity (Breijyeh & Karaman, 2020; Lopez et al., 2019; Wenk, 2003). This affects the individual's independence, reasoning, behaviour, motor system, visuospatial orientation, and speech (Breijyeh & Karaman, 2020; DeTure et al., 2019). Individuals with AD experience loss of episodic memory which leads to repeating conversations and questions, deficit in storing and making new memories, and difficulties in remembering names of people and objects (Lopez et al., 2019). Neurodegenerative diseases that have shared characteristics with AD include vascular cognitive impairment, Lewy body dementia and frontotemporal

dementia (Srivastava et al., 2021). While the individuals with AD also often suffer from other neurodegenerative diseases (DeTure et al., 2019), it has been proposed that difficulties with naming people and objects, semantic fluency, and storing information can help distinguish AD from other conditions (Lopez et al., 2019).

Breijyeh and Karaman (2020) describe the following 4 phases of AD: (1) pre-clinical or pre-symptomatic stage, (2) mild or early stage, (3) moderate stage, and (4) severe or late stage. AD has a long pre-clinical phase (DeTure et al., 2019), with neurodegeneration thought to start 20 or more years before clinical symptoms (Alzheimer's Association, 2019; Masters et al., 2015; Goedert & Spillantini, 2006). In this stage, the individual's daily activities and functioning are not affected, but mild memory loss and changes in the ability to store new memories can appear (Breijyeh & Karaman, 2020; Lopez et al., 2019). In the mild stages of AD, the individuals remain independent but can experience decline in executive function, followed by impaired language abilities and visuospatial skills (Kumar et al., 2022). Other symptoms in the mild phase can include disorientation, loss of concentration, mood changes and depression (Breijyeh & Karaman, 2020). With the progression of the disease, the changes in behaviour and the impairment of memory, cognition, social abilities, and motor skills become more severe (DeTure et al., 2019; Lopez et al., 2019). In the moderate stage of AD, individuals can experience difficulties with recognising their families, as well as with reading, writing, and speaking (Kumar et al., 2022; Breijyeh & Karaman, 2020). Other symptoms in this stage can include wandering, social withdrawal, apathy, and psychosis (Kumar et al., 2022). In the latest stage of AD, the patients may experience sleep disturbances and difficulties in motor tasks, followed by becoming bedridden and completely dependent on the caregiver, with difficulties with swallowing and urination, eventually leading to death (Kumar et al., 2022; Breijyeh & Karaman, 2020). The clinical phase of AD usually lasts for 8-10 years (Masters et al., 2015).

While the cause of AD-related pathological changes remains a topic of discussion and research, with no universally accepted theory (Breijyeh & Karaman, 2020; Lopez et al., 2019), there are two main hypotheses: the impairment of cholinergic function, and the changes in the production and processing of amyloid β-protein (Breijyeh & Karaman, 2020). While the exact cause of the condition is not fully understood, it has been found that the development of AD is

strongly dependent on heritable factors, such as changes in gene expression (Scheltens et al., 2021; Breijyeh & Karaman, 2020; Armstrong, 2019).

Age is the most important risk factor in AD (Breijyeh & Karaman, 2020; DeTure et al., 2019; Goedert & Spillantini, 2006). Based on the age at onset, AD can be divided into early-onset AD (EOAD) (before age 65) and late-onset AD (LOAD) (after age 65) (DeTure et al., 2019; Breijyeh & Karaman, 2020). EOAD is rare and makes up approximately 1-6% of the cases, while LOAD is more common (Breijyeh & Karaman, 2020). The risk of AD is higher in individuals over 65 years of age, growing year by year. In the US, 3% of the people aged 65-74 have AD, rising to 17% in the population aged 75-84, and 32% in the population older than 85 (Alzheimer's Association, 2019). AD disproportionately affects minority populations (Lopez et al., 2019) and women (Kumar et al., 2022; Srivastava et al., 2021).

Identifying other risk factors for AD has been an active area of research for decades. However, understanding the full picture remains a major challenge as there are many potential and unrelated factors associated with the condition (Armstrong, 2019). Some known risk factors include cardiovascular disease, head injuries, hypertension, infection, diabetes, environmental factors such as air pollution and exposure to certain metals, nutrition, and lifestyle (Srivastava et al., 2021; Breijyeh & Karaman, 2020). The risk of AD and cognitive decline could potentially be reduced by increased physical and cognitive activity, and healthy diet (Scheltens et al., 2019; Armstrong, 2019).

While Mini Mental State Examination (MMSE) (Folstein et al., 1975) and Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2011) are often used in screening to evaluate the patterns of cognitive decline in AD (Lopez et al., 2019), these tests can lack sensitivity and specificity (Kumar et al., 2022). For a more comprehensive examination of the disease stage, and distinguishing it for other conditions, a complete physical, neurological, and mental status examination is needed (Kumar et al., 2022). While the definitive diagnosis of clinical AD can only be completed post-mortem, the criteria for diagnosis in living patients has been developed over decades, and it is getting more accurate with the use of imaging studies, bodily fluids, and clinical biomarkers (Breijyeh & Karaman, 2020; DeTure et al., 2019). The novel promising biomarkers include PET scans and cerebrospinal fluid (CSF) for amyloid β and tau (Scheltens et

al., 2021; Breijyeh & Karaman, 2020; Masters et al., 2015). Active area of research is also focussed on identifying non-invasive biomarkers, such as those based on the changes in language use.

There is currently no cure for AD, although there are some available drugs that offer modest symptomatic treatment (Kumar et al., 2022; Srivastava et al., 2021; Lopez et al., 2019). There are several phase 3 trials for pharmacological treatment against amyloid β pathology, but their lack of success has sparked scepticism (Scheltens et al., 2021; Lei et al., 2021; Srivastava et al., 2021).

There is an eminent need to better understand AD pathology and design effective drugs as the number of people and families affected by the condition as well as the personal and financial costs of the disease are growing rapidly (Breijyeh & Karaman, 2020; DeTure et al., 2019). Research efforts are also focussed on early detection of the condition, as it increases the possibility of successful treatment (Breijyeh & Karaman, 2020; Lopez et al., 2019).

2.2. A systematic literature review of automatic Alzheimer's disease detection from speech and language

2.2.1. Introduction

To have a comprehensive overview of the state-of-the-art in automatic AD detection from speech and language and identify the main challenges and areas that needed improvement, I completed a systematic literature review of the available articles in this field (Petti et al., 2020). This systematic literature review was published in JAMIA in 2020 and at the time of writing this thesis has been cited over 140 times, informing both future research and the directions taken in this thesis.

In this review, I searched the main databases using relevant keywords, systematically selected the articles to review based on an established inclusion criteria and analysed them following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Moher et al., 2015).

The aim of this review was to answer five key questions:

- What were the characteristics of the participant groups involved in the studies?
- What type of language data was collected and how?
- Which were the most informative language and speech features?
- What classification methods were used?
- What classification performance was achieved?

These questions are useful to clinicians and researchers because they help to identify best practices, summarise the state-of-the-art in automatic language processing for AD detection, and guide further research. The current subsection is based on the published article (Petti et al., 2020).

2.2.2. Materials and methods

The search of the articles was conducted in the following databases: PubMed, Web of Science, and Ovid, using the keywords (1) automatic Alzheimer's disease detection, (2) Alzheimer natural language processing, (3) Alzheimer speech processing. All articles were published between 2013 and 2019 to allow capturing the most recent literature at the time and focussing on the time period where NLP, SP, and ML have been increasingly used in disease detection from speech and language. The developments since the publication of this article, from 2019-2023, are addressed in the following section (2.3).

The following inclusion criteria was established for all studies:

1) AD or mild cognitive impairment (MCI) was the condition of at least one of the participants (MCI has been described as the stage between normal ageing and dementia (Greenaway et al., 2006); examining MCI could contribute to detecting the earliest signs of AD (Schneider et al., 2009; Markesbery et al., 2006), however, not all people with MCI develop AD (Lezak, 2004));
2) participants' language or speech was considered;

3) there was either an NLP, SP, or ML element;
4) the focus was on language or speech production, and not comprehension;
5) experimental data was included;
6) full articles were available in English.

The initial study selection was performed by 1 reviewer (Ulla Petti). To minimise the bias in selecting studies, a sample of 274 articles consisting of a random sample of 10% of the articles excluded by the first reviewer (n=241), and all the articles included by the first reviewer (n=33) were independently reviewed by a second reviewer (Simon Baker). The initial overall agreement between the two reviewers was 97%, with 100% agreement on the 33 included articles. Remaining disagreement was resolved in a discussion with the third author (Anna Korhonen).

The following data relevant to the 5 research questions was extracted from all included articles: participant information, the type of language data and the language tests used, the most informative language and speech features, classification methods, and classification performance.

2.2.3. Results

The number of articles retrieved from the initial search was 2447. The flow diagram displayed in Figure 2 details the selection process that resulted in 33 included articles.

Out of 33 studies, 18 focussed on AD, 9 on both AD and MCI, and 6 solely on MCI. Twenty-eight studies focussed on spontaneous speech (SS), and 7 on both verbal fluency tasks (VF) and other tasks (OT). On average, 92 participants were included in the studies, with the number of participants ranging from 3 to 484. One study only reported the number of recordings (Luz et al., 2018) and all but 2 studies (Konig et al., 2018; Garrard et al., 2017) had a healthy control group. The average size of the control group was 43, ranging from 2 to 242 and the average size of the AD group was 45, the MCI group was 30, and the dementia group was 27. A large majority of the studies were conducted in European languages: 10 studies in English, 4 in French and Hungarian, 3 in Greek and Turkish, and 1 in Spanish and Italian. One study was

carried out with Taiwanese speakers, 5 studies used a dataset consisting of several languages, and 4 studies did not specify the language used. The number of studies grew year by year; 3 studies were published in 2013, 5 studies in 2014, 2015 and 2016, 6 studies in 2017, and 9 studies in 2018. This shows that research in the area is growing.

The information extracted from the 33 studies is summarised in Table 1.

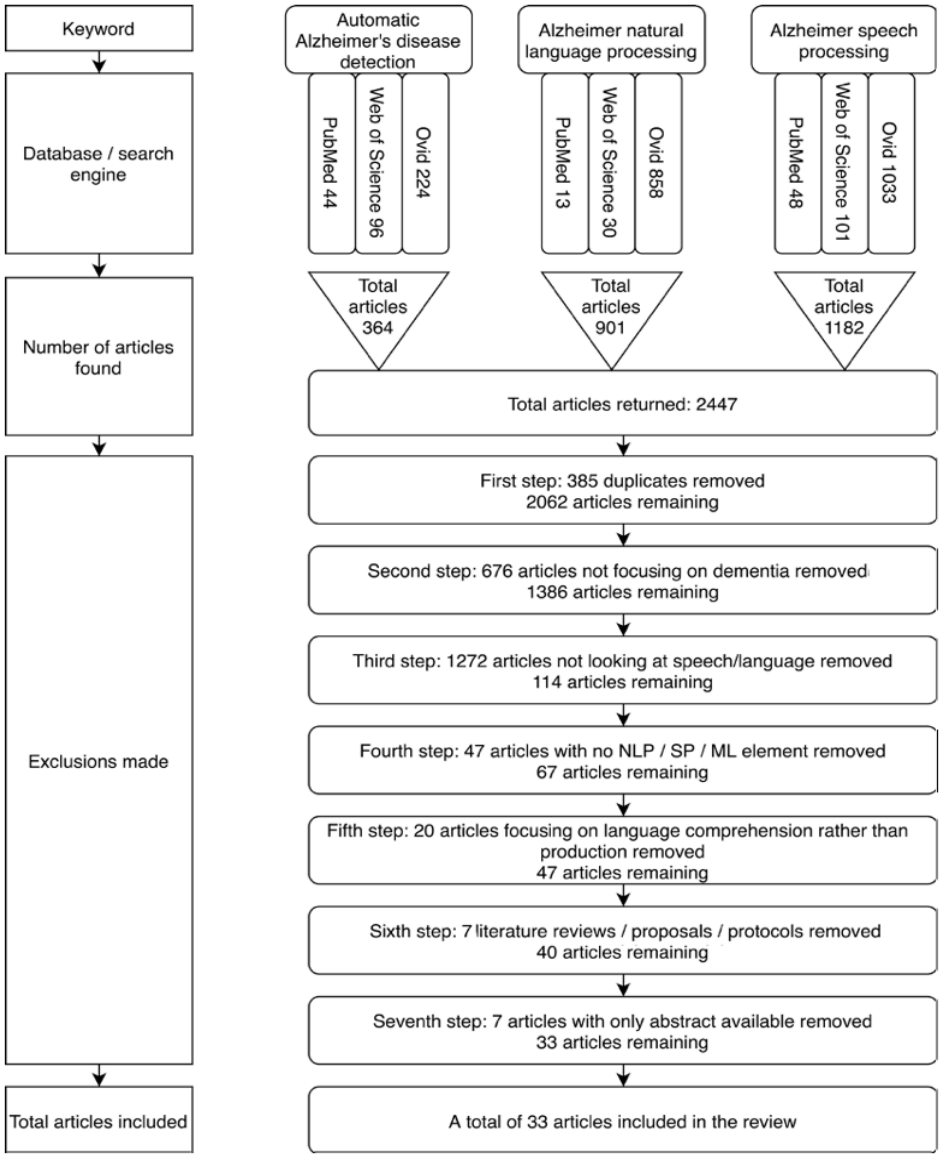


Figure 2: Flow diagram of study selection *(NLP: natural language processing; SP: signal processing; ML: machine learning)*

No	Study	Impaired group size (s)	Control group size	AD/MCI	Data collection method	Most informative language and speech features	Number of samples used to train the model	Classification algorithm	Classification performance
1	Ammer and Ayed 2018	AD n = 242	n = 242	AD	SS	Repetition, word errors, MLU morphemes, POS	–	SVM, NN, DT	Precision = 79%
2	Beltrami et al 2018	aMCI n = 16, mdMCI n = 16, early dementia n = 16	n = 48	MCI	SS	Acoustic, lexical, syntactic	–	–	–
3	Boye et al 2014	AD n = 5	n = 5	AD	SS	Lexical and semantic deficit, reduced conversation	–	–	–
4	Chien et al 2018	1) AD n = 15 2) AD n = 30	1) n = 15 2) n = 30	AD	SS	Speech length, non-silence tokens	150 samples from 60 participants	RNN	AUC = 0.956
5	Clark et al 2014	MCI n = 23, AD n = 10	n = 25	AD, MCI	SVF, PVF	Semantic similarity of words	–	–	–
6	Clark et al 2016	MCI-con n = 24, MCI-non n = 83	n = 51	MCI	SVF, PVF	Coherence, lexical frequency, graph theoretical measures	158	Random forest, SVM, NB, MLP	Acc = 81–84%
7	Fang et al 2017	MCI-con n = 1	n = 2	MCI	SS	Unique and specific words, grammatical complexity	–	–	–
8	Fraser et al 2016	AD n = 167	n = 97	AD	SS	Semantic, acoustic, syntactic, information	473 samples from, 264 participants	Logistic regression	Acc = 81%
9	Garrard et al 2017	Probable AD n = 5	n = 0	AD	SS	–	–	–	–
10	Gosztolya et al 2016	MCI = 48	n = 36	MCI	SS	Filled pauses	84	SVM	Acc = 88.1%
11	Gosztolya et al 2019	MCI n = 25, early AD n = 25	n = 25	AD, MCI	SS	Semantic, morphological, acoustic attributes	75	SVM	Acc = 86%
12	Guinn et al 2014	AD n = 28	n = 28	AD	SS	Ratio, POS, lexical, pauses, fillers	56	NB, DT	precision = 80%
13	Hernandez-Dominguez et al 2018	AD n = 169 MCI n = 19	n = 74	AD, MCI	SS	Information coverage, auxiliary verbs, hapax legomena	236 training, 26 testing	SVM, Random Forest	Acc = 87–94%
14	Khodabakhsh et al. 2014a	AD n = 27	n = 27	AD	SS	Log voicing ratio, average absolute delta formant and pitch	54	SVM, DT	Acc = 88–94%

Table 1: Information extracted from the studies *(continued)*

No	Study	Impaired group size (s)	Control group size	AD/MCI	Data collection method	Most informative language and speech features	Number of samples used to train the model	Classification algorithm	Classification performance
15	Khodabakhsh et al. 2014b	AD n = 20	n = 20	AD	SS	Fillers, unintelligibility, no of words, confusion, pause & no answer rate	40	SVM, DT	Acc = 90%
16	Khodabakhsh et al 2015	AD n=28	n=51	AD	SS	Ratio, POS, speech rate features	79	SVM, NN NB, CTree	Acc = 84%
17	Konig et al 2015	MCI n=23, AD n=26	n=15	AD, MCI	SS, SVF, OT	Speech continuity, ratio	64	SVM	EER = 13–21%
18	Konig et al 2018	AD n=27, mixed dementia n=38, MCI n=44, SCI n=56	n=0	AD, MCI	SS, SVF, PVF, OT	Location of first word, words' distribution in time	165	SVM	Acc = 86%
19	Lopez-de-Ipina et al 2013a	AD n = 20	n = 20	AD	SS	Impoverished vocabulary, limited replies	40	MLP	Acc = 75–94.6%
20	Lopez-de-Ipina et al 2013b	Early AD n = 1 Intermediate AD n = 2 Advanced AD n = 2	n = 5	AD	SS	Fluency, acoustic	10	SVM, MLP, kNN, DT, NB	Acc = 93.79%
21	Lopez-de-Ipina et al 2015	AD n = 20	n = 20	AD	SS	Duration, time, frequency	40	MLP, KNN	Acc = 95%
22	Lopez-de-Ipina et al 2018	1) AD n = 6, 2) AD n = 20, 3) MCI n = 38	1) n = 12, 2) n = 20 3) n = 62	AD, MCI	SS, SVF	Voicing, pauses, F0, harmonicity	18, 40, 100	MLP, CNN	Acc = 73–95%
23	Luz 2018	Nr of recordings reported AD n = 214 recordings	n = 184 recordings	AD	SS	Vocalisation, speech rate, number of utterances across discourse event	398	NB	Acc = 68%
24	Martinez de Lizarduy et al 2017	1) AD n = 6, 2) AD n = 20, 3) MCI n = 38	1) n = 12, 2) n = 20 3) n = 62	AD, MCI	SS	Voicing, pauses, F0, harmonicity	18, 40, 100	kNN, SVM, MLP, CNN	Acc = 80–95%
25	Martinez-Sanchez et al 2016	Possible AD n = 45	n = 82	AD	OT	Syllable intervals and their variation	–	–	AUC = 0.87
26	Mirzaei et al 2018	Early AD n = 16, MCI n = 16	n = 16	AD, MCI	OT	HNR, voice length, silences	48	kNN, SVM, DT	–
27	Rentoumi et al 2017	AD n = 30	n = 30	AD	OT	–	–	NB, SVM	Acc = 89%

Table 1: Information extracted from the studies *(continued)*

No	Study	Impaired group size (s)	Control group size	AD/MCI	Data collection method	Most informative language and speech features	Number of samples used to train the model	Classification algorithm	Classification performance
28	Sadeghian et al 2017	AD n = 26	n = 46	AD	SS	Long pauses, pause and speech duration	65 training, 7 testing	MLP	Acc = 94.4%
29	Satt et al 2013	MCI n = 43, AD n = 27	n = 19	AD, MCI	SS, OT	Verbal reaction time, voiced segments	89	SVM	EER = 15.5–18%
30	Toth et al 2015	MCI n = 32	n = 19	MCI	SS	Pauses, tempo	153 samples from 51 participants	SVM, Random Forest	Acc = 82.4%
31	Toth et al 2018	MCI n = 48	n = 36	MCI	SS	Pausation, tempo and duration	84	NB, Random Forest, SVM	Acc = 75%
32	Warnita et al 2018	AD n = 169	n = 98	AD	SS	Feature set from Interspeech 2010	488 samples from 267 participants	GCNN	Acc = 73.6%
33	Zimmerer et al 2016	AD n = 48	n = 38	AD	SS	Semantic errors, bigram and trigram proportions	–	Logistic regression	–

Table 1: Information extracted from the studies (continued) (*Acc: accuracy; AD: Alzheimer`s disease; aMCI: amnestic mild cognitive impairment; AUC: area under curve; CNN: convolutional neural networks; CTree: classification tree; DT: decision tree; ERR: equal error rate; GONN: gated convolutional neural networks; HNR: harmonics-to-noise-ratio; kNN: k-nearest neighour; MCI: mild cognitive impairment; MCI-con: mild cognitive impairment later converted into AD; MCI-non: mild cognitive impairment not converted into AD; MD: mixed dementia; mdMCI: multiple domain mild cognitive impairment; MLP: multilayer perceptron; MLU: mean length of utterance; NB: Naïve Bayes; OT: other tasks; POS: part-of-speech; SCI: subjective cognitive impairment; SS: spontaneous speech; SVF: semantic verbal fluency; SVM: support vector machine)*

Study examples

In this section I briefly describe 2 studies to provide the reader with a better understanding of what was examined. These 2 studies are chosen to cover different condition groups, data collection and analysis methods.

Fraser and colleagues (2016) used the recordings of 264 participants describing the Cookie Theft picture available on DementiaBank corpus. Cookie Theft picture is a commonly used test in language and cognitive disorder assessment because it features a complex scene and describing it triggers diverse language. DementiaBank is a corpus available for research purposes that gathers speech and language data from people with AD and other forms of dementia. The two participant groups in Fraser and colleagues' study were the AD and the healthy control group. A total of 370 language and speech features related to part-of-speech (POS), syntactic complexity, grammatical constituents, psycholinguistics, vocabulary richness, information content, repetitiveness, and acoustics categories were extracted. The dataset was divided into test and training data, and ML techniques were applied to explore the accuracy of automatic classification between healthy and AD groups. Standard accuracy of over 81% was achieved.

Clark and colleagues (2016) included both semantic verbal fluency (SVF) and phonemic verbal fluency (PVF) tasks from 107 MCI patients and 51 healthy control group participants. The tests were transcribed, and language features, such as the raw count of words, intrusions, repetitions, clusters, switches, mean word frequency, mean number of syllables, algebraic connectivity, and many more were captured. The study paired linguistic measures with the information from MRI scans which allowed creating novel scores. The study concludes that the classifiers trained on novel scores outperformed those trained on raw scores.

2.2.4. Research questions

The research questions were grouped into 5 categories.

What were the characteristics of control and impaired groups?

In 33 studies, 32 different datasets were used. While some studies included up to 3 different datasets for different experiments, a few datasets were used more than once across the studies. The conditions considered in this study were AD and MCI. Although MCI did not feature in the search terms, I decided not to exclude the studies focussing solely on MCI because while MCI patients do not meet the diagnostic criteria of dementia, they can sometimes convert to AD. The studies may therefore provide an insight into the early stages of the disease as well as capture the characteristics of those MCI patients who develop AD and of those who do not. To address the heterogeneity this approach creates, the studies focussing on MCI are looked at separately from the studies concerned with AD detection. Two studies also included other dementia groups (early dementia and mixed dementia) but as both groups only appeared once in the dataset, these groups were not included in further analyses.

64% of all studies reported participants' gender and age. The average number of male participants was 35, and of female participants was 50. The number of male and female participants was stated to be balanced in 13 studies and notable differences in the number of male and female participants appeared in 15 studies. There were significant differences in participants' average age between healthy control (66.94 +/- 5.75) and AD group (74.75 +/- 4.36), $t(30) = -4.223$, $P = .000$, and between MCI (70.21 +/- 5.64) and AD group (74.75 +/- 4.36), $t(25) = -2.351$, $P = .027$. Participants' education level was considered in 45% of the studies. The control group participants had spent on average more years in education than the impaired group in all but 1 study where the participants' education level was considered. Handedness was controlled for in 2 studies, and all but 4 studies mentioned the language the participants' spoke. See Table 2 for participant information.

Participant groups (total number of datasets including the group)	Information variable (number of datasets including this information)	Mean (SD)	Min	Max
Control group (30)	*Number of participants (29)*	**42.69 (46.34)**	2	242
	Age (18)	**66.94 (5.747)**	57	76
	Years of education (11)	**13.44 (2.274)**	9	18
MCI group (16)	*Number of participants (15)*	**30.07 (19.41)**	1	83
Including MCI, aMCI,	*Age (13)*	**70.21 (5.637)**	57	78
mdMCI, MCI-con, MCI-non	*Years of education (7)*	**13.15 (2.353)**	11	16
AD group (31)	*Number of participants (27)*	**45.04 (62.68)**	1	242
Including AD, early	*Age (14)*	**74.75 (4.360)**	66	80
AD, intermediate AD, advanced AD, probable AD, possible AD	*Years of education (8)*	**11.80 (2.006)**	8	15
Dementia group (2)	*Number of participants (2)*	**27.00 (15.56)**	16	38
Including early	*Age (2)*	**72.74 (8.990)**	66	79
Dementia, mixed dementia	*Years of education (1)*	**9.380**	9	9

Table 2: Participant information *(AD: Alzheimer's disease; aMCI: amnestic mild cognitive impairment; MCI-con: mild cognitive impairment later converted into AD; MCI-non: mild cognitive impairment not converted into AD; MD: mixed dementia; mdMCI: multiple domain mild cognitive impairment; SD: standard deviation)*

What kind of language data was collected and how?

From the 33 studies, 28 included at least one spontaneous speech (SS) task, 7 studies included a verbal fluency (VF) task, and 7 studies other tasks (OT).

The aim of SS tasks is to trigger spontaneous speech. This was most often attempted by asking the participants to describe a picture or by engaging in a conversation with the participants. Other tasks used to induce SS included recalling a movie, a day, an event, or a dream. In one study, the transcripts from press conferences were used as a source of SS. SS tasks allow analysing a variety of language attributes, such as word retrieval processes, syntactic, semantic, and acoustic impairment, lexical complexity, and communication errors (Ahmed et al., 2013; de Lira et al., 2014; Garrard et al., 2005b).

There are 2 types of VF tasks: phonemic verbal fluency (PVF) and semantic verbal fluency (SVF) task. In the PVF task, the participants are instructed to name as many words as possible in 1 minute that start with the same letter, such as the letter F. In the SVF task, the participants are instructed to name as many words from the same semantic category as possible in 1 minute, such as animals. Traditionally, the measure most commonly used to evaluate performance in fluency tests is the number of total or correct words produced in 1 minute. These tasks also

allow assessing how lexical and semantic information is accessed, as well as the prevalence of repetitions and memory dysfunction. NLP tools have been used for automatic analysis of semantic clusters and SP for the analysis of temporal and acoustic measures.

OT include all the tests that were not concerned with SS or VF, for example, repeating a sentence, reading out a paragraph, writing a story, counting down numbers, pronunciation, or denomination test. These tasks allow for the examination of different aspects of memory, semantic processing, and acoustic and phonetic measures.

In all tests, the language data was audio or video recorded and/or transcribed.

Figure 3 provides a summary of the methods and tasks used to collect language and speech data.

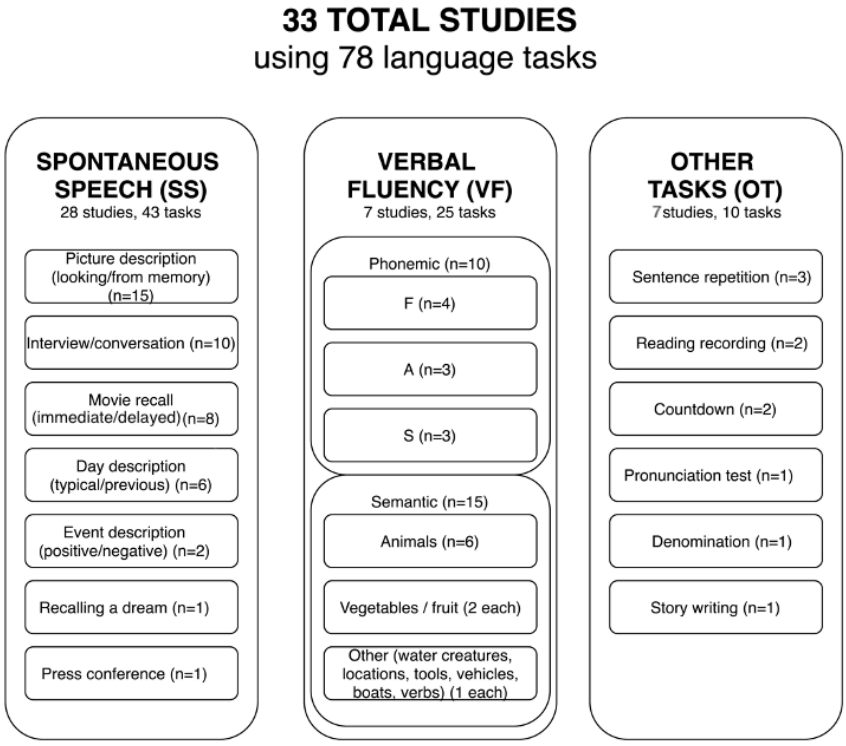


Figure 3: Division of language tests used to differentiate between the conditions

What language and speech features were the most informative?

The 33 studies included experiments from 21 individual research groups. Out of the individual research groups, 18 included SS tasks and 5 VF and OT tasks. The most informative language and speech features are looked at in 2 categories: those characteristic to AD, and those to MCI.

The number of the language and speech features used in the analyses ranged from 4 to 920. As the studies with a large number of features did not report all the features considered, it was difficult to examine what features were studied the most extensively. To avoid synthesis bias towards the features that have been studied more (Papaioannou et al., 2016), and multiple publication bias of over-representing one study or research group with significant results (Song et al., 2010), each feature that has been reported the most informative by at least one research group is reported on equal basis. See Tables 3-5 for the most informative language features from SS, VF, and OT tasks.

What methods were used to classify healthy people and the people with dementia?

Out of 33 studies, 27 used ML to distinguish between healthy people and the people with different medical conditions. Different ML algorithms were used across studies: Neural Networks (NN) were used in 17 studies, Support Vector Machines (SVM) in 16, Decision Trees (DT) in 11, Naive Bayes in 7, and logistic regression in 2 studies. See Table 6 for details and definitions.

What classification performance has been achieved?

The studies reviewed in this paper tend to use different measures to report classification performance (accuracy, precision, Area Under Curve Receiver Operating Characteristics (AUC – ROC), making the comparison of performance difficult. Standard accuracy refers to the level of agreement between the reference value and the test result, and precision refers to the level of agreement between independent test results obtained under stipulated conditions (ISO, 1994). The ROC curve shows the relationship between clinical sensitivity and specificity for every possible decision threshold. AUC measures the ability of the model to distinguish between the groups for all decision thresholds.

The heterogeneity of the performance measures, as well as the participant groups, data collection, and analysis methods does not allow for a direct comparison of classification accuracy. I aim to tackle this issue in 2 steps. First, I provide a table with qualitative information about the methods that each study concluded to have worked best. Second, as standard

accuracy was the most widely used performance measure, I compare the results and the methods used to achieve them in the studies that reported standard accuracy.
Table 7 presents the settings and the approaches that were used when top performance was achieved in each study.

				Alzheimer's disease (AD)	both AD & MCI	Mild cognitive impairment (MCI)
spontaneous speech	syntactic complexity			inflected verbs reported speech interpolated clauses depth of parsing tree width of parsing tree ratio depth:width parsing tree go-ahead utterances past tense past particles	speech or utternace duration or length	
	acoustic impairment			harmonicity voicing activation frequency average absolute delta pitch average absolute delta formant	voiced segment percentage	short-time energy spectral centroid
	word retrieval	lexical	POS	difficulty in word finding imporverished vocabulary and language limited answers Honere's statistic words indicating quantities Brunet's index unique words number, frequency, and rate of specific words type:token ratio pronoun:noun ratio pronoun frequency prepositions auxilliaries determiners adverbs conjunctions adverbs verb frequency and rate number of nouns / noun rate adjective rate		ratio features content density vocabulary richness open:closed class words ratio information coverage
		fluency and pausation		fraction of pauses greater than 10 seconds silence rate average continuous word count logarithm voicing rate bigrams trigrams number of turns	speech rate articulaiton rate number of pauses (total, silent, filled) length of pauses length of silent pauses pause:duration ratio filled pause:duration ratio	total length of filled pauses speech tempo speech continuity SD of speech segment duration SD of (un)voiced segment duration temporal regularity of voiced segments
	errors			number of stutters self-corrections incomplete sentences unitelligible word rate confusion rate no answer rate repetitions		
	semantic processing			semantic error rate semantic attributes		

Table 3: Most informative language and speech features in spontaneous speech tasks *(AD: Alzheimer's disease; MCI: mild cognitive impairment; POS: part-of-speech; SD: standard deviation)*

			Alzheimer's disease (AD)	both AD & MCI	Mild cognitive impairment (MCI)
verbal fluency	graph theory				algebraic connectivity radius transativity
	acoustic impairment				harmonicity voicing activation
	word retrieval	fluency and lexical pausation		duration of silent and voiced segments	lexical frequency position of words in time
	semantic processing		coherence semantic similarity between words		

Table 4: Most informative language and speech features in verbal fluency tasks *(AD: Alzheimer's disease; MCI: mild cognitive impairment)*

			Alzheimer's disease (AD)	both AD & MCI	Mild cognitive impairment (MCI)
other tasks	ratio features				harmonics to noise ratio other ratio features
	acoustic impairment				energy envelope periodicity
	word retrieval	fluency and pausation		token duration and total number syllable duration and intervals speech and articulation rate duration of silent and voiced segments	position of words in time pauses per token number of silences estimate of insertions and pauses
	errors				errors per token
	reaction				average verbal reaction

Table 5: Most informative language and speech features in other tasks *(AD: Alzheimer's disease; MCI: mild cognitive impairment)*

	Classification	AD vs healthy control				CD vs healthy control			
	performance measure	acc (n)	AUC (n)	precision (n)	EER (n)	acc (n)	AUC (n)	precision (n)	EER (n)
Neural Nets (NNs) (n = 17)	average of all reported outcomes	86% (17)	–	0.64 (4)	–	65% (4)	–	–	–
NNs are computer systems that are similar in structure to biological neural networks and mimic the way animals learn	average of best reported outcomes	88% (6)	0.96 (1)	0.69 (1)	–	69% (2)	–	–	–
Support Vector Machines (SVMs) (n = 16)	average of all reported outcomes	81% (33)	–	0.68 (4)	14% (2)	78% (13)	–	–	19% (3)
	average of best reported outcomes	88% (9)	–	0.79 (1)	14% (2)	77% (6)	–	–	19% (2)
SVMs are supervised models that use training data belonging to 1 or another category, and assign a category to test data based on the training data									
Decision Trees (DTs) (n = 11)	average of all reported outcomes	82% (24)	–	0.63 (5)	–	78% (9)	–	–	–
DTs take several input variables and, based on the observations about them, predict the value of a target variable	average of best reported outcomes	90% (4)	–	0.96 (2)	–	80% (3)	–	–	–
Naïve Bayes (NB) (n = 7)	average of all reported outcomes	81% (8)	–	0.81 (1)	–	67% (1)	–	–	–
NB classifiers are simple probabilistic classifiers where each variable contributes independently to the assigning of class label	average of best reported outcomes	81% (4)	–	0.81 (1)	–	67% (1)	–	–	–

Table 6: Details of machine learning (ML) methods used and the performance achieved. *"Average of all reported outcomes" refers to the average of all measures reported across studies using the ML algorithm and performance measure. "Average of best reported outcomes" takes the average measure of the best performance reported in each study (1 measure per study) using the ML algorithm and performance measure.* (*Acc: accuracy; AD: Alzheimer's disease; AUC: area under curve; CD: cognitive decline; ERR: equal error rate)*

ID	Study	Most effective technologies	Classification performance
1	Ammer and Ayed 2018	feature selection: kNN; classifier: SVM	precision = 79%
2	Beltrami et al 2018	Acoustic features	–
3	Boye et al 2014	–	–
4	Chien et al 2018	bidirectional LSTM RNN	AUC = 0.956
5	Clark et al 2014	Semantic similarity features	–
6	Clark et al 2016	Classifiers with novel scores including MRI data	Acc = 81–84%
7	Fang et al 2017	length of sentence, unique words, non-specific, and specific words	–
8	Fraser et al 2015	Using 35 features	Acc = 82%
9	Garrard et al 2017	Certain scripts and motives	–
10	Gosztolya et al 2016	automatically selected feature set, correlation-based feature selection technique	Acc = 88.1%
11	Gosztolya et al 2019	AD: combination of linguistic and acoustic features; MCI: semantic and acoustic features	Acc = 86%
12	Guinn et al 2014	go-ahead utterances and certain fluency measures	precision = 80%
13	Hernandez-Dominguez et al 2018	AD detection: RFC with coverage and linguistic features; decline detection: RFC with a combination of features with P-value <.001 when correlating with cognitive impairment	Acc = 87–94%
14	Khodabakhsh et al 2014a	SVM, logarithm of voicing ratio, average absolute delta feature of the first formant, and average absolute delta pitch feature	Acc = 88–94%
15	Khodabakhsh et al 2014b	SVM, DT	Acc = 90%
16	Khodabakhsh et al 2015	SVM classifier with the silence ratio feature	Acc = 84%
17	Konig et al 2015	–	EER = 13–21%
18	Konig et al 2018	Fluency tasks	Acc = 86%
19	Lopez-de-Ipina et al 2013a	Including fractal dimension sets	Acc = 75–94.6%
20	Lopez-de-Ipina et al 2013b	SVM and features from 3 datasets: spontaneous speech, emotional response and energy features	Acc = 93.79%
21	Lopez-de-Ipina et al 2015	MLP for Katz's and Castiglioni's algorithm with a window-size of 320 points	Acc = 95%
22	Lopez-de-Ipina et al 2018	SS task and AD patients: the recording environment within a relaxing atmosphere; the presence of subtle cognitive changes in the signal due to a more open language; and the use of AD patients instead of MCI subjects.	Acc = 73–95%
23	Luz 2018	–	Acc = 68%
24	Martinez de Lizarduy et al 2017	spontaneous speech task; CNN	Acc = 80–95%
25	Martinez-Sanchez et al 2016	The standard deviation of the duration of ΔS	AUC = 87%
26	Mirzaei et al 2018	kNN with 18 features	–
27	Rentoumi et al 2017	–	Acc = 89%
28	Sadeghian et al 2017	using all the potential features, including and choosing the 5 most informative ones: 1) MMSE, 2) race, 3) fraction of pauses greater than 10s, 4) fraction of speech length that was pause, 5) words indicating quantities	Acc = 94.4%
29	Satt et al 2013	Using 20 features	EER = 15.5–18%
30	Toth et al 2015	SVM with manually extracted features	Acc = 82.4%
31	Toth et al 2018	RFC with automatic and significant feature set	Acc = 66.7–75%
32	Warnita et al 2018	10-layer CNN with Interspeech 2010 feature set	Acc = 73.6%
33	Zimmerer et al 2016	connectivity, closed-class words, semantic error rate	–

Table 7: Most effective technologies (*Acc: accuracy; AD: Alzheimer's disease; CNN: convolutional neural networks; DT: decision tree; ET, emotional temperature; kNN: k-nearest neighbour; LSTM RNN: long short-term memory recurrent neural network; MLP: multilayer perceptron; MMSE: mini-mental state examination; MRI: magnetic resonance imaging; RFC: random forest classifier; SVM: support vector machine)*

Standard accuracy was used as a classification performance measure in 17 tasks across 15 studies that aimed to distinguish the people with AD from the people without AD and in 8 studies looking at MCI. The average classification accuracy was significantly lower when detecting MCI (81.7% +/- 5.3%) than when detecting AD (88.9% +/- 8.0%), t (14) = 2.40, P = .031.

Top result in AD detection (95% classification accuracy) was achieved using an SS task to collect information about voiced and unvoiced segments and other acoustic and phonetic features. Lopez-de-Ipina and colleagues (2015, 2018) used NN to distinguish the people with AD from those without AD.

Top result in MCI detection (86% classification accuracy) was reached by Konig and colleagues (2018) using SVF and PVF to collect language data, SP to analyse the data, and SVM to discriminate between the people with and without MCI.

2.2.5. Discussion

This review shows that the sociodemographic variables, especially age, often differ between the healthy and the impaired group. The language data was usually collected using SS tasks, with the most informative language features falling under lexical, syntactic, semantic, and acoustic impairment. NNs, SVMs, and DTs performed well as classifiers; 89% average classification accuracy was reached in AD detection and 82% in MCI detection.

Synthesis

The majority of the studies reviewed in this review demonstrate promising results in identifying AD or MCI based on speech and language data. While the results are promising, there is also room for improvement. For example, age, gender, education level, and handedness can affect speech and the outcome of language tests. However, there were significant differences in participants' ages between healthy and AD groups, more female than male participants were included in the studies, people with a clinical condition tended to be less educated than the control group, and only 6% of the studies considered whether the participants were right- or

left-handed, even though language behaviour tends to differ in the right- and-left-handed individuals: greater left hemisphere reliance is thought to emphasise syntactic relations when it comes to language processing, and greater right hemisphere reliance tends to emphasise meaning (Townsend et al., 2001). Similarly, the majority of the participants spoke European languages, leading to very few non-European languages being considered.

Two popular and well-performing language tasks were SS and VF. Promising results were achieved using language features relating to word retrieval, semantic and acoustic impairment, and error rate.

Various ML algorithms were used to classify between different condition groups. The best performing models were NNs, SVMs, and DTs.

The measures used to report performance were heterogeneous, making the comparison of the technologies difficult. Focussing only on the studies that used accuracy as a metric, the highest classification accuracy was achieved using SS task, SP method, and NN classifiers when distinguishing between AD and healthy groups, and VF task, SP method, and SVM classifier when detecting MCI. Average classification accuracy was 89% in the AD and healthy group distinction, and 82% in MCI detection.

Recommendations for future research

Based on the findings of this study, I propose the following directions for future research:

- constructing demographically and socioeconomically balanced datasets to minimise the effect of age and other factors on the results;
- including a larger number of participants to allow more data to be used when training a machine learning model;
- including non-European languages in future studies as the vast majority of the studies so far have been conducted in European languages;
- conducting longitudinal studies concerned with MCI to examine the language of those participants who convert from MCI to AD and of those who do not (this approach was taken in Clark and colleagues (2016));

- integrating linguistic analysis and signal processing to achieve maximum accuracy. Most studies focus on either SP and acoustic features or NLP and linguistic features. However, most language tasks are audio recorded which would allow collecting both acoustic and linguistic data (using both audio samples and transcripts). I suggest that adding linguistic variables (lexical, semantic, syntactic) to SP approach, and vice versa, adding SP measures (acoustic, voiced and unvoiced segment analysis) to studies mainly focussing on linguistic features. This will allow for the expansion of the set of variables beneficial in ML approach and could lead to more accurate classification results. An example of a study that has used both acoustic and linguistic measures was conducted by Fraser and colleagues (2016);
- using the 4 standard measures to report the performance: Accuracy, Precision, Recall, F1-score (AUC can be used in addition to those four).

The studies reviewed in this article include 19 additional suggestions for future research:

(1) ensure standardised recordings and language samples;
(2) add new and challenging tasks;
(3) calibrate audio measurements;
(4) add new features;
(5) couple speech analysis with neuroimaging;
(6) include follow-up studies;
(7) conduct longitudinal studies;
(8) add linguistic and acoustic features;
(9) automate feature selection;
(10) include voice onset time;
(11) extend the number of MCI samples;
(12) research the effect of sample size in healthy control groups;
(13) perform cross-linguistic studies;
(14) use automatic transcription of language tasks;
(15) include nonverbal communication (gestures);

(16) include syllable-timed and low-resource languages;

(17) replicate the results of currently available studies;

(18) evaluate the temporal change and the severity of the disease;

(19) include more forms of dementia, such as vascular dementia.

Study limitations

To evaluate the limitations and establish the confidence level of the outcomes, I adapt GRADE guidelines (Guyatt et al., 2011a). There are 5 main limitations, 4 of which contributed to the decision to rate down the outcome confidence level from high to moderate.

- **Potential publication bias**

  This refers to the possibility of only the studies with more significant results being published (Rosenthal, 1979; Guyatt et al., 2011b). Although publication bias was undetected in the current review, it is especially common in literature reviews written in the early stages of the specific research area due to negative studies being delayed (Guyatt et al., 2011b) and should therefore be mentioned. Potential publication bias was not used to decrease the confidence level.

- **Potential synthesis bias**

  This refers to only articles written in English being included in the current analysis (Papaioannou et al., 2016; Song et al., 2010), not allowing the data available in other languages to be considered, limiting the dataset, and possibly contributing to the small number of non-European languages being included. Language bias can especially affect the outcomes relating to the most informative language features, as these are directly dependent on the language used.

- **Age difference between AD and HC groups**

  This refers to the risk of bias in the outcome of studies focussing on AD detection because the AD group was very often significantly older than the control group. This increases the chance of the most informative language features being characteristic to older age instead of AD, as well as the classification algorithms differentiating between older and younger, and not necessarily detecting AD.

- **MCI not being included in the search terms**

  This refers to the risk of bias when reporting the outcomes of the studies concerned with MCI. The fact that the search terms did not include MCI is likely to have led to a situation where additional studies did exist—but were inaccessible —and therefore did not get included in the analysis.
- **Reporting classification performance**

  This refers to the potential risk of bias in reporting the classification performance, as often only the best outcomes are included, potentially leading to skewed understanding of how well the algorithms worked.

The last 4 limitations (potential synthesis bias, age difference between AD and control group, MCI not being included in search terms, and potential bias in reporting classification performance) contributed to the confidence levels of the outcomes concerned with informative language features, classification algorithms, and classification performance to decrease from high to moderate.

2.2.6. Conclusion

In this systematic review on automatic AD detection from speech and language, I report the characteristics of healthy and impaired groups, summarise the language tests that have been used, present the language and speech features that have shown to be the most informative, and identify the ML algorithms used and the classification performance achieved.
The findings show that the balance in demographic variables across AD and healthy groups could be improved, and that the SS-based studies have achieved top accuracy in distinguishing between AD and healthy conditions. Informative language and speech features capture problems with word retrieval, semantic processing, acoustic impairment, and errors in speech and communication. From ML algorithms, NNs and SVMs were the most widely used, and top accuracy was also achieved with these models. Standard accuracy was the most common metric used to report the classification performance, with the average accuracy in AD detection being 89%, and in MCI detection 82%.

In future studies, I suggest constructing larger, balanced, and more diverse datasets, focussing on the earliest markers of AD and longitudinal change, standardising the metrics used to report classification performance, combining linguistic features with other types of data and constructing novel, more informative and universal features.

2.3. Further developments

The findings of the systematic literature review have informed the research directions taken during the rest of the PhD as well as contributed to further developments in the field in other studies.

Based on the outcomes of the review and the recommendations from the studies included in the review, the rest of the thesis sets out to address some of the main challenges in the current practice:

(1) the lack of longitudinal datasets and studies;
(2) the replicability of previous results and the standardisation of research methods;
(3) the issues with fairness, accountability and transparency, including data balance, potential bias, and other ethical considerations that arise in the process of deploying AI and NLP in speech-based AD detection.

I will discuss each of these challenges in the next three sub-sections, including the recent literature addressing these issues and the developments that have been made after the publication of the systematic literature review.

2.3.1. Lack of longitudinal datasets

As the literature review showed, language-based classification between those already diagnosed with AD and the healthy individuals has achieved high accuracy. Therefore, the focus of current studies is increasingly moving towards longitudinal changes and the earliest signs of potential cognitive decline (Meltzer, 2020; Khoury & Ghossoub, 2019). This shift can also be seen in the studies published after the publication of the systematic literature review, with many more studies focussing on MCI. For example, Vetrab and colleagues (2022) employ sequence-to-sequence deep autoencoders to analyse spontaneous speech and report

competitive performance in differentiating 25 MCI subjects and 25 healthy controls. Liang and colleagues (2022) look at home-based cognitive assessment for detecting early markers of cognitive decline by examining voice commands using Voice-Assistant Systems. Nagumo and colleagues (2020) performed a classification task between healthy controls and MCI, global cognitive impairment and MCI with global cognitive impairment based on the acoustic features extracted from sentence reading task. Several recent studies (Sadeghian et al., 2021; Bertini et al., 2021; Al-Atroshi et al., 2022; Chau et al., 2022) have focussed on developing fully automatic systems to assess the speech of those with cognitive impairment and differentiate them for healthy individuals using an easy, cheap, and non-invasive tool.

However, due to the limited availability of language data from the early stages of the disease as well as longitudinal data, there are still relatively few studies conducted on the earliest changes in language, and the changes over time (Calza et al., 2021; Luz et al., 2021a; Lopez-de-Ipina et al., 2018), even though these changes could be promising indicators for early detection and a reflection of the regression of cognitive abilities.

Earlier studies have tackled the issue of limited availability of longitudinal language data by comparing the writings of authors some of whom eventually develop AD (Le et al., 2011), or the transcripts of press conferences (Berisha et al., 2015; Fang et al., 2017). There are also some available longitudinal datasets, such as the ADReSS$_o$ Challenge corpus (Luz et al., 2021a) and Pitt corpus (Becker et al., 1994), but these mostly span over a short time period and are based on picture description or semantic verbal fluency (SVF) tasks. While these datasets are extremely valuable and allow for controlling for the content of the speech data by using structured speech tasks, collecting spontaneous speech data that spans over a longer period of time and is not based on visual cues has its advantages. For example, it has been proposed that the earliest manifestations of cognitive decline can appear years or even decades before the diagnosis (Fox et al., 1998; Grundman et al., 2006; Ringman, 2017) - having a longitudinal corpus that consists of speech data from several decades can contribute to detecting these earliest pre-diagnostic changes. Using spontaneous conversational speech also allows for data collection in a more naturalistic setting, reflecting the everyday challenges that the AD sufferers face in communicative situations (Sabat & Harre, 1994) while causing less stress for the participants

which is often experienced during cognitive testing (Lopez-de-Ipina et al., 2018; Sabat & Harre 1994; Chien et al., 2018). Studies focussing on spontaneous speech also outperformed other tasks in the systematic literature review.

Although collecting spontaneous speech data in an everyday situation is becoming more feasible as speech can easily be gathered using mobile devices, and the process requires minimal instructions and equipment (Chien et al., 2018; Robin et al., 2021; Yamada et al., 2021), there is a lack longitudinal datasets spanning over a longer time period as these are time consuming and expensive to collect (Yancheva et al., 2015; Meltzer, 2020; Robin et al., 2021). In this thesis, I aim to tackle this issue by constructing a novel corpus of transcripts of public interviews with famous individuals recorded over several decades, half of whom will eventually be diagnosed with AD (Chapter 3). This corpus allows looking at the changes that appear in spontaneous speech over time as well as assess the generalisability of the results (Chapter 4) and the robustness of the features (Chapter 5). The motivation behind constructing the corpus builds primarily on two outcomes of the literature review: (1) spontaneous speech tasks being the most informative in language data analysis, and (2) the need to understand longitudinal changes and identify the earliest markers.

2.3.2. Replication of previous studies and standardisation of methods

Previous research analysed for the literature review recommends repeating the results of the published studies and increasing overall replicability by standardising the data collection and analysis methods, as well as the reporting standards. I incorporate these suggestions in Chapters 4 and 5 using the novel longitudinal corpus I created.

Chapter 4 focuses on the generalisability of the language features. Among other positive impacts, understanding the generalisability of language change would contribute to reducing dimensionality in the models (Berisha et al., 2021). Decreased dimensionality is desirable for more robust, accurate and interpretable models. Therefore, it is important to identify a small set of features that change consistently with the disease based on domain knowledge and existing theory, rather than using hundreds of features (Berisha et al., 2021). When analysing

the data in the novel corpus I created (Chapter 3), I follow this approach (Chapters 3 and 4) and make use of the list of the most informative language features identified in the literature review (section 2.2).

To tackle the issues of replicability and generalisability, in Chapter 4, I start by replicating the methods used by Berisha and colleagues (2015) on a larger sample size, examining the generalisability of the changes in language. Due to relatively low comparability of the results, I give suggestions for advancing the data analysis methods and conduct relevant experiments, looking at language change in relation to time before diagnosis rather than age, experimenting with a different set of language features that have shown to be informative in previous studies, and constructing aggregate scores.

Generalisability and standardisation are related challenges. Luz and colleagues (2021a) point out that the lack of standardisation "has hindered the benchmarking of the various approaches proposed to date". They claim that this has also contributed to the slow translation of the technologies into clinical practice. To tackle standardisation issues, Luz and colleagues (2021a) have developed a standardised dataset and the ADReSS$_o$ challenge that allows developing approaches of cognitive decline detection from standardised language data. Other approaches to examine the robustness of the language features have included looking at how the stability of the extracted features changes from one recording to another within an individual, finding notable variation (Stegmann et al., 2020), and comparing the performance between manual and automatic transcription concluding that human verification of automatic transcripts would increase the performance in spontaneous speech tasks (Soroski et al., 2022).

In the current thesis, I address the issue of standardisation in Chapter 5, focussing on data collection methods, specifically on the optimal length of the recordings. I aim to examine how much speech data is needed for informative analysis, and how sample length impacts the robustness of the extracted language features.

2.3.3. Ethical considerations

The ethics of early dementia detection in general have been discussed extensively in previous literature. One argument for early detection is the potential earlier start of treatment, however,

others argue that there is currently no approved treatment, although advances in treatment development are being made. Similarly, early detection could allow access to support and services, however Brayne and Kelly (2019) argue that there is no trial to show how and whether it would work and claim that early detection might instead overstretch the services. Other arguments for early detection include positive change for families and individuals and improved life quality, however, Brayne and Kelly (2019) and Ford and colleagues (2023) state that there is limited evidence for that in dementia.

Another area of discussion is the use of automatic, potentially app-based tools: who should have access to them, should they be freely available for at-home screening, or accessed only by professionals in a clinical setting? The arguments for free use include the opportunity to self-check for those who cannot or do not want to see a clinician, as well as having a relaxed setting, less anxiety and therefore more accurate outcome (Mirheidari et al., 2019) while the arguments against highlight the risk of dual use and potential discrimination, for example, in acquiring health insurance or applying for jobs. Several advantages for automatic speech-based in-clinic AD detection have been pointed out in previous work, such as the potential to:

(1) help clinicians detect the early changes in language that may otherwise go unnoticed but could aid further referrals;
(2) discriminate between anxiety or stress impact on language and dementia-related memory problems;
(3) monitor disease progression and response to treatment;
(4) identify people in the early stages for drug trials contributing to drug development.

(Mirheidari et al., 2019; Ford et al., 2023)

On the contrary, Brayne and Kelly (2019) argue that the drive for early diagnosis can lead to delayed referrals and have a negative impact on timely diagnosis.

Using speech and AI for early AD detection raises novel ethical questions which I will discuss in Chapter 6. While language-based detection could provide a cost-effective, accessible, fast, and non-invasive solution for AD screening, potential issues of autonomy, privacy, data and model bias and transparency arise. The findings of the literature review also highlight these issues by

revealing unbalanced datasets and small sample sizes which can lead to unfair distribution of research benefits as the model will work only on the population it has been trained on. Berisha and colleagues (2021) conducted a meta-analysis of this systematic literature review (Petti et al., 2020) and a similar review by de la Fuente Garcia and colleagues (2020), highlighting the issue of the best performing models having been trained on smaller datasets, leading to low applicability or failure in a real world setting where the data is more diverse. Figure 4 from Berisha and colleagues (2021) shows the meta-analysis of these two literature reviews, comparing model accuracy in classifying between the individuals with AD and healthy individuals (blue) and healthy individuals and those with cognitive impairment (red) in relation to the reported sample size.

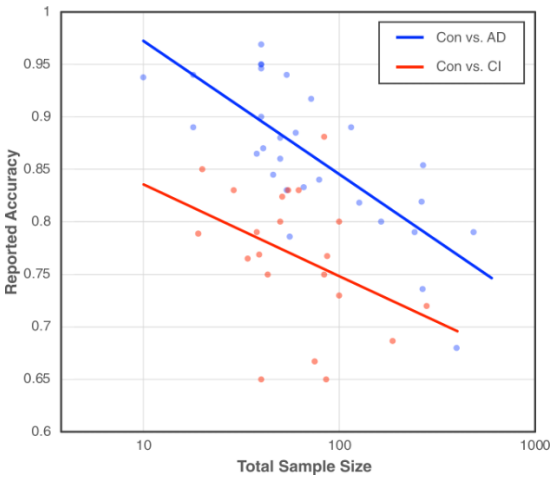


Figure 4: The relationship between model accuracy and sample size in classifying between healthy controls and the people with Alzheimer's disease (blue), and healthy controls and the people with cognitive impairment (red) *(Con: control group; AD: Alzheimer's disease; CI: cognitive impairment) (Berisha et al., 2021)*

The literature review in section 2.2 (Petti et al., 2020) also highlighted the issue of the majority of the studies being conducted on English language. Since the publication of the literature review, several studies have addressed this issue. For example, Kalman and colleagues (2022) tested an automatic speech recognition (ASR) -based system on Hungarian language and compared it to the performance on English, concluding that changes in acoustics that help

distinguish between MCI and HC participants might be present in different languages. Similarly, Schäfer and colleagues (2023) looked at Dutch and Swedish populations and reported the algorithm using acoustic features being relatively robust and language-agnostic. Notable work has also been done on adapting different cognitive and language tests to Thai language and culture since 2020. For example, Munthuli and colleagues (2021) used cognitive assessment tests that had been adapted to Thai culture and looked at the language-based classification between healthy controls, the individuals with MCI and AD. Similarly, Sangchocanonta and colleagues (2021) used a picture description task that is relatable to Thai culture, and a POS tagger that lines with Thai word types. Metarugcheep and colleagues (2022) adapt verbal fluency tests to Thai and perform language-specific feature extraction. All of these studies report promising results that could contribute to non-invasive and cheap screening tools that could aid the detection of cognitive decline in Thai speakers.

I will discuss these and other novel ethical considerations that arise in using NLP and AI to detect AD from speech in more detail in Chapter 6, including the issues with data balance, potential bias, and distribution of research benefits.

2.4. Summary

In this Background chapter, I first provided a more detailed description of AD (2.1), and conducted a systematic literature review on the state-of-the-art of automatic AD detection from speech and language (2.2).

I identified 3 main areas of improvement based on the literature review - the lack of:

(1) longitudinal datasets and studies;

(2) replicability of previous results and the standardisation of research methods;

(3) attention on ethical considerations that arise in deploying AI and NLP in speech-based AD detection.

I discussed how these questions have been addressed in the literature, and how this thesis sets out to tackle these challenges (2.3).

3. LoSST-AD: A longitudinal corpus for tracking Alzheimer's disease related changes in spontaneous speech

3.1. Introduction

In this chapter, I will introduce the corpus that I created during the PhD, and show how it can inform us about the longitudinal speech changes in AD. This chapter was motivated by the findings of the literature review, according to which, while the changes in language can appear years or even decades before AD diagnosis (Ringman, 2017; Grundman et al., 2006), there is a lack of studies focussing on the earliest AD-related language changes due to the limited availability of large-scale longitudinal datasets (Luz et al., 2021a; López-de-Ipiña et al., 2018). Detecting the early changes in language is critical for deeper linguistic understanding of AD and the development of technologies for its early detection that would allow early interventions, but research into the earliest signs is challenging, as it requires large longitudinal datasets. Finding recordings of spontaneous speech that have been conducted over longer periods of time, ideally several decades, can be difficult as this data is unlikely to exist. These datasets are also time-consuming and expensive to collect. Also, the process has numerous ethical issues, contributing to data sparsity in this domain (Luz et al., 2021a). Therefore, there is a need for alternative methods for monitoring longitudinal language change, including NLP and speech recognition technologies.

As a way around this challenge, I have chosen to focus on famous people who have TV and radio interviews of them available over a long period of time, and who are known to have died of Alzheimer's disease. I present a novel corpus of Longitudinal Spontaneous Speech Transcripts for tracking Azheimer's Disease related changes in language (LoSST-AD). The corpus consists of 135 public interviews recorded over several decades with 20 famous individuals, half of whom will eventually be diagnosed with AD. Unlike previous studies in the domain that have either focussed on a few individuals (such as Le et al., 2011; Berisha et al., 2015), used picture description task, or spanned over a shorter time period (such as Luz et al., 2021a; Becker et al., 1994), LoSST-AD consists of several decades of transcribed public spontaneous speech data

from a larger group of individuals, allowing us to examine longitudinal AD-related changes in language use.

However, this data also comes with several limitations. For example, public figures who perform on media are likely to have had media training which makes comparing their spontaneous speech produced under spotlight to the general population challenging. Similarly, it is impossible to know how much of the interviews are scripted or if there are any other factors impacting the participants' speech, such as other medical conditions or intoxication. Additionally, as I do not have access to the speakers' medical records, the time of diagnosis is based on media entries. However, this date is more representative of when the speakers made their condition public, and lacks accuracy in terms of the actual onset of the disease. Nevertheless, this data provides a unique insight into the longitudinal Alzheimer's disease related changes in language use that can inform future research.

I evaluate the corpus by validating the patterns of language change known from Alzheimer's literature, focussing on vocabulary richness. I show that such data can provide valuable insights into longitudinal language changes in AD, and help to develop non-invasive screening tools such as those based on NLP and speech technologies. When interpreting the results, the limitations of this data, as well as the size of the sample must be considered.

I collected, transcribed, and processed the data, performed all the analysis and conducted all the experiments in this chapter. This chapter is based on an article published in LREC-Coling 2024.

3.2. The language changes I expect to capture

AD-related changes have been documented in various speech and language domains such as lexicon, semantics, syntax, discourse, and acoustics (Bayles, 1982; Lima et al., 2014; Gosztolya et al., 2019). Due to the scope of this chapter and the nature of the data (varying audio quality, uncontrolled content due to secondary data, inconsistent turn-taking), I will focus on the changes in vocabulary richness and demonstrate that lexical diversity features can provide comprehensive results in tracking AD-related language change. Changes in vocabulary richness mostly fall under lexical-semantic domain, which in AD is affected by changes in cognitive

function, semantic, procedural, and declarative memory (Dijkstra et al., 2004; Ullman, 2004; Orange & Purves, 1996). These changes manifest in impoverished vocabulary and word finding difficulties. The two are connected – if the speaker cannot access the correct word, a simpler word, usually of higher frequency and decreased length, tends to be used instead (Saito & Takeda, 2001; Hodges et al., 1992). This results in producing less unique words, affecting type:token ratio (referring to the number of unique words used divided by the number of total words used) and vocabulary richness indices (such as the Brunet index (Brunet, 1978)) (Guinn et al., 2014; Hernández-Dominguez et al., 2018; Fang et al., 2017). The speakers with AD often struggle to find nouns, and therefore both noun and adposition frequency can be affected (Jarrold et al., 2014; Bucks et al., 2000; Ammar & Ayed, 2018). Other strategies to cope with not remembering a word include repeating the last word uttered (Croot et al., 2000), stuttering (Boyé et al., 2014), using fillers, and laughter (Sidnell & Stivers, 2012). Figure 1 illustrates the underlying AD-related problems that affect different domains in language, including the lexical/semantic domain and vocabulary richness, as well as how these changes manifest in speech, and which respective language features could be extracted from the transcripts.
The aim of this chapter is to:

- introduce and make available a new language resource – a longitudinal corpus of transcripts of interviews with public figures, half of whom will eventually develop AD;
- and evaluate the corpus by validating the patterns of AD-related vocabulary richness changes known from the literature. I report an evaluation that shows that this new resource can provide a highly valuable starting point for the development of NLP and speech -based early detection tools for AD.

3.3. Corpus creation

I collected publicly available TV and radio interviews from 20 famous individuals, half of whom will eventually be diagnosed with AD. I chose to focus on famous individuals, as they are more likely to have publicly available spontaneous speech recordings over several decades available than the people who are not public figures. This approach contributes to tackling the data availability challenge. It is important to look at how individuals' speech changes over decades,

as the language changes can be heterogeneous, and this kind of longitudinal data allows to use the participants' younger selves' speech as a point of comparison to their speech later in life. It also allows to investigate individual changes as well as spot potentially generalisable patterns across individuals to give us a better understanding of individual variabilities and the language changes that can provide more universal information about the manifestations of cognitive decline in AD.

The inclusion criteria for the AD group was based on:

1) known AD diagnosis - the individuals with AD were identified based on internet searches, using websites such as the Wikipedia page 'Deaths from Alzheimer's disease';
2) the largest number of publicly available interviews in English language on YouTube;
3) gender balance (5 male and 5 female speakers).

The participants with coexisting medical conditions that potentially also affect speech, such as Parkinson's disease, were excluded.

Each AD participant is paired with a control participant (HC) based on demographic data, such as year of birth (within a 5-year range), gender, place of residence or growing up, and occupation when possible. The control group participants had no known diagnosis of AD. Handedness, education level, and the number of languages spoken could not be controlled for in either group as this data was often not available in public domain.

I identified and transcribed all publicly available interviews of the 20 individuals. A total of 135 interviews are included in the corpus. The number of interviews per speaker ranges from 3 to 25, with the earliest AD group interview being recorded 37 years before the diagnosis, and the latest 2 years after the diagnosis. See Figure 5 for the distribution of interviews over time, and Table 8 for the participant group demographic information.

This research project was approved by the Ethics Committee of the School of Humanities and Social Sciences of the University of Cambridge. All transcripts and datasets – along with the datasheet with standardised data description (Gebru et al., 2021; Papakyriakopoulos et al., 2023) – are available online at: https://www.losst-ad.com/.

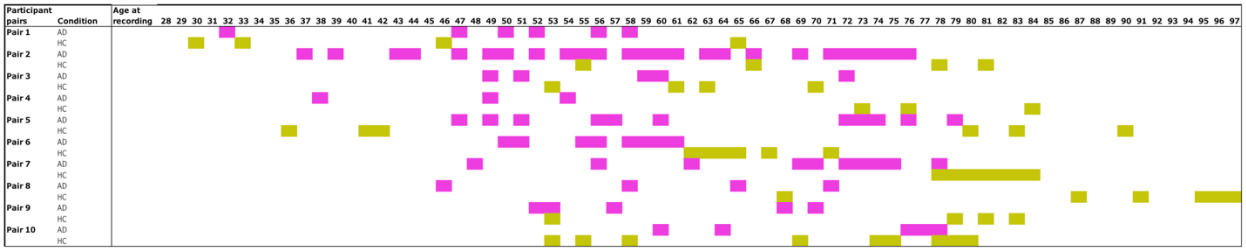


Figure 5: Distribution of interview recordings (*AD: Alzheimer's disease; HC: healthy control)*

	AD group (10 participants, 82 recordings)	**HC group** (10 participants, 53 recordings)
Average age over all recordings (years)	60 (SD: 11, range: 32-79)	70 (SD: 16, range: 30-97)
Average time before diagnosis (years)	13 (SD: 11, range: 37 years before diagnosis - 2 year after diagnosis)	-
Sex	5 female, 5 male	5 female, 5 male

Table 8: Participants' demographic information (*AD: Alzheimer's disease; HC: healthy control; SD: standard deviation)*

3.4. Data processing

I transcribed all interviews manually. Speaker diarisation was applied, excluding the speech of the interviewer. Direct quotes, such as song lyrics, were also excluded from the transcripts to allow more accurate analysis of vocabulary richness in spontaneous speech. Transcripts were anonymised using the guidance from Saunders and colleagues' (2015) paper regarding people's names, places, cultural background, occupation, family relations and other identifying information.

Given the variation in audio quality (some interviews were recorded decades ago, some in noisy environments, using different settings and microphones) acoustic features were not analysed in the current study. The transcripts included filled pauses (such as "uh"), false starts, and stutter, but did not include speech tempo related features, such as the length of pauses. While the

syntactic and semantic changes could also be explored based on this data, I have focussed specifically on vocabulary richness features in the current chapter.

SpaCy and NLTK were used to automatically extract language features from every transcript. One transcript refers to one interview, and all features were extracted on interview level. The extracted vocabulary richness features include Brunet index, hapax legomena frequency, type:token ratio, word frequency (referring to higher use of more frequent words), word length, the number of words used once and twice, noun and adposition frequency, and uni- and bigram repetitions. All extracted features have been proposed to be highly informative by previous studies (Ammar & Ayed, 2018; Guinn et al., 2014; Hernández-Dominguez et al., 2018; Yeung et al., 2021).

I controlled for text-length-sensitivity of the extracted feature values by conducting Pearson correlations between the feature value and transcript length. Since the majority of the extracted features were dependent on text length, I constructed a capped sub-corpus to minimise text length impact on the analysis. Building on the methods used to tackle text-length-sensitivity in Le and colleagues (2011), I capped all transcripts at the same length within each participant pair (the AD speaker and their matched control), keeping at least three transcripts per participant to allow for tracking longitudinal change. I aimed to keep as much speech data as possible, resulting in excluding the shortest samples from the capped corpus. The capped corpus consists of 99 transcripts in total. The lengths of both, the transcripts in the full and in the capped corpus, are given in Table 9.

Both corpora are available online at: https://www.losst-ad.com/. I have provided only the transcripts, and do not include links to the interviews or audio for ethical reasons.

Speaker pairs (AD and HC speaker)	The length of transcripts in the full corpus in each pair. Average (range)	The cut-off points in the capped corpus in each pair
Pair 1	880 (125-2561) words	417 words
Pair 2	1723 (260-6792) words	390 words
Pair 3	553 (183-1134) words	216 words
Pair 4	1110 (350-2848) words	350 words
Pair 5	1452 (159-4424) words	251 words
Pair 6	887 (141-2170) words	740 words
Pair 7	1510 (275-6385) words	574 words
Pair 8	3702 (197-8621) words	3065 words
Pair 9	1614 (67-5653) words	116 words
Pair 10	2693 (301-6327) words	3554 words

Table 9: The length of transcripts in the full corpus and in the capped corpus (in words) *(AD: Alzheimer's disease, HC: healthy control)*

3.5. Tracking changes in vocabulary richness

I evaluate the corpus against what is known about the development of AD in the medical literature. I conduct two experiments. First, I compare the earliest and latest recordings across the AD and HC group, hypothesising that based on previous literature, the change in the AD group should be more severe. Second, I investigate longitudinal change in vocabulary richness in relation to the time before diagnosis. In both experiments, I use the capped corpus if the language feature of interest is text-length-sensitive, and the full-length corpus if the language feature is not text-length-sensitive.

3.5.1. Comparison of the earliest and the latest samples

I created a subset of samples, consisting of the earliest and latest available recording from each AD participant. The control group samples were chosen by matching the ages of each participant as closely as possible to the respective AD group participant, or when the same age range was not available, by matching the recording intervals. I compared the difference in vocabulary richness features in the earliest samples to the latest recordings between the AD and the HC group.

To compare the values of the earliest and the latest samples, I used a two-tailed paired t-test for the features that were normally distributed, and a two-tailed Wilcoxon signed rank test for those that were not. The results indicate that the average noun frequency, word length and

word frequency differed significantly between the earliest and latest sample in the AD group, but not in the HC group. The statistical details are provided in Table 10.

Figure 6 visualises the data and illustrates the findings: the first column shows the difference between the feature values in the earliest and latest recordings in the AD and the HC group; the second column compares the average change in the feature values between the first and last recording; the third column shows how the feature values of each individual change from the earliest to the latest time point.

Feature	**AD early** (mean unless specified)	**AD late** (mean unless specified)	**AD group difference**	**HC early** (mean)	**HC late** (mean)	**HC group difference**
Noun frequency	0.130	0.098	Paired t-test p=0.028*	0.138	0.123	Paired t-test p=0.062
Word length	3.752	3.552	Paired t-test p=0.002**	3.852	3.833	Paired t-test p=0.832
Word frequency	12.621 (median)	12.963 (median)	Wilcoxon p=0.004**	12.828	12.759	Paired t-test p=0.466

Table 10: Statistical details of the comparison of noun frequency, word length, and word frequency values between the earliest and the latest recordings (*AD: Alzheimer's disease; HC: healthy control)*

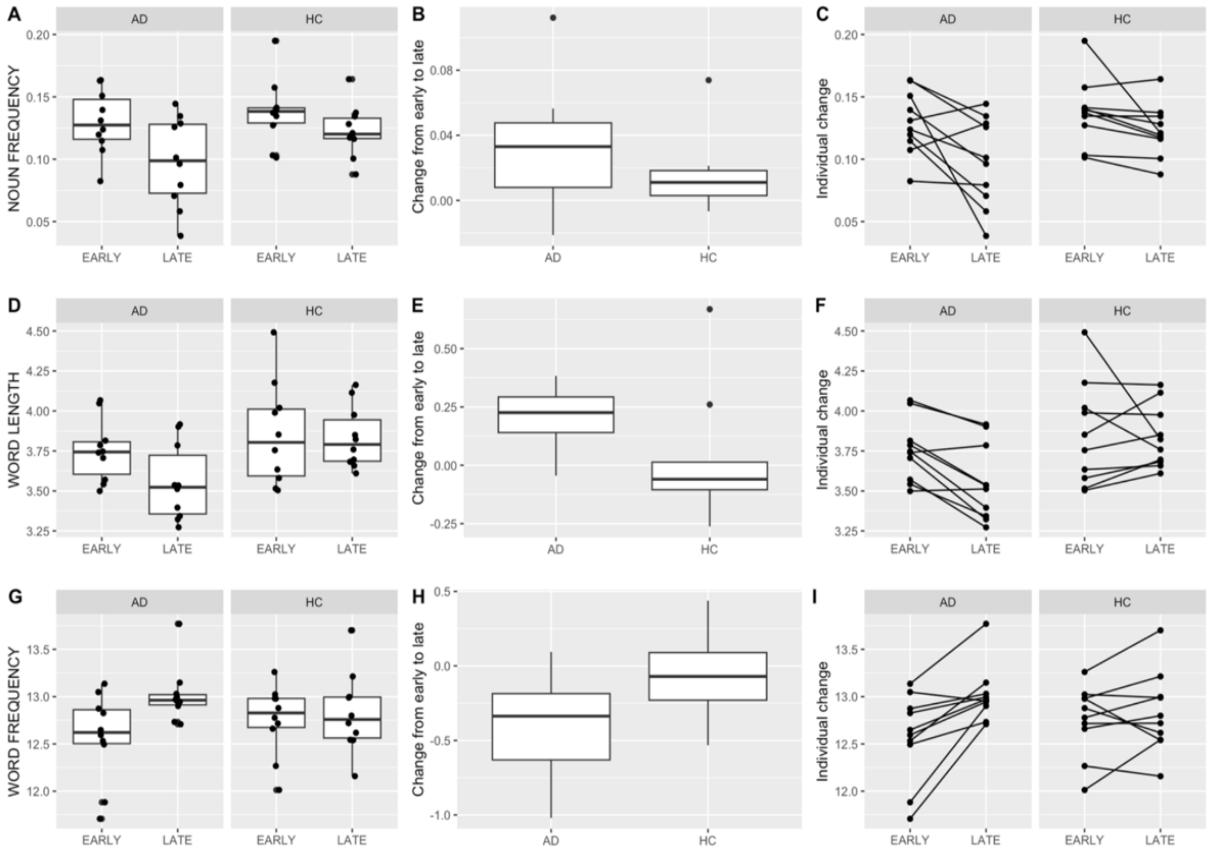


Figure 6: Comparison of noun frequency, word length, and word frequency values between the earliest and latest recordings (*AD: Alzheimer's disease; HC: healthy control)*

3.5.2. Longitudinal change

In this experiment, I investigated whether longitudinal change in vocabulary richness in relation to the time before diagnosis can be detected from the corpus, using all the available recordings and time points. Time before diagnosis was established based on available public sources and media entries. As the control group participants did not have a date of diagnosis, their recordings were mapped to the respective AD participants based on age at any given time point.

A simple linear regression showed that the number of years before the diagnosis in the AD group was a significant predictor of noun frequency (p=0.022), hapax legomena (words used once) (p=0.030), words used once or twice (p=0.035), Brunet index (p=0.031), type:token ratio (p=0.023), adposition frequency (p=0.025), word frequency (Zipf p=0.0005, Subtl p=0.0004),

and the interval of the uni- and bigram repetitions (p=0.003). Significant changes were not detected in the HC group. Noun frequency, hapax legomena, words used once or twice, Brunet index, type:token ratio, and adposition frequency were extracted from the capped corpus, and word frequency measures and the interval of the uni- and bigram repetitions were extracted from the full corpus. The change in vocabulary richness values in relation to the year before diagnosis is shown on Figure 7.

3.6. Discussion and conclusion

In this chapter, I presented a novel language resource - a longitudinal corpus of 135 spontaneous speech transcripts of public interviews with 20 famous individuals, half of whom will eventually be diagnosed with AD, recorded over several decades. This corpus could be highly valuable for research on AD as well as to train a system to automatically detect the risk for AD in speech. The corpus is available online at: https://www.losst-ad.com/.

I demonstrate that public data, the collection of which does not cause extra discomfort for the participants, can carry important information and has a great potential to contribute to developing language-based, fast, cheap, accessible, and non-invasive tools that could aid clinicians and help detect signs of AD early, as well as broaden our understanding of language change in AD in general.

Limited data availability is one of the most challenging issues in tracking language-based cognitive decline in AD, mostly due to data collection being time-consuming and expensive (Luz et al., 2021a). I show that collecting secondary, already publicly available data can help tackle this issue and capture AD-related changes in speech comprehensively, demonstrating the potential of such an approach, and encouraging the collection of more large-scale and controlled datasets to allow for more detailed analysis that could help understand the language changes in AD better.

Using secondary data as in the current study can help reduce participant burden, as no new recordings are conducted. This is not to say that ethical considerations can be overlooked when working with secondary data - researchers are still expected to handle data with care and make decisions with the best interest of the participants in mind. In the current study, I have aimed to

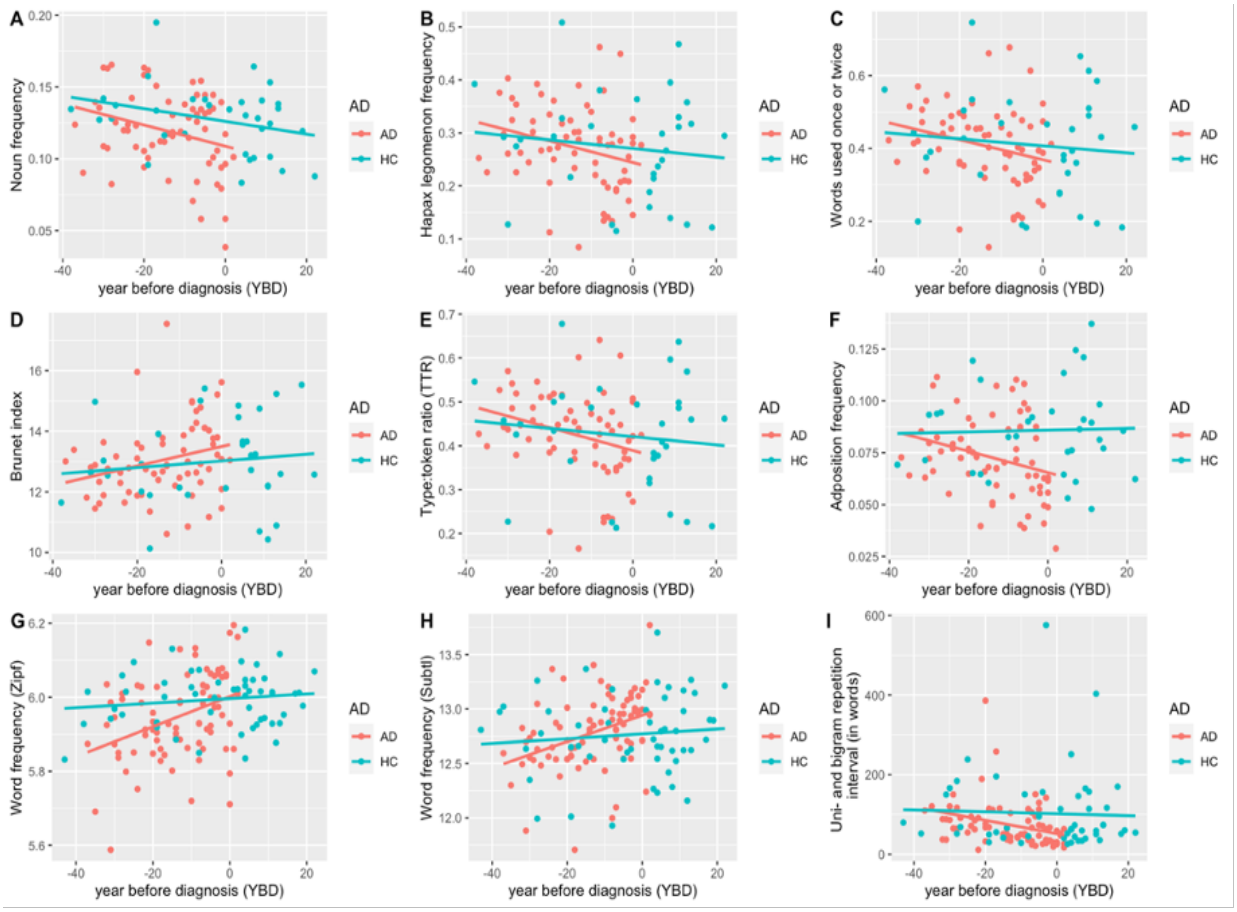


Figure 7: Change in vocabulary richness features in relation to the year before diagnosis (*AD: Alzheimer's disease; HC: healthy control)*

protect the participants' identity by only providing anonymised transcripts, and not include the names of the speakers, links to the interviews or audio. The ethical considerations that arise in the process of speech- and language-based AD detection using NLP and AI are addressed in more detail in Chapter 6.

I conducted two experiments to evaluate the corpus and investigate its ability to track AD-related longitudinal changes in vocabulary richness, and found that significant changes in noun and adposition frequency, word length and frequency, unique words, Brunet index, and repetitions can be captured.

Noun frequency is expected to decline as AD gets more severe (Jarrold et al., 2014; Bucks et al., 2000). In line with previous literature, I found a significant difference in noun frequency between the earliest and the latest recording in the AD group – a change that does not

manifest in the healthy group (Figure 6A-C). Similarly, looking at how noun frequency changes over time reveals an expected declining pattern closer to the time of diagnosis (Figure 7A). As adpositions in English tend to appear together with nouns, the decrease in adposition frequency is also detected as expected (Figure 7F). Supporting these findings, Guinn and colleagues (2014) and Ammar and Ayed (2018) conclude that noun frequency and adposition frequency are among the most informative features in distinguishing those with AD from healthy individuals. My findings suggest that not only can these features contribute to distinguishing the speakers already diagnosed with AD from their healthy peers, but that the significant longitudinal change captured in the current study could indicate that these changes could potentially be captured years before the diagnosis, encouraging more large-scale data collection for in-depth analysis into when these changes manifest, and how reliably they can be captured.

Changes in word length have also been identified as one of the most informative manifestations of AD (Yeung et al., 2021). For example, Balagopalan and colleagues (2021) found that using shorter words can be associated with lower MMSE scores. In line with these findings, the current corpus also captures significantly shorter word length in the late recordings compared to the early ones in the speakers who will go on to develop AD, but not in those who remain healthy (Figure 6D-F).

Previous literature also highlights the importance of word frequency (Yeung et al., 2021) and points out that the people with AD tend to start using more general and frequent words instead of specific ones (for example, “animal” instead of “dog”) (Saito & Takeda, 2001; Hodges et al., 1992). The current study supports these findings: word frequency (measured using either Zipf or Subtl libraries) increases significantly closer to the time of diagnosis in the AD group (Figure 7G-H), and the average values in the early and the late samples also differ significantly (Figure 6G-I).

Hapax legomena, words used once and twice, Brunet index, and type:token ratio are all dependent on the number of unique words used. The decline in unique words, and the importance of these features in AD has been addressed by many (Guinn et al., 2014; Hernández-Dominguez et al., 2018; Fang et al., 2017). In the current study, all four features

reflect a significant decrease in unique words, and therefore point to a decline in vocabulary richness (Figure 7B-E).

I also looked at how well the corpus captures repetition frequency (measured by the number of words between every occasion of a one- or two-word-repetition). These repetitions can reflect false starts (Croot et al., 2000) or stutter (Boyé et al., 2014), both of which are common in AD. I found that the repetitions get significantly more frequent closer to the diagnosis (Figure 7I). In support, Guinn and colleagues (2014) and Ammar and Ayed (2018) propose that repetitions and related errors are among the most informative features of AD-related language changes.

The primary aim of this study was to demonstrate the usefulness of this type of corpus data. I emphasise that this study serves as an indication for future research and does not aim to generalise based on 2x10 individuals. Similarly, the type of data and the lack of medical information allows this study to only be descriptive in nature, and more in-depth medical information and expertise would be needed to argue for a causation between AD and the language changes.

Future work could collect more data with regular intervals to allow for more complex analysis and precise representation of longitudinal changes, explore the AD-related syntactic changes (such as shorter utterance length and lower depths of structure, less frequent use of reported speech and interpolated clauses, and unfinished phrases closer to AD diagnosis) and semantic complexity (such as lower content density, use of information units, and cohesiveness, more frequent topic shifts, revisions, aborted phrases and indefinite words closer to AD diagnosis), and replicate the experiments with automatic transcription and anonymisation.

The main limitations of this study are the small size of the dataset and the lack of medical information, such as the year of diagnosis, stage or severity of the disease, or any co-existing conditions. The information related to the year of diagnosis was based on media entries, but it was not validated by a health professional, and might be untrue - for example, the public figures may not have disclosed their diagnosis immediately. Similarly, there is a lack of medical information about the control group speakers, and while their public information did not include AD-related entries, I did not have access to their validated health records. Unknown medical conditions or potential AD in the control group could have had an impact on the

language changes. For example, Figure 6 shows one HC group participant whose slope differs from the rest of the HC group and is more similar to the AD group participants. It is important to consider the potential impact of the lack of medical information when interpreting these results.

Additionally, the speech data was uncontrolled: using secondary data, it was not possible to control for the content of the questions, or how scripted the interviews were, contributing great variation in the data. Similarly, it was impossible to know if any other factors, such as intoxication, were affecting the participant's speech during the interview. The variability of the data should be considered when interpreting the results, especially the linear regression in Figure 7.

All in all, even with some limitations, the constructed corpus demonstrates that changes in vocabulary richness can be comprehensively captured using public data. These findings are promising and encourage future work in collecting large-scale datasets and developing spontaneous-speech-based tools for early AD detection.

3.7. Summary

In this chapter, I introduced a novel corpus that I collected, consisting of transcripts of public figures' interview recordings, half of whom will eventually receive AD diagnosis. The aim of collecting the corpus was to tackle the issues with data availability, and to build a resource that would allow tracking longitudinal AD-related changes in language. I discussed the language changes that I expect to capture based on previous literature (3.2), and provided details of how I collected the corpus (3.3) and processed the data (3.4). I validated the corpus based on what is known about language changes in AD in the medical literature focussing on vocabulary richness, and showed that this kind of data can be useful in understanding language changes in AD, as vocabulary richness indeed declined significantly over time in the speakers with AD as expected (3.5, 3.6).

4. The generalisability of longitudinal changes in speech before AD diagnosis

4.1. Introduction

In Chapter 3, I showed that longitudinal interview recordings can capture AD-related changes in language known from previous literature. In this chapter, I will focus on the generalisability of these language changes and explore whether they manifest similarly across multiple speakers. The aim of this chapter is to:

- replicate previous findings from Berisha and colleagues (2015) on a larger group of individuals, as suggested by the literature review in section 2.2 (Petti et al., 2020);
- and to address the issue of generalisability and explore the similarities and differences in language changes across multiple speakers who eventually develop AD.

Understanding how language use changes in a larger group of individuals would provide a better understanding of AD-related linguistic changes and help identify the most reliable language features that could capture cognitive decline. This knowledge would contribute to developing more robust, low-dimensional, and interpretable non-invasive tools for detecting the early signs of AD as well as analysing response to treatment and changes over time.
As previously discussed, extensive research into the long-term AD-related changes in language has been challenging due to the lack of available longitudinal language data. Therefore, previous studies have taken alternative approaches, for example, using secondary data from book authors (Le et al., 2011) or comparing press conference transcripts (Berisha et al., 2015). Comparing the writings of Iris Murdoch, Agatha Christie, and Phyllis Dorothy James, Le and colleagues (2011) found that Murdoch showed signs of impoverished vocabulary and syntax in her novels associated with her later dementia diagnosis and proposed that based on the change in Christie's writings, she too was likely to suffer from dementia. Berisha and colleagues (2015) showed that President Reagan, who was later diagnosed with AD, used less unique words and more low imageability verbs than President Bush, and that the number of unique words, fillers

and nonspecific words changed significantly over time in President Reagan's press conferences, but not in Bush's, suggesting that the early signs of dementia were evident in Reagan's speech prior to the diagnosis. Fang and colleagues (2017) found differences in the length of sentences, unique, non-specific, and special words, and the ratio of depth to width of the sentences' parsing tree when comparing the news conference transcripts of Reagan and Bush.

While these studies provide an insight into the individual cases of language changes in dementia, they still represent small, specialised samples, and the need to identify the changes that are generalisable and representative of many cases remains (Sabat & Harre, 1994; Bucks et al., 2000).

To tackle this issue, in this chapter I will analyse the data from two corpora that consist of the transcripts of interviews with public figures, half of whom eventually developed AD. In addition to the LoSST-AD corpus introduced in the previous chapter, I will include a corpus constructed by Winterlight Labs in this analysis to allow including more participants and validate the results. Both corpora consist of the transcripts of public interviews that span over several decades and include 10 and 9 AD-healthy participant pairs.

I aim to understand the extent to which the longitudinal changes in language that manifest prior to AD diagnosis are generalisable in a group of individuals with AD, and which linguistic features, if any, show consistent patterns across the individuals who later in life receive AD diagnosis compared to the matched healthy controls. I first replicate the methods from Berisha and colleagues (2015), and then explore several different approaches to improve the generalisability of the patterns of language change: using an alternative set of language features, age correlations, and compiling single features into aggregate scores.

This chapter is based on an article published in the Journal of Alzheimer's Disease (Petti et al., 2023a).

4.2. Materials and methods

I used two separate corpora of interview transcripts, featuring public figures who eventually develop AD, and their paired controls. I first replicated the methods from Berisha and colleagues (2015), and second, proposed alternative approaches for analysing longitudinal

speech data. The two corpora were collected independently: the first one is the LoSST-AD corpus that I collected as part of this thesis and described in the previous chapter, and the second one was collected by Winterlight Labs. As the corpora were collected separately by different researchers at different times, the methods used to gather, process, and analyse the data varied, further illustrating the need for standardisation in this field, addressed in more detail in Chapter 5. I carried out the same experiments on both corpora to maximise the validity of the results. The methods used to construct LoSST-AD corpus were described in the previous chapter. In the following sections, I will give an overview of the Winterlight Labs Famous People corpus, discuss the extracted language features, and explain the procedure used in this study.

4.2.1. Winterlight Labs corpus

This corpus included 9 individuals and their matched controls (paired based on the same criteria as in the LoSST-AD corpus). The individuals were identified through internet search, and clips of them speaking (e.g. interviews, public appearances, press conferences) that were found on YouTube were used as a data source. This corpus consisted of 405 manually transcribed interviews and monologues that were recorded over a period starting from 46 years before the diagnosis and up to 13 years after the diagnosis. Two participant pairs were female and seven were male. All data was publicly available on YouTube. The two corpora had 5 overlapping AD participants, resulting in the use of some of the same speech samples. See an overview of participants' demographic information in both corpora in Table 11, and Figure 8 for an overview of the distribution of available recordings over time in the two corpora.

4.2.2. Extracted language features

Two datasets of language feature values extracted from the transcripts were constructed based on the two corpora. The approaches to feature extraction differed in the two datasets. In the first dataset, I used the theory-based approach recommended by Berisha and colleagues (2021) and extracted 34 linguistic features that had been identified as the most informative by previous literature. The motivation behind the feature selection was to investigate whether the

	LoSST-AD corpus (Dataset 1)		Winterlight Labs corpus (Dataset 2)	
	AD group (10 participants, 82 recordings)	**HC group** (10 participants, 53 recordings)	**AD group** (9 participants, 217 recordings)	**HC group** (9 participants, 188 recordings)
Average age over all recordings (years)	60 (SD: 11, range: 32-79)	70 (SD: 16, range: 30-97)	65 (SD: 12, range: 31-97)	60 (SD: 16, range: 28-88)
Average time before diagnosis (years)	13 (SD: 11, range: 37 years before diagnosis - 2 year after diagnosis)	-	11 (SD: 12, range: 46 years before diagnosis – 13 years after diagnosis)	
Sex	5F, 5M	5F, 5M	2F, 7M	2F, 7M

Table 11: Participants' demographic information in the two corpora (*AD: Alzheimer's disease; HC: healthy control; SD: standard deviation; F: female; M: male*)

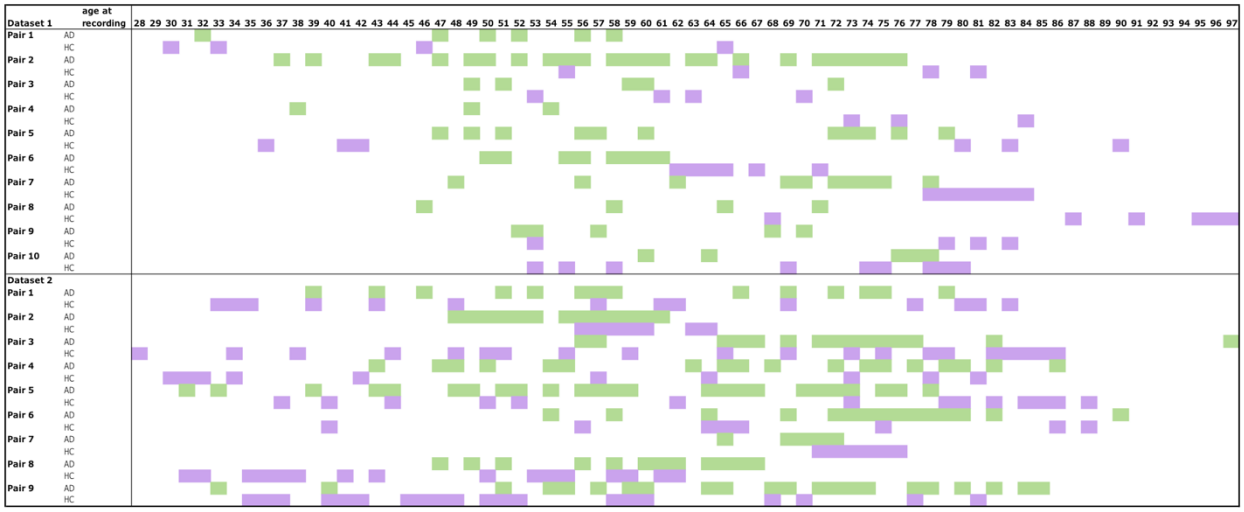


Figure 8: Distribution of recordings over time in the two corpora (*AD: Alzheimer's disease; HC: healthy control)*

features that have shown the clearest difference between the AD and the control group in prior research also produce measurable and generalisable patterns of language change in longitudinal speech data. Berisha and colleagues (2021) argues that rather than using a large set of automatically extracted features, research should focus on identifying a small set of consistently changing features, as this would contribute to lowering the dimensionality of the models and boost robustness.

As some feature values in this dataset were dependent on transcript length (significant Pearson correlation between the feature value and the number of tokens), I adopted a method of capping the transcripts at an even number of words for extracting only the text-length-sensitive features - a method also used by Le and colleagues (2011), and described in the previous chapter. I established a separate word limit for each participant pair that would allow keeping the longer samples while also maintaining at least 3 samples per participant to allow investigating longitudinal change. This resulted in each participant pair having a different length of the capped transcripts that were used for calculating the text-length-sensitive features only but allowed within-pair comparison as both the HC and the AD participant samples were the same length. Details of the lengths of the capped transcripts are given in Table 9.

See Table 12 for details of the features extracted from the LoSST-AD corpus, including the studies that have previously reported these features to be informative, and the information about whether the feature was text-length-sensitive and therefore extracted using the capped transcripts.

The dataset based on Winterlight Labs corpus included 300 linguistic features extracted from the transcripts, including lexical, syntactic, and semantic features, such as the proportions of different parts-of-speech (POS) tags, vocabulary richness statistics, syntax tree features, coherence features measured using cosine distances, and sentiment scores. The length of the speech samples in this corpus was not limited.

category	variable	informative according to	sensitive to text length
vocabulary richness	Brunet's index	Guinn et al., 2014	YES
	Honoré's statistic	Fraser et al., 2016, Yancheva et al., 2015, Guinn et al., 2014	YES
	indefinite nouns	Le et al., 2011	NO
	open:closed class ratio	Beltrami et al., 2018, Zimmerer et al., 2016	NO
	proportion of frequent verbs	Le et al., 2011	NO
	hapax legomena frequency	Hernandez-Dominguez et al., 2018	YES
	type:token ratio	Guinn et al., 2014	YES
	word frequency	Yeung et al., 2021	NO
	word length	Yeung et al., 2021	YES
	words used once and twice	Hernandez-Dominguez et al., 2018, combined with Sichel's S	YES
POS-related	adjective frequency	Guinn et al., 2014	YES
	adjective type frequency	Le et al., 2011	NO
	adposition frequency	Ammar & Ayed, 2018	YES
	adverb frequency	Beltrami et al., 2018, Fraser et al., 2016, Ammar & Ayed, 2018	YES
	adverb type frequency	Le et al., 2011	NO
	auxiliary frequency	Hernandez-Dominguez et al., 2018, Ammar & Ayed 2018	YES
	be - get difference	Le et al., 2011	YES
	be:auxiliary ratio	Le et al., 2011	YES
	conjunction frequency	Ammar & Ayed, 2018	NO
	determiner frequency	Ammar & Ayed, 2018	YES
	get:auxiliary ratio	Le et al., 2011	NO
	noun frequency	Guinn et al., 2014, Ammar & Ayed, 2018	YES

category	variable	informative according to	sensitive to text length
	noun type frequency	Le et al., 2011	YES
	noun:pronoun ratio	Fraser et al., 2016, Khodabakhsh et al., 2015	YES
	pronoun frequency	Khodabakhsh et al., 2015, Ammar & Ayed, 2018, Yancheva et al., 2015	YES
	verb frequency	Guinn et al., 2014	NO
	verb type frequency	Le et al., 2011	YES
	words indicating quantities	Sadeghian et al., 2017	NO
grammatical	inflected verbs	Fraser et al., 2016; Zimmerer et al., 2016	NO
	past	Ammar & Ayed, 2018	YES
	reported speech	Boye et al., 2014	YES
errors	number of repeated open class words in 10 consecutive open class words	Le et al., 2011	NO
	unigram + bigram repetitions	Guinn et al., 2014, Ammar & Ayad, 2018	YES
	filler frequency	Boye et al., 2014, Gosztolya et al., 2016	NO

Table 12: Features extracted from the LoSST-AD corpus

While I acknowledge that there have been recent developments in detecting AD from acoustic speech signals, such as using the extraction of higher order spectra features (Nasrolahzadeh et al., 2022), non-linear features (Lopez-de-Ipina et al., 2020), and emotional factors (Lopez-de-Ipina et al., 2018), there are several reasons why I have focussed only on the linguistic features in the current study:

- The datasets are based on YouTube videos with extensively varying audio quality, making the extraction of the acoustic features challenging.
- Sophisticated acoustic methods can be hard to interpret or relate back to clinically observable phenomena, which make clinicians hesitant to use them.
- This work is aiming to replicate the methods from Berisha and colleagues (2015) who have used written text and focussed on linguistic features in their analysis.

4.2.3. Procedure

I conducted 5 experiments. The first two experiments replicated the methods from Berisha and colleagues (2015) on two larger datasets to understand whether their findings are generalisable to a larger group of people who develop AD. The following three experiments proposed alternative ways of data analysis by looking at different language features, age instead of transcript index, and compiling single language features into aggregate scores.

Experiment 1

The aim of this experiment was to investigate whether the AD and HC groups show differences in the use of the language features reported by Berisha and colleagues (2015), as well as to find out how many participant pairs independently show differences between the AD and the HC participant.

I first conducted independent t-tests in both datasets using all the available samples to compare the performance of the AD and HC groups in the features similar to unique words, low imageability (LI) verbs, fillers, and non-specific words that were used by Berisha and colleagues.

In the LoSST-AD corpus, these features were type:token ratio (TTR), frequent verbs, fillers, and indefinite nouns and pronouns respectively. In the Winterlight Labs corpus, these features were TTR and moving average TTR (MATTR), light verbs, interjections, and pronouns respectively. Then, I looked at the differences at the participant pair level.

I hypothesised that if the language changes reported in the case study (lower number of unique words, and higher number of LI verbs, fillers, and nonspecific words in the AD group) are generalisable to a larger group of speakers, the group and participant level t-tests should reflect that.

Experiment 2

In the second experiment, I replicated Berisha and colleagues' approach of analysing the Pearson correlation between the transcript index (assigned based on the order in which the samples were recorded) and the feature value, using the same language features as in Experiment 1 and both group and single participant pair analyses. The aim of this experiment was to explore the patterns of change in the language features over time.

I hypothesised that if the language changes reported in the case study (decline in the number of unique words, and rise in the number of LI verbs, fillers, and nonspecific words in the AD group over time) are generalisable to a larger group of speakers, the AD group samples should show a more significant correlation between the transcript index and language feature value than the healthy group samples.

Experiment 3

In Experiment 3, I used the same methodology as in Experiment 2, but introduced new language features that have been identified as informative in previous studies: average word length, noun:pronoun ratio, particle frequency, noun frequency, word frequency and constituency average depth. The aim was to understand whether using different language features improves the generalisability of the patterns of language change.

I hypothesised that the AD group speakers should have a significantly lower word length, noun:pronoun ratio, and constituency average depth, and higher particle value and word frequency, compared to the control group speakers.

Experiment 4

In Experiment 4, I investigated the Pearson correlations between participants' age and the values of the language features that had been used in the previous experiments. The aim was to investigate whether using participants' age instead of transcript index improves the results by accounting for the different time intervals of the recordings. I hypothesised that the language changes over time would be more significant in the AD group speakers.

Experiment 5

In Experiment 5, I introduced an approach of compiling single language features into aggregate scores based on the previous literature. The aim of this experiment was to tackle the issue of single language features failing to show highly generalisable patterns of language change across the participants, and to understand whether there is a benefit to using a sum score of a group of features relating to a certain type of impairment instead of single language features. Using aggregate scores also improves the interpretability and accessibility of the results as the single language features are often technical, specific, and difficult to interpret.

The following five aggregate scores were computed:

(1) lexical diversity;
(2) word finding difficulty;
(3) discourse;
(4) syntactic;
(5) parts-of-speech (POS) scores.

The lexical diversity aggregate score consisted of vocabulary richness indices (Brunet index (Brunet, 1978), Honoré statistic (Honore, 1979)), pronoun proportions, type:token ratio (TTR) features, hapax legomena, indefinite nouns, total number of words, speech rate, and maximum

utterance length. This aggregate was based on Fang et al. (2017), Pistono et al. (2016), Luz et al. (2021b), Beltrami et al. (2018), Boye et al. (2014), Baldas et al. (2011), and Dijkstra et al. (2004).
The word finding difficulty aggregate score consisted of the proportion of function and open class words, fillers, repetitions and stutters, pronouns, indefinite nouns, word frequency, filled and unfilled pauses, and their ratio to words, and hesitations. This aggregate was based on Appell et al. (1982), Hodges et al. (1992), Obler and Albert (1981), Pistono et al. (2016), Guinn et al. (2014), Khodabakhsh et al. (2015), Boye et al. (2014), Baldas et al. (2011), Gosztolya et al. (2019), Rousseaux et al. (2010), Singh et al. (2001), and Szatloczki et al. (2015).
The syntactic aggregate score consisted of utterance length and constituency features and was based on Beltrami et al. (2018), Boye et al. (2014) and Ullman (2004).
The discourse aggregate score consisted of local coherence features calculated using cosine distances, TTR features, adpositions, pronouns, conjunctions, hapax legomena, indefinite nouns and open:closed class ratio. This aggregate was based on Lima et al. (2014), Taler and Phillips (2008), Beltrami et al. (2018), and Dijkstra et al. (2004).
The POS aggregate score consisted of the proportions and ratios of nouns, pronouns, verbs, and adjectives, and was based on Appell et al. (1982), Khodabakhsh et al. (2015), Boye et al. (2014), and Baldas et al. (2011).

To calculate the aggregate scores, all single features belonging to the aggregate were z-scored, the polarity of the features was considered, and the average of the polarised z-scores was used as the aggregate score. The syntactic aggregate score was not constructed for Dataset 1 due to the nature of the data and the extracted features.
I compared the AD and HC groups as well as the independent participant pairs in both corpora using independent t-tests. Transcript index and age correlations with aggregate scores to analyse longitudinal change were conducted only in Dataset 2, as in Dataset 1, some of the features in each aggregate were text-length-sensitive and therefore calculated based on capped transcripts, resulting in often having only three data points per participant for which the aggregate score was available. I hypothesised that the aggregate scores would change more significantly over time in the AD group and show a clearer pattern that the single features.

4.3. Results

4.3.1. Experiment 1

In this experiment I replicated the independent t-tests from Berisha and colleagues (2015), focussing on the features relating to unique words, low imageability (LI) verbs, fillers, and nonspecific words. I conducted t-tests on group level, comparing all samples from the AD and HC groups, and on the participant pair level, comparing the samples of the matched individuals in each pair. To follow the methods from Berisha and colleagues (2015), when comparing the results on participant pair level, I did not lower the p-value for significance testing although multiple pairs were tested. Therefore, the potential occurrence of false positives must be acknowledged. Details of results are given in Table 13.

I found no significant differences between the HC and AD participant groups in either corpus when looking at unique words, measured in type:token ratio (TTR) and moving average TTR (MATTR) features, or LI verbs, measured in frequent and light verbs.

When looking at the participant pairs separately, I found that 3 out of 9 AD participants differed significantly in the expected direction from the matched HC participants in the moving average TTR (MATTR) values in Corpus 2 (MATTR was measured in 10-, 20-, 30,- 40-, and 50-word windows but as all MATTR features acted similarly, I only report the findings of the 20-word window in Table 13 to avoid repetition). For TTR and LI verb features, 0 or 1 participant pair differed in the expected direction in both corpora.

While the use of fillers and nonspecific words did not differ significantly in Berisha and colleagues' study (2015), both datasets in the current study showed significant differences when comparing the AD and HC groups in fillers and pronouns, but not in indefinite nouns. However, fillers in Dataset 1 differed in unexpected direction in group comparison and no differences were found in participant pair comparison, and only 1 pair differed significantly in the expected direction in Dataset 2.

The use of pronouns was significantly higher in the AD group in both datasets, with 2 and 3 participant pairs showing expected patterns in Dataset 1 and 2 respectively.

Feature	Dataset	Expected direction in AD	AD and HC group comparisons					Pairwise comparisons				
			Significant	*t*	*p*	AD mean (SD)	HC mean (SD)	Change in expected direction	Total significantly different pairs	Significantly different pairs in expected direction	Expected direction AD mean (SD)	Expected direction HC mean (SD)
Unique words	*B*	*lower*	*yes*	–	–	–	–	*yes*	*1/1*	*1/1*	–	–
TTR	1	lower	no	t(97) = 0.063	0.950	0.435 (0.098)	0.423 (0.124)	–	0/10	0/10	–	–
TTR	2	lower	no	t(403) = –0.954	0.341	0.351 (0.129)	0.339 (0.132)	–	4/9	1/9	0.335 (0.094)	0.450 (0.149)
MATTR_20	2	lower	no	t(401) = 0.529	0.565	766 (0.046)	0.768 (0.040)	–	4/9	**3/9**	1) 0.745 (0.028) 2) 0.760 (0.023) 3) 0.761 (0.014)	1) 0.761 (0.025) 2) 0.788 (0.021) 3) 0.817 (0.018)
LI verbs	*B*	*higher*	*yes*	–	–	–	–	*yes*	*1/1*	*1/1*	–	–
Frequent verbs	1	higher	no	t(133) = –0.206	0.837	0.448 (0.101)	0.452 (0.117)	–	3/10	1/10	0.581 (0.121)	0.372 (0.135)
Light verbs	2	higher	no	t(403) = –0.274	0.784	0.034 (0.016)	0.034 (0.015)	–	1/9	0/9	–	–
Fillers	*B*	*higher*	*no*	–	–	–	–	–	–	–	–	–
Filler frequency (words between fillers)	1	lower	yes	t(124) = 2.828	0.005	92.91 (98.274)	54.67 (55.113)	no	1/10	0/10	–	–
Interjection proportion	2	higher	yes	t(267) = –2.666	0.008	0.142 (0.014)	0.020 (0.028)	**yes**	3/9	1/9	0.007 (0.004)	0.004 (0.002)
Nonspecific words	*B*	*higher*	*no*	–	–	–	–	–	–	–	–	–
Indefinite nouns	1	higher	no	t(133) = –0.079	0.937	0.054 (0.054)	0.055 (0.049)	–	1/10	1/10	0.041 (0.029)	0.092 (0.040)
Pronouns	1	higher	yes	t(97) = 2.767	0.007	0.166 (0.033)	0.145 (0.041)	**yes**	2/10	**2/10**	1) 0.173 (0.022) 2) 0.113 (0.007)	1) 0.102 (0.037) 2) 0.097 (0.006)
Pronouns	2	higher	yes	t(403) = 1.990	0.047	0.147 (0.038)	0.139 (0.033)	**yes**	3/9	**3/9**	1) 0.145 (0.014) 2) 0.157 (0.019) 3) 0.164 (0.027)	1) 0.103 (0.015) 2) 0.115 (0.031) 3) 0.146 (0.020)

Table 13: Independent samples t-test between AD and HC participants (*AD: Alzheimer's disease; HC: healthy controls; SD: standard deviation; dataset B: Berisha et al. (2015) results comparison; TTR: type:token ratio; MATTR_20: moving average type:token ratio in 20-word window; LI: low imageability)*

4.3.2. Experiment 2

In Experiment 2, I replicated the transcript index and feature value correlations from Berisha and colleagues (2015) using Pearson correlation. I compared the number of AD and HC participants that showed significant correlations.

While the use of unique words differed significantly in the AD participant in Berisha and colleagues' study (2015), TTR values only differed in the expected direction in 1 or 2 AD participants in the current datasets, and in 1 HC participant in both datasets. However, MATTR values showed promising results with more than half of the AD participants showing significant correlation between the feature value and transcript index in Dataset 2 (as MATTR features act similarly, only the 20-word window results presented in Table 14 to avoid repetition).

In line with Berisha and colleagues' study (2015), features related to LI verbs did not show significant correlation in the expected direction with transcript index in the AD participants. While fillers showed significant change with transcript index in the single tested participant pair

in Berisha and colleagues (2015), only one AD participant differed in the expected direction in the two larger corpora of the current study.

From the nonspecific word category, pronoun use changed significantly in the expected direction in 3/9 participants in Dataset 2.

See Table 14 for details.

When interpreting these results, it must be kept in mind that the participant-level correlations of the text-length-sensitive features (shown in Table 12) in Dataset 1 can be based on very few (minimum 3) observations, and only serve as an indication. I have not included the r value for these features in Table 14 as due to the low number of observations, the correlation can be exaggerated.

Feature	Dataset	Expected direction in AD	Total AD participants with significant correlation	AD participants with expected correlation	r in expected correlations in AD	Total HC participants with significant correlation	r in significant HC participants
unique words	*B*	*lower*	*1/1*	1/1	–	0/1	–
TTR	1	lower	1/10	1/10	–	1/10	–
TTR	2	lower	3/9	2/9	1) $r = -0.567$ 2) $r = -0.761$	1/9	$r = 0.529$
MATTR_20	2	lower	5/9	**5/9**	1) $r = -0.775$ 2) $r = -0.623$ 3) $r = -0.376$ 4) $r = -0.567$ 5) $r = -0.403$	2/9	1) $r = -0.389$ 2) $r = -0.800$
LI verbs	*B*	*higher*	*0/1*	0/1	–	0/1	–
Frequent verbs	1	higher	0/10	0/10	–	1/10	$r = -0.883$
Light verbs	2	higher	2/9	0/9	–	1/9	$r = 0.529$
fillers	*B*	*higher*	*1/1*	1/1	–	0/1	–
Filler frequency (words between fillers)	1	lower	2/10	1/10	–	0/10	–
Interjection proportion	2	higher	1/9	1/9	$r = 0.430$	1/9	$r = 0.650$
Nonspecific words	*B*	*higher*	*1/1*	1/1	–	0/1	–
Indefinite nouns	1	higher	0/10	0/10	–	0/10	–
Pronouns	1	higher	1/10	1/10	–	0/10	–
Pronouns	2	higher	3/9	**3/9**	1) $r = 0.422$ 2) $r = 0.417$ 3) $r = 0.637$	2/9	1) $r = 0.870$ 2) $r = 0.475$

Table 14: Transcript index correlations between AD and HC participants (*AD: Alzheimer's disease; HC: healthy controls; dataset B: Berisha et al. (2015) results comparison; TTR: type:token ratio; MATTR_20: moving average type:token ratio in 20-word window; LI: low imageability)*

4.3.3. Experiment 3

As the features used in Experiment 1 and 2 did not show consistent patterns across all pairs, in this experiment I looked at an alternative set of features that have been informative in previous studies (average word length, noun:pronoun ratio, particle frequency, noun frequency, word frequency and constituency average depth) (Robin et al., 2022) and tested their correlation with the transcript index. The results were as promising as those of the previous experiments, with the strongest results emerging in the ratio of pronouns to the sum of pronouns and nouns in Dataset 2 where 4 out 9 AD participants changed significantly in the expected direction. 3/9 of the AD participants showed significant correlation in the expected direction between the transcript index and the average word length, word frequency and the constituency average depth in Dataset 2. See Table 15 for details.

As the average word length and noun:pronoun ratio in Dataset 1 were text-length-sensitive, resulting in very few observations on participant-pair-level due to the capped transcripts, the r values of these correlations have not been included in the table, and the number of correlations in these two features should be treated with caution.

Feature	Dataset	Expected direction in AD	Total AD participants with significant correlation	No of AD participants correlating in expected direction	r in expected correlations in AD	Total HC participants with significant correlation	r in significant HC participants
Average word length	1	lower	0/10	0/10	–	1/10	–
Average word length	2	lower	3/9	**3/9**	1) $r = -0.566$ 2) $r = -0.677$ 3) $r = -0.438$	1/9	$r = -0.818$
Noun:pronoun ratio	1	lower	1/10	1/10	–	1/10	–
Pronouns: nouns+pronouns ratio	2	higher	4/9	**4/9**	1) $r = 0.422$ 2) $r = 0.393$ 3) $r = 0.680$ 4) $r = 0.520$	2/9	1) $r = 0.663$ 2) $r = 0.896$
Particles	2	higher	0/9	0/9	–	0/9	–
Noun frequency	1	lower	0/10	0/10	–	0/10	–
Noun frequency	2	lower	2/9	2/9	1) $r = -0.494$ 2) $r = -0.709$	2/9	1) $r = -0.647$ 2) $r = -0.854$
Word frequency	1	higher	2/10	1/10	$r = 0.672$	1/10	$r = 0.983$
Word frequency	2	higher	3/9	**3/9**	1) $r = 0.391$ 2) $r = 0.693$ 3) $r = 0.484$	1/9	$r = 0.586$
Constituency average depth	2	lower	3/9	**3/9**	1) $r = -0.414$ 2) $r = -0.450$ 3) $r = -0.516$	1/9	$r = -0.468$

Table 15: Alternative features correlating with transcript index in AD and HC (*AD: Alzheimer's disease; HC: healthy control)*

4.3.4. Experiment 4

To account for the fact that the interviews were not recorded with consistent time intervals, in this experiment I used participants' age instead of transcript index to measure Pearson correlation with the features from previous experiments. However, no significant improvements in generalisability were achieved. See Table 16 for details.

As the correlations for TTR, fillers, pronouns, average word length, noun:pronoun ratio and noun frequency were calculated on very few data points in Dataset 1 due to the text-length-sensitivity of these features, the r values for these features have not been included in Table 16, and the number of correlations for these features in this dataset should be interpreted with caution.

4.3.5. Experiment 5

To address the issue of generalisability demonstrated by the single features, I compiled aggregate scores of lexical diversity, word finding difficulty, discourse building, syntactic complexity and POS-related features in this experiment.

First, I looked at the differences between the AD and HC groups, and the single participant pairs using independent t-tests as I did in Experiment 1 with the single features. I found that word finding difficulty scores differed significantly between the AD and the HC group in both datasets, however, only Dataset 2 showed greater word finding difficulty in the AD group, and this tendency did not carry on to the individual participant pair level. A maximum of 2 participant pairs per dataset showed significant difference in the expected direction across aggregate scores (lexical diversity, discourse, and POS aggregates). See Table 17 for details.

Second, I looked at Pearson correlations between the aggregate scores, and age and transcript index in the participant pairs in Dataset 2. Dataset 1 was excluded from this experiment due to the limited data availability because of transcript capping.

I found that lexical diversity scores correlated with age and transcript index in the majority of the AD participants. While the number of AD participants that showed a significant correlation between the aggregate score with both transcript index and age was higher in all aggregates

compared to the number of HC participants, it did not exceed half the group for any other aggregate scores, with between 1 and 4 significant correlations. See Table 18 for details.

Feature	Dataset	Expected direction	Total AD participants with significant correlation	No of AD participants correlating in expected direction	r in expected correlations in AD	Total HC participants with significant correlation	r in significant HC participants
TTR	1	lower	0/10	0/10	–	1/10	–
TTR	2	lower	2/9	1/9	$r = –0.503$	2/9	1) $r = –0.386$ 2) $r = 0.499$
MATTR_20	2	lower	4/9	**4/9**	1) $r = –0.736$ 2) $r = –0.520$ 3) $r = –0.410$ 4) $r = –0.485$	3/9	1) $r = –0.417$ 2) $r = –0.418$ 3) $r = –0.812$
Frequent verbs	1	higher	0/10	0/10	–	2/10	1) $r = –0.883$ 2) $r = –0.983$
Light verbs	2	higher	1/9	0/9	–	0/9	–
Filler frequency (words between fillers)	1	lower	1/10	1/10	–	0/10	–
Interjection proportion	2	higher	0/9	0/9	–	1/9	$r = –0.459$
Indefinite nouns	1	higher	0/10	0/10	–	0/10	–
Pronouns	1	higher	0/10	0/10	–	2/10	–
Pronouns	2	higher	3/9	3/9	1) $r = 0.478$ 2) $r = 0.474$ 3) $r = 0.495$	2/9	1) $r = 0.900.$ 2) $r = 0.525$
Average word length	1	lower	0/10	0/10	–	1/10	–
Average word length	2	lower	2/9	2/9	1) $r = –0.554$ 2) $r = –0.457$	2/9	1) $r = 0.390$ 2) $r = –0.831$
Noun:pronoun ratio	1	lower	0/10	0/10	–	2/10	–
Pronouns: nouns+pronouns ratio	2	higher	4/9	**4/9**	1) $r = 0.422$ 2) $r = 0.416$ 3) $r = 0.516$ 4) $r = 0.520$	2/9	1) $r = 0.675$ 2) $r = 0.923$
Particles	2	higher	0/9	0/9	–	0/9	–
Noun frequency	1	lower	1/10	1/10	–	0/10	–
Noun frequency	2	lower	3/9	3/9	1) $r = –0.461$ 2) $r = –0.560$ 3) $r = –0.495$	2/9	1) $r = –0.631$ 2) $r = –0.861$
Word frequency	1	higher	2/10	1/10	$r = 0.659$	0/10	–
Word frequency	2	higher	3/9	3/9	1) $r = 0.363$ 2) $r = 0.509$ 3) $r = 0.490$	1/9	$r = –0.388$
Constituency average depth	2	lower	3/9	3/9	1) $r = –0.466$ 2) $r = –0.466$ 3) $r = –0.430$	0/9	–

Table 16: All features correlating with participants' age in AD and HC (*AD: Alzheimer's disease; HC: healthy control; TTR: type:token ratio; MATTR_20: moving average type:token ratio in 20-word window)*

Aggregate	Dataset	Expected direction in AD	Group comparisons: Significant	t	p	AD mean (SD)	HC mean (SD)	Pair comparisons: Change in expected direction	Total significantly different pairs	Significantly different pairs in expected direction	Expected direction AD mean (SD)	Expected direction HC mean (SD)
Lexical diversity	1	lower	no	t(97) = –1.142	0.256	0.363 (4.173)	0.726 (5.031)	–	2/10	1/10	–0.388 (2.312)	6.539 (3.392)
	2	lower	no	t(403) = –1.083	0.281	0.335 (6.837)	0.387 (6.568)	–	3/9	2/9	1) 1.840 (3.913) 2) 2.353 (1.763)	1) 5.051 (5.946) 2) 6.833 (3.236)
Word finding difficulty	1	higher	yes	t(97) = –2.859	0.005	0.518 (2.620)	1.036 (2.398)	no	3/10	1/3	3.434 (1.094)	–0.694 (1.104)
	2	higher	yes	t(403) = 2.263	0.024	0.493 (4.319)	–0.568 (5.116)	**yes**	0/9	0/9	–	–
Discourse	1	higher	no	t(97) = –2.859	0.306	1.990 (2.442)	–0.398 (3.211)	–	3/10	2/10	1) 1.498 (1.561) 2) –2.127 (.876)	1) 3.754 (0.784) 2) 0.085 (0.576)
	2	higher	no	t(403) = 1.269	0.206	0.789 (13.564)	0.911 (13.362)	–	3/9	2/9	1) –8.171 (11.256) 2) 2.144 (5.557)	1) 1.131 (10.541) 2) 10.930 (5.951)
Syntactic	2	lower	no	t(403) = –0.139	0.889	0.040 (6.109)	0.047 (6.495)	–	1/9	0/9	–	–
POS	1	lower	no	t(97) = 0.730	0.467	0.991 (2.024)	–0.198 (1.649)	–	1/10	1/10	–0.259 (0.988)	–2.148 (1.865)
	2	lower	no	t(403) = 1.371	0.171	0.209 (3.491)	–0.242 (3.070)	–	2/9	2/9	1) 1.114 (1.443) 2) 0.905 (0.602)	1) 2.247 (1.443) 2) 3.319 (1.354)

Table 17: Independent t-test between AD and HC participants using aggregate scores (*AD: Alzheimer's disease; HC: healthy control; SD: standard deviation; POS: part-of-speech)*

Aggregate	Dataset	Expected direction	Transcript index: Total AD participants with significant correlation	AD in expected direction	r in expected correlations in AD	Total HC participants with significant correlation	r in significant HC participants	Age: Total AD participants with significant correlation	AD in expected direction	r in expected correlations in AD	Total HC participants with significant correlation	r in significant HC participants
Lexical diversity	2	lower	**6/9**	**6/9**	1) $r = -0.524$ 2) $r = -0.584$ 3) $r = -0.439$ 4) $r = -0.415$ 5) $r = -0.432$ 6) $r = -0.435$	2/9	1) $r = -0.800$ 2) $r = -0.459$	**5/9**	**5/9**	1) $r = -0.509$ 2) $r = -0.601$ 3) $r = -0.393$ 4) $r = -0.405$ 5) $r = -0.377$	3/9	1) $r = -0.847$ 2) $r = -0.399$ 3) $r = -0.500$
Word finding difficulty	2	higher	3/9	3/9	1) $r = 0.487$ 2) $r = 0.836$ 3) $r = 0.442$	2/9	1) $r = -0.525$ 2) $r = 0.658$	3/9	3/9	1) $r = 0.420$ 2) $r = 0.661$ 3) $r = 0.388$	1/9	$r = -0.542$
Discourse	2	lower	3/9	3/9	1) $r = -0.604$ 2) $r = -0.543$ 3) $r = -0.474$	1/9	$r = -0.672$	3/9	3/9	1) $r = -0.523$ 2) $r = -0.512$ 3) $r = -0.426$	1/9	$r = -0.710$
Syntactic	2	lower	2/9	2/9	1) $r = -0.426$ 2) $r = -0.423$	1/9	$r = -0.443$	2/9	2/9	1) $r = -0.451$ 2) $r = -0.344$	0/9	–
POS	2	higher	3/9	3/9	1) $r = 0.566$ 2) $r = 0.644$ 3) $r = 0.596$	2/9	1) $r = 0.848$ 2) $r = 0.644$	4/9	4/9	1) $r = 0.439$ 2) $r = 0.566$ 3) $r = 0.502$ 4) $r = 0.513$	2/9	1) $r = 0.840$ 2) $r = 0.666$

Table 18: Transcript index and age correlations with aggregate scores (*AD: Alzheimer's disease; HC: healthy control; POS: part-of-speech)*

4.4. Discussion and conclusion

The current chapter focussed on the generalisability of longitudinal language change in AD. I started by replicating the methods from Berisha and colleagues (2015), who compared the speech of President Bush (HC) and President Reagan (AD) looking at unique words, low imageability (LI) verbs, fillers, and nonspecific words. Instead of using just one participant pair, I included two similar corpora of spontaneous speech recordings with public figures, consisting of 10 and 9 AD-HC participant pairs. As I could not replicate the results of the previous study or

find other generalisable patterns using these methods, I proposed 3 alternative approaches to data analysis:

- using different single features;
- using age instead of transcript index;
- and compiling single features into aggregate scores.

While universal patterns of language change representative of many cases were challenging to capture, highlighting the relevance of individual variabilities during speech and use of language, moving average type:token ratio, pronoun-related features, and lexical diversity aggregate score showed the most promising results.

The flowchart on Figure 9 presents an overview of the methods and results of the study.

While Berisha and colleagues (2015) found that the use of unique words and LI verbs was significantly different between the two participants, I failed to replicate these results in larger groups in either corpus. Looking at the number of individual pairs where the AD and the HC participant's performance differed significantly in the expected direction, type:token ratio in 20-word window (MATTR_20) in Dataset 2 performed best, but still only 3 out of 9 participant pairs showed the pattern captured by Berisha and colleagues (2015) and Fang and colleagues (2017), suggesting that the results are generalisable to one third of the participants. When considering the potential effect of multiple comparisons and the chance of false positives, this number could be even smaller.

While Berisha and colleagues (2015) did not find a significant difference in the use of non-specific words and fillers when comparing the speech of Bush and Reagan, Fang and colleagues (2017) reported a difference in the use of non-specific words. The current study found significant group differences in the use of fillers and pronouns in both datasets, but the use of fillers differed in an opposite direction in the two datasets. The number of individual pairs with significant differences in the expected direction in these features was the highest in the pronoun category in Dataset 2, with one third of the participants with AD using significantly more pronouns than their matched controls. These findings suggest that although group differences between the HC and the AD participants may appear, the differences are often not

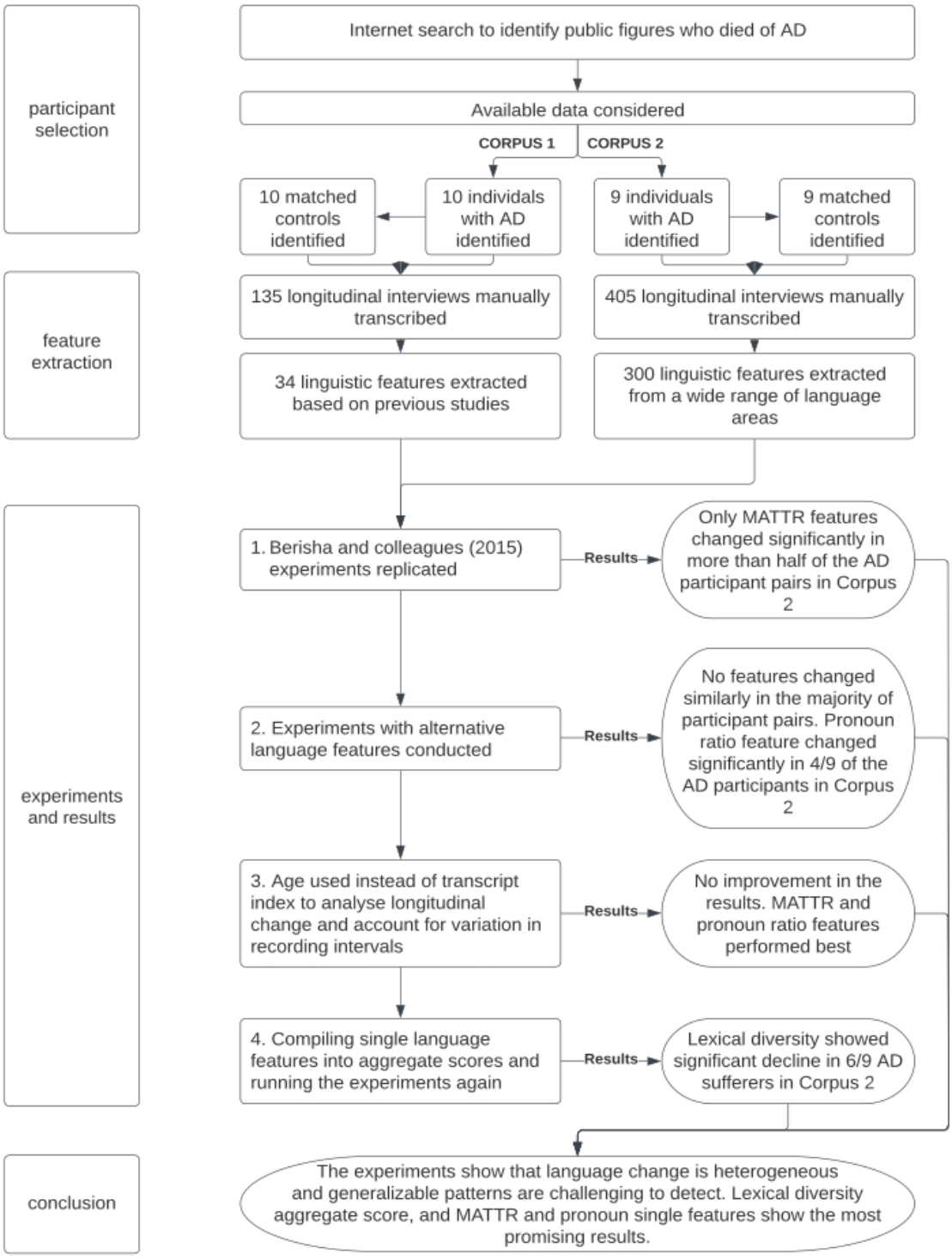


Figure 9: Flowchart of the overview of the study (*AD: Alzheimer's disease; MATTR: moving average type:token ratio)*

seen on individual participant pair level in more than half of the pairs. These findings support the critique by Sabat (1994), who argues that the statistically significant mean differences between the groups of individuals with AD and healthy controls should not be looked at as representative of the individual participants, as the standard deviations often overlap. Sabat (1994) also argues that there is no "typical" AD sufferer and that focussing the average AD participant's performance can overlook the actual abilities of the individual. Therefore, the individual variabilities of the way language and speech changes must be acknowledged.

Next, I replicated the transcript index and language feature correlations from Berisha and colleagues' study (2015) who found that the use of unique words, fillers, and nonspecific words changes significantly over time in Reagan's, but not in Bush's speech. Looking at the individual pairs, I failed to replicate these results in the majority of the pairs. The feature showing the most promising results was the moving average of type:token ratio in 20-word window (MATTR_20), changing in 5 out of 9 AD participants in Corpus 2. However, this feature also changed similarly in 2 healthy participants. Moving average type:token ratio features were proposed by Covington and McFall (2010) to overcome the text-length-sensitivity of type:token ratio.

Based on Experiment 1 and 2, the findings of the current study failed to replicate the results of the previous study, suggesting that either a) change in speech in AD is heterogeneous;
or b) different approaches to capture the change are needed. Next, I explored three alternative approaches:

(1) using different language features;
(2) using age instead of transcript index to analyse longitudinal change;
(3) and grouping the features into aggregate scores instead of using single features.

Switching to different single language features known as informative from previous studies or using age instead of transcript index to track longitudinal change did not improve the universality of the patterns of language change compared to the methods used in Experiments 1 and 2, with less than half of the AD participants showing statistically significant results. The

best performing feature in these experiments was the number of pronouns divided by the sum of nouns and pronouns in Dataset 2 where 4 out of 9 AD and 2 out of 9 HC participants changed in the expected direction. The lack of generalisability in language patterns across participants is also addressed in Sabat (1994) who stresses that there is a great variability from one AD sufferer to another, resulting from the differences in personal history, pathological process and social relationships potentially affecting the individual markers of cognitive abilities.

Due to the inconsistency of single language features showing generalisable changes in the speech of the AD participants, I compiled the single features into aggregate scores based on the literature on known language dysfunctions in AD and their manifestations in speech. I focussed on lexical diversity, word finding difficulty, discourse building, syntactic complexity, and POS-related aggregates. I found that only word finding difficulty showed significant differences between the AD and the HC group in both datasets, and even then, the direction of the change was inconsistent. The correlations between the transcript index or age and the aggregate scores were more promising, with all aggregates changing in the expected and consistent direction in a larger number of the AD than the HC participants. Lexical diversity aggregate score showed the best results, with two thirds of the AD participants declining significantly over time.

One of the limitations of this study is the uncontrolled nature of the speech data. While the included samples consisted of free, naturalistic speech, offering a more realistic reflection of the participants' condition, speech content and the potential scriptedness of interviews or speeches could not be controlled for as the samples were pre-recorded. One way to control the speech content in the future is using tasks like picture description or conducting structured interviews. However, this approach would compromise the naturalistic nature of the speech data, and its resemblance to an everyday conversational situation. Structured speech tasks are also known to cause more stress for the participants with dementia. Similarly, as the data was collected from YouTube videos, it was not possible to include the participants' medical history and clinical characteristics, or control for potential confounders. The recordings were also conducted in different time intervals, contributing to the data sparsity issue illustrated on Figure 8. The time period covered in the current study was also notably longer than in the

original study, potentially decreasing the comparability of the results. However, the longer timespan can also be seen as an improvement of the dataset as it covers a longer time period and allows investigating long-term changes in speech. The recordings also varied in audio quality due to the differences in set-up and available technology at the time as some of the samples were recorded decades ago which is one of the reasons I have not included the acoustic features in the analysis and focussed only on the transcript-based linguistic features like the study I was replicating, despite the recent developments and promising classification accuracy in studies using acoustic speech signal analysis (Lopez-de-Ipina et al., 2018; Nasrolahzaden et al., 2022; Lopez-de-Ipina et al., 2020). While the acoustic features have been promising, they are often more difficult to understand and raise the question of interpretability in a clinical setting. Another limitation was the sample size - although I included more participants than the original study, the datasets were still relatively small, with 10 and 9 AD-HC participant pairs in the two corpora. The length of the transcripts used in the current and the original study also differed - while Berisha and colleagues (2015) used 1400-word transcripts, the transcript lengths in the current study varied, and were mostly shorter due to data availability. Although the data in both corpora were challenging and at times noisy, the findings across the two corpora demonstrate consistency, contributing to the reliability of the results.
In the future, a controlled longitudinal dataset with consistent time intervals, length and audio quality could be collected. Larger and higher-quality datasets could contribute to detecting more universal patterns of language change in the people who eventually develop AD, which would aid developing a non-invasive, cheap, and accessible language-based screening tool.
As the current chapter found little generalisability in the patterns of language change in the individuals with AD, different analysis methods for longitudinal change could also be developed. The current chapter suggests that the significant group differences in the language features are often not evident on individual participant pair level, resulting in the majority of AD participants not replicating the difference suggested by group comparisons. Future studies should take this into account when interpreting the results based on mean values, as the mean may not be representative of the individual cases. Similarly, focussing on participant level analysis to understand language change in AD in more detail, and finding more generalisable language

features or feature groupings could be considered. In the current study, some aggregate scores showed promising results, and future studies could focus on developing comprehensive measures for monitoring change in lexical diversity and word finding difficulty in spontaneous speech. Compiling the features into group scores could also be improved by more detailed analysis into the importance of the individual single features in the sum score calculation, for example, by placing more weight on the moving average type:token ratio in 20-word window (MATTR_20) which showed the most promising results in the current study when calculating the lexical diversity aggregate score. It would also be relevant to analyse how the longitudinal nature of the data affects the computation of aggregate scores and feature weights, for example, how long before the diagnosis does the importance of a single feature become apparent.

In sum, this chapter contributes to understanding the long-term change in language use in the individuals who later in life develop AD by using longitudinal spontaneous speech datasets where speech data spans over several decades, and investigates whether the longitudinal linguistic change previously identified in a single participant is generalisable to 10 and 9 participant pairs in the two corpora, or whether any other generalisable patterns can be detected. I highlight the issues of generalisability of language change by comparing the individual trajectories of the participants and show that language decline in AD is not homogeneous. I propose that different methods to detect more universal patterns should be developed, and demonstrate the potential of using aggregate scores, especially lexical diversity. I also show that, out of the single language features used, the most universal patterns of language decline are captured by moving average type:token ratio (MATTR) and pronoun-related features. As little generalisability was found, this study encourages acknowledging that the manifestations of cognitive decline in language can vary from one individual to another. As the classification tasks are already quite accurate, understanding the longitudinal change and the uniqueness of each individual's language change is a crucial research direction as it contributes to developing tools for detecting the earliest signs of cognitive decline, disease prediction and tracking, taking the individual differences into account.

While the change in language in AD seems to be heterogeneous across participants, the current study found that the most informative single features were MATTR and pronoun-related features. Aggregate scores captured change over time better than the single features, with lexical diversity scores showing the most promising results.

4.5. Summary

In this chapter, I tackled the issues of replicability and generalisability. First, I replicated a previous case study on a larger group of individuals, using the corpus that I created (described in Chapter 3) and an additional corpus. As I found little generalisability of the results reported in the case study, I proposed several alternative methods for data analysis. I concluded that while aggregate scores tend to produce more generalisable results compared to the single language features, there seems to be a lot of individual variability in the way humans speak changes over time, both in AD and non-AD aging. When interpreting these results, it must be remembered that these findings are based on two small and specific corpora (2x10 and 2x9 famous people), and therefore provide a starting into the research on generalisability of longitudinal language changes, rather than aim to offer any conclusive statements.

5. How much speech data is needed for language-based AD detection? A comparison of random length, 5-minute and 1-minute spontaneous speech samples

5.1. Introduction

In addition to the lack of longitudinal data (addressed in Chapter 3) and generalisability (addressed in Chapter 4), there is also a lack of standardisation in data collection and analysis methods in the studies looking at AD-related language changes. For example, the methods of speech data collection, such as optimal sample length, have not been standardised (Voleti et al., 2019; Sajjadi et al., 2012), limiting the comparability across studies and the transferability of research methods, as well as their translation to clinical practice. Previous studies have often used random sample length which can lead to bias in feature values due to their text-length-sensitive, especially as people with AD tend to speak less (Lopez-de-Ipina et al., 2018; Ash et al., 2006; Knibb et al., 2009; Graham et al., 2004). The analysis in Chapter 3 and Chapter 4 also showed that many language features extracted from the transcripts tend to be dependent on text length.

The aim of this chapter is to:

- explore how much speech data is needed for informative analysis of speech in AD;
- and how the length of the speech sample impacts the robustness of the features.

Having a better understanding of the optimal sample length as well as the relationship between the language feature value and the length of transcript it was extracted from would contribute to the standardisation of the methods, boosting the replicability of the studies, as well as lead to more reliable analysis and robust models.

Different strategies have previously been proposed to cope with text-length-sensitivity, for example, extracting language features as ratios (Beltrami et al., 2018; Bucks et al., 2000; Gosztolya et al., 2019; Guinn et al., 2014), however, ratio features can be unreliable as the proportion of specific language features does not change linearly with text length (Duran et al., 2004; Le et al., 2011). Another strategy is capping the samples at a maximum length (Duran et al., 2004), however, the minimum amount of speech data needed to spot AD-related changes

remains debated (Sajjadi et al., 2012). Previous studies have proposed that 150-word interviews can reflect language impairment in dementia realistically (Sajjadi et al., 2012), used 1-minute- (Chien et al., 2018), 4-minute- (Romero & Kurz, 1996), or 1400-word-samples (Berisha et al., 2015), capped the transcripts only when analysing text-length-sensitive features (Le et al., 2011), or removed the acoustic features that correlate with duration (Haider et al., 2019). All in all, there is no standard method for dealing with text-length-sensitivity, and very different sample lengths have been used in previous research.

Due to limited availability of longitudinal datasets, previous work on standardising the duration of the samples has not, to the best of my knowledge, considered the ability of different length speech samples to capture language change over time. While longitudinal changes in speech have received little attention compared to classification studies (Lopez-de-Ipina et al., 2018; Luz et al., 2021b) as the datasets are expensive and time-consuming to collect (Yancheva et al., 2015; Meltzer, 2020; Robin et al., 2021), previous literature suggests that changes in language use, such as vocabulary richness (Fox et al., 1998; Simon et al., 2018; Berisha et al., 2015), syntax (Le et al., 2011; Fang et al., 2017), and pausation patterns affected by word retrieval difficulties (Pistono et al., 2016; Meilan et al., 2012) begin before the clinical diagnosis and worsen over time.

This chapter aims to provide an insight into the effect of sample length in analysing longitudinal recordings of spontaneous speech in AD by comparing the original random length, 5- and 1-minute-long samples. The length of the samples is determined based on common sample or transcript lengths in previous studies or language tests. I hope to understand whether capping the audio improves the accuracy of the analysis, and whether an extra 4 minutes conveys necessary information. Using shorter samples would promote data collection and analysis by requiring less effort from vulnerable subjects and minimising computation time, but the samples must be long enough to be informative. This study aims to provide a starting point to exploring the trade-off between more data and practicality in the context of speech-based AD assessment, which future studies could investigate further using, for example, a larger variety of sample lengths.

This chapter is based on a journal article published in Digital Biomarkers (Petti et al., 2023b).

5.2. Materials and methods

5.2.1. Materials

The data used in this study is based on a subset of the Winterlight Labs corpus described in Chapter 4. This dataset consists of transcripts of public interviews with famous individuals, some of whom eventually develop AD. The original dataset consisted of 405 recordings from 9 AD - healthy control (HC) participant pairs of public figures. As the design of the current study required the length of the samples to be at least 5 minutes long, all shorter samples were excluded from the original dataset, resulting in a subset of 110 speech samples from 17 individuals. See Table 19 for details of the individual participants and the recordings included in the final dataset. As the clinical information of the public figures was not accessible, the time of diagnosis was based on Wikipedia and media entries.

Participant	Number of recordings	Age range over recordings	Sex	Education	Age at diagnosis
AD_1	12	43–76	M	12	75
HC_1	9	44–88	M	12	–
AD_2	7	46–74	M	17	80
HC_2	4	34–77	M	16	–
AD_3	7	50–77	M	14	79
HC_3	8	35–81	M	14	–
AD_4	3	65–75	M	18	80
HC_4	7	32–81	M	13	–
AD_5	3	68–71	M	14	83
HC_5	8	44–86	M	16	–
AD_6	8	64–72	M	16	78
HC_6	9	70–75	M	18	–
AD_7	2	47–59	F	17	58
HC_7	7	55–63	F	17	–
AD_8	1	63	F	12	63
HC_8	7	37–62	F	12	–
AD_9	8	75–81	M	16	83

Table 19: Individual participant and recording information in the final dataset *(AD: Alzheimer's disease; HC: healthy control; F: female; M: male)*

Three different length datasets were constructed from the same voice samples by cutting the transcripts and audio: original random length, 5-minute, and 1-minute dataset. All included speech samples in the original random length dataset were at least 5-minutes long after

diarisation, but their length was not capped, ranging from 5 to 21 minutes (mean=9.75), and 479 to 6665 words (mean=1893).

To create the 5-minute dataset, a cut-off point was applied at the 300-second mark of these samples. The number of words across the 110 samples in the 5-minute dataset ranged from 479 to 1339 (mean=908).

To avoid re-transcription and potential between-transcriber differences, the 1-minute cut-off point was established at the nearest utterance boundary of the 60-second mark, resulting in +/- 5-second variation in sample length. The transcript length in the 1-minute dataset ranged from 102 to 278 words (mean=179).

This process resulted in three different length datasets consisting of the same speech samples, so that the first minute was the same in all datasets, and the first 5 minutes was the same in the 5-minute and original random length dataset.

456 identical linguistic and acoustic features were automatically extracted from each dataset. The linguistic features were related to coherence (such as cosine similarity between utterances), syntactic structures (such as t-units - the shortest grammatical sentences into which text can be split), lexical features (such as age of acquisition (AoA) or level of arousal), and vocabulary richness (such as type:token ratio (TTR) – the number of different words divided by the number of total words; moving average type:token ratio (MATTR) (Covington & McFall, 2010) – TTR in different moving window sizes of consecutive words; Brunet index (Brunet, 1978) - based on text length and vocabulary size (Bucks et al., 2000); Honoré statistic (Honore, 1979) - based on hapax legomena and an assumption that growth in their use is constant to the logarithm of text size (Tweedie & Baayen, 1998)).

Acoustic features included speech tempo- and pause-related features, Mel-frequency cepstral coefficient (MFCC) measures and zero-crossing rate (ZCR) statistics. Acoustic features were extracted using Praat and Parselmouth software (Boersma & Weenink; Jadoul et al., 2018).

The data collection, transcription, the cutting of audio and feature extraction for this dataset was conducted by Winterlight Labs employees.

While this dataset cannot replace clinical studies, it has several advantages:

- collecting longitudinal data from pre-recorded interviews in the public domain provides an insight into language change in AD without expert labelling and extensive privacy concerns, unlike clinical data (Voleti et al., 2019), and allows to explore how speech duration affects the robustness of the extracted features;
- while cognitive tasks conducted in laboratory setting can have methodological advantages, spontaneous speech is considered to provide a more realistic reflection of cognitive abilities and promotes longitudinal data collection due to natural setting (de la Fuente Garcia et al., 2020);
- using multiple speech samples recorded over several decades from a number of participants allows to explore longitudinal change instead of the difference at a single time point, contributing to finding the earliest markers and monitoring disease progress;
- using over 450 language features contributes to detailed analysis of the changes in speech patterns in a wide range of language areas, and automated feature calculation improves the objectivity of the analysis (Voleti et al., 2019).

5.2.2. Procedure

I conducted 3 experiments to explore the text-length-sensitivity and the robustness of the language features, and to understand how much speech data is needed for comprehensive analysis of language changes in AD over time.

Experiment 1

The aim of this experiment was to explore the text-length-sensitivity of the language features in three different datasets by measuring the Spearman correlations between the language feature values and the number of words. This was done separately in each dataset (n=3), between each feature (n=456) and text length, resulting in 3x456 correlations using 110 data points. Bonferroni correction was applied to account for multiple comparisons, resulting in a significant p-value of p=0.0001. For informative analysis of speech in AD, I would expect the feature values not to be extensively affected by transcript length.

Experiment 2

The aim of this experiment was to investigate the comparability of the feature values across the different length datasets using Kruskal-Wallis test to identify the number of features that differed in the three datasets, and Dunn test for post-hoc analysis to identify the datasets where these differences occurred. Spearman correlation was used to analyse whether the values of the features that differed in the two datasets correlated with each other. I hypothesised that if the datasets of different lengths capture the same information, the feature values across the datasets should be comparable or strongly correlated.

Experiment 3

The aim of this experiment was to analyse how well the datasets of different length capture language change over time. To do this, I investigated the relationship between participant age, dataset length, and language features. I first explored the number of significant Spearman correlations between participant's age and the language feature values in each dataset, comparing the AD and HC groups. I expected an informative dataset to capture the largest number of significant correlations with age in the AD group as a representation of a more rapid decline in language due to progressing cognitive difficulties, and the correlations between age and language feature values in the HC group to be less significant. Second, I used a linear mixed effect model (LMER) to better understand whether the age impact on the language features differs depending on dataset length in the AD group.

5.3. Results

5.3.1. Experiment 1

I found 120 text-length-sensitive features in the original random length, 30 in the 5-minute, and 22 in the 1-minute dataset. 104 text-length-sensitive features in the original random length dataset were acoustic MFCC features, compared to 0 and 1 in the other two datasets. Figure 10 shows the details of the text-length-sensitivity of the features, with the MFCC features removed due to strong correlations with each other. See Table 20 for examples of feature categories and their text-length-sensitivity.

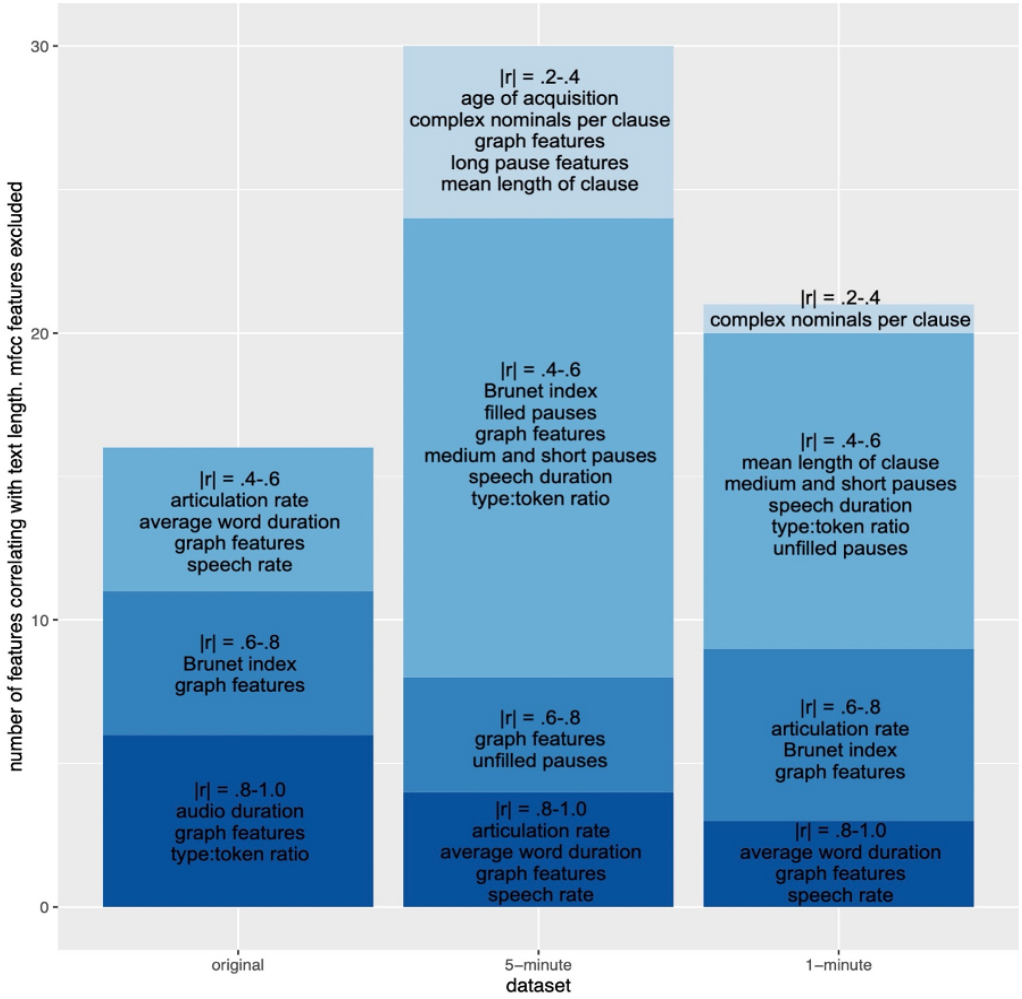


Figure 10: The number of features correlating with text length in each dataset, with Mel-frequency cepstral coefficient (MFCC) features excluded

Category	Not text-length-sensitive	Text-length-sensitive
Discourse	Cosine similarity between utterances	Graph features
Syntax	Phrase and sentence constructions	–
Lexical	Familiarity, imageability	Age of acquisition
Vocabulary richness	Honoré statistic, moving average type:token ratio	Type:token ratio, brunet index
Acoustic	Zero-crossing rate features	Mel-frequency cepstral coefficient features
Speech timing	–	Unfilled pauses, hesitation, pause count and duration, word duration, articulation rate

Table 20: Examples of language features according to their category and text-length-sensitivity

5.3.2. Experiment 2

226 language features showed significant differences across datasets in Kruskal-Wallis test (118 acoustic, 55 syntactic, 18 vocabulary richness and POS proportions, 15 coherence, 11 graph, and 9 fluency features). Table 21 shows a breakdown of post-hoc analysis exploring the difference between dataset pairs using Dunn test, as well as the number of features correlating significantly between the datasets using Spearman correlation, and the number of features that remained incomparable (significantly different and not correlated).

	1-min versus 5-min dataset	5-min versus original random length dataset	1-min versus original random length dataset
Number of significantly different features (Dunn test)	176	154	214
Out of the different features, the number of features that correlate across the two datasets (Spearman correlation)	159	71	81
Number of critical features that differ and do not correlate	17	83	133
Number of critical acoustic features	9	64	64
Number of critical linguistic features	8	19	69
Breakdown of critical linguistic features	6 syntactic 2 coherence	8 syntactic 8 coherence 3 graph	44 syntactic 14 coherence 7 graph 3 vocabulary richness 1 fluency

Table 21: Dataset comparability – number of features that show significant differences in the Kruskal Wallis tests, Dunn test, and Spearman correlation across different length datasets

5.3.3. Experiment 3

The 5-minute dataset captured a significant correlation with age in six language features in the AD group: AoA of words, AoA of nouns, average word duration, articulation rate, speech rate and total words uttered. The original random length dataset captured 3 features that correlated significantly with age in the AD group: non-words and incomprehensible words, AoA

of words, and AoA of nouns. The 1-minute dataset did not capture significant change in any features in the AD group. No feature correlated significantly with age in any dataset in the HC group.

LMER showed that the effect of age in the AD group significantly differed in 6 language features depending on whether the dataset was 1-minute-long or of random length: the number of complex nominals, the length of t-units, verb familiarity, function words, average shortest path in a graph, and sentiment arousal. The 5-minute dataset did not differ significantly.

As an illustration, Figures 11 and 12 show examples of average group and individual language feature values at different timepoints in different datasets. The feature choices are based on the results of Experiments 1-3; the one individual AD participant was chosen based on most available samples.

5.4. Discussion and conclusion

This study examined the role of audio duration in capturing the changes in language in AD. Understanding the role of sample length contributes to standardising the methods of language-based cognitive decline analysis, making them more transferrable and increasing the comparability in future studies, as well as the efficiency of clinical applications.

I compared 1-minute, 5-minute and original random length spontaneous speech samples recorded over several decades, focussing on text-length-sensitivity of the language features, the comparability of the information captured by different length datasets, and language change with time.

As expected, capping the audio eliminated the text-length-sensitivity of the acoustic features, supporting the standardisation of sample length in acoustic analysis. The impact on linguistic features was less clear, as a similar, relatively small number of features correlated with sample length in all three datasets (see Figure 10). While most of these correlations were straightforward, vocabulary richness has been discussed more in previous literature (Voleti et al., 2019; Tweedie & Baayen, 1998; Duran et al., 2004). Many studies suggest that as a function

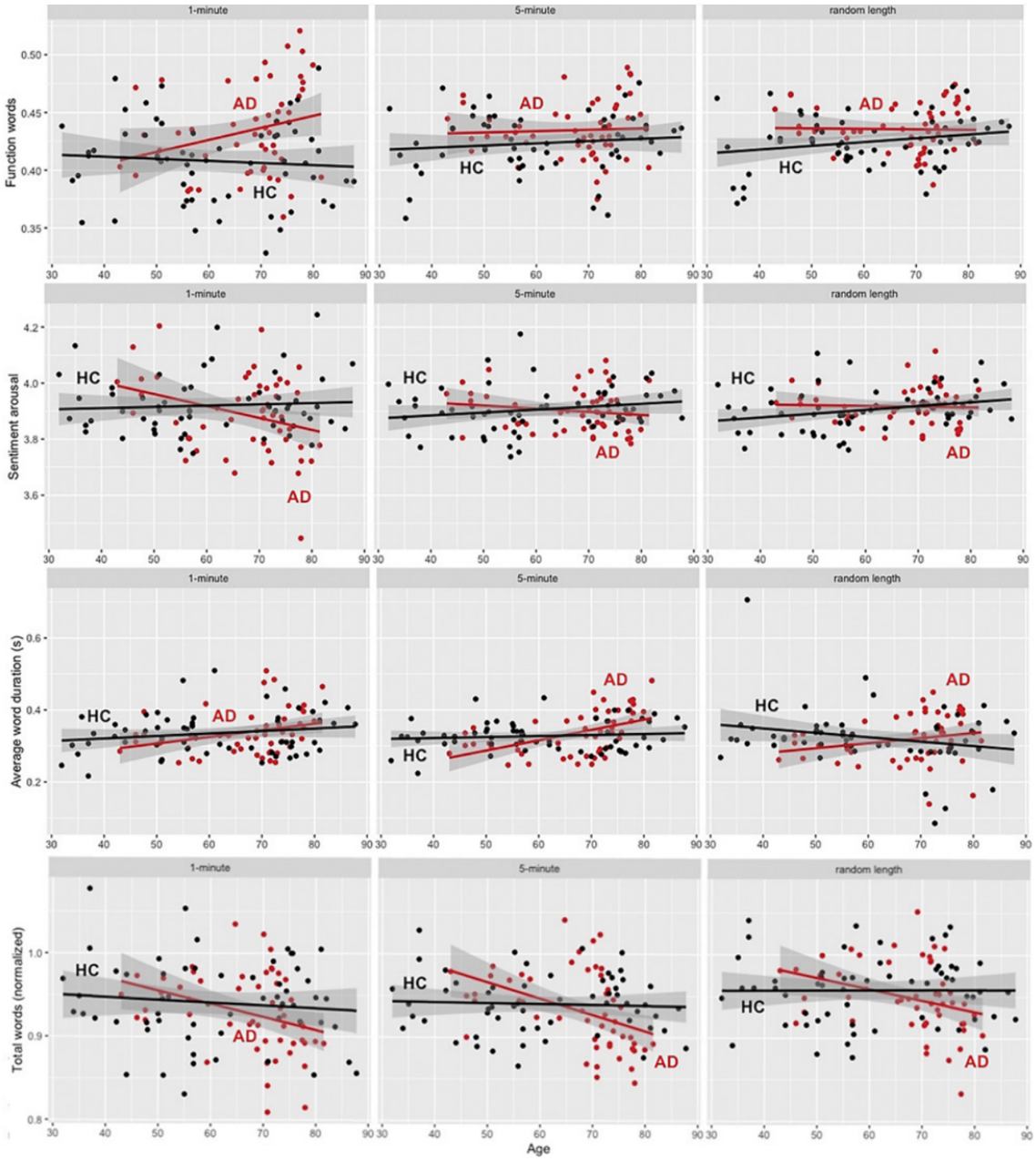


Figure 11: Average group values of function words, arousal, word duration, and the number of words at different time points across 3 datasets and AD and HC participant groups *(AD: Alzheimer's disease - red; HC: healthy control - black)*

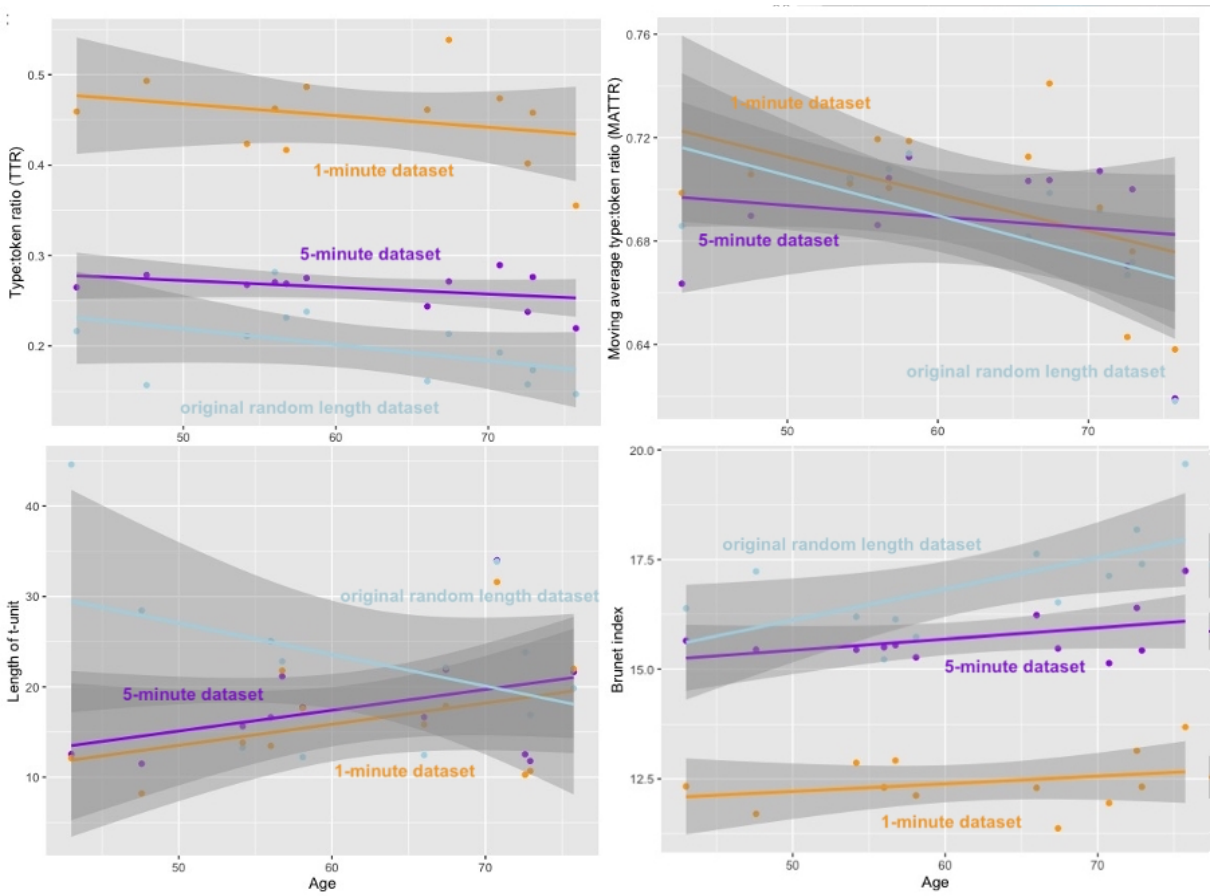


Figure 12: Individual participant's feature values of type:token ratio, moving average type:token ratio, average length of t-units, and Brunet index at different time points across 3 datasets *(1-minute dataset – orange, 5-minute dataset – purple, original random length dataset – light blue)*

of the total tokens, TTR is likely to be affected by sample length (Voleti et al., 2019; Bucks et al., 2000; Guinn et al., 2014; Duran et al., 2004), and MATTR has been proposed as an alternative, less text-length-dependent metric (Covington & McFall, 2010). The results of Experiment 1 (Figure 10) and the illustration on Figure 12 support these claims. The first row of Figure 12 compares the TTR and MATTR features values of the same speaker in different sample lengths over time, demonstrating more stable values across sample lengths in MATTR values.

While numerous sources suggest that Brunet index is independent of sample length (Calza et al., 2021; Voleti et al., 2019; Guinn et al., 2014; Hernandez-Dominguez et al., 2018), the results of Experiment 1 showed significant correlations with text length in all 3 datasets, in line with the illustration on Figure 12 where Brunet index values extracted from identical samples that only differ in length are notably different. Both Brunet index and TTR values suggest poorer vocabulary in longer datasets, explainable by the number of unique words not increasing linearly with the number of tokens (Voleti et al., 2019; Le et al., 2011; Covington & McFall, 2010). As vocabulary richness is often used in studies concerned with speech in AD, it is important to keep the potential effect of sample length in mind and apply appropriate preventative measures, such as using MATTR instead of TTR, controlling for (the correlation with) sample length (Tweedie & Baayen 1998; Duran et al., 2004), and reporting the length of the samples to increase comparability across studies.

In line with previous studies arguing that there are no major differences between the first and the second 150 words of an interview (Saffran et al., 1989), or the first 150 and 600 words (Sajjadi et al., 2012), I found that the information captured by 1-minute and 5-minute dataset was mostly comparable (see Table 21). A few acoustic and linguistic features suggested differences in the content captured: some complex syntactic structures may not have appeared in the shorter 1-minute samples (Saffran et al., 1989), and some coherence metrics suggest that 5-minute samples understandably convey more diverse content. As expected, the random length samples differed the most from the other two datasets.

Earlier literature suggests that changes in vocabulary richness and verbal memory (Le et al., 2011; Romero & Kurz, 1996; Fox et al., 1998; Simon et al., 2018), syntax complexity (Fang et al., 2017; Le et al., 2011), and fluency (Pistono et al., 2016; Meilan et al., 2012) manifest before AD diagnosis and continue declining as the condition worsens. In the current study, the 5-minute dataset suggested significant age correlation in six language features related to vocabulary complexity (AoA) and speech fluency (speech rate, word duration, total words) in the AD group. It has been proposed that familiarity-based memory is likely to be preserved in early stages of AD (Simon et al., 2018), potentially explaining the changes in AoA features. Similarly, word

retrieval difficulties have been linked to pause proportion in speech, affecting speech rate, word duration, and the number of words (Pistono et al., 2016) - the percentage of voiceless segments has been proposed to be one of the most informative features of the decline in language ability in AD with a potential to provide a low-cost solution to early AD detection (Meilan et al., 2012).

The impact of age on six language features related to vocabulary complexity and familiarity, and syntax differed significantly between the 1-minute and original random length dataset, suggests that dataset length plays an important role in capturing language change in AD. While the 1-minute dataset seems to capture changes in some language features in AD most clearly (see, for example, function words on Figure 11, clearly showing an increase in proportion in the first minute of speech as AD gets more severe, however this difference seems to flatten in longer samples) and could therefore have a potential to differentiate between the AD and HC participants from early on, the feature values tend to be more randomly distributed and further from the predicted values than the same feature values in the 5-minute samples, potentially indicating that the first minute is too short to provide reliable and stable results. The downside of the random length samples seems to be the appearance of the odd outliers, potentially due to some drastically longer or shorter samples. The age impact of the 5-minute dataset did not differ significantly from the other two datasets, potentially suggesting that 5 minutes is long enough to capture the content comparable to the longer random length samples while, potentially due to standardised length, also remaining comparable to the 1-minute samples.

Based on these findings, the 5-minute dataset seems to provide most stable feature values and capture the expected decline in language most consistently. However, when interpreting the results, it must be remembered that this study is based on a small sample and the findings are descriptive in nature and could only act as an indication suggesting directions for further research. Dataset size and heterogeneity have been identified as the main limitation of the studies exploring speech in AD (Luz et al., 2021b), contributing to potential overfitting. To avoid overfitting, it is also important to consider the number of necessary features extracted from the limited amount of data (see Berisha et al., 2021 for more detailed discussion).

Other limitations include the lack of clinical and confounder information, potential scriptedness of the recordings, uncontrolled content, intervals, and settings which limit the comparability of the samples. Further research using larger sample size, real participants, standardised recording conditions and intervals, and reliable clinical information is needed to make stronger conclusions. It must also be acknowledged that while the current study uses 5-minute or longer speech samples, it can be challenging to find or collect such data in naturally occurring environments. Additionally, free speech and interviews might place different cognitive demand on the participants, and different tasks and lengths of speech can capture different changes. Therefore, the research design decisions on the type and length of the data should be made based on the research question.

Future studies could investigate whether the amount of speech data needed to capture change in language reduces as the participant's condition worsens, for example, whether less data could be used for more severe cases due to the impairment being more easily detectable. This would help reduce testing-related stress of the participants with more severe dementia and encourage participation. Another direction could be exploring the trade-off between the number and the length of the samples and analyse whether it is best to collect more shorter samples or fewer longer samples.

To sum, the current study demonstrates that the length of the recordings plays an important role in analysing speech changes in AD. The findings suggest that capped audio files have advantages over the random length ones, and while the 1-minute and 5-minute dataset convey largely comparable information, the stability of the feature values and the ability to track language change over time in AD advocate for using 5-minute samples.

5.5. Summary

In this chapter, I investigated the optimal length of a speech sample that is needed to analyse AD-related language changes. The aim was to contribute to the standardisation of data collection methods.

I compared the 1-minute, 5-minute, and random length samples of the same recordings, focussing on the text-length sensitivity of the extracted features, the comparability of the information that the samples capture, and their capacity to provide information about longitudinal changes.

I found that capping samples at an even length produced more stable feature values, and while the information captured by 1- and 5-minute recordings was largely comparable, 5-minute samples were more stable and more informative of longitudinal changes.

6. Ethical considerations in the early detection of Alzheimer's disease using speech and AI

6.1. Introduction

Research on using AI in the early speech-based detection of AD has made significant progress in recent years due to the increased availability of speech and language data, and the development of technology, methods, and tools that allow for the efficient collection, storage, and processing of vast amounts of data. Numerous studies have shown that changes in language can act as a promising biomarker of cognitive decline, and that being able to detect those changes can contribute to early diagnosis and monitoring disease progress from an early stage. Language-based tools have a potential to provide a non-invasive, accessible, fast, and cheap tool for early AD detection. While there is no approved treatment for AD, early detection is desirable for several reasons:

- early interventions could slow disease progress;
- preventative lifestyle changes could be introduced;
- access to services that could help with adjusting to the condition and lessen the economic and emotional burden of the patients and families could be acquired;
- patients could be empowered to make decisions about the future as self-determining agents.

(Calza et al., 2015; Schicktanz et al., 2014; Brock, 1993; Porteri et al., 2017; Ursin et al., 2021)

As a benefit for the broader community, population-based cognitive screening studies help generate evidence-based information that would aid the healthcare system and policy makers (Calza et al., 2015). Screening for cognitive decline can also help identify individuals with non-neurological disorders that might be responsive to treatment, providing obvious benefits to the participant (Mattsson et al., 2010). As AD treatment and medication to slow cognitive decline becomes available, the need for better tools to identify AD-related changes early and at scale increases.

ML-based NLP has matured to a point where it can offer effective support for clinical practice (Velupillai et al., 2018; Wu et al., 2020; Hasan & Farri, 2019). NLP techniques are used to represent text computationally by converting written text into interpretable data (Wu et al., 2020; Harrison & Sidey-Gibbons, 2021), commonly including features related to lexicon (parts-of-speech tagging), syntax (dependency trees), and semantics (word embeddings, sentiment analysis, dictionary features). Current NLP techniques are based on ML, reducing the burden of manual feature engineering common in the traditional approaches, and increasing the efficiency of the analysis and the accuracy of the results (Wu et al., 2020). Common NLP tasks in the clinical domain include Text Classification, Named Entity Recognition (NER), and Relation Extraction, which can be applied on texts such as health records, social media posts, drug reviews, and speech transcripts (Wu et al., 2020).

While there are numerous benefits to using NLP in the clinical domain, and studies have shown that early signs of AD could be detected from speech using these techniques, several ethical considerations arise and must be addressed before this technology can be developed and applied in a real-world setting (Calza et al., 2015; Wu et al., 2020). However, the specific ethical considerations that arise when using NLP techniques for early AD detection have not yet received sufficient attention.

In this chapter, I aim to address this gap. I map out the most challenging issues and provide suggestions that could be incorporated into ethical guidelines and best practices for researchers and clinicians working in this area.

I group the ethical concerns into five broad categories. Part 1 of the chapter focusses on autonomy, discussing the issues related to informed consent in AD, depersonalisation, and the disclosure of research outcomes. Part 2 focusses on privacy and data protection, exploring the issues that arise when working with personal speech data. Part 3 focusses on welfare, identifying the harms related to distress, discrimination and stigmatisation, and research reliability. Part 4 focusses on transparency, discussing the interpretability of language features and diagnostic decisions and highlighting the importance of intelligibility to clinicians and developers. Part 5 focusses on fairness, asking who will benefit from such research and addressing issues such as bias and the inclusion of minority populations.

This chapter is based on an article published in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (Petti et al., 2023c).

6.2. Autonomy

The principle of autonomy has been central to bioethics for many decades (O'Neill, 2002). While the precise meaning of "autonomy" is contested, it generally refers to respect for a person's right to self-govern, that is, to make decisions based on their own judgement and values, free from external coercion or manipulation (Christman, 2003). In AD, the issues of autonomy can arise in day-to-day affairs, legal representation, consenting to treatment and participating in research (Gauthier et al., 2013). In this section, I will discuss:

- informed consent in AD;
- risk of depersonalisation in language-based testing;
- concerns with disclosing the research outcome.

6.2.1. Informed consent in AD

Informed consent is essential for any research involving human subjects. According to the Declaration of Helsinki, consent needs to be 'voluntary, competent, informed and comprehensive'. Concerns with consent can arise in AD where these capacities are compromised, potentially affecting the research participant's ability to protect themselves, and putting them in a vulnerable position (Rogers & Lange, 2013; Nickel, 2006).

The consent-related ethical considerations depend on the stage of the disease. Schicktanz and colleagues (2014) discuss the issues of consent in pre-symptomatic, early, and advanced stages. For the pre-symptomatic stage, they stress that the participants must be made aware of the difference between biomarker research and clinical trials to help them better evaluate the potential benefits they get from participating in the study. They suggest carefully considering the questions of disclosure, data storing and sharing to minimise the psychological distress, issues with personal relationships, discrimination, and stigmatisation.

For the early stages of the disease, Schicktanz and colleagues (2014) point out that the subjects' decision-making capacity might be compromised and understanding the risks and benefits of

the research as well as the impact on one's own life might be limited. This is largely due to the compromised ability to process complex verbal information and impaired semantic knowledge (Gauthier et al., 2013), which is why the consent form for the early stages of AD should use simple language, repetitions, and extra explanations when necessary, as well as visual aids to demonstrate the risks and benefits (Schicktanz et al., 2014). It is also recommended to consider the cultural differences when evaluating one's ability to consent to participating in research (Schicktanz et al., 2014).

In the advanced stages of the disease, the individuals' "ability to comprehend and encode information will be significantly impaired due to increasing deficits in semantic knowledge, verbal learning and recall, receptive aphasia and attentional deficits" (Gauthier et al., 2013). This stage requires higher safeguarding to protect the participants' best interests and integrity, and the ethical issues of proxy decision-making must be considered (Schicktanz et al., 2014). However, it is also important to acknowledge that AD diagnosis and incompetency are not synonyms (Marson, 2001). Many AD volunteers remain capable of consenting, and those who can consent should be recruited before those who cannot (Illes et al., 2007).

When preparing a study and a consent form, the severity of AD as well as the individual's competence to make decisions (Marson, 2001) must be considered, and the consent should be re-evaluated as the disease progresses (Schicktanz et al., 2014). For longitudinal studies, considering "gradual informed consent transfer" which allows transferring the consent-giving capacity from the participant to the accompanying others is recommended (Schicktanz et al., 2014). Considering the issues of consent in longitudinal studies is especially important in biomarker research, as the classification models differentiating between those who already have AD diagnosis and the healthy population is already quite accurate (Petti et al., 2020), and the current research focus is mainly on the earliest signs and prevention (Meltzer, 2020) which usually uses longitudinal data to analyse cognitive decline over time. It has been suggested that the population-based longitudinal studies looking at cognitive decline would contribute to the development of the most appropriate screening tools (Calza et al., 2015).

In any stage of the disease, the participants need to be made aware of the benefits and risks of the research prior to taking part in the study (Illes et al., 2007), understand what is being consented to, and how their data will be used.

Additional ethical questions regarding consent arise when using secondary speech and language data that is available in the public domain. Due to the lack of available longitudinal speech datasets in AD, many studies have used public speech and language samples from famous individuals, such as writings (Le et al., 2011), press conferences (Berisha et al., 2015), and public interviews (Petti et al., 2023a). While this data is accessible to anyone, and the individuals have been aware of their speech or writing being made public, they have not been aware that their language data will be used for research looking at the early signs of AD. This poses a risk of the subjects' recordings being used for purposes they did not consent to and may consider inappropriate research (Meyer, 2018). Many of these participants are deceased, and their families can be challenging to reach, making obtaining the consent to use their language data for researching cognitive decline challenging. The research on secondary data must be conducted carefully, with the ethical considerations and the best interests of the individuals in mind, and guidelines with best practices should be developed.

6.2.2. Language tests and depersonalisation

Analysing changes in speech in AD is often done using different cognitive and language tests. However, going through language testing can be stressful and highlight the subject's deficits, leading to depersonalisation and humiliation, and compromising the person's autonomy. Communication is often linked to personhood, and nonstandard communication can put an individual at risk of being treated as a nonperson, experiencing damaged self-esteem and discouragement (Sabat & Harre 1994; Kitwood, 1990). It is often assumed that the absence of communication means that the internal mental processes have stopped (Foley, 1992). While some of the earlier literature suggested that the actions of the individuals with AD are not guided by meaning, others argue that these individuals retain their sense of self and act from reasons relevant to account for their behaviour (Sabat & Harre, 1994). However, their conduct may be dependent on their own local rules or meanings that the interlocutors may struggle to

engage with (Sabat & Harre, 1994). Using insensitive psychometric tests highlight the communicative deficits and can create an image of a wandering, inappropriate, labile, and confused person, contributing to not seeing the disease as a separate entity from the person (Sabat & Harre, 1994; Lipowski, 1969). In addition to being disrespectful in its own right, this form of depersonalisation threatens autonomy by construing persons with AD as incapable of meaningful decision-making. This can reinforce paternalistic attitudes and behaviour towards persons with AD, whereby decisions are made for them, without their involvement. Furthermore, by undermining the person's confidence in their own decision-making abilities, it can make them more likely to accept such paternalistic treatment.

To avoid depersonalisation, researchers have suggested treating individuals with AD as semiotic subjects, presuming their actions to be guided by meaningful reasons and looking at any communication deficits as similar to unfamiliar conventions from another culture (Sabat & Harre, 1994), as well as focussing research on the aspects of communication that are preserved in addition to the ones that are lost (Lipowski, 1969).

6.2.3. Disclosure of research outcomes

Disclosure of the signs of AD has been an ongoing debate for decades. Some of the arguments against disclosure include diagnostic uncertainty, the absence of treatment, the inability of individuals in the later stages of AD to understand the diagnosis, and the potential psychological reaction, while the arguments for disclosure are largely based on patient autonomy (Ursin et al., 2021; Gauthier et al., 2013).

Gauthier and colleagues (2013) state that the most pressing issue now is disclosing the results of biomarker research in asymptomatic persons, and while researchers are not currently obligated to share the biomarker results due to the uncertainty of their clinical utility, the current consensus is increasingly leaning towards informing the participants of their diagnosis. Similarly, Kaduszkiewicz and colleagues (2008) report an evident change in attitudes towards disclosure among clinicals. The information sheet and consent form should clearly state the procedures of informing the participants of the research outcomes, and their right to know and not to know must be respected (Schicktanz et al., 2014; Porteri et al., 2017).

Introducing AI-aided AD detection has raised additional questions about model accuracy and predictive power which also affect the communication on research outcomes (Ursin et al., 2021). Consistently with the literature on using AI for detecting the risk of psychiatric conditions (McKernan & Clayton, 2018), it is important to inform the participant prior to participating in the research that the model can be wrong, acknowledge the accuracy of the models and clearly communicate related concerns. This information can affect participants' decision on whether or not they choose to know the research outcome. For example, the participants might make a different choice depending on whether the model is 60% or 95% accurate.

6.3. Privacy and data protection

Natural everyday speech can capture relevant information about cognitive decline in AD with the aid of ML-based NLP. NLP requires sufficient amount of data, and the collection of speech and language data has become increasingly feasible with the development and availability of smart devices, high-quality microphones, voice-based interaction systems, digital texts, and the growth in data storage capacities (Robin et al., 2021; Yamada et al., 2021; Cychosz et al., 2020; Le Franc et al., 2018; Mehl, 2017), contributing to the development of language-based tools. The practices of collecting language data can include remote recordings in everyday environments, speech recordings in clinical setting, or gathering secondary data from available sources, such as public writings, transcripts, or recordings. However, the development of technology for collecting and analysing language data to detect signs of AD poses significant concerns regarding privacy and data protection. While many of these concerns overlap with other research domains involving speech and language data, working with the speakers with cognitive decline raises additional risks related to both revealing private speech content, and identifying the individuals via speech when cognitive function is being assessed and sensitive health-related information involved. I will discuss the issues arising in working with personal language data and medical information, and the potential strategies to tackle these issues in the following sections.

6.3.1. Privacy concerns

Natural speech can carry personal and identifiable information. When collecting such data remotely in an everyday situation, the researcher has little control over whose speech is being recorded and what is being said (Kumar et al., 2015). Longform recordings are particularly risky, especially from the speakers with potential cognitive decline, as they introduce a greater possibility of the participant exposing information that they would not have willingly shared (Cychosz et al., 2020). The importance of privacy concerns can be illustrated by the finding that only 8% of the questioned participants would consider wearing a continuously recording microphone (Klasnja et al., 2009).

Privacy and data protection issues in collecting and sharing spontaneous speech data contribute to the lack of publicly available longitudinal datasets focussing on cognitive decline in AD. To overcome this, researchers have constructed alternative datasets using publicly available secondary speech and language data (such as public interviews, press conferences, books, comments) from individuals who will eventually develop AD. While this data is publicly available, constructing a coherent dataset using this information poses additional concerns, as it can highlight the aspects that the subject did not think about, and make this information permanent (Meyer, 2018). This is especially relevant when the research focuses on sensitive health information, such as cognitive decline or AD. For example, while the people writing public comments, publishing books, giving press conferences or TV interviews are aware of the recording taking place, and can reasonably expect these performances to be observed and made public, they may not be expecting their language data to be used for investigating and potentially making public early signs of AD and their language decline.

Additionally, storing and sharing voice data can pose a risk of identification, especially when the speaker is known to the listener. Therefore, sharing voice data with a new group of researchers increases the possibilities of the participants being recognised (Cychosz et al., 2020). The extent of harm caused by potential identification and link to sensitive health data is difficult to estimate and must be acknowledged.

6.3.2. Approaches to protect participant privacy

Due to the digital nature of the data, the issues of privacy are much harder to regulate than in the analogue world (Müller, 2020). There is a need to promote discussion between science and law across the different legal systems and diverse populations, and the scientific community must promote high standards of research, especially in countries with low levels of informed consent and data protection (Schicktanz et al., 2014).

Several strategies to tackle the issues of privacy when working with language data have been proposed. For example, to ensure that the researcher does not have access to the speech content that the participant does not wish to share, the participant could be given control over the recording device, allowing them to choose the time they are being recorded, pause and remove the recorder at any point, or delete the parts of the recording they do not want to be listened to (Cychosz et al., 2020). A potential limitation of this approach is that the participants may forget to turn the recording device back on (Cychosz et al., 2020).

Additionally, to protect the content of the recordings from being disclosed, automatic feature extraction and the reduction of speech intelligibility could be applied. Automatic feature extraction uses algorithms to extract and store the features of speech instead of raw audio, contributing to retaining participants' privacy (Wyatt et al., 2007; Lane et al., 2012). It could be applied on stored audio data after which the audio could be deleted, or in real time on the recording device, which would allow never storing the raw audio at all (Cychosz et al., 2020). A potential limitation of this approach is that the feature extraction would rely on the current knowledge about AD, and if the original recording is not stored, new emerging knowledge cannot be applied, and new features cannot be extracted. Automatic feature extraction tools are still quite limited, and future work should focus on manually annotating the data to train NLP and ML algorithms for automatic feature extraction (Cychosz et al., 2020).

Reducing speech intelligibility could protect the privacy of everyone in microphone reach (Wyatt et al., 2007). Some techniques to do so include:

- using Principal Component Analysis (PCA) of audio spectrogram as it allows to detect non-speech sounds (such as pauses, hums, coughs, breaths, laughter), while preventing intelligible speech reconstruction (Kumar et al., 2015; Larson et al., 2011);

- using linear prediction for replacing the vowels in speech (Chen et al., 2008);
- focussing on voiced and unvoiced segments, pitch, energy and speech rate instead of formant information (Wyatt et al., 2007);
- recording bone-constructed acoustic energy fluctuations from vocal apparatus using an accelerometer instead of raw speech (Garrard et al., 2017) (the authors claim that this methodology enables capturing repetitions in the speech of the individuals with AD in their home environment without invading their privacy).

It is also important to note that the MFCC features, which are often used in analysing speech in AD, are considered poor from a privacy perspective as they reveal private speech information such as content, identity, and prosody (Kumar et al., 2015; Larson et al., 2011). When applying the strategies of making the intelligible speech non-reconstructable, it must be kept in mind that audio modifications limit the available analysis, and that mobile devices have limited capacity for advanced audio processing (Cychosz et al., 2020).

Another way to reduce the amount of content accessible to the listener is subsampling, that is, using shorter subsamples of speech while retaining high-fidelity audio (Mehl, 2017; Mehl & Pennebaker, 2003). Shorter samples cause less risk to participant privacy and therefore do not require as strong ethics training by the staff accessing the data (Cychosz et al., 2020). However, there are also several limitations:

- shorter samples may still include sensitive information;
- less context may lead to misinterpretation;
- not all relevant information may be captured (Cychosz et al., 2020).

To overcome these issues, future studies could combine the algorithms that first detect the events of interest, and second record in short windows (Cychosz et al., 2020).

To protect the identity of the participants, anonymisation or pseudonymisation should be applied, even when using publicly available secondary data (Meyer, 2018; Müller, 2020). Anonymising the content of speech by eliminating the names or places mentioned reduces the risk of identification of the speaker. However, even if the data has been anonymised, many

argue that participants should still be given a choice to be excluded from the dataset (Mohammad, 2022).

While the transcripts can be anonymised, “the anonymity of the voice recordings cannot be completely guaranteed” (Cychosz et al., 2020). It is encouraged to use encryption for storing audio data as it allows hiding speech while retaining the information about prosody and conversation (Klasnja et al., 2009; Müller, 2020). Being able to apply audio encryption on mobile devices has been dependent on state of technology - while earlier studies state that the encryption techniques are not suitable for mobile phones due to limited power (Kumar et al., 2015, Klasnja et al., 2009), more recent studies have shown fast development in applying audio encryption on mobile devices (Farsana et al., 2023; Eltengy 2021).

Researchers should consider the level of data sharing and vetting, and the type of data being shared with the best interests of the participants in mind, depending on the obtained consent from the speakers, the risk of harm, and the ethical training of those accessing the recordings (Meyer, 2018; Cychosz et al., 2020). The different levels of data sharing include public, research purposes, and individual permissions, and the different types of data could include quantified metrics, transcripts, audio, or video recordings (either in long form or in snippets) (Meyer, 2018; Cychosz et al., 2020). The considerations of the level and type of data sharing must be addressed and made clear to the participant before obtaining the consent (Meyer, 2018), including the information about who will have access to the data, and examples on how exactly the data will be used (Low et al., 2020). It is generally advised not to promise to not share the data, but carefully consider how and with whom the data will be shared (Meyer, 2018). It must be kept in mind that “even when data is shared consensually, understanding what consent actually implies” can be challenging (Low et al., 2020). This applies especially strongly in conditions like AD. While it is generally agreed that the participants do have the right to control access to their information (Cychosz et al., 2020; Müller, 2020), several suggestions have been made to address the question of data sharing on consent forms. For example, Meyer (2018) proposes that the participants could be given a choice how much of their data will be shared but argue that they should not be given an option that would eliminate the chance of reproducing the analysis, while Low and colleagues (2020) propose that the participants could

have an option to deactivate their consent. It is generally not advised to promise to destroy the data, and the short- and long-term storage options and the use of data repositories should be considered (Meyer, 2018; Cychosz et al., 2020). However, Kumar and colleagues (2015) suggest not storing complete audio recorded in a real-world setting to protect the privacy of those involved. While complete real-world audio recordings could come close to capturing natural spontaneous speech which is often desired from a researcher perspective, they are also most likely to reveal personal speech content of both the participant, as well as of other speakers in microphone reach. To minimise this risk, applying privacy protecting storage and analysis approaches described above is advised.

6.3.3. Benefits of data sharing

While there are risks involved in data sharing that need to be considered, having more available data would contribute to reproducibility of the research as well as developing higher quality tools for AD detection. Here, the benefits and risks to the individual and the scientific community need to be considered. While sharing the data publicly is generally preferred from a research perspective, researchers need to keep their promises to the participants, possibly only make the data available upon request (Meyer, 2018), and ensure the participants retain the right to protect their identifiable information (Cychosz et al., 2020). At the same time, they should not assume that all participants oppose data sharing. Although many participants may prioritise the protection of their privacy, others may be motivated to participate if they know that data will be widely shared and contribute to more research (Meyer, 2018). From a researcher's perspective, data sharing can also have several advantages, such as feeling part of the community who are dealing with similar questions or contributing to the advances of the research and development of new tools and innovations by allowing others to reanalyse the data, enrich the analysis, and increase accuracy and efficiency (Cychosz et al., 2020). However, as data collection can be time consuming and expensive, researchers also often disagree on sharing the data or combining datasets (Meyer, 2018).

6.4. Welfare

In addition to the questions of autonomy and privacy, several concerns related to participant welfare arise when using ML-based NLP and speech technology for detecting the early signs of AD. In this section, I will discuss three possible sources of welfare harms: psychological distress, discrimination and stigmatisation, and incorrect diagnoses.

6.4.1. Distress

Participating in cognitive testing and learning the research outcome can cause distress not only in the research subjects, but also in other people involved in the process, such as families, clinicians, and researchers.

Cognitive testing

Going through language testing can be challenging to the individuals with AD due to limited cognitive abilities and difficulties in speech production (Chien et al., 2019; Lopez-de-Ipina et al., 2018). While the researcher might see testing or recording simply as gathering information of someone's cognitive state, the individuals with AD have reported that being put in a position that highlights their deficit is humiliating, and they often find the tests pointless (Mattsson et al., 2010; Sabat & Harre, 1994). Indeed, when listening to the recordings of verbal fluency tasks (for example, in Pitt Corpus (Becker et al., 1994)), frustration and confusion can be spotted when the participants struggle to remember a correct word. Such frustration and anger resulting from being asked to do something that one cannot do but could have done in a healthier stage can cause severe emotional reactions, which can contribute to processing-related difficulties and impact test performance (Sabat & Harre, 1994).

Lessening the burden of the participants is crucial to encourage them to undergo testing, and care must be taken to avoid causing distress. Potential approaches to reducing participants' stress include collecting language data through spontaneous speech samples recorded in a friendly and relaxed conversational environment instead of formal testing, giving the participants a right to choose when to start and stop the recording, and respecting the

participant's wish to stop the process at any point (Cychosz et al., 2020; Chien et al., 2019; Lopez-de-Ipina et al., 2018).

Learning the research outcome

It is also important to acknowledge that being informed about the signs of progressive neurological condition that does not have a cure is distressing and can lead to the development of negative self-image or even suicide (Schicktanz et al., 2014; Ursin et al., 2021; Mattsson et al., 2010; Gauthier et al., 2013). The reactions to receiving early diagnosis can be affected by the availability of treatment. If treatment is available, detecting the earliest signs of the disease could lead to more effective treatment outcome, and therefore the reaction to the diagnosis may not be as negative (Mattsson et al., 2010). However, even without available treatment, the subjects may value the information on early signs of AD, as this allows them to make informed decisions about their future and find the tools to cope with progressing cognitive decline (Porteri et al., 2017; Ursin et al., 2021; Mattsson et al., 2010). It is also important to recognise that testing negative for early signs of AD can be upsetting for participants who were expecting to find an explanation for the memory problems they were experiencing (Smith & Beattie, 2001). It must also be remembered that there is always a risk of false positives and false negatives (Mattsson et al., 2010; Illes et al., 2007), contributing to the concerns and potential distress related to disclosure of the early diagnosis (Ursin et al., 2021).

To reduce disclosure-related distress of the participants, it is important to develop highly accurate models to minimise the chance of false positives and false negatives, and to be clear about the process of delivering the research outcome as well as the procedures that follow and the support that is available in case signs of AD are detected. This information would help the participants make informed decisions about participating in the research or using the language-based tools in the first place and choosing whether or not they wish to know the outcome. Porteri and colleagues (2010) recommend conducting a personal ethical evaluation and including the subject in the process before deciding on disclosing the early diagnosis, and Schicktanz and colleagues (2014) and Mattsson and colleagues (2010) stress the importance of

considering the cultural context when delivering the research outcomes and evaluating the participants' ability to handle prognostic information.

Distress in others involved

In addition to the research subjects, research into the early markers of AD can also cause distress in their families, clinicians, and researchers. While families may be relieved to know the cause of the problems they have been witnessing (Smith & Beattie, 2001), this knowledge can also increase worrying about the relative, as well as their own health due to genetic risk factors which can cause tensions in family relations (Cammen et al., 2004; Schicktanz et al., 2014). Clinicians can feel conflicted due to the uncertainty of the results, especially in early biomarker research, and the uncomfortable process of disclosing the research outcome and deciding on how much information to disclose (Mattson et al., 2010; Smith & Beattie, 2001). Researchers working with potentially emotionally triggering speech data witnessing cognitive decline can also experience distress, and they should have the right to stop the work when needed (Cychosz et al., 2020). If the content is highly emotional, the annotators should be informed prior to the process and the amount of content per one annotator should ideally be limited (Mohammad, 2022).

6.4.2. Discrimination, stigmatisation, and the role of technology

One of the major risks involved in the research into early signs of AD from speech is the potential stigmatisation and discrimination, even in the pre-symptomatic stages of the disease (Calza et al., 2015; Ursin et al., 2021; Gauthier et al., 2013; Meyer, 2018; Cychosz et al., 2020). Stigmatisation can be experienced in personal and professional environments, largely due to the myths and misconceptions about AD (Alzheimer's Association, 2008) which can portray the people with AD as non-persons (Gauthier et al., 2013). While effort has been put into reducing anxiety about having AD and dementia, Foley (1992) argues that these efforts might have a negative impact and contribute to stigmatisation of the condition even more, making the lives of the person with dementia and their families less tolerable, and sabotage the potential to attract social resources. To reduce stigmatisation, Schicktanz and colleagues (2014) recommend

avoiding “framing dementia only as a threat to the social and healthcare system”. More studies into the actual experience of pre-symptomatic AD experience could encourage discussion, raise public awareness, and potentially reduce stigmatisation.

Discrimination of those with early or pre-symptomatic AD can manifest in issues with legal status, job opportunities, access to health insurance, driving, or access to healthcare such as organ transplantation (Schicktanz et al., 2014; Ursin et al., 2021; Illes et al., 2007). Developing technologies that predict health conditions can contribute to these issues (Cychosz et al., 2020). For example, Low and colleagues (2020) discuss the issues in detecting early signs of psychiatric disorders and argue that even when the models have not been clinically validated, researchers developing these technologies need to be aware of the risk of the insurance companies and employers turning down the applicants if there is a possibility that a psychiatric disorder is present or will develop. The same is applicable to neurodegenerative diseases such as AD. Similarly, Illes and colleagues (2007) discuss the ethical implications of detecting the early markers of AD from imaging, and state that this research introduces a possibility for the health insurance companies to ask the individuals to take a test or provide the results of one prior to agreeing on insurance. This is also a risk in speech-based early biomarker detection, where the test is even easier to conduct than in imaging, especially for technologies capable of analysing everyday speech. On the other hand, when developed responsibly and used in a clinical setting, this technology could have a huge positive impact. There is an imminent need to develop regulations to avoid commercialisation of this technology and ensure ethical behaviour of third parties, such as insurance companies and employers.

The risk of unethical and dual use is an important aspect of developing new technologies, making it extremely important to consider the ethical implications of the research (Low et al., 2020). Dual use was originally used to refer to technologies that had both civilian and military uses. However, it has now acquired a broader meaning, referring to technology that can be used in ways that were not intended by the researchers that first developed it, especially in cases where a technology that was developed for a benefit purpose can also be used unethically.

While a proportion of the research is focussed on developing tablet-based tools that could be used for initial at-home AD screening, several questions on how and by whom the tools would be used arise. Having an app that anyone with a mobile phone could use to detect signs of AD from speech would introduce the opportunity to test anyone within the microphone range. This would mean that anyone could access the private clinical information and gain insight into other people's private mental lives (Low et al., 2020), which could contribute to stigmatisation and discrimination, and potentially pose a risk of presenting the recorded person with information about their health that they did not wish to know. To avoid this, the developed applications should not be in free use, and only used by clinicians on the participants who have initiated the visit or have concerns about their health, and already consent to the process (Ursin et al., 2021).

In addition to legislation, researchers must apply the highest professional standards to protect the privacy and confidentiality of the participants and consider the ethical implications and potential consequences of developing new technology (Schicktanz et al., 2014; Veliz, 2019; Karlawish, 2011).

6.4.3. Reliability of research outcomes

Inadequate data and model design can lead to unreliable outcomes and cause harm to the participants. In this section, I will discuss the adequacy of the information captured by language tests, and the reliability of the model outcome and reported results in language-based AD detection.

Adequacy of language data

While language and psychological tests can be useful in identifying some AD-related changes, these tests often show low sensitivity and struggle with capturing the challenges that the AD sufferers face in everyday conversations (Calza et al., 2021; Boye et al., 2014). Additionally, the repetitiveness of the tasks can result in habituation, creating an inaccurate image of cognitive abilities and compromising screening accuracy (Ujiro et al., 2018). Tests that encourage reporting mean scores can contribute to the understanding that every participant in the

condition group behaves as an average. Psychological tests can also have integrated biases, such as the lack of validation for different education levels, and the difference in quality and quantity of published research (Calza et al., 2015). Furthermore, cultural background can affect cognitive processes and language use, and therefore introduce the risk of misinterpreting communicative performance. For example, the likelihood of confusing people in the same relationship-category (Fiske, 1993), and the ability to recall objects depending on the background (Masuda & Nisbett, 2001) can vary depending on one's cultural background. As both confusion and object recall are often used to detect signs of AD, it is important to keep the potential effect of cultural background in mind when recruiting participants and analysing their language use. Good examples of adapting the tests and analysis methods to a certain culture can be found in the research on Thai. For example, Munthuli and colleagues (2021) use the Thammasat-NECTEC-Chula's Thai Language and Cognition Assessment (TLCA) where many parts of the tests are made more appropriate for Thai culture, and report promising results in classifying between healthy Thais, and those with MCI and AD. In addition, picture description tasks that are relatable to Thai culture were used by Sangchocanonta and colleagues (2021) who also used a POS tagger that aligns with the recommendations for Thai word types to analyse the data, and report good performance in classifying between healthy controls, MCI and AD. Metarugcheep and colleagues (2022) adapt verbal fluency tests and the extracted features so that they are suitable for analysing cognitive decline based on Thai language.

To overcome the limitations of traditional language tests, different methods, such as collecting spontaneous speech data instead of formal testing, have been proposed. Collecting spontaneous speech data causes less stress for the participants (Chien et al., 2019; Lopez-de-Ipina et al., 2018), allows for a more naturalistic setting and a more realistic reflection of the impairment than structure tasks (Sabat & Harre, 1994; Gosztolya et al., 2016), can easily be conducted in everyday situations using mobile devices requiring minimal instructions and equipment (Robin et al., 2021), and allows assessing patients who may not tolerate formal testing (Blanken et al., 1987). In addition, the developments in NLP and AI contribute to automated analysis of speech, making data analysis more feasible and scalable (Robin et al., 2021; Yamada et al., 2021; Chien et al., 2019; Luz et al., 2021a).

While spontaneous speech samples have many advantages, some limitations relating to the adequacy also need to be considered. For example, it is difficult to control whether the speech is semantically correct - the speech can be coherent, but potentially describe events that are untrue, or include confusing the names of family members which is challenging for the outside investigator to detect. Overcoming this issue would require someone who knows the subject to go over the transcripts and evaluate the truthfulness of the statements. However, this raises additional ethical concerns, such as privacy.

Similarly, it is important to note that while the studies looking at spontaneous speech often include some analysis of emotion, AD affects people's emotional lives. Current automatic emotion recognition systems cannot handle these variabilities and should therefore not be used, or the limitations of emotion analysis should be clearly communicated (Mohammad, 2022). Additional potential biases related to language and speech data include the multivariate nature of the data (Cychosz et al., 2020) and the variations in audio quality (Balagopalan et al., 2021).

Despite these limitations, spontaneous speech tasks have performed well in the past: 95% classification accuracy in discriminating between the healthy individuals and those with AD has been achieved, outperforming the more structured tasks (Lopez-de-Ipina et al., 2018). The authors argue that the main reasons behind the strong performance of the model trained on spontaneous speech are the relaxed atmosphere and more open language, which allows for the manifestation of subtle cognitive changes (Lopez-de-Ipina et al., 2018).

Adequacy of the models

The adequacy of a given model has different sides to it. First, there is a question of how well the model performs in general, and whether the results can be trusted. Second, there is a question of whether the model performance varies between groups within the population (e.g., between majority and minority groups). In this section, I will discuss issues related to general model performance, such as sample size, overfitting, and the use of suitable metrics. I will discuss issues of bias and disproportionate harm to disadvantaged groups in the next section on Fairness (6.6).

While speech-based AD classification models have achieved high test-accuracy, the reliability of these models in a real-world setting must be considered. Unreliable models can lead to biased outcomes and contribute to making decisions that are harmful to the research subject based on false positives and false negatives. False negative diagnosis may lead to exclusion from clinical trials or treatment as well as false reassurance, while false positive diagnosis can result in over-diagnosis, over-treatment, and inappropriate inclusion in clinical trials (Ursin et al., 2021).

There can be many possible issues contributing to the potential (un)reliability of these models. One common problem is sample size and overfitting. Although large sample size is crucial in ML, the models achieving high classification accuracy in discriminating between the people with and without AD based on language data are often trained on a small, specialised sample due to the lack of available data (Petti et al., 2020; Voleti et al., 2019; de la Fuente Garcia et al., 2020). Based on two systematic literature reviews in this area, including a meta-analysis of dozens of studies, Berisha and colleagues (2021) demonstrates that models trained on smaller datasets consistently achieve higher test accuracy (Figure 4) but tend to fail in the wild. This lack of generalisability can lead to catastrophic outcomes in a clinical setting (Berisha et al., 2021).

The lack of large available datasets contributes to the issue of overfitting. Having more available speech data could lead to better health outcomes, benefit the individuals and communities (Cychosz et al., 2020), and help avoid re-using language data that has previously been collected for a different purpose (Voleti et al., 2019) which may result in the ML models exaggerating the historical bias (Müller, 2020). To increase the amount of available language data and push the state-of-the-art forward, collaborations between researchers from speech neuroscience, neuropsychology and language technology are encouraged (Voleti et al., 2019). As alternative approaches, data synthesis and adjusting patient consent models with regards to data sharing have been proposed (Velupillai et al., 2018).

Publication bias also contributes to overfitting. Due to positive and clear results being more likely to be published (Munafo et al., 2017), the studies that achieve higher accuracies are preferred, potentially encouraging the researchers to overlook the issues of sample size. Moreover, researchers themselves often tend to only report studies that “work” (Simonsohn et al., 2014) and assume that the findings based on a small proportion and diversity of the

population are generalisable to a wider group (Henrich et al., 2010). Cognitive bias (the tendency to interpret information as something they already believe to be true (Müller, 2020)) must also be considered here. Some approaches to reduce cognitive bias include blinding the researcher by masking the experimental conditions and key parts so that their relation to the hypothesis is not interpretable and conducting integrative training (Munafo et al., 2017). However, Müller (2020) also raises a question of whether AI systems could or should have the cognitive biases that are characteristic to humans.

Several methods, such as using cross-validation and held-out sets have been recommended to avoid overfitting. However only 10% of the studies reviewed in de la Fuente Garcia and colleagues' (2020) paper reported holdout sets, and only one study validated their model on an entirely different dataset. The evaluation metrics must be carefully considered, especially when using small, unbalanced datasets. While accuracy is not a robust metric and should not be used for imbalanced datasets, 35% of the studies only report accuracy in de la Fuente Garcia and colleagues (2020) review. Instead of accuracy, metrics such as F1 (suggested by de la Fuente Garcia et al., 2020), UAR (Unforeseen Attack Robustness), and ROC (receiver operating characteristic) (suggested by Gebru et al., 2021) are recommended. When introducing new evaluation metrics for using longitudinal speech data for diagnostic purposes, Velupillai and colleagues (2018) recommend considering time sensitivity, personalisation of the models, and model interpretability.

The lack of standardisation of the data, model characteristics and evaluation metrics result in limited comparability of the models across studies, hindering the contextualisation of results (Petti et al., 2020; Balagopalan et al., 2021; Voleti et al., 2019; de la Fuente Garcia et al., 2020). To illustrate this issue, only 61% of the studies contextualised their results by directly and quantitatively comparing the findings to previous work or provided a baseline against which the results can be compared in de la Fuente Garcia and colleagues (2020). Standardising the data collection and analysis for cognitive assessment would contribute to more comparable data being available for different types of tasks, increasing the comparability of the findings of different studies (Voleti et al., 2019). This would increase the reliability of research outcomes and reduce the risk of making harmful decisions based on false positives and false negatives.

6.5. Transparency

In AD detection from speech using AI and NLP, ML algorithms are used to draw inferences about the subject's health. It is essential that clinical decision-making is transparent and interpretable (Ursin et al., 2021; Voleti et al., 2019; Tu et al., 2017). While classification models have achieved high accuracy in differentiating between people with and without AD, these models can be difficult to interpret as it can remain unclear which aspects of language contribute to the classification due to the unintelligibility of the language features and the multivariate nature of the data (Robin et al., 2021). There are two sides to interpretability: interpretability for the developers of NLP and AI, and interpretability for users, such as clinicians. The former helps us develop better AI as we can understand the predictions and address errors; the latter is important for users: doctors who are using AI to support diagnosis need to understand why the model made a prediction before deciding whether to trust it, as well as to explain treatment recommendations to patients for the purpose of informed consent and shared decision-making (Ursin et al., 2021; Tjoa & Guan, 2020; Keeling & Nyrup, 2022).
The topic of explainable AI is an active area of research (Tjoa & Guan, 2020). Several techniques have been developed to boost the interpretability of AI systems, including verbal, visual, and mathematical approaches (Tjoa & Guan, 2020). Researchers working on explainability must consider who the explanations are meant for, how easy they are to comprehend, and to what extent do the people trust the explanations (Mohammad, 2022; Keeling & Nyrup, 2022).
In detecting signs of AD from speech, the issues with interpretability can arise on two levels: understanding the extracting language features, and the processes underlying automatic decision-making. These two issues are connected - understanding the extracted language features would contribute to understanding why a certain decision about a participant's health has been made by the AI system.
As more speech data becomes available and more data-driven AI systems are used in speech analysis for clinical purposes, it is expected that the domain-expert features would be replaced by the features that perform better when applied to a specific task or condition diagnosis, compromising the interpretability of the features (Voleti et al., 2019). Performance gains in speech and NLP applications have been achieved using deep neural networks (DNNs), but these

gains come at an expense of interpretability. Therefore, it is often the case that the extracted language features remain difficult to interpret to both developers and healthcare professionals. Interpretability is especially relevant in analysing acoustic speech features as these are high-dimensional and require expert knowledge to be intelligible, such as the MFCC features (Tu et al., 2017). These features often lack clinical interpretability, making clinicians hesitant to use them, even though they contain a lot of information about the speaker and the disease (Tu et al., 2017; Jiao et al., 2017). Jiao and colleagues (2017) argue that rather than making automatic decisions based on acoustic signals, the features and outcomes should be interpretable to aid clinicians' decisions, and propose a visualisation tool of spider plots that would allow assessing the difference between the phonological signals of the patients and the expected pronunciations of the healthy population.

The growing need for health literacy in general, and the technical understanding among clinicians using AI has been addressed by many (Ursin et al., 2021; Wu et al., 2020; Harrison & Sidey-Gibbons, 2021; McKernan & Clayton, 2018; Tjoa & Guan, 2020; Le Glaz et al., 2021). It is, however, stressed, that ML and NLP should only act as a tool for the clinicians, and not disempower or replace them, making it especially important to understand why a certain decision has been made, and whether or not to trust it (Wu et al., 2020; Le Glaz et al., 2021). For example, clinical knowledge is needed to understand when the model outcome does not fit the clinical picture, and technical knowledge about why an algorithm might have made the decisions can aid clinical decisions about further investigations (Mattsson et al., 2010). When clinicians cannot explain the predictions of a black-box algorithm due to its opacity, it also significantly affects obtaining informed consent (Ursin et al., 2021). The development of clinical NLP systems needs to consider the input from the clinicians and be developed with clinical experts in mind (Velupillai et al., 2018).

Regulations regarding transparency of the AI algorithms vary from country to country. For example, EU regulation allows one to ask for an explanation of algorithm-based decisions (Wachter et al., 2017), while in the US, there is no such right (Garcia, 2016). However, EU regulation also raises questions about the extent to which it can be enforced (Wachter et al., 2017), and the double standards in asking the machines to be highly explainable while human

decisions themselves may be hard to explain (Müller, 2020). Algorithmic transparency is challenging to achieve as simply having access to the code would not be informative or intelligible (Garcia, 2016), contributing to the decisions often remaining opaque to both the affected person and the expert (Müller, 2020). Transparency and bias can go hand in hand, as bias can be exacerbated by opacity, and the response needs to tackle both together (Müller, 2020). Issues with transparency underlie the issues of trust. It is often thought that scientists, medical professionals, biotechnology companies pursue their own, rather than the patients' or the public interests (O'Neill, 2002). If the decisions and the contributing factors are not transparent, and the methods and data are inaccessible for replication, the widespread distrust and the lack of confidence in science are difficult to change (Meyer, 2018). It must be kept in mind that science is a social enterprise, working towards public good, and the credibility and trust comes from the openness and transparency of the process (Munafo et al., 2017).

6.6. Fairness

Many studies suggest that developing language-based tools to detect the early signs of AD would help improve the quality of life of the people with dementia and promote equality due to the increased accessibility and affordability of the screening tools. It is important to ask whose quality of life will be improved and how, and consider the questions related to recruitment and treatment availability, the advanced age of the subjects, and the bias contributing to disproportionate distribution of risks and benefits in disadvantaged groups. These considerations are especially relevant as AD disproportionately affects minority populations and women (Babulal et al., 2019; Lopez et al., 2019), and while the population of minorities continues to rise, the gaps in the clinical knowledge of AD in these populations remain (Mayeda et al., 2016).

As there is currently no approved treatment for AD, it must be borne in mind that people who participate in the current research are themselves unlikely to directly benefit from the research outcome, as both the screening tool and treatment development take time. Having no immediate benefit to participating in an experiment might make the subjects hesitant to do so. For example, a participant in Sabat and Harre (1994) describes their experience of undergoing

testing as humiliating and not helpful for them in any way - however, they also claim that if the research did help them, they would be happy to participate. Although some may be happy to participate in research for purely altruistic reasons, alternative methods for compensating research participants should also be considered. In either case, what kinds of benefits—if any—people may expect from participating in research should be clearly explained.

When developing models that would contribute to the early diagnosis in the future, it is important to consider who will eventually benefit from such research, and keep in mind that the design of current studies has a great impact on the future outcome. The models will perform adequately only when used on the same type of population as they were trained on. Therefore, applying "responsible design" (Müller, 2020) and deciding who to recruit in current research and whose data to use for building the models are crucial decisions to develop fair and inclusive screening tools. Having a balanced dataset is an essential part of research design because imbalanced datasets contribute to the risk of biased outcome as the model will reproduce the bias it is introduced to in the data (Cychosz et al., 2020; Gebru et al., 2021). Biases like gender, race, nationality, and age affect both speech (Eichhorn et al., 2018; Hagiwara, 1997) and NLP performance. These parameters need to be considered in the training data to build robust models (Maley et al., 2020). If these biases remain unaddressed, the bias in data will lead to the AI systems performing better for majority groups, known as allocational harm, as the picture of a "typical" behaviour is based on the characteristics of a privileged group, leading to perceiving nonstandard groups as atypical (Cychosz et al., 2020). This leads to disproportionate risk of harm to the disadvantaged groups and poses a risk of false outcomes and stigmatisation.

While balanced datasets are essential, the literature reviews on ML-based AD detection from speech and language show a small proportion of studies using balanced datasets in terms of age, gender, race, and education level. For example, de la Fuente Garcia and colleagues (2020) report that only 25% of the studies looking at language-based AD detection used datasets with balanced age, gender, and education, and few of them present statistical values showing the balance between the groups. In the review described in the Background chapter of the current thesis (section 2.2, Petti et al., 2020), only 64% of the studies reported the age and gender of

the participants with 40% of the datasets being gender-balanced, and 45% of the studies reported the education level of the participants, with the control group participants having spent on average more years in education in all but one study. As age, gender and education level can impact both speech and AD manifestations, as well as the performance of the NLP models, it is important to consider these factors when constructing a dataset.

In addition, as many studies in this area use automatic speech recognition (ASR) tools, it must be acknowledged that age affects the performance of these models. ASR tools struggle with elderly voices due to the changes in fundamental frequency, increased jitter, shimmer and breathiness, and slower speaking rate, potentially leading to higher word error rates (WER) (Vipperla et al., 2008). The mismatch between the acoustic model and the user is the main contributor to the limited ASR performance (Baba et al., 2004), adding to the reasons for using balanced and appropriate datasets.

Another pressing issue is the availability of language resources in different languages, and the European-centric focus of the current studies. While active research is also being conducted on other languages, the literature reviews summarising the recent studies on detecting the signs of AD in language still show that a large proportion of the models are trained on English speaking populations (41% of the studies in de la Fuente Garcia and colleagues (2020) and 30% of the studies in Petti and colleagues (2020)). This means that these models are applicable only to a similar population and should not be applied to other languages, leading to largely benefiting the rich Western countries. While it could be argued that studying language change in English first would help us understand how AD develops in the first place, and then transfer the technology to other languages, it poses questions of equality and fairness. It must also be kept in mind that the language models often fail to account for dialectal variation (Jurgens et al., 2017). Overlooking dialect impact can also become a problem in speech-based AD detection. For example, Fukuda and colleagues (2022) studied speech-based dementia detection in elderly Japanese speakers, and found a strong dialect influence on the results, demonstrating that by removing dialect-related features (formants, MFCC features), the accuracy of dementia detection improved by almost 15%. Acknowledging dialect impact is especially relevant in studies concerned with the older generation as they tend to use stronger regional dialects

(Fukuda et al., 2022). To avoid disadvantaging the speakers of non-European languages, or those with regional dialects, it is important to develop NLP tools for different languages and dialects, and include a range of speakers of different languages and dialects in the training data. Collecting data concerned with AD detection in various languages and making it available would promote tool development and advance research in different languages (de la Fuente Garcia et al., 2020). As an example, to address the lack of multilingual datasets, Lopez-de-Ipina and colleagues (2018) constructed a database for detecting the signs of AD including 8 different languages, called AZTIAHO.

Sampling bias is not unique to studies in AD detection from speech - extreme sampling bias has been recognised across behavioural sciences that have to a large extent included the WEIRD (Western, Educated, Industrialised, Rich, Democratic) population, leading to calls for more inclusive geographic, linguistic, and cultural settings (Cychosz et al., 2020). While there could be many potential reasons to why minority groups have been underrepresented in research (such as distrust in health research or lack of flexibility in terms of childcare or employment), Rogers and Lange (2013) argue that in contrast to the wide-spread belief that minority groups are less willing to participate in health research, they are simply often not given a chance to participate, making the lack of opportunity the main barrier to participation. Some of the reasons why the researchers might exclude minority populations include using occupation-based or geographic recruitment methods, lack of resources, access to minority populations, expertise, experience, or interest, or the assumption that an individual would not be interested in participating in health research, which in itself can be seen as stereotyping (Rogers & Lange, 2013). Underrepresenting minorities can lead to lack of information about the group's health, poor health outcomes, lack of applicability of the research findings and the models, and decreased access to interventions and treatment. This contributes to the injustice as the benefits of the research are only available to a certain group of people. More inclusive sampling would allow the findings to be generalisable to a broader population and "reveal the limits of applicability of research findings to different communities" (Cychosz et al., 2020).

Several methods have been proposed to ensure data balance and identify potential biases. The imbalanced dataset could be made more balanced by sub-sampling (Henrich et al., 2010).

However, in some cases, the pre-processing of previously balanced datasets by removing samples can also lead to imbalanced data (de la Fuente Garcia et al., 2020). The identification of bias could be made more explicit by including a standard description of a dataset with the studies (Henrich et al., 2010). There are also several tools available for risk of bias assessment, such as QUADAS-2 (Quality Assessment of Diagnosis Studies checklist 2) (Whiting et al., 2011), and PROBAST (Prediction model Risk Of Bias ASsessment Tool) (Wolff et al., 2019).

6.7. Conclusion

Although using AI in detecting the early signs of AD in speech is a promising area of research, several ethical concerns arise in the process that need to be discussed, considered, and addressed. I have summarised the issues related to autonomy, privacy and data protection, welfare, transparency, and fairness. I emphasise the need to develop ethical guidelines and best practices, and promote discussion between different fields, research areas, and communities. I stress the importance of respecting the choices of the research participants, making decisions on data protection with the best interests of the participants in mind, applying research methods that cause minimal amount of discomfort to the participants, considering the ethical aspects and the potential impact and consequences when developing new technologies, focussing on the explainability of the models, including diverse populations in the research, and ensuring fair distribution of research benefits. I provide a more detailed list of suggestions in Table 22.

6.8. Summary

In this chapter, I focussed on the ethical questions that arise in AI-based AD detection using language and speech data. I discussed the questions related to autonomy (6.2), privacy (6.3), welfare (6.4), transparency (6.5), and fairness (6.6), and conclude with a list of suggestions (Table 22) that could be incorporated when developing ethical guidelines and best practices.

Ethical considerations in AI-based AD detection from speech

Autonomy

Consent

1. Ensure participants are aware of the risks and benefits, what is being consented to and how the data will be used prior to participating in research

2. Consider the stage of the disease when obtaining consent

Pre-symptomatic stage:	2.1 Ensure participants understand the difference between biomarker research and clinical trial to best evaluate the potential benefits
	2.2 Consider questions of disclosure, data sharing and storing
Early stage:	2.3 Use simple language, repetition, extra explanations, and visual aid to communicate the risks and benefits of participation when needed
	2.4 Consider cultural background when evaluating the ability to consent
Advanced stage:	2.5 Apply higher safeguarding
	2.6 Consider the ethical issues of proxy decision-making
	2.7 Remember that AD does not equal incompetence

3. When using longitudinal data, consider using gradual informed consent transfer

4. When using secondary data where consent cannot be obtained, give additional consideration to the best interests of the participants

5. Develop best practices and guidelines for using secondary data

Avoiding depersonalisation

6. Treat the persons with AD as semiotic subjects

7. Treat communication deficit and the disease as separate from the person

8. Focus research on the aspects of communication that are preserved in addition to those that are lost

Disclosure of research outcome

9. When seeking consent, clearly communicate the procedures regarding the disclosure of research outcomes

10. Respect the participant's right not to know the outcome

11. Explain that the model can be wrong

Privacy and data protection

12. Promote high standards of self-regulation, especially in countries where informed consent and data protection rules are less stringent

13. Consider giving the participants control over the recording device, with a chance to start and stop the recording

14. Work on the development of automatic feature extraction algorithms by manually annotating the data to train NLP models

15. Reduce speech intelligibility to protect participants' privacy by using, for example:

- 15.1 PCA of audio spectrum
- 15.2 linear prediction for replacing the vowels in speech
- 15.3 voiced and unvoiced segments, pitch, energy, and speech rate
- 15.4 bone-constructed acoustic energy fluctuations from vocal apparatus

16. Reconsider the use of language features that can compromise privacy, such as the MFCC features

17. Consider the necessary length of the audio samples to protect participants' privacy while retaining the important information

18. Develop algorithms that would first detect the event of interest and then record in short windows

19. De-identify or pseudonymise and anonymise the data (including public data)

20. Apply encryption on audio data

21. Avoid storing the full audio

22. Consider the level of data sharing and vetting, and the type of data that is shared depending on the obtained consent, risk of harm, and the ethical training of those accessing the recordings

23. When seeking consent, clearly communicate the level of data sharing and the type of data being shared, as well as who will have access to the data and how the data will be used

24. Rather than promising not to share, or destroy the data, carefully consider the data sharing and storing process

25. Consider using data repositories

26. Consider giving the participants the choice of how much and what type of data will be shared

Welfare

Reducing distress

27. Favour friendly conversational speech over stressful cognitive testing

28. Respect the participants' right to stop the recording

29. Be clear about the process of delivering the research outcome, including the procedures that follow and the support that is available

30. Explain the issues related to the accuracy of the models and the reliability of the research outcome

31. Conduct personal ethical evaluation including the subject in the process when deciding on the disclosure process

32. Consider cultural context when making decisions on the disclosure of the research outcome

33. Consider the potential distress for families, clinicians, and researchers

34. Give prior warning to researchers working on potentially emotionally triggering content, allow them to stop the work when needed, limit the amount of data per annotator, and provide access to support when needed

Avoiding stigmatisation and discrimination

35. Avoid framing dementia as a threat to social and healthcare system

36 Study pre-symptomatic and early AD experience to prompt discussions and raise public awareness to reduce stigmatisation

37. Be aware of the risk of dual use and potential consequences when developing new technologies

38. Develop restrictions for the use of the speech-based tools to avoid unethical and dual use

39. Apply highest research and ethical standards when developing new technologies

Improving the reliability of research outcome

40. Use free speech that captures the abilities of the subjects more adequately

41. Consider factors that affect language use, such as cultural background and education level when interpreting the outcome

42. Be mindful of sample size and avoid overfitting, using, for example, cross validation and held-out sets

43. Promote using large sample size

44. Increase the availability of the data by promoting collaborations between researchers from different fields

45. Be aware of publication bias

46. Be aware of cognitive bias of the researcher - consider masking experiment conditions and key parts of data

47. Use appropriate evaluation metrics, especially when working with small, unbalanced samples

48. Standardise data collection, reporting of the results, model characteristics and evaluation metrics to increase comparability and contextualisation of the studies

Transparency

49. Improve the explainability of the models and the language features

50. Consider who the explanations are meant for, how easy they are to interpret, and to what extent do people trust the explanations

51. Include the input of the clinicians when developing clinical NLP

52. Reconsider using language features that are difficult to interpret, such as the MFCC features

53. Promote technical training among clinicians working with AI

54. Promote general health and technological literacy
55. Acknowledge the role of transparency in overall trust in science and working towards public good
56. Promote transparency in research

Fairness

57. Consider who will benefit from the research
58. Clearly explain the risks and benefits available for the participant prior to the research
59. Collect and use balanced datasets that are applicable to the population that is being tested
60. Include different languages and dialects when developing the models
61. Reconsider using language features that are sensitive to dialects, such as the MFCC features
62. Develop language resources for low-resource languages
63. Include minority groups and diverse populations
64. Include standard description of datasets with the studies
65. Use risk of bias assessment toolkits

Table 22: Ethical considerations in AI-based AD detection from speech

7. Conclusions

In this concluding chapter, I will (1) recap the motivations for the current thesis, (2) summarise the findings, contributions, and implications of this work, and (3) propose directions for future research. The summary of the chapter is provided at the end.

7.1. Recap of motivations

In this thesis, I focussed on the automatic detection of early signs of AD from speech and language. As the prevalence of AD is growing due to the ageing population, research into both treatment development and early detection is becoming more and more relevant. Detecting AD early would allow early interventions and have the potential to slow disease progression. However, as many of the approaches to detect AD are invasive or expensive, there is an increasing need for robust, non-invasive, and cheap screening tools. Changes in speech and language can provide relevant information about AD-related cognitive decline and therefore have a great potential to act as early markers and aid in developing these tools. While promising results have been achieved in differentiating between healthy individuals and those with AD using AI- and NLP-based tools for language analysis, several areas of improvement have been identified.

In this thesis, I first filled the gap in the existing literature by conducting a systematic literature review (section 2.2, Petti et al., 2020). This literature review was the first in the area and was needed to understand what research has been done and what still needs to be done. The review has informed the directions taken in further research and has been valuable to the broader research community, which can be demonstrated by the article having been cited over a hundred times during the writing of this thesis.

This literature review also acted as a basis and starting point of the current thesis and informed its motivations and aims. Based on the findings of the literature review, the rest of this thesis set out to address three main challenges:

1. Lack of longitudinal data
2. Lack of generalisability, standardisation, and replicability
3. Lack of ethical guidelines

7.1.1. Lack of longitudinal data

The lack of longitudinal speech data has compromised the amount and depth of research being conducted into the earliest signs of the disease. Having large amounts of longitudinal language data would allow exploring the changes over time and provide a more accurate picture of how cognitive changes manifest in language use, contributing to developing more accurate screening tools. However, this data is expensive and time-consuming to collect, which makes limited data one of the major challenges in this research area. To address this limitation, in this thesis I presented a novel corpus of transcripts of longitudinal spontaneous speech recordings and demonstrated the usefulness of this kind of data.

7.1.2. Lack of generalisability, standardisation, and replicability

Another limitation in automatic language-based AD detection, partly fuelled by limited data availability, is the lack of replicability, generalisability and standardisation across studies and research methods. Understanding how universal the changes in language are across different individuals and standardising the methods of data collection and analysis would promote research reliability, increase replicability, and help develop robust models applicable to wider populations. To address these challenges, I replicated a previous case study on a larger group of individuals, compared the changes in language over time in 20 speakers, and explored the optimal length of a speech sample necessary for informative analysis.

7.1.3. Lack of ethical guidelines

Lastly, while AI- and NLP-aided AD detection could have great potential in improving the life quality of the elderly, the process faces various ethical dilemmas that have received little attention. To address this issue, I discussed the ethical considerations in speech-based early AD detection, related to autonomy, data protection, welfare, transparency, and fairness, and established a list of suggestions that could be implemented in the development of ethical guidelines for researchers and clinicians working in the area.

7.2. Findings and contributions

Based on the systematic literature review, I established three main areas of improvement that motivated the current thesis, outlined in 7.1. In the following sections, I will lay out the findings and contributions related to these motivations in chapter order, and summarise the implications of this work.

7.2.1. LoSST-AD corpus

While language impairment in Alzheimer's disease (AD) has been widely studied, due to limited data availability, relatively few studies have focussed on the longitudinal change in language in the individuals who later develop AD. I addressed this issue in Chapter 3 and constructed a novel corpus of transcripts of public interviews with 20 famous figures, half of whom will eventually develop AD, recorded over several decades. I evaluated the corpus by validating patterns of vocabulary richness changes known from literature, such as decline in noun frequency, word length, and several other features. While this data provides a unique viewpoint to understanding longitudinal changes in language in AD, it must be remembered that the public interviews of famous individuals may not be representative of the wider population when interpreting the results.

Contributions and findings:

- I developed and made available a novel language resource of 135 spontaneous speech transcripts.
- I showed that public data could be used to collect longitudinal datasets without causing extra stress for the participants.
- I demonstrated that this data can adequately reflect longitudinal AD-related changes in vocabulary richness features.
- I ran two experiments to investigate the change in lexical diversity features over time and compare the change in healthy and AD groups, and found significant changes in noun and adposition frequency, word length and frequency, unique words, Brunet index, and repetitions in the AD group.

7.2.2. Replicability and generalisability

In Chapter 4, I focussed on understanding how universal the patterns of AD-related language changes are across individual speakers. Significant differences in speech changes over time have previously been found by comparing individual speakers, for example, the press conference transcripts of President Bush and President Reagan, who was later diagnosed with AD. However, these findings have not been replicated on a larger number of individuals. I explored whether the language change patterns previously established in the single AD-HC participant pair apply to a larger group of individuals who later receive AD diagnosis, and investigated which language features change the most consistently in a group of speakers.

Contributions and findings:

- I replicated a previous case study on a larger group of individuals.
- I failed to find generalisable patterns of language change using previous methodology.
- I proposed alternative methods for data analysis, investigating the benefits of using different language features and their change with age, and compiling the single features into aggregate scores.
- I found that the language features that change the most consistently are moving average type:token ratio and pronoun-related features.
- I found that the aggregate scores performed better than the single features, with lexical diversity capturing a similar change in two thirds of the participants.

7.2.3. Standardisation of sample length

In Chapter 5, I concentrated on the standardisation of research methods in automatic speech-based AD detection, focussing on the optimal length of a speech sample. While shorter speech samples would promote data collection and analysis, the minimum length of informative speech samples remains debated. To investigate the effect of the sample length on the language features, I compared three different sample lengths: original random length, 5- and 1-minute-long samples. I analysed the information captured by the three datasets to investigate

the robustness of the features, understand whether capping the audio improves the accuracy of the analysis, and whether the extra 4 minutes convey necessary information.

Contributions and findings:

- I argued that sample length plays an important role in extracting the language features from speech and should be considered when studying language changes in AD.
- I found that capped audio files have advantages over the random length ones.
- I argued that the 4 extra minutes do convey necessary information for tracking longitudinal changes.

7.2.4. Ethical considerations

While recent studies indicate that AI could play an important role in detecting early signs of Alzheimer's disease in speech, this use of data from individuals with cognitive decline raises numerous ethical concerns. In Chapter 6, I identified and explained the concerns related to autonomy (including consent, depersonalisation and disclosure), privacy and data protection (including the handling of personal content and medical information), welfare (including distress, discrimination and reliability), transparency (including the interpretability of language features and AI-based decision-making for developers and clinicians), and fairness (including bias and the distribution of benefits).

Contributions and findings:

- I outlined the ethical concerns posed by the use of AI in speech-based AD detection.
- I identified ways in which these concerns might be addressed.
- I established a list of suggestions that could be incorporated into ethical guidelines for researchers and clinicians working in this area.

7.2.5. Implications

The corpus constructed in Chapter 3 and the findings based on this data provide a valuable starting point for the development of early detection tools and enhance our understanding of how AD affects language over time. Creating knowledge of long-term language changes in AD contributes to bridging the gap (described by Sadeghian et al., 2021) between the scientific understanding and the creation of diagnostic tests. While capturing universal patterns of language change prior to AD diagnosis can be challenging, as demonstrated in Chapter 4, the research in this area is of significant importance. Identifying the language features that change similarly across larger groups could help lower the dimensionality of the models and contribute towards building more generalisable models (Berisha et al., 2021). In Chapter 4, I identified changes in lexical diversity, MATTR and pronoun-related features as showing the most consistent patterns in the speakers with later AD diagnosis. This work provides practical knowledge of language changes that is applicable in tool development and clinical setting. Replicability could also be improved by the standardisation of research and data collection methods, which I examined in Chapter 5. The findings of this chapter can be applied in the standardisation of sample length and inform further research about the optimal informative recording time necessary to capture longitudinal language changes due to cognitive decline. The list of suggestions addressing the ethical considerations developed in Chapter 6 can be used to inform the development of guidelines and best practices for researchers and clinicians working on AI-based early AD detection from speech and language.

7.3. Future directions

Based on the work completed in this thesis, there are several directions that future research could take. In what follows, I will outline ideas for future work.

Collecting a larger controlled dataset including real participants from diverse populations, reliable clinical information, regular time intervals, audio length and quality. As emphasised throughout this thesis, data availability and quality is one of the major challenges in the field. A higher quality dataset would allow for creating more in-depth knowledge about longitudinal

language changes in AD that are generalisable to a broader population. Such a dataset would contribute to developing more robust methods for data analysis, discovering more reliable patterns, and developing screening tools that are applicable in a real-world clinical setting.

Developing methods to analyse longitudinal language change in AD. Based on the findings of the current thesis, previous methods for analysing longitudinal change could be improved. This research would be supported by having a larger, more representative, and less sparse dataset. Some directions for improving the methods of longitudinal analysis could include grouping the individual language features into aggregate scores, focussing on the different stages of AD and exploring the best ways for analysing each, and exploring the role of the number and length of voice recordings or transcripts. I will elaborate on these approaches in the following three paragraphs.

Constructing reliable and interpretable aggregate scores. Current work showed notable advantages of using aggregate scores, such as lexical diversity and word finding difficulty, over the single language features. However, the methods of compiling the scores could be improved, for example, by weighing the relevance of the individual features at different points in time. Similarly, while the current work used literature-based approach to construct the aggregate scores, future work could compare this method to the data-driven approach.

Investigating the relationship between disease stage and amount of speech data needed. Based on the findings of the current work, I hypothesise that less speech data might be needed to identify more severe stages of AD as the cognitive decline will be more apparent than in the earlier stages. Collecting less data from more severely affected individuals would reduce the burden placed on the speaker and encourage participation, as well as reduce computation time. This hypothesis could be tested in future work by comparing the informativeness of longer and shorter samples in different stages of the disease.

Exploring the trade-off between the number and the length of samples. While the current work focussed on finding the optimal length of voice recordings, this work could be extended by analysing whether the AD-related language changes are better captured by a larger number of shorter samples per speaker or a smaller number of longer samples. This would enhance our understanding of how cognitive decline can be captured in speech and contribute to the

standardisation of research methods which is one of the major areas of improvement in this research area.

Focussing on participant level analysis. As the current work suggests that the group patterns are often not reflective of the individual's language change, I encourage future work to focus the efforts on developing the understanding of AD-related language changes in individual speakers. Understanding how the changes in individual speakers resemble and differ would contribute to improving generalisability, drawing more reliable conclusions, and developing models that take individual differences into account, making them applicable to diverse real-world populations. When interpreting the results, I recommend future work to consider that the mean values may not be representative of the individuals.

Developing ethical guidelines and best practices for automated speech-based early AD detection. While detecting early signs of AD from speech using AI has shown promising results, numerous ethical concerns arise in the process, and there are currently no universally accepted ethical standards or guidelines. In the current work, I outlined the ethical considerations related to autonomy, privacy and data protection, welfare, transparency, and fairness, and developed a list of suggestions that could provide a useful starting point for establishing widely accepted ethical guidelines and best practices. Future research should focus on developing these guidelines to ensure the models are developed ethically and with the best interests of the speakers in mind.

Promoting discussions between different stakeholders, research areas, and communities. As one of the next research projects, I plan to focus on patient involvement as well as developing a better understanding of the experiences of the families, clinicians, and developers and researchers in both private and academic settings. I aim to identify any challenges, worries, or tensions that the different groups face, and the motivations, hopes and potential solutions by interviewing the individuals who have a different connection to AI- and speech-based AD detection. This work will contribute to understanding the needs of the affected individuals and guide further research and tool development, taking the perspectives of different stakeholders into consideration.

Focussing on model transparency and explainability. The importance of transparency in clinical decision making has been discussed in the current work, and based on that I suggest that one of the directions of future research should focus on developing explainable models and including clinician input in the process. Clinician input is important as they would be the users of this technology, and it is important that the outcome is understandable, and any arising concerns noticed and addressed. This work could include looking at the interpretability of both language features and model decisions, the trade-off between transparency and performance, as well as comparing the transparency and accuracy of the decisions made by clinicians and machines.

Addressing the risks of privatising automatic speech-based AD detection. While research in this area could significantly improve the lives of the elderly, privatising automatic speech-based AD detection technology poses risks of misuse and discrimination. Potential harms could include compromising eligibility to health insurance, job opportunities and legal status. These concerns are especially relevant in speech-based cognitive decline detection, as speech data is easy to collect and link to sensitive health information. Future work should assess these risks, consider how to avoid them, and focus on developing responsible technology and appropriate legislation.

Exploring privacy trade-offs. As speech data can carry personal and identifiable information, it is important to explore the ways to maximise speakers' privacy and anonymity, especially when working with sensitive medical information, such as cognitive decline. Future research could explore the extent to which the data can be modified to increase privacy (for example, by voice modification, transcript anonymisation, changing the mapping of demographic data or collecting less of it) without damaging the model.

Developing data and sample effective models. As discussed in the current work, data availability is one of the major limitations in this research area. Therefore, current models are often based on small, specialised samples, including the WEIRD population. Even when these models achieve high performance accuracy, they are not generalisable to a wider population, reproducing the bias in data and leading to overfitting. While collecting large amounts of real-world data would be ideal, it is also challenging, expensive, and time-consuming. Future research could investigate the pros and cons of alternative data collection methods, such as

social media entries, data synthesis, or consent modifications for data sharing purposes, and focus on developing data and sample effective models.

Expanding the research into different languages. While recent years have shown an increase in the number of studies focussing on early AD detection from languages other than English and comparing cognitive decline manifestations in different languages, there is still room for improvement. Therefore, another future direction could be constructing multilingual datasets. This would help develop an understanding of how language-dependent the AD-related language changes are, as well as extend the technology to different languages which would make the tools applicable to wider populations.

Distinguishing between different forms of dementia. While the current thesis focussed specifically on AD detection, this work could be expanded by including different forms of dementia, and exploring whether language analysis could be used to distinguish between different conditions.

7.4. Summary

In this chapter, I recapped the motivation of this thesis, summarised the findings, contributions, and implications of the research I conducted, and provided directions for future research.

To conclude, while AI-based early AD detection from speech is a promising area of research, there are several challenges that I aimed to address in this thesis. I presented a systematic literature review in automatic early AD detection from speech and language to summarise the state-of-the-art and identify the main areas of improvement. I tackled the challenges of the lack of longitudinal data by presenting a novel corpus, replicability, generalisability, and standardisation by exploring the consistency of AD-related language changes across participants and investigating the role of sample length, developed a list of suggestions for the development of ethical guidelines and best practices, and provided recommendations for further research. I hope that the insights of the current thesis will contribute to developing robust, fair, and ethical methods for automatic early AD detection from speech and language.

Bibliography

1. Ahmed, S., Haigh, A. M. F., de Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*, *136*(12), 3727-3737.
2. Al-Atroshi, C., Rene Beulah, J., Singamaneni, K. K., Pretty Diana Cyril, C., Neelakandan, S., & Velmurugan, S. (2022). Automated speech based evaluation of mild cognitive impairment and Alzheimer's disease detection using with deep belief network model. *International Journal of Healthcare Management*, 1-11.
3. Alzheimer's Association. (2008). 2008 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *4*(2), 110-133.
4. Alzheimer's Association. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's & dementia*, *15*(3), 321-387.
5. Ammar, R. B., & Ayed, Y. B. (2018). Speech processing for early Alzheimer disease diagnosis: machine learning based approach. In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.
6. Appell, J., Kertesz, A., & Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain and language*, *17*(1), 73-91.
7. Armstrong, R. A. (2019). Risk factors for Alzheimer's disease. *Folia neuropathologica*, *57*(2), 87-105.
8. Ash, S., Moore, P., Antani, S., McCawley, G., Work, M., & Grossman, M. (2006). Trying to tell a tale: Discourse impairments in progressive aphasia and frontotemporal dementia. *Neurology*, *66*(9), 1405-1413.
9. Baba, A., Yoshizawa, S., Yamada, M., Lee, A., & Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan (Part II: Electronics)*, *87*(7), 49-57.
10. Babulal, G. M., Quiroz, Y. T., Albensi, B. C., Arenaza-Urquijo, E., Astell, A. J., Babiloni, C., Bahar-Fuchs, A., Bell, J., Bowman, G.L., Brickman, A.M., Chételat, G., Ciro, C., Cohen, A. D., Dilworth-Anderson, P., Dodge, H. H., Dreux, S., Edland, S., Esbensen, A., Evered, L., Ewers, M., & O'Bryant, S. E. (2019). Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: update and areas of immediate need. *Alzheimer's & Dementia*, *15*(2), 292-312.
11. Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., & Novikova, J. (2021). Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech. *Frontiers in aging neuroscience*, *13*, 635945.
12. Baldas, V., Lampiris, C., Capsalis, C., & Koutsouris, D. (2011). Early diagnosis of Alzheimer's type dementia using continuous speech recognition. In *Wireless Mobile Communication and Healthcare: Second International ICST Conference, MobiHealth 2010, Ayia Napa, Cyprus, October 18-20, 2010. Revised Selected Papers 1* (pp. 105-110). Springer Berlin Heidelberg.

13. Bayles, K. A. (1982). Language function in senile dementia. *Brain and language*, *16*(2), 265-280.
14. Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of neurology*, *51*(6), 585-594.
15. Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., & Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline?. *Frontiers in aging neuroscience*, *10*, 369.
16. Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. *NPJ digital medicine*, *4*(1), 153.
17. Berisha, V., Wang, S., LaCross, A., & Liss, J. (2015). 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.
18. Bertini, F., Allevi, D., Lutero, G., Montesi, D., & Calzà, L. (2021). Automatic speech classifier for mild cognitive impairment and early dementia. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1-11.
19. Blanken, G., Dittmann, J., Haas, J. C., & Wallesch, C. W. (1987). Spontaneous speech in senile dementia and aphasia: Implications for a neurolinguistic model of language production. *Cognition*, *27*(3), 247-274.
20. Boersma P, Weenink D. Praat. Doing phonetics by computer (Version 5.2. 20)[Software].
21. Boyé, M., Tran, T. M., & Grabar, N. (2014). NLP-oriented contrastive study of linguistic productions of Alzheimer's and control people. In *Advances in Natural Language Processing: 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings 9* (pp. 412-424). Springer International Publishing.
22. Brayne, C., & Kelly, S. (2019). Against the stream: early diagnosis of dementia, is it so desirable?. *BJPsych bulletin*, *43*(3), 123-125.
23. Breijyeh, Z., & Karaman, R. (2020). Comprehensive review on Alzheimer's disease: causes and treatment. *Molecules*, *25*(24), 5789.
24. Brock, D. (1993). Quality of life measures in health care and medical ethics. *The quality of life*, *1*, 95-133.
25. Brunet, É. (1978). *Jean Giraudoux's vocabulary structure and evolution* . Slatkin.
26. Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, *14*(1), 71-91.
27. Calzà L, Gagliardi G, Favretti RR, Tamburini F (2021) Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language* **65**, 101-113.
28. Calzà, L., Beltrami, D., Gagliardi, G., Ghidoni, E., Marcello, N., Rossini-Favretti, R., & Tamburini, F. (2015). Should we screen for cognitive decline and dementia?. *Maturitas*, *82*(1), 28-35.

29. Cammen, T. J. V. D., Croes, E. A., Dermaut, B., Jager, M. C. D., Cruts, M., Van Broeckhoven, C., & Van Duijn, C. M. (2004). Genetic testing has no place as a routine diagnostic test in sporadic and familial cases of Alzheimer's disease. *Journal of the American Geriatrics Society*, *52*(12), 2110-2113.
30. Casserly, I. P., & Topol, E. J. (2009). Convergence of atherosclerosis and Alzheimer's disease: Cholesterol, inflammation, and misfolded proteins. *Discovery medicine*.
31. Chau, H. H. H., Chau, Y., Wang, H. L., Chuang, Y. F., & Lee, C. C. (2022). MCI Detection Based on Deep Learning with Voice Spectrogram. In *2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)* (pp. 212-216). IEEE.
32. Chen, F., Adcock, J., & Krishnagiri, S. (2008). Audio privacy: reducing speech intelligibility while preserving environmental sounds. In *Proceedings of the 16th ACM international conference on Multimedia* (pp. 733-736).
33. Chien, Y. W., Hong, S. Y., Cheah, W. T., Fu, L. C., & Chang, Y. L. (2018). An assessment system for Alzheimer's disease based on speech using a novel feature sequence design and recurrent neural network. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 3289-3294). IEEE.
34. Chien, Y. W., Hong, S. Y., Cheah, W. T., Yao, L. H., Chang, Y. L., & Fu, L. C. (2019). An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network. *Scientific Reports*, *9*(1), 19597.
35. Christman, J. (2003). Autonomy in moral and political philosophy.
36. Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R. M., Kapur, P., DeRamus, T. P. & Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *2*, 113-122.
37. Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of quantitative linguistics*, *17*(2), 94-100.
38. Critchley, M. (1964). The neurology of psychotic speech. *The British Journal of Psychiatry*, *110*(466), 353-364.
39. Croot, K., Hodges, J. R., Xuereb, J., & Patterson, K. (2000). Phonological and articulatory impairment in Alzheimer's disease: a case series. *Brain and language*, *75*(2), 277-309.
40. Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., Casillas, M., De Barbaro, K., Bang, J.Y. & Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best practices. *Behavior research methods*, *52*, 1951-1969.
41. de Ajuriaguerra, J., & Tissot, R. (1975). Some aspects of language in various forms of senile dementia (comparisons with language in childhood). In *Foundations of language development* (pp. 323-339). Academic Press.

42. de la Fuente Garcia, S., Ritchie, C. W., & Luz, S. (2020). Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, *78*(4), 1547-1574.
43. de Lira, J. O., Minett, T. S. C., Bertolucci, P. H. F., & Ortiz, K. Z. (2018). Evaluation of macrolinguistic aspects of the oral discourse in patients with Alzheimer's disease. *International Psychogeriatrics*, 1–11.
44. de Lira, J. O., Minett, T. S. C., Bertolucci, P. H. F., & Ortiz, K. Z. (2014). Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dementia & neuropsychologia*, *8*, 260-265.
45. DeTure, M. A., & Dickson, D. W. (2019). The neuropathological diagnosis of Alzheimer's disease. *Molecular neurodegeneration*, *14*(1), 32.
46. Dijkstra, K., Bourgeois, M. S., Allen, R. S., & Burgio, L. D. (2004). Conversational coherence: Discourse analysis of older adults with and without dementia. *Journal of Neurolinguistics*, *17*(4), 263-283.
47. Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, *25*(2), 220-242.
48. Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice*, *32*(5), 644-e1.
49. Eltengy, A. H. (2021, July). Encryption Of Voice Calls Using CryptoBin Algorithm. In *2021 International Telecommunications Conference (ITC-Egypt)* (pp. 1-5). IEEE.
50. Fang, C., Janwattanapong, P., Martin, H., Cabrerizo, M., Barreto, A., Loewenstein, D., Duara, R., & Adjouadi, M. (2017) Computerized neuropsychological assessment in mild cognitive impairment based on natural language processing-oriented feature extraction. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 543-546). IEEE.
51. Farsana, F. J., Devi, V. R., & Gopakumar, K. (2023). An audio encryption scheme based on Fast Walsh Hadamard Transform and mixed chaotic keystreams. *Applied Computing and Informatics*, *19*(3/4), 239-264.
52. Fiske, A. P. (1993). Social errors in four cultures: Evidence about universal forms of social relations. *Journal of Cross-Cultural Psychology*, *24*(4), 463-494.
53. Foley, J. M. (1992). The experience of being demented. *Dementia and aging: Ethics, values, and policy choices*, 30-43.
54. Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, *12*(3), 189-198.
55. Forbes-McKay, K. E., & Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological sciences*, *26*, 243-254.

56. Ford, E., Milne, R., & Curlewis, K. (2023). Ethical issues when using digital biomarkers and artificial intelligence for the early detection of dementia. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *13*(3), e1492.
57. Fox, N. C., Warrington, E. K., Seiffer, A. L., Agnew, S. K., & Rossor, M. N. (1998). Presymptomatic cognitive deficits in individuals at risk of familial Alzheimer's disease. A longitudinal prospective study. *Brain: a journal of neurology*, *121*(9), 1631-1639.
58. Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*(2), 407-422.
59. Fukuda, M., Nishimura, R., Umezawa, M., Yamamoto, K., Iribe, Y., & Kitaoka, N. (2022). Elderly Conversational Speech Corpus with Cognitive Impairment Test and Pilot Dementia Detection Experiment Using Acoustic Characteristics of Speech in Japanese Dialects. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 1016-1022).
60. Garcia, M. (2016). Racist in the Machine. *World Policy Journal*, *33*(4), 111-117.
61. Garrard, P., & Carroll, E. (2006). Lost in semantic space: a multi-modal, non-verbal assessment of feature knowledge in semantic dementia. *Brain*, *129*(5), 1152-1163.
62. Garrard, P., Maloney, L. M., Hodges, J. R., & Patterson, K. (2005b). The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author. *Brain*, *128*(2), 250-260.
63. Garrard, P., Nemes, V., Nikolic, D., & Barney, A. (2017). Motif discovery in speech: application to monitoring Alzheimer's disease. *Current Alzheimer Research*, *14*(9), 951-959.
64. Garrard, P., Ralph, M. A. L., Patterson, K., Pratt, K. H., & Hodges, J. R. (2005a). Semantic feature knowledge and picture naming in dementia of Alzheimer's type: a new approach. *Brain and language*, *93*(1), 79-94.
65. Gauthier, S., Leuzy, A., Racine, E., & Rosa-Neto, P. (2013). Diagnosis and management of Alzheimer's disease: past, present and future ethical issues. *Progress in neurobiology*, *110*, 102-113.
66. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.
67. Goedert, M., & Spillantini, M. G. (2006). A century of Alzheimer's disease. *science*, *314*(5800), 777-781.
68. Gosztolya, G., Tóth, L., Grósz, T., Vincze, V., Hoffmann, I., Szatlóczki, G., Pákáski, M., & Kálmán, J. (2016). Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection.
69. Gosztolya, G., Vincze, V., Tóth, L., Pákáski, M., Kálmán, J., & Hoffmann, I. (2019). Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. *Computer Speech & Language*, *53*, 181-197.
70. Graham, N. L., Emery, T., & Hodges, J. R. (2004). Distinctive cognitive profiles in Alzheimer's disease and subcortical vascular dementia. *Journal of Neurology, Neurosurgery & Psychiatry*, *75*(1), 61-71.

71. Greenaway, M. C., Lacritz, L. H., Binegar, D., Weiner, M. F., Lipton, A., & Cullum, C. M. (2006). Patterns of verbal memory performance in mild cognitive impairment, Alzheimer disease, and normal aging. *Cognitive and Behavioral Neurology*, *19*(2), 79-84.
72. Grundman, M., Petersen, R. C., Bennett, D. A., Feldman, H. H., Salloway, S., Visser, P. J., Thal, L. J., Schenk, D., Khachaturian, Z., & Thies, W. (2006). Alzheimer's association research roundtable meeting on mild cognitive impairment: what have we learned?. *Alzheimer's & Dementia*, *2*(3), 220-233.
73. Guinn, C., Singer, B., & Habash, A. (2014). A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)* (pp. 98-103). IEEE.
74. Guyatt, G. H., Oxman, A. D., Montori, V., Vist, G., Kunz, R., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011b). GRADE guidelines: 5. Rating the quality of evidence—publication bias. *Journal of clinical epidemiology*, *64*(12), 1277-1282.
75. Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P., & Schünemann, H. J. (2011a). GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology*, *64*(4), 383-394.
76. Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America*, *102*(1), 655-658.
77. Haider, F., De La Fuente, S., & Luz, S. (2019). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 272-281.
78. Hamberger, M. J., Friedman, D., Ritter, W., & Rosen, J. (1995). Event-related potential and behavioral correlates of semantic processing in Alzheimer' s patients and normal controls. *Brain and Language*, *48*(1), 33-68.
79. Harrison, C. J., & Sidey-Gibbons, C. J. (2021). Machine learning in medicine: a practical introduction to natural language processing. *BMC medical research methodology*, *21*(1), 158.
80. Hasan, S. A., & Farri, O. (2019). Clinical natural language processing with deep learning. *Data Science for Healthcare: Methodologies and Applications*, 147-171.
81. Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, *33*(2-3), 61-83.
82. Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*, 260-268.
83. Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261-266.

84. Hodges, J. R., Salmon, D. P., & Butters, N. (1992). Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge?. *Neuropsychologia*, *30*(4), 301-314.
85. Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, *12*(1), 29-34.
86. Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, *7*(2), 172-177.
87. Illes, J., Rosen, A., Greicius, M., & Racine, E. (2007). Prospects for prediction: ethics analysis of neuroimaging in Alzheimer's disease. *Annals of the New York Academy of Sciences*, *1097*(1), 278-295.
88. International Organization for Standardization. (1994). *ISO 5725-1: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions*. International Organization for Standardization.
89. Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1-15.
90. Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 27-37).
91. Jiao, Y., Berisha, V., & Liss, J. (2017). Interpretable phonological features for clinical applications. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5045-5049). IEEE.
92. Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 51-57).
93. Kaduszkiewicz, H., Bachmann, C., & van den Bussche, H. (2008). Telling “the truth” in dementia—Do attitude and approach of general practitioners and specialists differ?. *Patient education and counseling*, *70*(2), 220-226.
94. Kálmán, J., Devanand, D. P., Gosztolya, G., Balogh, R., Imre, N., Tóth, L., Hoffmann, I., Kovacs, I., Vincze, V., & Pákáski, M. (2022). Temporal speech parameters detect mild cognitive impairment in different languages: validation and comparison of the Speech-GAP Test® in English and Hungarian. *Current Alzheimer Research*, *19*(5), 373-386.
95. Karlawish, J. (2011). Addressing the ethical, policy, and social challenges of preclinical Alzheimer disease. *Neurology*, *77*(15), 1487-1493.
96. Keeling, G., & Nyrup, R. (2021). Explainable machine learning, patient autonomy, and clinical reasoning.

97. Khodabakhsh, A., Kuscuoglu, S., & Demiroglu, C. (2014a). Detection of Alzheimer's disease using prosodic cues in conversational speech. In *2014 22nd signal processing and communications applications conference (SIU)* (pp. 1003-1006). IEEE.
98. Khodabakhsh, A., Kuşxuoğlu, S., & Demiroğlu, C. (2014b). Natural language features for detection of Alzheimer's disease in conversational speech. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 581-584). IEEE.
99. Khodabakhsh, A., Yesil, F., Guner, E., & Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*, 1-15.
100. Khoury, R., & Ghossoub, E. (2019). Diagnostic biomarkers of Alzheimer's disease: a state-of-the-art review. *Biomarkers in Neuropsychiatry*, *1*, 100005.
101. Kitwood, T. (1990). The dialectics of dementia: with particular reference to Alzheimer's disease. *Ageing & Society*, *10*(2), 177-196.
102. Klasnja, P., Consolvo, S., Choudhury, T., Beckwith, R., & Hightower, J. (2009). Exploring privacy concerns about personal sensing. In *International Conference on Pervasive Computing* (pp. 176-183). Berlin, Heidelberg: Springer Berlin Heidelberg.
103. Knibb, J. A., Woollams, A. M., Hodges, J. R., & Patterson, K. (2009). Making sense of progressive non-fluent aphasia: an analysis of conversational speech. *Brain*, *132*(10), 2734-2746.
104. Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, *15*(2), 120-129.
105. Kramer, J. H., Mungas, D., Reed, B. R., Schuff, N., Weiner, M. W., Miller, B. L., & Chui, H. C. (2004). Forgetting in dementia with and without subcortical lacunes. *The Clinical Neuropsychologist*, *18*(1), 32-40.
106. Kumar, N., Kumar, V., Anand, P., Kumar, V., Dwivedi, A. R., & Kumar, V. (2022). Advancements in the development of multi-target directed ligands for the treatment of Alzheimer's disease. *Bioorganic & Medicinal Chemistry*, *61*, 116742
107. Kumar, S., Nguyen, L. T., Zeng, M., Liu, K., & Zhang, J. (2015). Sound shredding: Privacy preserved audio sensing. In *Proceedings of the 16th international workshop on mobile computing systems and applications* (pp. 135-140).
108. Lane, N., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T. & Campbell, A. (2012). Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*.
109. Larson, E. C., Lee, T., Liu, S., Rosenfeld, M., & Patel, S. N. (2011). Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 375-384).

110. Le Franc, A., Riebling, E., Karadayi, J., Wang, Y., Scaff, C., Metze, F., & Cristia, A. (2018). The ACLEW DiViMe: An Easy-to-use Diarization Tool. In *Interspeech* (pp. 1383-1387).
111. Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., & Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, *23*(5), e15708.
112. Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). 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.
113. Lei, P., Ayton, S., & Bush, A. I. (2021). The essential elements of Alzheimer's disease. *Journal of Biological Chemistry*, *296*.
114. Lezak, M. D. (2004). *Neuropsychological assessment*. Oxford University Press, USA.
115. Liang, X., Batsis, J. A., Zhu, Y., Driesse, T. M., Roth, R. M., Kotz, D., & MacWhinney, B. (2022). Evaluating voice-assistant commands for dementia detection. *Computer Speech & Language*, *72*, 101297.
116. Lima, T. M., Brandão, L., Parente, M. A. D. M. P., & Peña-Casanova, J. (2014). Alzheimer's disease: cognition and picture-based narrative discourse. *Revista CEFAC*, *16*, 1168-1177.
117. Lipowski, Z. J. (1969). Psychosocial aspects of disease. *Annals of Internal Medicine*, *71*(6), 1197-1206.
118. Lopez-de-Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., Travieso, C. M., Ecay-Torres, M., Martinez-Lage, P., & Eguiraun, H. (2013b). On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, *7*, 44-55.
119. López-de-Ipiña, K., Alonso, J. B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., Ezeiza, A., Barraso, N., Ecay-Torres, M., Martinez-Lange, P., & de Lizardui, U. M. (2013a). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, *13*(5), 6730-6745.
120. Lopez-de-Ipina, K., de-Lizarduy, M. U., Calvo, P. M., Mekyska, J., Beitia, B., Barroso, N., Estanga, A., Tainta, M., & Ecay-Torres, M. (2018). Advances on automatic speech analysis for early detection of Alzheimer disease: a non-linear multi-task approach. *Current Alzheimer Research*, *15*(2), 139-148.
121. López-de-Ipiña, K., Martinez-de-Lizarduy, U., Calvo, P. M., Beitia, B., García-Melero, J., Fernández, E., Ecay-Torres, M., Faundez-Zanuy, M., & Sanz, P. (2020). On the analysis of speech and disfluencies for automatic detection of Mild Cognitive Impairment. *Neural Computing and Applications*, *32*, 15761-15769.
122. López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J. B., Travieso, C. M., Ezeiza, A., Barroso, A., Ecay-Torres, M., Martinez-Lage, P., & Beitia, B. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, *30*(1), 43-60.

123. Lopez, J. A. S., González, H. M., & Léger, G. C. (2019). Alzheimer's disease. *Handbook of clinical neurology*, *167*, 231-255.
124. Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, *5*(1), 96-116.
125. Luz, S., de la Fuente, S., & Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*
126. Luz, S., Haider, F., de la Fuente Garcia, S., Fromm, D., & MacWhinney, B. (2021b). Alzheimer's dementia recognition through spontaneous speech. *Frontiers in computer science*, *3*, 780169.
127. Luz, S., Haider, F., de la Fuente, S., Fromm, D., & MacWhinney, B. (2021a). Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.
128. Maley, J. H., Wanis, K. N., Young, J. G., & Celi, L. A. (2020). Mortality prediction models, causal effects, and end-of-life decision making in the intensive care unit. *BMJ Health & Care Informatics*, *27*(3).
129. Markesbery, W. R., Schmitt, F. A., Kryscio, R. J., Davis, D. G., Smith, C. D., & Wekstein, D. R. (2006). Neuropathologic substrate of mild cognitive impairment. *Archives of neurology*, *63*(1), 38-46.
130. Marson, D. C. (2001). Loss of competency in Alzheimer's disease: conceptual and psychometric approaches. *International Journal of Law and Psychiatry*, *24*(2-3), 267-283.
131. Martinez-de-Lizarduy, U., Salomón, P. C., Vilda, P. G., Torres, M. E., & de Ipiña, K. L. (2017). ALZUMERIC: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, *4*(1), e037-e037.
132. Martínez-Sánchez, F., Meilán, J. J., Vera-Ferrandiz, J. A., Carro, J., Pujante-Valverde, I. M., Ivanova, O., & Carcavilla, N. (2016). Speech rhythm alterations in Spanish-speaking individuals with Alzheimer's disease. *Aging, Neuropsychology, and Cognition*, *24*(4), 418-434.
133. Masters, C. L., Bateman, R., Blennow, K., Rowe, C. C., Sperling, R. A., & Cummings, J. L. (2015). Alzheimer's disease. *Nature reviews disease primers*, *1*(1), 1-18.
134. Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: comparing the context sensitivity of Japanese and Americans. *Journal of personality and social psychology*, *81*(5), 922.
135. Mattsson, N., Brax, D., & Zetterberg, H. (2010). To know or not to know: ethical issues related to early diagnosis of Alzheimer's disease. *International journal of Alzheimer's disease*, *2010*.
136. Mayeda, E. R., Glymour, M. M., Quesenberry, C. P., & Whitmer, R. A. (2016). Inequalities in dementia incidence between six racial and ethnic groups over 14 years. *Alzheimer's & Dementia*, *12*(3), 216-224.
137. McKernan, L. C., & Clayton, E. W. (2018). Protecting life while preserving liberty: ethical recommendations for suicide prevention with artificial intelligence. *Frontiers in psychiatry*, *9*, 424191.
138. Mehl, M. R. (2017). The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior. *Current directions in psychological science*, *26*(2), 184-190.

139. Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, *84*(4), 857.
140. Meilán, J. J., Martínez-Sánchez, F., Carro, J., Sánchez, J. A., & Pérez, E. (2012). Acoustic markers associated with impairment in language processing in Alzheimer's disease. *The Spanish journal of psychology*, *15*(2), 487-494.
141. Meltzer, J. A. (2020). Towards early prediction of Alzheimer's disease through language samples. *EClinicalMedicine*, *29*.
142. Metarugcheep, S., Punyabukkana, P., Wanvarie, D., Hemrungrojn, S., Chunharas, C., & Pratanwanich, P. N. (2022). Selecting the Most Important Features for Predicting Mild Cognitive Impairment from Thai Verbal Fluency Assessments. *Sensors*, *22*(15), 5813.
143. Meyer, M. N. (2018). Practical tips for ethical data sharing. *Advances in methods and practices in psychological science*, *1*(1), 131-144.
144. Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., & Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, *58*(2), 373-387.
145. Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., & Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, *53*, 65-79.
146. Mirzaei, S., El Yacoubi, M., Garcia-Salicetti, S., Boudy, J., Kahindo, C., Cristancho-Lacroix, V., Kerherve, H., & Rigaud, A. S. (2018). Two-stage feature selection of voice parameters for early Alzheimer's disease prediction. *Irbm*, *39*(6), 430-435.
147. Mitchell, T. M. (1997). Machine Learning. 15-17. New York: McGraw-Hill Science/Engineering/Math.
148. Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, *48*(2), 239-278.
149. Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & Prisma-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews*, *4*, 1-9.
150. Müller, V. C. (2020). Ethics of artificial intelligence and robotics.
151. Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.
152. Munthuli, A., Vongsurakrai, S., Anansiripinyo, T., Ellermann, V., Sroykhumpa, K., Onsuwan, C., Chutichetpong, P., Hemrungrojn, S., Kosawat, K., & Tantibundhit, C. (2021). Thammasat-NECTEC-Chula's Thai language and cognition assessment (TLCA): The Thai Alzheimer's and mild cognitive impairment

screening test. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 690-694). IEEE.

153. Nagumo, R., Zhang, Y., Ogawa, Y., Hosokawa, M., Abe, K., Ukeda, T., Sumi, S., Kurita, S., Nakakubo, S., Lee, S., & Shimada, H. (2020). Automatic detection of cognitive impairments through acoustic analysis of speech. *Current Alzheimer Research*, *17*(1), 60-68.
154. Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2011). Montreal cognitive assessment. *The American Journal of Geriatric Psychiatry*.
155. Nasrolahzadeh, M., Rahnamayan, S., & Haddadnia, J. (2022). Alzheimer's disease diagnosis using genetic programming based on higher order spectra features. *Machine Learning with Applications*, *7*, 100225.
156. Nickel, P. J. (2006). Vulnerable populations in research: the case of the seriously ill. *Theoretical medicine and bioethics*, *27*, 245-264.
157. O'Neill, O. (2002). *Autonomy and trust in bioethics*. Cambridge University Press.
158. Obler, L. K. (1983). Language and brain dysfunction in dementia. In *Language functions and brain organization* (pp. 267-282). Academic Press.
159. Obler, L. K., & Albert, M. L. (1981). Language and aging: A neurobehavioral analysis. *Aging: Communication processes and disorders*, 107-121.
160. Oppenheim, A. V., Schafer, R. W., & Buck J. R. (1998). Discrete-Time Signal Processing. 2nd ed. 1-5. Upper Saddle River, NJ: Prentice Hall.
161. Orange, J. B., & Purves, B. (1996). Conversational discourse and cognitive impairment: Implications for Alzheimer's disease. *Journal of Speech Language Pathology and Audiology*, *20*, 139-139.
162. Papaioannou, D., Sutton, A., & Booth, A. (2016). Systematic approaches to a successful literature review. *Systematic approaches to a successful literature review*, 1-336.
163. Papakyriakopoulos, O., Choi, A. S. G., Thong, W., Zhao, D., Andrews, J., Bourke, R., Xiang, A., & Koenecke, A. (2023). Augmented datasheets for speech datasets and ethical decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 881-904).
164. Petti, U. & Korhonen, A. (forthcoming, accepted to LREC-COLING 2024). LoSST-AD: A Longitudinal Corpus for Tracking Alzheimer's Disease Related Changes in Spontaneous Speech
165. Petti, U., Baker, S., & Korhonen, A. (2020). A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, *27*(11), 1784-1797.
166. Petti, U., Baker, S., Korhonen, A., & Robin, J. (2023a). The Generalizability of Longitudinal Changes in Speech Before Alzheimer's Disease Diagnosis. *Journal of Alzheimer's Disease*, 1-18.
167. Petti, U., Baker, S., Korhonen, A., & Robin, J. (2023b). How Much Speech Data Is Needed for Tracking Language Change in Alzheimer's Disease? A Comparison of Random Length, 5-Min, and 1-Min Spontaneous Speech Samples. *Digital Biomarkers*, *7*(1), 157. (Chapter 5)

168. Petti, U., Nyrup, R., Skopek, J. M., & Korhonen, A. (2023c). Ethical considerations in the early detection of Alzheimer's disease using speech and AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*
169. Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., Köpke, B., Puel, M., & Pariente, J. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *Journal of Alzheimer's disease*, *50*(3), 687-698.
170. Porteri, C., Albanese, E., Scerri, C., Carrillo, M. C., Snyder, H. M., Martensson, B., Baker, B., Giacobini, E., Boccardi, M., Winblad, B., & Frisoni, G. B. (2017). The biomarker-based diagnosis of Alzheimer's disease. 1—ethical and societal issues. *Neurobiology of aging*, *52*, 132-140.
171. Porteri, C., Galluzzi, S., Geroldi, C., & Frisoni, G. B. (2010). Diagnosis disclosure of prodromal Alzheimer disease-ethical analysis of two cases. *Canadian Journal of Neurological Sciences*, *37*(1), 67-75.
172. Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., & Ferri, C. P. (2013). The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, *9*(1), 63-75.
173. Rascovsky, K., Salmon, D. P., Hansen, L. A., Thal, L. J., & Galasko, D. (2007). Disparate letter and semantic category fluency deficits in autopsy-confirmed frontotemporal dementia and Alzheimer's disease. *Neuropsychology*, *21*(1), 20.
174. Rentoumi, V., Paliouras, G., Danasi, E., Arfani, D., Fragkopoulou, K., Varlokosta, S., & Papadatos, S. (2017). Automatic detection of linguistic indicators as a means of early detection of Alzheimer's disease and of related dementias: A computational linguistics analysis. In *2017 8th IEEE international conference on cognitive infocommunications (CogInfoCom)* (pp. 000033-000038). IEEE.
175. Ringman, J. M. (2017). Update on Alzheimer's and the Dementias: Introduction. *Neurologic clinics*, *35*(2), 171-174.
176. Robin, J., Xu, M., Balagopalan, A., Novikova, J., Kahn, L., Oday, A., Hejrati, M., Hashemifar, S., Negahdar, M., Simpson, W., & Teng, E. (2022). Characterizing progressive speech changes in prodromal-to-mild Alzheimer's disease using natural language processing. *Alzheimers Dement*, *18*, e063244.
177. Robin, J., Xu, M., Kaufman, L. D., & Simpson, W. (2021). Using digital speech assessments to detect early signs of cognitive impairment. *Frontiers in digital health*, *3*, 749758.
178. Rodríguez-Aranda, C., Johnsen, S. H., Eldevik, P., Sparr, S., Wikran, G. C., Herder, M., & Vangberg, T. R. (2016). Neuroanatomical correlates of verbal fluency in early Alzheimer's disease and normal aging. *Brain and language*, *155*, 24-35.
179. Rogers, W., & Lange, M. M. (2013). Rethinking the vulnerability of minority populations in research. *American journal of public health*, *103*(12), 2141-2146.
180. Romero, B., & Kurz, A. (1996). Deterioration of spontaneous speech in AD patients during a 1-year follow-up: homogeneity of profiles and factors associated with progression. *Dementia and Geriatric Cognitive Disorders*, *7*(1), 35-40.

181. Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

182. Rousseaux, M., Sève, A., Vallet, M., Pasquier, F., & Mackowiak-Cordoliani, M. A. (2010). An analysis of communication in conversation in patients with dementia. *Neuropsychologia*, *48*(13), 3884-3890.

183. Sabat, S. R. (1994). Excess disability and malignant social psychology: A case study of Alzheimer's disease. *Journal of community & applied social psychology*, *4*(3), 157-166.

184. Sabat, S. R., & Harré, R. (1994). The Alzheimer's disease sufferer as a semiotic subject. *Philosophy, Psychiatry, & Psychology*, *1*(3), 145-160.

185. Sadeghian, R., Schaffer, J. D., & Zahorian, S. A. (2017). Speech processing approach for diagnosing dementia in an early stage.

186. Sadeghian, R., Schaffer, J. D., & Zahorian, S. A. (2021). Towards an automatic speech-based diagnostic test for Alzheimer's disease. *Frontiers in Computer Science*, *3*, 624594.

187. Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: Procedure and data. *Brain and language*, *37*(3), 440-479.

188. Saito, A., & Takeda, K. (2001). Semantic cueing effects on word retrieval in aphasic patients with lexical retrieval deficit. *Brain and language*, *77*(1), 1-9.

189. Sajjadi, S. A., Patterson, K., Tomek, M., & Nestor, P. J. (2012). Abnormalities of connected speech in semantic dementia vs Alzheimer's disease. *Aphasiology*, *26*(6), 847-866.

190. Sangchocanonta, S., Vongsurakrai, S., Sroykhumpa, K., Ellermann, V., Munthuli, A., Anansiripinyo, T., Onsuwan, C., Hemrungrojn, S., Kosawat, K., & Tantibundhit, C. (2021). Development of Thai picture description task for Alzheimer's screening using part-of-speech tagging. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2104-2109). IEEE.

191. Satt, A., Sorin, A., Toledo-Ronen, O., Barkan, O., Kompatsiaris, I., Kokonozi, A., & Tsolaki, M. (2013). Evaluation of speech-based protocol for detection of early-stage dementia. In *Interspeech* (pp. 1692-1696).

192. Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: Challenges and compromise in practice. *Qualitative research*, *15*(5), 616-632.

193. Schäfer, S., Mallick, E., Schwed, L., König, A., Zhao, J., Linz, N., Bodin, T. H., Skoog, J., Possemis, N., Ter Huurne, D., & Tröger, J. (2023). Screening for mild cognitive impairment using a machine learning classifier and the remote speech biomarker for cognition: evidence from two clinically relevant cohorts. *Journal of Alzheimer's Disease*, *91*(3), 1165-1171.

194. Scheltens, N. M. E., Tijms, B. M., Koene, T., Barkhof, F., Teunissen, C. E., Wolfsgruber, S., & Amsterdam Dementia Cohort (2017). Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *13*(11), 1226–1236.

195. Scheltens, P., Blennow, K., Breteler, M. M., De Strooper, B., Frisoni, G. B., Salloway, S., & Van der Flier, W. M. (2016). Alzheimer's disease. *The Lancet*, *388*(10043), 505-517.

196. Scheltens, P., De Strooper, B., Kivipelto, M., Holstege, H., Chételat, G., Teunissen, C. E., Cummings, J., & van der Flier, W. M. (2021). Alzheimer's disease. *The Lancet*, *397*(10284), 1577-1590.

197. Schicktanz, S., Schweda, M., Ballenger, J. F., Fox, P. J., Halpern, J., Kramer, J. H., Micco, G., Post, S. G., Thompson, C., Knight, R. T., & Jagust, W. J. (2014). Before it is too late: professional responsibilities in late-onset Alzheimer's research and pre-symptomatic prediction. Frontiers in Human Neuroscience, 8, 921.

198. Schneider, J. A., Arvanitakis, Z., Leurgans, S. E., & Bennett, D. A. (2009). The neuropathology of probable Alzheimer disease and mild cognitive impairment. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *66*(2), 200-208.

199. Sidnell, J., & Stivers, T. (Eds.). (2012). *The handbook of conversation analysis*. John Wiley & Sons.

200. Silagi, M. L., Bertolucci, P. H. F., & Ortiz, K. Z. (2015). Naming ability in patients with mild to moderate Alzheimer's disease: what changes occur with the evolution of the disease?. *Clinics*, *70*, 423-428.

201. Simon, J., Bastin, C., Salmon, E., & Willems, S. (2018). Increasing the salience of fluency cues does not reduce the recognition memory impairment in Alzheimer's disease!. *Journal of neuropsychology*, *12*(2), 216-230.

202. Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

203. Singh, S., Bucks, R. S., & Cuerden, J. M. (2001). Evaluation of an objective technique for analysing temporal variables in DAT spontaneous speech. *Aphasiology*, *15*(6), 571-583.

204. Smith, A. P., & Beattie, B. L. (2001). Disclosing a diagnosis of Alzheimer's disease: patient and family experiences. *Canadian Journal of Neurological Sciences*, *28*(S1), S67-S71.

205. Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess*, *14*(8), 1-193.

206. Soroski, T., da Cunha Vasco, T., Newton-Mason, S., Granby, S., Lewis, C., Harisinghani, A., Rizzo, M., Conati, C., Murray, G., Carenini, G., & Jang, H. (2022). Evaluating web-based automatic transcription for Alzheimer speech data: transcript comparison and machine learning analysis. *JMIR aging*, *5*(3), e33460.

207. Srivastava, S., Ahmad, R., & Khare, S. K. (2021). Alzheimer's disease and its treatment by different approaches: A review. *European Journal of Medicinal Chemistry*, *216*, 113320.

208. Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S. B., Kawabata, K., Bhandari, K. S., Duncan, J. C., & Berisha, V. (2020). Repeatability of commonly used speech and language features for clinical applications. *Digital biomarkers*, *4*(3), 109-122.

209. Stengel, E. (1964). Speech disorders and mental disorders. In *Ciba Foundation Symposium-Bioassay of Anterior Pituitary and Adrenocortical Hormones (Colloquia on Endocrinology)* (pp. 285-298). Chichester, UK: John Wiley & Sons, Ltd..
210. Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in aging neuroscience*, *7*, 195.
211. Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, *30*(5), 501-556.
212. Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., & Nakamura, S. (2017). Detecting dementia through interactive computer avatars. *IEEE journal of translational engineering in health and medicine*, *5*, 1-11.
213. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, *32*(11), 4793-4813.
214. Tóth, L., Gosztolya, G., Vincze, V., Hoffmann, I., & Szatlóczki, G. (2015). Automatic detection of mild cognitive impairment from spontaneous speech using ASR. ISCA.
215. Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatlóczki, G., Bánréti, Z., Pákáski, M. & Kálmán, J. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, *15*(2), 130-138.
216. Townsend, D. J., Carrithers, C., & Bever, T. G. (2001). Familial handedness and access to words, meaning, and syntax during sentence comprehension. *Brain and Language*, *78*(3), 308-331.
217. Tu, M., Berisha, V., & Liss, J. (2017). Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Interspeech* (pp. 1849-1853).
218. Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323-352.
219. Ujiro, T., Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T., & Nakamura, S. (2018). Detection of Dementia from Responses to Atypical Questions Asked by Embodied Conversational Agents. In *Interspeech* (pp. 1691-1695).
220. Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1-2), 231-270.
221. Ursin, F., Timmermann, C., & Steger, F. (2021). Ethical implications of Alzheimer's disease prediction in asymptomatic individuals through artificial intelligence. *Diagnostics*, *11*(3), 440.
222. Van Dyck, C. H., Swanson, C. J., Aisen, P., Bateman, R. J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., & Iwatsubo, T. (2023). Lecanemab in early Alzheimer's disease. *New England Journal of Medicine*, *388*(1), 9-21.
223. Véliz, C. (2019). Three things digital ethics can learn from medical ethics. *Nature Electronics*, *2*(8), 316-318.

224. Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., Osborn, D., Hayes, J., Stewart, R., Downs, J., Chapman, W., & Dutta, R. (2018). Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, *88*, 11-19.
225. Venneri, A., Jahn-Carta, C., De Marco, M., Quaranta, D., & Marra, C. (2018). Diagnostic and prognostic role of semantic processing in preclinical Alzheimer's disease. *Biomarkers in medicine*, *12*(6), 637-651.
226. Vetráb, M., Egas-López, J. V., Balogh, R., Imre, N., Hoffmann, I., Tóth, L., Pakaski, M., Kalman, K., & Gosztolya, G. (2022). Using spectral sequence-to-sequence autoencoders to assess mild cognitive impairment. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6467-6471). IEEE.
227. Vipperla, R., Renals, S., & Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices.
228. Voleti, R., Liss, J. M., & Berisha, V. (2019). A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, *14*(2), 282-298.
229. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International data privacy law*, *7*(2), 76-99.
230. Warnita, T., Inoue, N., & Shinoda, K. (2018). Detecting Alzheimer's disease using gated convolutional neural network from audio data. *arXiv preprint arXiv:1803.11344*.
231. Weintraub, S., Wicklund, A. H., & Salmon, D. P. (2012). The neuropsychological profile of Alzheimer disease. *Cold Spring Harbor perspectives in medicine*, *2*(4), a006171.
232. Wenk, G. L. (2003). Neuropathologic changes in Alzheimer's disease. *Journal of Clinical Psychiatry*, *64*, 7-10.
233. Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M.M., Sterne, J.A., Bossuyt, P.M., & QUADAS-2 Group*. (2011). QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of internal medicine*, *155*(8), 529-536.
234. Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J.B., Kleijnen, J., & Mallett, S. (2019). R ESEARCH AND R EPORTING M ETHODS PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies.
235. World Health Organisation (2023). *Dementia.* World Health Organisation. https://www.who.int/news-room/fact-sheets/detail/dementia/?gad_source=1&gclid=CjwKCAjwte-vBhBFEiwAQSv_xSkNA7TP7D5u-vPoe89HOe8Wc1yhu3Ln-lXetA-74EGhkpWND8KBGxoClw4QAvD_BwE (Accessed: 21 March 2024)
236. Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., Soni, S., Wang, Q., Wei, Q., Xiang, Y., Zhao, B., & Xu, H. (2020). Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, *27*(3), 457-470.

237. Wyatt, D., Choudhury, T., & Bilmes, J. A. (2007). Conversation detection and speaker segmentation in privacy-sensitive situated speech data. In *Interspeech* (pp. 586-589).
238. Yamada, Y., Shinkawa, K., Kobayashi, M., Nishimura, M., Nemoto, M., Tsukada, E., Ota, M., Nemoto, K. & Arai, T. (2021). Tablet-based automatic assessment for early detection of Alzheimer's disease using speech responses to daily life questions. *Frontiers in Digital Health*, *3*, 653904.
239. Yancheva, M., Fraser, K. C., & Rudzicz, F. (2015). Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies* (pp. 134-139).
240. Yeung, A., Iaboni, A., Rochon, E., Lavoie, M., Santiago, C., Yancheva, M., Novikova, J., Xu, M., Robin, J., Kaufman, L.D., & Mostafa, F. (2021). Correlating natural language processing and automated speech analysis with clinician assessment to quantify speech-language changes in mild cognitive impairment and Alzheimer's dementia. *Alzheimer's research & therapy*, *13*(1), 109.
241. Zimmerer, V. C., Wibrow, M., & Varley, R. A. (2016). Formulaic language in people with probable Alzheimer's disease: A frequency-based approach. *Journal of Alzheimer's Disease*, *53*(3), 1145-1160.