

A Multidisciplinary Investigation of Conversation and Disfluencies in Cognitive Decline

Megan Thomas



Supervisors: Dr. Traci Walker & Prof. Heidi Christensen

The University of Sheffield
School of Allied Health Professions, Nursing, and Midwifery
Division of Human Communication Science

A report submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

December 2024

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations which are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

I hereby declare that this dissertation is of my own work, except where I specifically referred to the works done by the other authors in the text. The contents of the study are original and have not been submitted for any other awards, qualifications, or degrees in universities. Parts of the findings of this study have already been published as journal or conference papers.

Megan Thomas

To Nita, John, and Milly. I wish you could have seen this.

Acknowledgements

In the first instance I must say the biggest of thank yous to my supervisors Traci and Heidi. Heidi- I wouldn't have had the opportunity to complete this work without your support, your kindness, and your belief in me. Thank you for not letting me give up. Traci- thank you for bearing with me and helping me find direction when I was lost, I doubt anyone else could have helped me to the finish line.

Next, my utmost thanks to Rob Gaizauskas and Thomas Hain for giving me the chance to join the CDT. I have experienced things I never thought I would have the chance to experience. Stu and Lizzie, you are lifesavers and your endless patience in answering my (many) stupid questions will never be forgotten. Thank you for making me feel so welcome.

To the SpandH Health members, Heidi's research team, and the CognoSpeak group, thank you for the countless discussions, ideas, and help! I am particularly thankful to Bahman Mirheidari, Nathan Pevy, Dan Blackburn, Ronan O'Malley, and Hend Elghazaly.

To Sam Hollands, my collaborator and conference partner, thank you for your wisdom and knowledge, not just of everything speech technology-based but also the best spots for cheap beer. I still owe you a Blue Moon!

To Mum and Dad, for helping me believe I could do this and calming me when no-one else could. Sorry for all of the tears!

To Katie and Mathilde, my sisters and my truest friends. Look at us!! Thank you for embracing me as your eternal third wheel. You are the purest souls and I am so grateful for everything that you have done for me.

To Seb and Kuba. Whilst submitting this thesis is pretty cool, you are by FAR the best things to have come out of this journey. Thank you for being my best friends, for the countless laughs, and for lifting me up in my darkest moments.

To Tom; for being an amazing cheerleader, for keeping me sane by organising countless Sherlocks and puzz clubs, and for your invaluable LaTeX help! Thank you for brightening up all those days in the office.

To Robbie, Tas, Sylvie, and Tymon, thank for all the laughs and the dinner parties. You're my Sheffield family, and I don't know how I got so lucky.

To Le Sparky-Spark, for all the wine!

And finally to Ziggy, for being by my side.



Figure 1: Ziggy: An accurate portrayal of how the author felt after finishing this thesis.

This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]. We acknowledge IT Services at The University of Sheffield for the provision of the High Performance Computing Service. This work was also supported by Apple and I owe a big thanks to Panayiotis Georgiou for his helpful advice.

Abstract

Cognitive Decline (CD) encompasses a spectrum of conditions affecting millions globally, manifesting in cognitive impairments such as memory and language deficits. Neurodegenerative Dementias (NDs), including Alzheimer's Dementia (AD), represent a group of degenerative disorders contributing to progressive CD. The early stages of CD often exhibit language disturbances, and research indicates that early diagnosis can improve patient outcomes. Speech has emerged as a prominent, non-invasive biomarker for CD assessment, offering potential insights into disease progression. Studies investigating how speech is affected by CD have frequently reported that, as cognition decreases, the presence of disfluencies such as unfilled pauses increases. This thesis explores the diagnostic utility of disfluency analysis, as well as investigating which tasks may elicit the most useful speech for analysis.

In CD detection, advancements in machine learning have led to the development of Automatic Cognitive Decline Classification (ACDC) systems, which demonstrate remarkable accuracy in distinguishing dementia patients from healthy controls based on speech samples. However, ACDC methodologies often struggle to generalise across diverse demographics and lack transparency in their classification rationale. This thesis presents evidence that integrating disfluency features into ACDC systems enhances classification accuracy and addresses issues of generalisation and transparency.

Additionally, Conversation Analysis (CA) has been employed to develop conversational profiles that could assist doctors in differentiating between patients with Neurodegenerative Dementia (ND) and those with Functional Memory Disorder (FMD), a non-neurodegenerative psychological condition. This thesis further investigates whether CA can be utilised to create conversational profiles that help differentiate between ND and Mild Cognitive Impairment (MCI), an early stage of CD.

List of Acronyms and Abbreviations

CA Conversation Analysis

AD Alzheimer's Dementia

CD Cognitive Decline

FMD Functional Memory Disorder

MCI Mild Cognitive Impairment

ND Neurodegenerative Dementia

NN Neural Network

ACDC Automatic Cognitive Decline Classification

TOFFA Taxonomy of Fluency for Forensic Analysis

DisCo Disfluencies in Cognition

WHO World Health Organisation

LBD Lewy Body Dementia

PDD Parkinson's Disease Dementia

VaD Vascular Dementia

MRI Magnetic Resonance Imaging

FTD Frontotemporal Dementia

GP General Practitioner

CSF Cerebrospinal Fluid

GDS Global Deterioration Scale

MMSE Mini Mental State Examination

MoCA Montreal Cognitive Assessment

PVF Phonemic Verbal Fluency

SVF Semantic Verbal Fluency

TCU Turn Constructional Unit

TRP Transition Relevance Place

IVA Intelligent Virtual Agent

RIAS Roter Interaction Analysis System

ASR Automatic Speech Recognition

LM Language Model

LLM Large Language Model

NNLM Neural Network-based Language Model

AM Acoustic Model

MFCC Mel Frequency Cepstral Coefficient

LLD Low-level Descriptor

F0 Fundamental Frequency

ComParE Computational Paralinguistics Evaluation

GeMAPS Geneva Minimalistic Acoustic Parameter Set

eGeMAPS Extended Geneva Minimalistic Acoustic Parameter Set

PoS Part of Speech

TPT Total Phonation Time

TLT Total Locution Time

TTR Type to Token Ratio

KNN k-Nearest Neighbour

DNN Deep Neural Network

SVM Support Vector Machine

ROC Receiver Operating Characteristic

AP Accompanying Person

HC Healthy Control

pwMCI people with Mild Cognitive Impairment

pwFMD people with Functional Memory Disorder

pwND people with Neurodegenerative Dementia

pwAD people with Alzheimer's Dementia

AI Artificial Intelligence

MCID Minimal Clinically Important Difference

VAD Voice Activity Detection

ML Machine Learning

Contents

Contents	x
List of Figures	xiv
List of Tables	xv
1 Introduction	1
1.1 Thesis Overview	2
1.2 Motivation	3
1.3 Thesis Contributions	5
1.4 Publications	7
1.5 Thesis Structure	7
2 Fundamentals	8
2.1 Introduction	10
2.2 Cognitive Decline	11
2.2.1 Dementia	13
2.2.2 Alzheimer’s Dementia	15
2.2.3 Mild Cognitive Impairment	18
2.2.4 Functional Memory Disorder	21
2.2.5 Effects of Cognitive Decline on Language and Communication	22
2.3 Language and Memory Tests for Cognitive Decline	24
2.3.1 Mini Mental State Examination	25
2.3.2 The Montreal Cognitive Assessment	25
2.3.3 Picture Description Tasks	26
2.3.4 Verbal Fluency Tests	26
2.4 Disfluencies	27
2.4.1 A Brief Introduction to Disfluency Categories	28
2.4.2 Disfluencies in the Healthy Ageing Population	30

2.4.3	Disfluencies in Cognitive Decline	31
2.4.4	Disfluencies in Human-Computer Interaction	31
2.5	Automatic Cognitive Decline Classification	32
2.5.1	ACDC Systems	33
2.5.2	System Evaluation	45
2.5.3	CognoSpeak and the Intelligent Virtual Agent	47
2.6	Conversation Analysis	49
2.6.1	Turn-Taking	50
2.6.2	Adjacency Pairs	51
2.6.3	Repairs and Trouble Sources	51
2.6.4	Preference Structure	52
2.6.5	Analysis of Medical Dialogues	53
2.6.6	Conversation Analysis of Human-Computer Interactions	55
2.6.7	Disfluencies in Conversation Analysis	56
2.7	Summary	56
3	First Manual Disfluency Analysis	58
3.1	Introduction	60
3.2	Background	60
3.3	Methodology	68
3.3.1	The Disfluencies in Cognition Schema	69
3.3.2	Data	75
3.3.3	Disfluency Collection Process	82
3.3.4	Statistical Methods	83
3.4	Results	84
3.4.1	Total Disfluency Rates	84
3.4.2	Number of Fluent Words	84
3.4.3	Unfilled Pauses	85
3.4.4	Filled Pauses	87
3.4.5	Pause-to-Speech Ratio	89
3.4.6	Repetitions	90
3.4.7	Prolongations	90
3.4.8	Speech Errors	93
3.4.9	Summary	94
3.5	Discussion	95
3.5.1	Total Disfluency Rates	95
3.5.2	Unfilled Pauses	95
3.5.3	Filled Pauses	96

3.5.4	Pause-to-Speech Ratio	97
3.5.5	Repetitions	98
3.5.6	Prolongations	98
3.6	Conclusions	98
4	Second Manual Disfluency Analysis	100
4.1	Introduction	102
4.2	Background	102
4.2.1	Picture Description Tasks	103
4.2.2	Task Differences	107
4.2.3	Automatic Detection of Disfluencies	108
4.3	Methodology of the Second Manual Disfluency Study	111
4.3.1	Data	111
4.3.2	Adjustments to Disfluency Schema	112
4.4	Results from the Second Manual Disfluency Study	117
4.4.1	Number of Fluent Words	117
4.4.2	Unfilled Pauses	118
4.4.3	Filled Pauses	120
4.4.4	Repetitions	121
4.4.5	Prolongations	122
4.4.6	Speech Errors	122
4.4.7	Discussion of Results from the Second Manual Disfluency Study . .	123
4.5	Comparison of Results from Manual Disfluency Studies One and Two . . .	125
4.5.1	Total Disfluencies	125
4.5.2	Pauses	126
4.5.3	Repetitions	126
4.5.4	Prolongations	127
4.5.5	Speech Errors	127
4.5.6	Discussion	127
4.6	Automatic Cognitive Decline Classification	129
4.6.1	Background	129
4.6.2	Data	130
4.6.3	Methodology	131
4.6.4	Baseline System	132
4.6.5	Results	133
4.6.6	Discussion	133
4.7	Conclusion	134

5	A Conversation Analysis of Human-Avatar Data	135
5.1	Introduction	137
5.2	Conversation Analysis in Medical Interactions	137
5.2.1	Problem Presentation Phase	138
5.2.2	Communicating with People with Neurodegenerative Dementia . . .	145
5.2.3	Using Conversation Analysis to Study Human-Computer Interactions	146
5.2.4	Conversation Analysis as a Diagnostic Tool	148
5.3	Methodology	150
5.3.1	Participants	151
5.3.2	Procedure	152
5.4	Analysis	153
5.4.1	Comparison of Problem Presentation Phases Initiated by a Human Doctor vs an Intelligent Virtual Agent	154
5.4.2	Case Study: Participant 096/0221's Problem Presentations	177
5.5	Conclusion	184
6	Conclusions and Further Work	185
6.1	Conclusions	186
6.1.1	Assessment of Contributions	188
6.2	Limitations and Further Work	192
6.2.1	Disfluency Analysis	193
6.2.2	Automatic Cognitive Decline Classification	194
6.2.3	Conversation Analysis	195
6.3	Concluding Remarks	196
A	Table of all group-based results from the first manual disfluency study	197
B	Individual counts and normalised rates per 100 fluent words for part word, whole word, and phrase repetitions in the interview task.	199
C	Variations in the Intelligent Virtual Agent questions	203
D	Table of all group-based results from the second manual disfluency study	205
E	Transcription conventions for conversation analysis	207
	References	209

List of Figures

2.1	Automatic Cognitive Decline Classification System Architecture	34
2.2	Support Vector Machine Hyperplanes	45
2.3	The Original Intelligent Virtual Agent	48
3.1	Praat Text Grid - Manual Disfluency Analysis I	82
3.2	Number of Fluent Words	85
3.3	Rate of Unfilled Pauses	86
3.4	Average Lengths of Unfilled Pauses	87
3.5	Number of Filled Pauses	88
3.6	Duration of Filled Pauses	89
3.7	Number of Whole-word Repetitions	91
3.8	Number of Prolongations	92
3.9	Length of Prolongations	93
4.1	The Cookie Theft Picture	103
4.2	Differences in Diagnoses Across Memory Services	106
4.3	Disfluency Boundaries	110
4.4	Praat Text Grid - Manual Disfluency Analysis II	115
4.5	Number of Fluent Words	118
4.6	Number of Unfilled Pauses	119
4.7	Average Length of Unfilled Pauses	120
4.8	Number of Filled Pauses	120
4.9	Average Length of Filled Pauses	121
4.10	Number of Prolongations	122
4.11	Length of Prolongations	122
5.1	The Structure of Primary Care Visits	138
5.2	Categorisation of Patients' Responses	158

List of Tables

2.1	The Four Main Types of Neurodegenerative Dementia	17
2.2	The Global Deterioration Scale	19
2.3	The Roter Interaction Analysis System vs. Conversation Analysis	54
3.1	The DisCo Taxonomy.	72
3.2	Disfluency Schema Comparison.	76
3.3	Participant Information - Disfluency Analysis I.	78
3.4	Total Disfluency Rates	84
3.5	Median Number of Filled Pauses	88
3.6	Pause-to-Speech Ratios	90
3.7	Median Number of Filled Pauses	91
3.8	Speech Errors	93
4.1	Participant Information (Manual Disfluency Analysis II	112
4.2	The DisCo2.0 Taxonomy	116
4.3	Number of Repetitions	122
4.4	Number of Speech Errors	123
4.5	Total Disfluency Rates across Tasks and Cognitive Groups	125
4.6	Pause Differences across Tasks and Cognitive Groups	126
4.7	Repetition Rates across Tasks and Cognitive Groups	126
4.8	Summary of Prolongations per 100 Fluent Words	127
4.9	Speech Error Rates across Tasks and Cognitive Groups	128
4.10	Linguistic Features	132
4.11	Automatic Classification Results	133
5.1	Participant Information for the Conversation Analysis.	153
5.2	Frequently Used Transcription Symbols.	153
5.3	Breakdown of Response Types According to Dataset	158
5.4	Breakdown of Response Types According to Diagnostic Group	159

5.5	Extract 1 - Doctor's Questions	178
5.6	Extract 1 - Patient's Responses	179
5.7	Extract 3 - Patient's Responses	180
5.8	Extract 4 - Patient's Responses	181
5.9	Extract 5 - Patient's Responses	181
5.10	Extract 6 - Patient's Responses	182

Chapter 1

Introduction

Contents

1.1	Thesis Overview	2
1.2	Motivation	3
1.3	Thesis Contributions	5
1.4	Publications	7
1.5	Thesis Structure	7

1.1 Thesis Overview

This thesis employs three distinct methodologies to investigate speech from people diagnosed with different levels of Cognitive Decline (CD), and examines the potential role of technology in the diagnostic pathway in the U.K. The first study employs a quantitative analysis of speech to investigate the presence of disfluencies in individuals with varying stages of CD. The speech analysed is directed at an intelligent virtual agent rather than a human doctor, and was collected as part of the larger CognoSpeak project (introduced in Section 2.5.3 of this thesis). Alongside healthy controls, the study examines people diagnosed with three conditions affecting cognition and language: functional memory disorder, mild cognitive impairment, and neurodegenerative dementia. This study is motivated by two primary considerations. Firstly, having data from healthy controls allows comparison with cognitively impaired groups and provides insight into disfluency rates in healthy aging adults interacting with a virtual system. This enables comparison with previously reported disfluency rates and suggests that the previous claim of lower disfluency rates in computer-directed speech may not hold in conversational settings. Secondly, by including participants at different stages of CD, we are able to investigate whether disfluency information is useful for distinguishing among these conditions — a task known to be challenging for clinicians using traditional diagnostic tools. This study addresses our first and second research questions:

RQ 1: How do the frequency and duration of speech disfluencies differ when participants engage in an interview-style task with a digital avatar in a simulated medical interview scenario, compared to similar interviews conducted with human clinicians?

RQ 2: Can an analysis of speech disfluencies be used to discriminate between different levels of cognitive decline?

After demonstrating that disfluency information *can* be used to discriminate between different levels of CD, we shift our focus to another task commonly used to assist in diagnosing CD: a picture description task. Given the elevated rates of disfluencies ob-

served in the interview-style task, our third research question aims to investigate potential differences across tasks when analysing disfluencies:

RQ 3: How do patterns of disfluency vary between an interview-style task and a picture description task?

We subsequently adopt a Machine Learning (ML) approach, utilising the insights gained from the manual disfluency studies to address our fourth research question:

RQ 4: Can disfluency information improve the accuracy of an automatic cognitive decline classification system?

Our results from a proof-of-concept study indicate that incorporating disfluency information in this way can enhance the accuracy of Automatic Cognitive Decline Classification (ACDC) systems. Moreover, disfluency information is more interpretable than many traditional features typically used for such tasks. Finally, we use Conversation Analysis (CA) as a methodology to conduct an in-depth qualitative assessment of the differences and similarities in how patients describe their memory issues to either a human doctor or an intelligent virtual agent, addressing our final research question:

RQ 5: How do patients construct their problem presentation phases in a medical interview with a human doctor versus a digital avatar?

By examining the distinct conversational actions used when speaking to a computer rather than a human, we aim to understand how these systems could be designed to elicit the maximum amount of useful information from patients. Results from this analysis indicate that people interacting with a computer in this context display more pausing behaviours, which contradicts previous research suggesting that humans exhibit fewer disfluencies when interacting with a machine.

1.2 Motivation

The prevalence of CD is increasing alongside the increase in the number of people aged 65 and above. This increase is set to reach over 10 million people in the U.K alone over the

next 40 years, with the category of people aged 80 and above being the fastest growing subset, which is set to double in number [Centre for Better Ageing, 2020]. Dementia, a range of syndromes and diseases, is especially pervasive in the ageing population. Dementia affects a person's cognitive functioning in numerous different domains including memory, attention, language, and orientation. This can severely impact a person's daily living, and people with more severe CD require frequent help from professional carers or family members, placing considerable strain on both caregivers and healthcare systems worldwide.

Accurately detecting and diagnosing CD is a particularly difficult task due to its heterogeneity and the countless different potential causes. Even more difficult is the task of differentiating between different *levels* of CD, where differences are often subtle and vary greatly from person to person. Recently there has been a growing interest in using the analysis of speech as a potential solution to these challenges. Speech is complex and can reflect various cognitive processes including language comprehension and production, and executive functioning skills. Researchers have been exploring the use of speech as a non-invasive and cost effective tool for assessing the presence and severity of cognitive impairment. A range of different approaches have been investigated for this purpose. Qualitative analyses can help researchers understand the underlying cognitive processes and linguistic patterns associated with CD. This can offer valuable insights into early detection of CD and intervention strategies. Quantitative (including ML) based approaches can help us understand the statistical patterns and computational features that are indicative of cognitive decline, often providing insight that humans alone would be unable to access.

The advancement of ACDC systems marks a promising field of research. These systems use Artificial Intelligence (AI) and ML algorithms to analyse speech and quickly extract meaningful information which can be used to indicate levels of cognitive ability in many different areas, including speech and language. By automating the process of speech analysis these systems could prove to be an efficient method of aiding doctors and clinicians in diagnostic tasks. There are numerous benefits of using ACDC systems. There is evidence

that such systems can help detect CD at a much earlier stage than current diagnostic tests, and early detection has been proven to improve the outcomes of people with CD, allowing them to make earlier lifestyle adjustments or start pharmacological treatments sooner. These systems are also cheaper and more scalable; theoretically if a system can be developed that works on a broad range of populations then it can be used on an infinite number of people. However, a combination of both qualitative and quantitative approaches will provide the most thorough analysis and result in a broad range of data. These systems have the potential to revolutionise the way CD is diagnosed and managed which will ultimately enhance the lives and wellbeing of people affected by CD.

1.3 Thesis Contributions

1. The DisCo Taxonomy of Disfluency.

As detailed throughout this thesis, there are numerous different ways of measuring and classifying speech disfluencies. This makes it difficult to directly compare results from different disfluency studies, as such work is frequently not explicit in how exactly the researchers are classifying each disfluency. This is particularly prevalent across fields that are not directly related to linguistics or speech pathology, such as computer science. In order to address this, this thesis proposes a novel way of classifying and measuring disfluencies in speech. The Disfluencies in Cognition (DisCo) schema is specifically designed to uncover various different types of disfluency in speech affected by different levels of cognitive decline, but at a level that facilitates a (relatively) straightforward approach to automating the process of analysing disfluencies. The ability to automatically perform this analysis means that disfluency information can be included in automatic cognitive decline detection systems.

2. Accuracy Improvements in Automatic Cognitive Decline Classification Systems.

This thesis demonstrates that disfluency information can be a valuable addition to ACDC systems. Research into such systems is becoming increasingly popular as the need for quick and accurate cognitive decline diagnosis increases. ACDC systems need

to be able to generalise to diverse populations, and in order for them to be used in the most effective way they should be able to be understood by doctors. Disfluency features provide an interpretable measure of different levels of cognitive decline, and the *DisCo* schema was created in such a way that enables the analysis of disfluencies across a broad range of accents and dialects of English. Because disfluencies have predictable language-specific patterns, the *DisCo* schema could be easily modified for use across a range of different languages.

3. Classification of Problem Presentation Phases.

Researchers employing conversation analysis as a method for investigating medical interviews have predominantly focussed on two main areas; how doctors design their questions and how patients design their responses. In terms of patient responses, research has found that these vary depending on how much prior knowledge of their condition they assume the doctor to have. Much of this work has focused specifically on check-ups or GP consultations, where the range of symptoms patients can discuss is very broad. Our work specifically investigates medical interviews in memory clinics, where we suggest that there is an underlying constraint that patients are aware of which suggests they should be primarily discussing their memory symptoms. We observe that when patients are asked about their memory problems there are two main types of response. These are dependent on whether the patient accepts their symptoms as memory related, or if they deny this link (or indeed that they are experiencing any memory-related symptoms at all). In addition, we observe three different approaches to discussing memory issues from the patients that fall into the “acceptee” group. The first group of participants are accepting of their memory concerns and are able to give specific examples of when their memory has let them down. In addition, these patients describe these symptoms in terms of the emotional distress they are causing. The second group does not attempt to minimise or deny their memory problems, but does not offer specific examples of their memory failing. Rather, these patients talk in very general terms with vague language. Whilst these patients signal that they are aware of their memory issues by naming them, the lack of detail suggests that they don’t feel these symptoms are severe enough to warrant investigation. Patients belonging to the

final group of acceptees can identify that they have some memory related issues, but offer no elaboration at all. These responses are short and lack even general descriptions of what exactly the patients have been experiencing.

1.4 Publications

1. **M. Thomas**, N. Pevy, and T. Walker (2023). Disfluencies in Cognitive Decline: An Investigative Study. *Proceedings of the biennial symposium of the International Clinical Phonetics and Linguistics Association (ICPLA 2023)*, University of Salzburg, 4-7 July 2023.
2. **M. Thomas**, S. Hollands, D. Blackburn, and H. Christensen (2023). Towards Disfluency Features for Speech Technology Based Automatic Dementia Classification. *Proceedings of the 20th International Congress of the Phonetic Sciences (ICPhS 2023)*, Prague Congress Centre, Czech Republic, 7-11 August 2023.

1.5 Thesis Structure

This thesis presents a multidisciplinary approach to the analysis of speech for the purpose of identifying cognitive decline. Each analysis chapter in this thesis uses a different methodology and as such, a broad range of background knowledge about each methodology is required. This thesis starts with a fundamentals chapter which lays out the general background information required to understand the basics of the three methodologies used in this thesis. Then, each analysis chapter begins with a more specific overview of the methodology relevant to the particular analysis. This ensures that the most important information for each analysis is presented alongside the analysis, in an attempt to avoid confusion among the three different methodologies.

Chapter 2

Fundamentals

Contents

2.1	Introduction	10
2.2	Cognitive Decline	11
2.2.1	Dementia	13
2.2.2	Alzheimer's Dementia	15
2.2.3	Mild Cognitive Impairment	18
2.2.4	Functional Memory Disorder	21
2.2.5	Effects of Cognitive Decline on Language and Communication	22
2.3	Language and Memory Tests for Cognitive Decline	24
2.3.1	Mini Mental State Examination	25
2.3.2	The Montreal Cognitive Assessment	25
2.3.3	Picture Description Tasks	26
2.3.4	Verbal Fluency Tests	26
2.4	Disfluencies	27
2.4.1	A Brief Introduction to Disfluency Categories	28
2.4.2	Disfluencies in the Healthy Ageing Population	30
2.4.3	Disfluencies in Cognitive Decline	31

2.4.4	Disfluencies in Human-Computer Interaction	31
2.5	Automatic Cognitive Decline Classification	32
2.5.1	ACDC Systems	33
2.5.2	System Evaluation	45
2.5.3	CognoSpeak and the Intelligent Virtual Agent	47
2.6	Conversation Analysis	49
2.6.1	Turn-Taking	50
2.6.2	Adjacency Pairs	51
2.6.3	Repairs and Trouble Sources	51
2.6.4	Preference Structure	52
2.6.5	Analysis of Medical Dialogues	53
2.6.6	Conversation Analysis of Human-Computer Interactions	55
2.6.7	Disfluencies in Conversation Analysis	56
2.7	Summary	56

2.1 Introduction

This chapter presents the general background knowledge of the main concepts and methodologies discussed in this thesis. The first section (2.2) provides an overview of the different types of Cognitive Decline (CD) that are investigated in this thesis, and details the symptoms of each. This section then discusses how the different levels of CD can affect speech and language, and presents an overview of the different language tests currently used by doctors to help ascertain levels of cognitive decline. Section 2.3 introduces some of the most commonly used tests for assessing levels of CD. Particular attention is paid to the tests that feature in the data analysed throughout this thesis. Section 2.4 then turns the focus to disfluencies. A brief introduction to the types of disfluencies that are paid particular attention to in this thesis is followed by an overview of disfluencies in cognitively healthy older adults compared to those with CD. Section 2.5 presents all the background knowledge required to understand the Artificial Intelligence (AI) and Machine Learning (ML) aspects of this thesis. We start with a general overview of Automatic Cognitive Decline Classification (ACDC) systems, and describe the main components involved in these. We then present the most commonly used metrics to assess the performance of ACDC systems. The next section (2.5.3) introduces the CognoSpeak system, an intelligent virtual agent designed to simulate a medical interview with a doctor in a memory clinic. This system provided the bulk of data for analysis in this thesis, and serves as motivation for investigating how disfluencies and disfluency information may be able to improve the performance of such systems. The final section (2.6) provides a general introduction to Conversation Analysis (CA) and covers the foundational principles of talk-in-interaction covering turn-taking, sequence organisation, repairs and trouble sources, and preference structure. After this we discuss another commonly used method for analysing medical discourse (the Roter Interaction Analysis System), and address why this thesis favours the CA approach.

2.2 Cognitive Decline

The term Cognitive Decline (CD) describes a decrease in different cognitive functions such as memory, visual-spatial processing, and executive functioning. CD can be a non-pathological effect of normal bodily ageing, the symptoms of which vary greatly from person to person [Deary et al., 2009]. For example, individual differences such as education level or brain physiology mean that no two cognitively healthy ageing people are the same. In fact, studies have shown that normal ageing is primarily person-specific [Wilson et al., 2002]. There are still large research gaps when considering the relationship between age and cognition in the absence of diseases such as dementia. As Salthouse [2019] states, without a thorough understanding of non-pathological cognitive ageing it is difficult to identify the earliest stages of pathological CD and therefore more difficult to identify the best time at which to start interventions intended to slow future decline. This in turn minimises the effects of any potential treatments, and contributes to the many difficulties faced by people living with CD.

However, it is important to understand the differences between language deficits as a result of CD, compared to those caused by normal cognitive ageing. There is no one model of normal cognitive ageing when considering language processing and production. One of the most specific models is the Transmission Deficit hypothesis Burke et al. [2000]. This hypothesis describes how the effect of normal ageing on language is asymmetric, where semantic representations and retrieval are relatively well preserved throughout the process of ageing compared to phonological and orthographic representations [Abrams and Farrell, 2011]. Broadly speaking, there are two main categories of theories of cognitive ageing; *information-universal* and *information-specific* [Burke and MacKay, 1997]. Information-universal theories suggest that a slowing of memory and cognitive functioning occurs with ageing, regardless of the kind of task or information that needs to be processed. Individual theories that fall under this banner include the “sensory decline” theory [Lindenberger and Baltes, 1994] which suggests that sensory functioning (specifically visual and auditory acuity) is an important correlate of cognitive functioning in old and very old age, or the “general slowing” theory posited by Salthouse [1996] which

suggests that ageing causes a general slowing of all types of processing.

Conversely, information-specific theories of cognitive ageing suggest that the effects of ageing *do* depend on the task or information, and the respective parts of the brain responsible for the processing of that information. Theories in this group explain the patterns of deficiency seen in patients with similar lesions or damage to specific parts of the brain (such as age-related lesions in the hippocampus leading to memory impairment but a preservation of general cognitive functions; as described in [Moscovitch and Winocur \[1992\]](#)), and suggest that the effects of ageing are not always balanced, with people often experiencing deficiencies with a particular cognitive function more than others.

The present work suggests that language problems could be used to differentiate between the general slowing of cognitive processes we would expect to see given a person's age, and cognitive decline caused by other factors. For example, it is well reported that language deficits in cognitively healthy ageing adults are often asymmetric, with production tasks being more severely affected by age than comprehension tasks (see [Abrams and Farrell \[2011\]](#) for a thorough overview of language in normal ageing). If, then, a patient presents to a memory clinic with particularly impaired speech and severe aphasia, it follows that something *other* than healthy ageing is having an effect on the patient. We take this assumption one step further and suggest that different aspects of pathological speech could be used to differentiate between different levels of [CD](#). Whilst differences in pathological language production and comprehension have been researched across different disorders, there is little work supporting any differences in disfluency frequency according to differing levels of pathological [CD](#). This thesis aims to address this gap in literature by investigating whether the presence of speech disfluencies could be an indicator of severity or type of [CD](#).

Whilst non-pathological cognitive ageing is under-researched, a much larger body of work has focused on [CD](#) with causes outside of standard ageing. The remainder of this thesis focuses on different levels of pathological [CD](#). The following section focuses on dementia and its various associated syndromes.

2.2.1 Dementia

Dementia is an umbrella term referring to a group of syndromes or diseases characterised by a loss of cognitive abilities severe enough to impact day-to-day life [Geldmacher and Whitehouse, 1996]. The World Health Organisation (WHO) estimates that more than 55 million people worldwide currently live with dementia, and every year there are almost 10 million new cases [WHO, 2023]. Focusing specifically on the U.K, dementia was the leading cause of death in 2022 and has been the leading cause of death for women since 2011, according to Alzheimer’s Research U.K [2023]. Surprisingly, dementia was still the biggest killer of women in the U.K throughout the recent Covid 19 pandemic.

The term dementia is typically used when talking about neurodegenerative dementias, although nondegenerative dementias do exist. Nondegenerative dementias typically occur at a younger age, and can be caused by numerous different factors from nutritional problems to tumours (for a full overview of known causes of nondegenerative dementia, please see Ghosh [2010]). Importantly, many types of nondegenerative dementia are treatable or preventable, unlike most types of neurodegenerative dementia.

Neurodegenerative Dementia (ND) is characterised by a gradual worsening of symptoms over time. Neurodegeneration has two pathological hallmarks; deposits of proteins into the brain tissue and/or cell death [Matej et al., 2019]. Other examples of neurodegenerative diseases include Parkinson’s Disease and Motor Neurone Disease. There are four main types of ND. The most common, Alzheimer’s Dementia (AD), is discussed in detail in the following section. The remaining three are discussed below. Although current research agrees in these four main types of ND, there is increasingly more work suggesting that these groups may not be so clear cut, and that there is potentially more overlap between the groups than had previously been thought [Matej et al., 2019]. Further complicating the matter of diagnosing different kinds of ND is the fact that each of the groups discussed below also have their own subcategories (for example, Frontotemporal Dementia (FTD) can be subdivided into a behavioural variant, a semantic variant, a non-fluent variant, etc). Table 2.1 at the end of this subsection presents a condensed comparison of the four most common types of ND.

2.2.1.1 Lewy Body Dementia (and Parkinson's Dementia)

Lewy Body Dementia (LBD) is a dementia caused by Lewy Body disease. This disease causes small clumps of proteins (the Lewy Bodies) to form in the brain. Lewy Body disease is also the cause of Parkinson's disease. LBD is diagnosed when the symptoms of dementia arise either before or alongside the onset of motor symptoms caused by Parkinson's Disease. If symptoms of dementia arise one year or more after a Parkinson's Disease diagnosis, the condition is diagnosed as Parkinson's Disease Dementia (PDD). Both LBD and PDD share the same pathophysiology [Walker et al., 2015], and both dementias are so similar that they are commonly referred to simply as LBD. Approximately 10-15% of all dementia cases in the U.K are diagnosed as LBD [Alzheimer's Research U.K, 2023].

2.2.1.2 Vascular Dementia

Vascular Dementia (VaD) is a type of ND responsible for around 15% of all dementia cases worldwide. Although many symptoms of VaD overlap with other dementias, the cognitive changes caused by VaD are more variable and depend on which neural substrates are affected by the disease [O'Brien and Thomas, 2015]. It is widely recognised that a diagnosis of VaD is dependent on the presence of white matter lesions in the brain, visible through Magnetic Resonance Imaging (MRI) scans [Iadecola, 2013]. However, there may be several causes of VaD including strokes and small vessel disease of the brain [ARUK, 2023], or other types of cerebrovascular disease [Jellinger and Attems, 2010].

2.2.1.3 Frontotemporal Dementia

FTD is a slow-progressing type of ND characterised by prominent personality change and expressive language problems [Welsh-Bohmer and Warren, 2006]. In terms of communication deficits, primary progressive aphasia is common but dysarthria is rarely observed [Horner et al., 2007]. FTD is the leading kind of early-onset dementia (affecting people under the age of 65), but is often misdiagnosed as a psychiatric disorder due to the personality and behavioural changes it can cause [Bang et al., 2015].

2.2.2 Alzheimer's Dementia

As previously mentioned, Alzheimer's Dementia (AD) is the most common type of ND and is frequently used in colloquial language as an umbrella term for all kinds of dementia, despite having many differentiating features. AD accounts for around 60% of all dementia cases in the U.K. While it is the most prevalent kind of dementia, AD tends to have a longer survival time than other kinds of ND [Alzheimer's Research U.K, 2023]. Such high instances of the disease results in huge economic burdens. Nationally, AD and related care costs £34.7 billion annually. Most people with AD in the U.K have to fund their own care, resulting in a typical annual cost of £32,250 per person [Alzheimer's Society, 2021]. This is particularly concerning given the average annual income of pensioners in the U.K is only £12,000 [Office for National Statistics, 2022]. However, the cost of AD is not purely felt by the patient. Families of AD patients often face indirect costs such as the loss of earnings whilst caring for their loved ones [Castro et al., 2010]. There are also countless emotional costs of living with or caring for someone with AD. Carers often face distressing scenarios arising from the behavioural changes resulting from the disease [Burns, 2000] alongside the usual difficulties that come with witnessing a loved one in ill health.

2.2.2.1 Cognitive Symptoms

The most prevalent cognitive symptoms of AD are those concerning problems with memory. At earlier stages of the disease, patients may experience things such as misplacing items or missing appointments. As the disease progresses these symptoms will become more severe, and can result in patients forgetting key events from their past or close family members. AD does not present the same in all patients. Some people may experience deficits in executive functioning before the onset of memory loss, and a small number of patients may not experience memory loss at all [López and DeKosky, 2008]. Anosognosia (a lack of awareness of one's own cognitive deficits) has been found to worsen along with the progression of AD [Clare, 2004].

Visuospatial difficulties are also commonly experienced by people with AD. This refers to the inability to picture parts in space, and difficulty in performing mental operations

on spatial concepts. People with visuospatial deficits may not be able to detect colour and motion, and will struggle to complete basic drawing tasks [Salimi et al., 2018].

Language difficulties complete the list of the most common cognitive symptoms of AD, and these are covered in more detail in Section 2.2.5.

2.2.2.2 Non-Cognitive Symptoms

Behavioural symptoms of AD include things such as increased violence, overactivity, and insomnia. A study by Margallo-Lana et al. [2001] found that up to 60% of people with AD living in care facilities exhibit some form of aggressive behaviour.

Psychological symptoms of AD include hallucinations, delusions, depression, and fearfulness. In a study of more than 2,200 AD patients, it was found that delusions occurred in up to 73% of all patients. Hallucinations were observed in up to 67% of patients, and up to 30% of patients had misidentified people or places [Molchan et al., 1995]. Although reported rates of depression in people with AD vary, depression is an important symptom especially in terms of patient care, as people who have depression and AD but are not experiencing anosognosia are often left in fear and grief when facing the fact that they have a progressive degenerative disease [Borson and Raskind, 1997].

2.2.2.3 Current Diagnostic Pathway

Recent figures from an audit of memory assessment services [Royal College of Psychiatrists, 2022] has found that the average time to a diagnosis of AD is 2.8 years. This includes months of worry before the patient decides to seek help from a GP, and a potential waiting time of 17 weeks to be referred to a specialist memory service. This situation is worse for people diagnosed with early onset AD (those aged under the age of 65), where the average waiting time is 4.4 years.

2.2.2.4 Biomarkers

At present, the most reliable way of diagnosing AD is by looking for in-vivo biomarkers such as beta-amyloid plaques in brain tissue [Matej et al., 2019]. Biomarkers for AD are present in Cerebrospinal Fluid (CSF). In order to assess the levels of biomarkers present

in CSF a patient must undergo a lumbar puncture, a procedure involving a needle being inserted into a patient's lower spine to allow the collection of the CSF [Doherty and Forbes, 2014]. It is also possible to investigate neuronal loss [Scheltens et al., 2016] which is visible through neuroimaging. Whilst the analysis of in-vivo biomarkers is currently the most accurate way of diagnosing AD, there are a number of potential issues with these methods. For example, these tests need to be administered by a trained doctor or clinician, and involve the use of expensive medical equipment. This contributes towards the high costs of dementia care. In addition, these tests must take place in a hospital which can cause problems for people who are less mobile. Recently, speech has been investigated for its usefulness as a biomarker for assessing the presence of AD and other kinds of CD (see, for example, Robin et al. [2020], Chakraborty et al. [2020], or Laguarda and Subirana [2021]). Speech is easy to collect compared to in-vivo biomarkers, does not require a doctor to be present during the collection process, and can (theoretically) be collected from any location.

Comparison of the Four Main Types of Neurodegenerative Dementia				
	Alzheimer's Dementia (AD)	Vascular Dementia (VaD)	Lewy Body Dementia (LBD)	Frontotemporal Dementia (FTD)
Onset and Course	insidious onset, likely from 65+, progressive, slow	abrupt onset, fluctuating course	early stages fluctuate between cognitively normal and abnormal, progressive, rapid (1-5 years)	insidious onset, likely before 65, progressive, slow
Profile	memory and cognition deficits, impaired daily life function	memory and cognition deficits, impaired daily life function	fluctuating attention, visual hallucinations, parkinsonism	varies from subtype but executive dysfunction, semantic deficits and aphasia are common
Communication Changes	aphasia is common, semantic system most affected, syntax and phonology affected later on, slow progression to mutism	motor speech disorder, grammar simplification	parkinsonian dysarthric features	primary progressive aphasia
Behavioural Changes	depression, insomnia, incontinence, delusions, agitation	frequent falls, dysarthria, visual deficits	parkinsonian features common	not a major feature of FTD

Table 2.1: Comparison of the four main types of neurodegenerative dementia, adapted from Horner et al. [2007].

2.2.3 Mild Cognitive Impairment

Mild Cognitive Impairment (MCI) is a type of CD that is often thought of as the early stages of dementia such as AD. According to the Diagnostic and Statistics Manual 5, [Sachs-Ericsson and Blazer, 2015], MCI should now be referred to as “Mild Neurocognitive Disorder”. However, as this thesis uses data from 2014-2018 that is labelled as MCI, we retain the name MCI to avoid confusion. The main difference between MCI and dementia is that patients with MCI show a decrease in cognitive functioning that is greater than average for their age and education level, but does not heavily interfere with their day-to-day lives. MCI is generally regarded as the point between normal cognitive ageing and very early CD [Petersen, 2016].

Attention was first brought to MCI (originally termed mild cognitive decline) in 1982 when Reisberg et al. published their Global Deterioration Scale (GDS). Table 2.2 provides an overview of the GDS. This work was important as it was the first dementia scale that clearly differentiated between MCI and other levels of CD.

More recently, the First Key Symposium for Mild Cognitive Impairment held in Stockholm, Sweden, saw a group of experts from a range of different disciplines produce a revised criterion for diagnosing MCI [Winblad et al., 2004]. These guidelines (the Stockholm Criteria) state that a patient can be diagnosed with MCI if:

- The patient is neither cognitively normal nor cognitively demented
- There has been a report of CD either from the patient or from an informant
- There is evidence of CD over time on objective cognitive tasks
- The basic activities of daily living are preserved

There is no consensus as to whether or not MCI always leads to dementia. In instances where this progression does take place, MCI can sometimes be referred to as prodromal dementia. Numerous different studies have unfortunately reached numerous different conclusions to this question. For example, Chertkow et al. [2001] found that even after 10 years since the onset of memory complaints, a quarter of their MCI patients had not converted to AD. Conversely, Morris et al. [2001] found that 100% of their participants

The Global Deterioration Scale	
Level	Clinical Characteristics
1 - No Cognitive Decline	No subjective complaints of memory
2 - Age Associated Memory Impairment	Subjective complaints of memory deficit but no objective evidence of memory deficit in clinical interview, and no objective deficits in employment or social situations
3 - Mild Cognitive Impairment	Objective evidence of memory deficit obtained only through intensive clinical interview, decreased performance in demanding employment and social settings, experiences events such as concentration deficits, forgetting names of new people, word and name finding errors
4 - Mild Dementia	May exhibit some deficit in memory of personal history, inability to perform complex tasks, decreased knowledge of current or recent events, decreased ability to travel and handle finances
5 - Moderate Dementia	Patient can no longer survive without assistance, inability to recall major relevant aspects of their lives such as their home address, frequent disorientation, but are usually able to remember their spouses' and children's names and don't require help with eating or using the bathroom
6 - Moderately Severe Dementia	Unaware of most recent events or momentous occasions in their lives, occasionally forgets spouses' and children's names, generally unaware of their surroundings, may undergo personality changes
7 - Severe Dementia	All verbal abilities will eventually be lost, require assistance with most, if not all, of their daily tasks, eventual loss of basic psychomotor skills

Table 2.2: The Global Deterioration Scale, adapted from [Reisberg et al. \[1982\]](#).

with [MCI](#) had progressed to [AD](#) over a span of 9.5 years. Yet another study [[Gauthier et al., 2006](#)] states that more than half of people diagnosed with [MCI](#) will progress to [AD](#) within five years.

There are numerous potential reasons for this disparity. Firstly, there is no single set of guidelines for aiding in the diagnosis of [MCI](#) [[Chertkow, 2002](#)]. The Stockholm Criteria highlighted above is only one example of several different [MCI](#) criteria, including the Pe-

tersen Criteria [Petersen et al., 2001b] and the Global Deterioration Scale [Reisberg et al., 1982]. It is also important to note that while these criteria exist, they should be thought of as rating scales rather than diagnostic tools [Petersen et al., 1999]. Secondly, there is a high level of interpersonal difference when it comes to the ways in which symptoms may present, and in what order. Petersen et al. [2001a] suggest three different subgroups of MCI which are categorised according to which cognitive functions are affected:

1. Amnesic MCI where only memory is affected
2. Multiple-Domain MCI where impairment is found across multiple cognitive domains but is not severe enough to constitute dementia
3. Single Non-Memory Domain MCI where impairment is found in a single domain that is not memory.

Expanding on this work, Petersen went on to suggest that specific subtypes of MCI may be more likely to progress to specific types of dementia [Petersen, 2003]. At the same time however, there is work to suggest that MCI subtype is a poor predictor of future dementia type [Fischer et al., 2007]. This contradiction serves to prove how closely linked MCI and dementia are, and that misdiagnoses between the two are not uncommon. In fact, some people that were previously diagnosed with possible AD are now being reclassified as in fact having MCI, thanks to advances in diagnostic criteria [McKhann et al., 2011].

Despite the various attempts to standardise the definition of what exactly constitutes MCI, problems still arise when it comes to diagnosing the condition. In their review of MCI clinical trials, Stephan et al. [2013] highlight the lack of consistency in how MCI was diagnosed from study to study. Whilst researchers agree that the earlier CD can be diagnosed the better, a lack of standardisation in diagnosis methods across the board makes this a difficult task. Although there is disagreement surrounding the best criteria for diagnosing MCI, this thesis uses the term to specifically refer to the level of CD represented at level three on the GDS, irrespective of the subgroup classification.

2.2.4 Functional Memory Disorder

Functional Memory Disorder (FMD) is a syndrome that causes many symptoms that overlap with those caused by neurodegenerative dementia. However, FMD is a non-organic, non-progressive psychological disorder that is caused by distress or psychosocial burden [Schmidtke et al., 2008]. FMD (also referred to in the literature as functional cognitive disorder or functional cognitive impairment) results in significant patient distress, and can greatly affect a person’s social life and employment [Pennington et al., 2015]. FMD also impacts a person’s internal consistency, resulting in periods of time where the person can function normally but other periods where a person is severely impaired [Ball et al., 2020]. One defining feature of FMD is the subjectivity of the memory complaints. Pennington et al. [2015] found that patients frequently rated their memory abilities as excessively low on self-reporting scales, resulting in many people with FMD often being misdiagnosed as having early stages of neurodegenerative dementia.

Not much research exists examining potential language and communication-related symptoms of FMD. Elsey et al. [2015] found that in their study of 30 patients (half with ND and half with FMD), the people with FMD interacted more confidently with doctors and clinicians and provided more detailed descriptions of their symptoms compared to patients with ND. 11 of their ND patients were diagnosed with an early stage of dementia, while the remaining four were diagnosed with amnesic MCI, described as being “highly likely to develop into dementia”. Cognitive scores for participants were not reported in this paper, so the comparison was based on the expectation that people with FMD are less severely affected by cognitive decline than those with ND.

A study by Jones et al. in 2016 obtained similar results, and demonstrated that people with FMD were better at following compound questions than those with ND, although they did take more time when responding to questions. A 2019 study from Alexander et al. found that FMD patients’ ability to provide detailed accounts of their symptoms in itself constitutes conversational evidence of their cognitive and memory capacity, which frequently contrasts with their subjective complaints.

The next section of this chapter examines the different language deficits caused by neu-

rodenerative dementia and mild cognitive impairment.

2.2.5 Effects of Cognitive Decline on Language and Communication

Researchers have always linked CD with a decline in linguistic ability. Even in the earliest studies, language deficits were noted as symptoms of cognitive impairment, such as in the seminal documentation of progressive mental deterioration observed in a 51 year old woman which was produced in 1901 and later published in 1911 by Alois Alzheimer [Möller and Graeber, 1998]. Numerous studies report that difficulties with language and speech are amongst the earliest symptoms of dementia [Snowdon et al., 1996; Stanyon et al., 2016; Tang-Wai and Graham, 2008]. Research also shows that as the disease progresses, language difficulties become more severe [Kempler, 2005].

A large bulk of this research has focussed solely on Alzheimer’s Dementia, primarily as it is the most common kind of dementia. Some of the most common language problems found in severe AD are temporal changes such as speech rate and tempo [Forbes-McKay and Venneri, 2005; Jarrold et al., 2014; Meilán et al., 2014], phonemic paraphasia (where words are produced with unintended sounds) [Croot et al., 2000; Wutzler et al., 2013], and word finding difficulties [Kempler and Goral, 2008; Santos et al., 2011; Taler and Phillips, 2008]. There is also work investigating lesser known speech problems associated with severe AD, such as a reduction in syntactic complexity and comprehension [Bickel et al., 2000] and problems with grammaticality [Small et al., 1997].

More recently, some work has investigated different *levels* of cognitive impairment to uncover differences in linguistic ability. Investigations into MCI (such as in Taler and Phillips [2008]), different types of ND (as in Jiskoot et al. [2023] and Klimova and Kuca [2016]), and prodromal stages of dementia (such as in Vincze et al. [2021] and Laske et al. [2015]) make up the majority of this research and support the theory that the worse a person’s dementia is, the more their speech and language abilities will be affected.

The remainder of this subsection describes some of the most commonly reported effects of cognitive decline on speech and language.

2.2.5.1 Aphasia

Aphasia is characterised by symptoms such as having difficulties speaking clearly or understanding speech, trouble remembering words, and trouble naming objects. Aphasia is particularly pervasive in AD, with research from Cummings et al. [1985] finding in their study of 30 AD patients that all patients were aphasic, and that language became more impaired as the severity of AD increased. Anomia, a type of aphasia that causes semantic impairments and affects the production of words such as names and numbers, is frequently observed as a symptom of AD, although the reported number of people with AD who experience anomic aphasia varies between studies [Aronoff et al., 2006]. Research has shown that aphasia can also be a symptom caused by other kinds of dementia, such as FTD [Kirshner, 2014] and LBD [Watanabe et al., 2020]. A study from Forbes et al. [2002] that examined the language differences between people with “minimal AD” (people that scored in the MCI range on neuropsychological tests) and healthy controls found that the people with minimal AD took longer to produce words, and often used incorrect words when describing a picture (for example saying “boy” instead of “girl”, a phenomenon called semantic paraphasia).

2.2.5.2 Reduced Syntactic Complexity

People with CD often exhibit language with a reduced syntactic complexity compared to age-matched healthy adults. Syntactic complexity refers to the range and diversity of sentences present in language. Numerous studies have examined syntactic complexity in patients with AD, including a study by Can and Kuruoglu [2018] which compared sentence construction between people with early onset AD and healthy controls that had been matched for age and education. This study used two different picture description tasks, a random speech task, and a story picture sequencing task and found that the AD patients produce fewer sentences than the healthy controls overall. The healthy controls were also found to use longer sentences that included more conjunctions and compound sentences in the cookie theft picture description task than the AD patients ($p = 0.001$). A smaller study investigating syntactic complexity in eight people with MCI compared to age and gender matched controls found that the patients with MCI produced speech

that was less descriptive than their healthy counterparts, although syntactic complexity did not differ between the two groups [Fleming and Harris, 2008].

2.2.5.3 Slower Speech and Articulation Rates

There is conflicting evidence as to whether or not CD causes slower speech and a lowered articulation rate. Linguistic studies investigating speech and articulation have not found any statistically significant differences in speech and articulation rate between people with AD and healthy controls (see Murray [2010] and Ahmed et al. [2013]). However, studies from speech and language technology that have automatically analysed speech in an attempt to detect early levels of CD have reported that measures of speech and articulation rate can help differentiate between demented and healthy speech when used in combination with other measures. For example, Meilán et al. [2014] found that their participants with AD had a mean speech rate of 2.55 syllables per second, compared to a mean of 3.59 syllables per second for their healthy controls. Another study [Luz et al., 2018] included syllables per minute in their analysis of patient dialogues for dementia detection, finding a lower mean number of syllables per minute for the AD group (168 syllables per minute) compared to the healthy controls (180 syllables per minute).

2.2.5.4 Disfluency

Many aspects of fluency can be affected by CD, such as the frequency and duration of pauses, repetitions, and hesitation phenomena. Disfluencies are further discussed in Section 2.4 of this chapter.

2.3 Language and Memory Tests for Cognitive Decline

Memory and cognitive tests are commonplace in both GP practices and specialised memory clinics. Whilst their diagnostic utility has been proven, there are currently no guidelines in place dictating which of the several tests should be used for specific disease identification, resulting in a number of different assessment scales being proposed [Burns et al.,

2002]. These tests also cannot be used on their own to diagnose **ND**, but would be used in conjunction with a combination of the tests for in-vivo biomarkers, and an overview of the patient's medical history. The following section details the language and memory tests used in this thesis, and discusses the pros and cons of each.

2.3.1 Mini Mental State Examination

First presented by [Folstein et al. \[1975\]](#), the Mini Mental State Examination (**MMSE**) remains one of the most commonly used cognitive tests for the elderly [[Burns et al., 2002](#)]. The **MMSE** takes an average of ten minutes for a patient to complete, which is particularly helpful in cases where patients may have limited concentration. In general, it is best to have the test conducted by someone who is familiar with the **MMSE**. However, administering the test is very simple. Practitioners follow a set of questions and instructions in order, and make judgements as to the patient's performance as they go. The questions in the **MMSE** cover a variety of different topics, from orientation and attention to language and recall. The language section of the test assesses patients on the basis of naming simple items, repeating a sentence, comprehension in the form of following simple instructions, reading, writing, and copying. The **MMSE** has been found to be highly sensitive for moderate to severe levels of cognitive decline, although this sensitivity level decreases at less severe levels of impairment. **MMSE** scores can be affected by age, level of education, and cultural background [[Tombaugh and McIntyre, 1992](#)]. The **MMSE** is scored out of 30, with a score of 25 or more being classed as unimpaired.

2.3.2 The Montreal Cognitive Assessment

The Montreal Cognitive Assessment (**MoCA**) was developed as an assessment tool specifically for **MCI**. Most individuals meeting the criteria for **MCI** score relatively highly on the **MMSE**, making it difficult to differentiate these individuals from people in the healthy range for the elderly using that scale [[Nasreddine et al., 2005](#)]. In comparison to the **MMSE**, which has a sensitivity of around 18% for detecting **MCI**, the **MoCA** is able to achieve a sensitivity of 90% for the same task. In terms of mild **AD**, the **MoCA** achieved

a sensitivity of 100% compared to the *MMSE*'s 78%. Studies such as [Smith et al. \[2007\]](#) have found that the *MoCA* is also helpful in determining which people with an existing diagnosis of *MCI* may progress to dementia at a six-month follow-up exam.

The assessment consists of a 30 point test which can be administered in 10 minutes by anyone who is able to follow the instructions on the website. However, only health professionals should interpret the results [[Hobson, 2015](#)]. A 2013 study found that the *MoCA* was a superior tool to the *MMSE*, but also commented that the most accurate diagnoses were made when these assessment tools were used in conjunction with other tests [[Roalf et al., 2013](#)].

2.3.3 Picture Description Tasks

The general instructions to participants completing a picture description task are to describe a given picture in as much detail as they can. The pictures that participants are presented with differ from study to study, although the Cookie Theft picture description task is perhaps the most well-known (a full overview of which is presented in Chapter 4). In asking participants to describe a picture, doctors and clinicians are eliciting speech that will contain quantifiable measures reflective of speech production abilities from four different perspectives: production, elaboration, and complexity; speech disfluency; consciousness; and the amount of information imparted [[Cooper, 1990](#)]. A comprehensive review of studies using picture description task data from [Mueller et al. \[2018\]](#) found that these tasks are able to accurately distinguish between healthy older adults and those diagnosed with *AD*. Additionally, the review noted that there was a weak correlation between picture description tasks and verbal fluency tasks (p.933). This suggests that there is valuable information to be gained about cognition and speech production through the analysis of connected speech which is not found in lists of singular words.

2.3.4 Verbal Fluency Tests

There are two main categories of verbal fluency tests used to assess levels of cognitive decline: Phonemic Verbal Fluency (*PVF*) tests and Semantic Verbal Fluency (*SVF*) tests.

PVF tests require a participant to produce as many words as they can that begin with a given letter in a set time frame (usually within one minute). **SVF** tests are similar, but require participants to produce words that all belong to a given semantic category, again within a set time frame. These tests are scored according to the number of correct and unique words they can produce for each category [Shao et al., 2014]. Studies have found that people with **AD** score significantly worse on both kinds of fluency tests, although **SVF** tests seem to be the worst affected (see Henry et al. [2004] for an analysis of studies investigating verbal fluency tests and their results). Although recordings of verbal fluency tests are not analysed as part of this thesis, the datasets used as part of this study do also contain such recordings which could be used as part of some further investigations.

2.4 Disfluencies

When discussing research into disfluencies across various different fields, there are numerous different terms that are frequently used interchangeably when they in fact refer to different and distinct phenomena. This is exemplified by the differentiation (or lack thereof) between the terms “disfluency” and “dysfluency”. The former should refer to a broad range of speech disfluencies that are considered normal in everyday conversation, and the latter used in instances of disruptions caused by a speech disorder. In practice, “dysfluency” is usually reserved for talking specifically about stuttering (Koller [1983], Horner and Massey [1983]). However, the two terms are often confused, especially in studies where the study of disfluencies forms a small part of the overall research objective. This thesis uses the term “disfluency” to refer to phenomena that interrupt the flow of speech (the prefix “dis” is key and denotes anything that is not, in this case, fluent). This therefore allows us to include phenomena such as repetitions and repairs in the category of disfluency.

It is important to note that disfluencies are a normal part of speech, particularly unrehearsed, spontaneous speech. Although exact figures differ between studies, for every 100 words of natural speech there can be up to six disfluencies [Fox Tree, 1995]. Research has shown that most of the time, people essentially “tune out” disfluencies in the speech of

their conversational partner [Brennan and Schober, 2001].

This thesis focuses on disfluencies present in speech rather than in text where authors have the opportunity to revise and edit what is being written. When people wish to edit something they have verbalised, this can result in a disfluency. Speech disfluencies may also serve specific functions in conversation. For example, filled pauses (sometimes referred to as hesitations or fillers depending on the field from which the research originates) can be used to signal to a conversational partner that the speaker is formulating their response and not willing to give up the floor of the conversation. It is a signal to the conversational partner that they should be patient, because a response is coming. Other disfluencies do not have such clear-cut functions in conversation but may be representative of the underlying processes of speech production.

This thesis adopts the approach first presented in Eklund [2004] in treating disfluencies as inherently multidisciplinary phenomena, from which different conclusions can be drawn according to how they are viewed. In Chapter 3 we present a novel disfluency categorisation system, designed to be simple enough to perform both manually and automatically, but detailed enough to capture potential differences between “normal” and pathological disfluencies (those caused by CD).

2.4.1 A Brief Introduction to Disfluency Categories

Due to the inherently multidisciplinary nature of disfluencies, there are various different ways of classifying, measuring, and naming the various phenomena referred to as disfluencies in this thesis. This section introduces the main categories of disfluencies investigated in this thesis to solidify the use of specific terminology. A more extensive discussion of these categories, as well as a discussion surrounding the issues associated with the use of inconsistent terminology across different fields of research, is presented in Chapter 3.

2.4.1.1 Pauses

This thesis differentiates between silent and filled pauses, rather than combining the two into one broader “pause” category. Silent pauses are breaks in conversation that interrupt

the flow of speech. This thesis does not consider turn-initial silent pauses as part of this category. Filled pauses appear in similar ways, but contain some vocalisation. They also can mark the beginning of a speaker's turn, or serve to "hold the floor" whilst a speaker is searching for a word, informing their conversational partner that they have not finished talking.

2.4.1.2 Repetitions

Repetitions in this thesis are divided into three different categories depending on which segment is being repeated; part-word repetitions, whole-word repetitions, or phrase repetitions.

2.4.1.3 Prolongations

Prolongations are sub-word segments that are extended in a manner that is not infitting with the rest of the speech, and are not consciously produced for emphatic reasons (for example, emphasis in a phrase such as "*oh my god*" often appears in the form of vowel prolongations). A similar approach is described in [Roberts et al. \[2009\]](#).

2.4.1.4 Speech Errors

Speech errors form a category of disfluencies in this thesis that consist of substitutions, additions, or deletions. This thesis does not differentiate between phonetic or phonological speech errors. This category also includes malapropisms, lexical retrieval errors, and circumlocutions.

2.4.1.5 Repairs

Repairs can be described as a speaker going back to fix a speech error they have made. As discussed further in [Chapter 5](#), repairs are often initiated by the speaker but may also be initiated by a conversational partner who notices an error that the speaker themselves may have overlooked. These errors are often semantic in nature, such as:

"The boy is next to the chai- the table"

2.4.2 Disfluencies in the Healthy Ageing Population

Much of the research into the presence of disfluencies has been based on data collected from young children or young adult participants. As a result, few studies have focused specifically on speech disfluencies amongst the healthy ageing population.

[Bortfeld et al. \[2001\]](#) presented a study of conversational differences amongst different pairs of speakers; 16 pairs were young with a mean age of 28, 16 pairs were middle aged with a mean age of 47, and 16 pairs were older with a mean age of 67. The disfluencies the authors investigated as part of this study were repeated words or phrases, restarts, and fillers. They found that overall the older participants had higher disfluency rates (6.65) compared to the middle-aged group (5.69) and the young group (5.55), although this difference was not found to be statistically significant.

[Horton et al. \[2010\]](#) conducted a large study of 336 participants aged between 17 to 68 in order to investigate age-related differences in the timing and content of spontaneous speech. They found that the older participants had slower speech and produced more filled pauses, particularly when the filled pauses were associated with lexical selection choices. The authors argue that the older participants were (consciously or unconsciously) slowing their speech in order to accommodate the increase in time taken for lexical retrieval.

[Arslan and Göksun \[2022\]](#) investigated disfluency and gesture production differences between younger and older adults. The group of younger adults consisted of 30 participants (17 female, 13 male) with a mean age of 21. The older adult group also consisted of 30 participants (16 female, 14 male) with a mean age of 65. In terms of disfluencies they investigated three different categories; filled pauses, word repetitions, and repairs. They presented both groups with a picture description task to elicit speech and gesture samples. The authors expected that the older group of participants would exhibit higher disfluency rates than the younger adults. However, their study actually found comparable disfluency rates across the two different age groups. Younger participants used more filled pauses in their picture description tasks, whereas older participants produced more repetitions and repairs.

More recently, [Beier et al. \[2023\]](#) presented a longitudinal study of 91 people who were recorded completing interviews several times throughout the course of their lives, from

ages 20-94. These people are well known public figures, and have careers such as singers, directors, actors, and politicians. As participants aged, the researchers observed that they exhibited more repetitions in their speech, as well as a reduced speaking rate. They did not observe an increase in filled pauses or repairs.

2.4.3 Disfluencies in Cognitive Decline

Amongst the research investigating the effects of CD on speech and language, a body of work looking specifically at disfluencies resulting from CD has started to emerge. These studies focus predominantly on pauses (see [Nasreen et al. \[2021\]](#), and [Pakhomov et al. \[2011\]](#)) and their ability to reliably differentiate between different levels of CD. Even the more comprehensive studies of speech in ageing, such as [Martínez-Nicolás et al. \[2022\]](#), pay little attention to speech disfluencies other than pauses. [Lee et al. \[2011\]](#) presented a study in which they compared the speech of healthy older adults with speech from older adults diagnosed with AD. These authors included an analysis of prolongations (termed “lengthenings” in their study) alongside an analysis of filled and unfilled pauses. They found that for all three disfluency categories the disfluencies spoken by the AD patients were longer in duration.

2.4.4 Disfluencies in Human-Computer Interaction

Early work investigating the speech of people communicating with computers (including robots, intelligent systems, etc) found that, in general, people exhibit fewer disfluencies. [Oviatt \[1995\]](#) presented a study specifically investigating speech disfluencies in human-computer interactions. 44 participants of ages from mid-20s to 65+ who were all native speakers of English were recruited and instructed to interact with a service transaction system. This simulated either an automatic system for verbal-temporal tasks (such as conference registrations, where the spoken content would be primarily proper nouns and scheduling information) or computational-numeric (such as personal banking, where the spoken content would be primarily digits). The authors also reanalysed previously collected human-human speech data, in order to provide directly comparable disfluency

rates. The study revealed substantially lower disfluency rates in the human-computer interactions. In the human-computer unconstrained verbal-temporal interaction, the mean disfluency rate was 1.74 disfluencies per 100 words. A one person non-interactive monologue yielded a disfluency rate of 3.60 disfluencies per 100 words, and a human-human telephone interaction yielded a disfluency rate of 8.83 disfluencies per 100 words. However, as the authors note, the mean utterance length in the human-computer interactions was very short (between 2-5 words) and the nature of the interaction was more structured, meaning that the participants had to do less planning of the conversation.

2.5 Automatic Cognitive Decline Classification

Before this section goes on to discuss Automatic Cognitive Decline Classification (ACDC) systems and the current approaches to building them, there must be a short discussion around the term “automatic cognitive decline classification” and why this thesis prefers this over any of the numerous other titles such systems are given in the research.

Firstly, this thesis aims to be as specific as possible. This thesis does not deal purely with data obtained from people with AD, or even just dementia, and therefore any reference to “dementia” classification systems would be inaccurate. This inaccuracy is common in speech technology fields, and it is not difficult to find examples. [Weiner et al. \[2017\]](#) describe a “dementia detection” system which in fact is actually a system that is trained and tested on speech data from three groups; healthy controls, patients with age-associated CD, or people with AD. In a similar vein, [Chlasta and Wolk \[2021\]](#) present a dementia screening tool which is designed to screen specifically for AD, with no mention of the numerous other kinds of dementia. If the term dementia is being used as an umbrella term to describe the range of different syndromes (as discussed in Section 2.2.1) then dementia screening systems should be able to screen for more than just one specific disease.

Secondly, this thesis will not refer to these automatic systems as “prediction” or “detection” systems and rather favours the term “classification” systems. The term *prediction* suggests that these systems are able to anticipate whether a person will go on to develop CD. *Detection* suggests that systems can identify levels of CD in a person, even if they

currently do not have any symptoms. Neither of the above statements are currently true. Instead, these systems analyse a range of features derived from speech and decide if those features are more likely to belong to one group or another based on how similar those features are to the data the systems have been trained on, making a classification. Therefore, such systems discussed throughout this thesis are termed **ACDC** systems, unless stated otherwise.

The remainder of this chapter discusses the different component parts of **ACDC** systems that are pertinent to the analyses presented in this thesis, and ends with an overview of the CognoSpeak Intelligent Virtual Agent (**IVA**) system that provided the majority of the data for these analyses.

2.5.1 ACDC Systems

Although there is a lot of variety in **ACDC** systems, they mostly share the same constituent parts. Firstly, some kind of Automatic Speech Recognition (**ASR**) system is used to transcribe the speech from selected recordings. Once the speech has been transcribed, a number of different features are extracted from both the audio and the generated transcript. A process called feature selection then determines which of these features are the most meaningful for the classification process, and features which are found to impede classification are discarded. In Neural Networks (**NNs**), this process can be done automatically during the training phase. The model is then trained on a subset of data, and learns to associate features with the presence or absence of **CD**.

Once the model has been trained it is tested on some unseen (excluded from the training) data which allows researchers to assess how well the model can generalise to data that it has not yet encountered. The model is also validated which gives researchers an overview of the model's performance. This performance can be assessed in many different ways through the use of metrics such as accuracy or precision which are usually automatically computed, although human evaluation of model performance is also possible. A system of sufficiently good performance can then be used to classify recordings, and the output could (theoretically) be used to aid doctors and clinicians in making diagnostic decisions.

Currently in the U.K these systems are not used by healthcare professionals, although researchers from groups such as CognoSpeak are working to change this. The following section describes the main parts of any typical **ACDC** system.

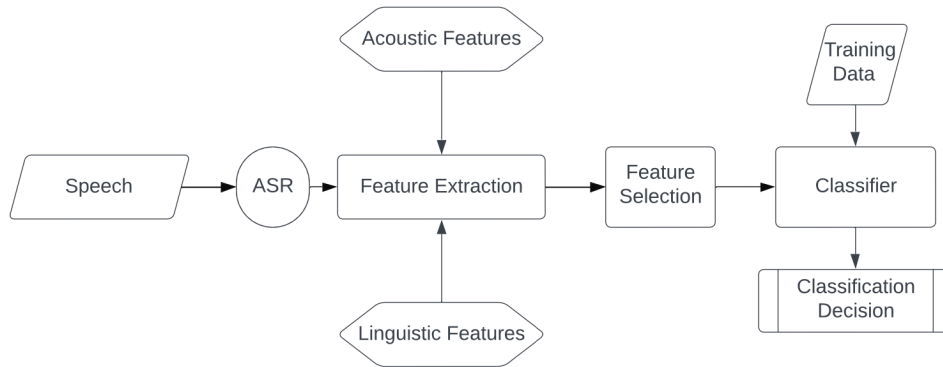


Figure 2.1: Simple Architecture for Automatic Cognitive Decline Classification Systems. Adapted from [Ammar and Ayed \[2018\]](#).

2.5.1.1 Automatic Speech Recognition

At the most basic level, the role of an **ASR** system is to find speech sounds within a recording and provide an appropriate transcription or label of those sounds. **ASR** systems contain many of the same constituent parts as an **ACDC** system. However, **ASR** systems use information from language and acoustic models during a decoding phase in which the language and acoustic models determine the most probable sequence of words to match an input recording [[Wang et al., 2019](#)]. Numerous open-source **ASR** toolkits are available for researchers, such as the Kaldi **ASR** toolkit [[Povey et al., 2011](#)] which was used to create the CognoSpeak system described below in Section 2.4.3.

In addition to **CD** classification, **ASR** has a number of different applications in fields such as automatic transcription or voice search.

Language Models

A Language Model (**LM**) trained on a dataset learns the probability of any set of words in that dataset occurring in a sequence, and serves as a probabilistic approximation of language as expressed in that dataset. **LMs** are used by **ASR** systems to verify the plausibility of transcription hypotheses. More recently, **LMs** scaled to billions of parameters

and trained on vast amounts of data (i.e., Large Language Models (LLMs)) have led to significant advancements in general-purpose language modelling, enabling the models to perform a variety of tasks such as question answering or summarisation without the need for task-specific adaptation.

Statistical LMs use n -grams to direct the search for the next word [Ghai and Singh, 2012]. “ n -gram” is a term used to name a sequence of items according to the number (n) of items in that sequence. These items are typically words, but may also be phonemes, characters, or syllables. A unigram (1-gram) will typically be a single word. Using this label, we can say that the sentence “*this is an apple*” contains four unigrams; “this”, “is”, “an” and “apple”. A bigram (2-gram) usually refers to a pair of words, three of which can be found in the above example; “this is”, “is an”, and “an apple”. The probability of a future n -gram is based on only the last n -gram given in a sequence. For example, in trying to predict the next word in a sentence starting with “*this is an*”, only the last n -gram in that sequence ($n-1$ gram, in this case “an”) affects the probability of the upcoming n -gram. Simple statistical LMs would not take into account the $n-2$ or $n-3$ grams (“is” and “this”, respectively).

However, n -gram models are too simple to take into account the overall context of a text, which can have a strong effect on the probability of the next word in a sequence. Another drawback of n -gram models is that they would assign a probability of 0 to n -grams that have not appeared in the training data, even if they are grammatically correct constructions [Jing and Xu, 2019]. This particular issue is referred to as data sparsity.

To combat these shortcomings, a number of different Neural Network-based Language Models (NNLMs) have been proposed. NNLMs use vector representations of words or sentences as their inputs, which allow the semantic relationships between words to be captured. NNLMs are deeply complex models, described in detail in Lebret [2016]. For the present thesis it is enough to know what LMs are used for, how they generally accomplish this task, and that most state of the art ASR systems will use some form of NNLM.

Acoustic Models

Acoustic Models (AMs) link particular features extracted from a sound file to a linguistic

unit [Bhatt et al., 2020]. Linguistic units could be phonemes, diphones, triphones, or other sub-word units. AMs learn the relationships between acoustic features and linguistic units using statistical methods or NNs, and are required to enable ASR systems to create accurate textual representations of speech. AMs are trained on datasets consisting of speech recordings and their textual or phonetic transcriptions. However, these datasets need to be large and very diverse as AMs can only recognise what has been present in the training data [Aggarwal and Dave, 2011]. This presents numerous problems when considering the complex relationship between acoustic features and linguistic units, and the inherent diversity of speech. For example, AMs that have been trained on a dataset that contains recordings of only Scottish-accented English speakers would likely not be accurate if used on recordings containing Standard Southern British English accented speech. This is particularly problematic in the case of non-native language speakers, as exemplified in Hollands et al. [2022], where mispronunciations and abnormal pause locations severely degrade the performance of ASR systems.

2.5.1.2 Feature Extraction

Following the production of a transcript of the speech in a recording by the ASR system, the next step of the ACDC pipeline is feature extraction. Two types of features are typically used in an ACDC system: acoustic and linguistic. Acoustic features include metrics such as Mel Frequency Cepstral Coefficients (MFCCs), measurements of energy at different frequencies, and other measurements that are extracted directly from the audio recordings (Thomas et al. [2023]; Pulido et al. [2020]). Conversely, linguistic features are those extracted from the transcripts of the speech produced by the ASR system as described in Section 2.4.1.1, such as the number of nouns or the ratio of nouns to pronouns. Below we delve into commonly extracted acoustic and linguistic features for the task of CD classification.

Acoustic Features

Acoustic features are typically extracted through the use of toolkits such as openSMILE [Eyben et al., 2010] that can extract specific sets of acoustic features. The combination of acoustic features used in ACDC can affect system performance. Three of the most

commonly used feature sets are described below.

emobase

The openSMILE toolkit provides a featureset consisting of 988 acoustic features under the name emobase. These features are all Low-level Descriptors (LLDs), or features that are closely related to the characteristics of the signal (rather than high-level descriptors which carry semantic or syntactic meaning [Amatriain et al., 2005]). The LLDs that makeup the features in emobase are: intensity, loudness, MFCCs, Fundamental Frequency (F0), probability of voicing, F0 envelope, line spectral frequencies, and zero-crossing rate. Delta regression coefficients are computed from these descriptors, and then functionals including standard deviation, arithmetic mean, and skewness are applied to those coefficients to produce the features. For more information about the emobase features and how these are calculated, see the openSMILE documentation (<https://audeering.github.io/opensmile/get-started.html>).

ComParE

The Computational Paralinguistics Evaluation (ComParE) feature set is a larger and enhanced version of the original emobase feature set [Eyben et al., 2013]. ComParE contains 6,373 suprasegmental features. These features are a combination of LLDs and functionals, and removes features that were found to frequently contain zero information such as the arithmetic mean of delta coefficients. There are new LLD features in ComParE; auditory model loudness, Viterbi algorithm smoothing, psychoacoustic sharpness, and harmonicity [Schuller et al., 2013]. These are used alongside LLDs that were included in emobase. The functionals remain largely the same, with the addition of linear predictive coding and more robust peak picking algorithms. A full overview of the LLDs and functionals in ComParE can be found in Weninger et al. [2013].

eGeMAPS

The original Geneva Minimalistic Acoustic Parameter Set (GeMAPS) was created in an attempt to refine the massive feature sets described above. Features were selected based on their theoretical significance, how valuable they had proven to be in previous feature sets, and how well they can index physiological changes in voice production [Eyben et al.,

2015]. The GeMAPS features are: 18 LLDs with applied arithmetic mean and coefficient of variation functionals which results in 36 parameters. Eight functionals are applied to loudness and pitch, resulting in an additional 16 parameters. The arithmetic mean of the Alpha Ratio, the Hammarberg Index, and spectral slopes are also included alongside six temporal features including rate of loudness peaks and the number of continuous voiced regions per second, bringing the total to 62 parameters [Eyben et al., 2015, p.193]. However, there were no cepstral and very few dynamic parameters included in GeMAPS, therefore an Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) was proposed to include seven additional LLDs and a range of different functionals, bringing the total number of parameters in eGeMAPS to 88 [Eyben et al., 2015, p.194]. GeMAPS and eGeMAPS are frequently used by researchers to provide baseline classification results which can be compared to results from more specific parameter sets, allowing for the replication of findings and comparison between studies [Eyben et al., 2015, p.191].

Linguistic Features

Linguistic features are intended to capture patterns of language impairment that can be observed in CD. The linguistic features used in ACDC systems vary from study to study depending on which specific levels of CD are being investigated, and typically there will be far fewer linguistic features than acoustic. Linguistic features can be grouped according to the linguistic phenomena they are attempting to capture. Below is an overview of some of the most common types of linguistic features used in ACDC tasks.

Part of Speech

Part of Speech (PoS) measures capture information surrounding the usage of different word classes in speech. Common PoS measures include noun rates, verb rates, and rates of pronoun usage. Bucks et al. [2000] measured pronoun, noun, verb, and adjective rates per 100 words as part of their study investigating spontaneous speech from people with AD, and found statistically significant differences between AD patients and healthy controls in terms of all PoS measures that were analysed. AD patients were found to have higher pronoun, verb, and adjective rates compared to the healthy controls, but lower noun rates.

PoS information can also be captured in the forms of ratios such as pronoun:noun or noun:verb. Automatic taggers, such as the one provided by the spaCy NLP library [Alti-nok, 2021] are used to identify which parts of speech each word in a text belongs to.

Temporal Rates

Articulation, phonation, and verbal rates are typically used as linguistic features, along with other temporal measures such as total locution time and syllable duration. The calculations for these measures are found below:

1. Total Phonation Time (TPT)

$$\text{TPT} = \text{Speech} - (\text{Filled Pauses} + \text{Silence})$$

2. Standardised Phonation Time

$$\text{Standardised Phonation Time} = \frac{\text{Number of Words}}{\text{TPT}}$$

3. Articulation Rate

$$\text{Articulation Rate} = \frac{\text{Number of Syllables}}{\text{TPT}}$$

4. Total Locution Time (TLT)

$$\text{TLT} = \text{Speech} - \text{Silence}$$

5. Verbal Rate

$$\text{Verbal Rate} = \frac{\text{Number of Words}}{\text{TLT}}$$

6. Speech Rate

$$\text{Speech Rate} = \frac{\text{Number of Syllables}}{\text{TLT}}$$

7. Average Syllable Duration

$$\text{Average Syllable Duration} = \frac{\text{TPT}}{\text{Number of Syllables}}$$

More information about these measures can be found in [Themistocleous et al. \[2020\]](#), and [Calzà et al. \[2021\]](#).

Lexical Richness

Measures of lexical richness are designed to capture information surrounding how broad a person's vocabulary is. [Thomas et al. \[2005\]](#) present an ACDC system for AD that pays particular attention to lexical analysis. Three commonly cited measures of lexical richness are described below:

1. Type to Token Ratio (TTR): A simple measure of richness calculated as

$$\text{TTR} = \frac{\text{Number of Unique Words}}{\text{Total Number of Words}}$$

2. Honoré's Statistic (HS): Measures lexical diversity and takes into account the total vocabulary size (number of unique words) and the proportion of words that occur only

once in a text

$$HS = 100 \frac{\log(V)}{1 - (\frac{V^1}{N})}$$

where V is the number of distinct words, V^1 is the number of words that appear only once, and N is the total number of words in the text. A higher HS indicates a more diverse vocabulary.

3. Brunét’s Index (W): A measure of lexical diversity that is independent of text length

$$W = N^{(V-a)}$$

where N is the length of the text, V is the number of different words, and $-a$ is a scaling constant usually set at -0.172 . A lower W value indicates a richer vocabulary.

Novel Feature Combinations

Whilst most [ACDC](#) systems will use a combination of both linguistic and acoustic features for their purpose, some work has demonstrated the possibility of using only acoustic features for classification [[Warnita et al., 2018](#)]. Likewise, there are studies that have focussed more on the linguistic features, such as a study by [Vincze et al. \[2016\]](#) which found that features based on morphology were able to discriminate between healthy controls and participants with [MCI](#). One issue with using “traditional features” is that they often are difficult to interpret and describe. This issue is discussed further in [Chapter 4](#).

A contribution of this thesis is the work investigating how a third set of features, disfluency features, could be used in [ACDC](#) systems. Current [ACDC](#) systems typically do not make use of disfluency information, other than that of unfilled pauses which can be extracted acoustically by looking for periods of silence in the speech signal. This is largely due to the fact that the [LMs](#) that [ASR](#) systems are trained on consist of large collections of texts, or speech that has come from text (read speech). Authored texts by their nature contain far fewer disfluencies than natural, spontaneous speech. This results in [LMs](#) that ideally assign a much higher probability to utterances that are free of disfluencies [[Zwarts and Johnson, 2011](#)]. In addition, disfluencies are typically regarded as problems in speech that are unwanted in a transcript, with numerous [ML](#) studies aiming to “clean” spontaneous

speech in an effort to remove disfluencies and improve LMs (Adda-Decker et al. [2003]; Rao et al. [2007]; Lou and Johnson [2020]). Chapter 5 details how different methods can be used to combat these issues and leverage disfluency information to increase the accuracy of ACDC systems.

2.5.1.3 Feature Selection

Feature selection allows for the removal of features that are not contributing to the performance of the ACDC system. It also helps to reduce the complexity of ACDC models and can help reduce the chance of overfitting, a problem that is of particular concern when working with small datasets. The objective of training a predictive model is to discover the underlying patterns in the observed data which may be true also of the general population the data is merely a sample of. Overfitting happens as a result of an ML model too closely capturing the patterns of the training dataset. Although the model exhibits excellent performance on the training data, that result cannot be matched on previously unseen data, meaning that the model does not generalise well [Ying, 2019].

There are two main types of feature selection techniques: filter methods and wrapper methods. Filter methods rank features in terms of how much they contribute to the classification. Once features have been ranked, a subset of the best N features can be chosen [Shardlow, 2016]. Filter methods work independently of the classifier being used. Conversely, wrapper methods allow the classifier to make decisions as to which are the best features. A wrapper model will select a subset of features from the original featureset, and evaluate the performance of an ML algorithm on that subset. If particular features are found to be enhancing the performance of the algorithm, these features are then included in the final feature subset. The wrapper then tests the performance of different combinations of features, selecting the highest performing features until the final feature subset contains the optimal combination of features for system performance. Wrapper models generally give better results than filter models, but are typically much more computationally expensive [Sánchez-Marroño et al., 2007].

2.5.1.4 Training and Testing

Data is required to train an ML model. Typically, a given data set will be split into two groups, the training and the test set. Researchers may also decide to split the data three ways, adding a validation set. In the training set, the information about each patient (the features extracted above) are the inputs, and the possible diagnoses are the classes. A binary AD classification system will have two classes; healthy control or Alzheimer's Dementia.

Typically, the training set will contain the majority of the original data ensuring that the model has as many examples of different inputs and classes as possible. The model then learns the patterns in the training data and maps this to the attributes [Alpaydin, 2014]. For example, the model could learn that the examples of data labelled as healthy all tend to have a higher Honoré's Statistic, and the data labelled AD have a lower. The model can then use this information to sort the recordings according to the different classes, but based on information from all of the inputs and their combinations, rather than a single statistic. A validation set can be used to fine-tune the model by adjusting the hyperparameters depending on the performance of the model during training.

The testing portion of the data has not been seen by the model before. The model then uses the patterns it uncovered during the training to help sort the new data according to what class the new data is most similar to. This gives researchers an idea of how well the model generalises to new data and how well it carries out the given task.

2.5.1.5 Classification

Different models will perform differently on classification tasks, and the model choice varies from study to study for a range of different reasons such as the computational resources required or the amount of data available. The section below details the most commonly found models used for ACDC tasks, as covered in the review paper by Petti et al. [2020].

Decision Trees

Decision trees are hierarchical tree-like structures that are composed of nodes, with each

node representing a rule based on the features of the input data. Decision trees are interpretable thanks to their structure, which makes it easier to understand how a particular classification decision has been reached [De Ville, 2013].

k-Nearest Neighbour

When unlabelled inputs are fed to a k-Nearest Neighbour (**KNN**) classifier, the **KNN** finds the most similar instances from the training data and uses that information to determine the appropriate label for the new inputs. Similarity between an unlabelled input and its closest known neighbours is measured by a distance metric, such as Euclidean distance or Hamming distance [Kramer, 2013].

Neural Networks

NNs are complex **ML** algorithms composed of nodes that are able to communicate with other nodes via *connections*. Each node is a mathematical processing unit that learns complex relationships between the input and the output. The first layer of nodes is the input layer that directly “sees” the input data. Information from the input layer is then passed to a set of hidden layers, and each node in the new layer receives input from each node in the previous layer [Georgevici and Terblanche, 2019].

Deep Neural Networks (**DNNs**) are a type of **NN** that have the ability to learn hierarchical representations of data. **DNNs** are often the models used in state-of-the-art systems for a range of different applications due to their ability to capture patterns that people and other **ML** models often are not able to capture. Whereas a traditional **NN** will have one-to-three hidden layers, **DNNs** have tens or hundreds of them [Choi et al., 2020]. **DNNs** are also reliably good at generalising these patterns to unseen data, enabling them to make accurate classifications. In the field of **CD** classification the main argument against the use of **DNNs** is that they require high volumes of data, something that is often unavailable to researchers. They also often come with high computational costs.

Support Vector Machines

Support Vector Machines (**SVMs**) are particularly well suited to classification tasks, as they work to find boundaries in data to define separate classes. **SVMs** use hyperplanes to separate data points into different classes. We want these hyperplanes to have the largest

margin possible to separate classes [Gil and Johnsson, 2011]. At its most basic level, an SVM can be thought of as a way of drawing a line through data points to separate the different data points into groups in a two-dimensional space. Figure 2.2 demonstrates the different ways in which two different data classes could be separated, alongside the optimal hyperplane which would be uncovered by the SVM.

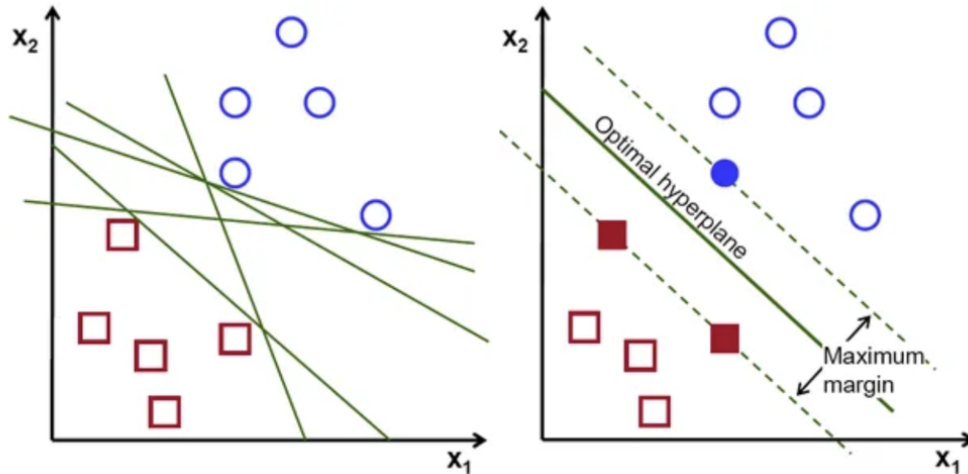


Figure 2.2: Possible Hyperplanes vs. the Optimal Hyperplane [Gandhi, 2018].

Although not as complex as DNNs, SVMs are still commonly used in ACDC systems as they work well with limited amounts of data. Early work from Orimaye et al. [2014] found that an SVM was more accurate in their classification task when compared to other ML models such as decision trees and NNs. More recent studies (Balagopalan et al. [2020]; Zolnoori et al. [2023]) still use SVMs for their classification tasks thanks to their high performance with small datasets and comparably high accuracy when compared to other models tested on the same task. For these reasons we chose to use an SVM for our own classification task, as described in Chapter 4.

2.5.2 System Evaluation

Many different metrics can be used to assess how well an ML model is working for a given task. These metrics will be selected according to the type of task being investigated. In classification tasks there are two possible outcomes: the instance is classified correctly or not. Correct classifications result in true positives or true negatives, and incorrect

classifications result in false positives or false negatives. Evaluation metrics that are commonly selected for classification tasks are described below [Vujović et al., 2021].

2.5.2.1 Accuracy, Precision, Recall, and Specificity

One of the most commonly reported metrics of ACDC system performance is accuracy. This is defined as:

$$\textit{Accuracy} = \frac{\textit{Number of Correct Predictions}}{\textit{Total Number of Predictions}}$$

The best accuracy score is 1.0, with the worst being 0.0. Accuracy can be misinterpreted when used with imbalanced datasets. A measure of precision can be used to assess how good the model is at avoiding false positives:

$$\textit{Precision} = \frac{\textit{True Positives}}{(\textit{True Positives} + \textit{False Positives})}$$

As with accuracy, the best precision score is 1.0 and the worst is 0.0.

Recall (or sensitivity) measures the rate of true positives, and is calculated as:

$$\textit{Recall} = \frac{\textit{True Positives}}{(\textit{True Positives} + \textit{False Negatives})}$$

A high recall (1.0) is particularly important in medical fields where false negatives (for example, diagnosing a person as healthy when they actually have a disease) could have severe repercussions.

Specificity is the rate of true negatives:

$$\textit{Specificity} = \frac{\textit{True Negatives}}{(\textit{True Negatives} + \textit{False Positives})}$$

The highest specificity score is 1.0, the lowest is 0.0.

2.5.2.2 F1-score

Precision and recall scores are often inversely related, where improving the score from one metric will downgrade the score from the other. In order to balance the precision and recall, the scores are combined with a weighted harmonic mean and report it as an F score [Derczynski, 2016]. This is calculated as:

$$F_{\beta} = (1 + \beta^2) \frac{PR}{\beta^2 P + R}$$

β is what determines the balance between precision and recall. Higher β values favour recall. The β value is commonly set to 1, hence the name “ $F1$ score” [Derczynski, 2016, p.262].

2.5.2.3 Area Under the ROC Curve

Receiver Operating Characteristic (ROC) curve graphs plot the true positive rate against the false positive rate at different thresholds. Deciding which threshold to use depends on the specific task, system, and application. For example, in some scenarios it may be reasonable to prefer a higher sensitivity at the cost of more false positives [Handelman et al., 2019]. The closer the ROC curve is to the upper left corner of the graph (where sensitivity is 1 and the false positive rate is 0), the better the classifier is. We measure the area under the ROC curve, and the closer this value is to 1, the better the model is performing [Vujović et al., 2021].

2.5.3 CognoSpeak and the Intelligent Virtual Agent

The original IVA that went on to form the basis of the CognoSpeak system was presented in Mirheidari [2018]. The idea was to use an IVA to elicit information from patients in a memory clinic, in a similar manner to how a human clinician would elicit information in the same scenario. Using an IVA to collect data to be used in an ACDC system would allow a fully automated approach to analysing speech from people with different levels of CD.

The first iteration of the IVA (the data from which is used for the analyses in Chapters 3, 4 and 5 of this thesis) consisted of an animated image of a human head. After obtaining ethical consent, participants were invited to take part in the IVA study. Participants were encouraged to bring an Accompanying Person (AP) with them on the day of recording. Once participants arrived at the memory clinic, they were seated in front of a laptop displaying the image of the IVA (Figure 2.3). When the study started, the laptop would play a recording of a question to the participant. The participant then has two options. If they experienced any trouble in understanding the IVA, participants were able to replay the question by pressing space bar on the keyboard. If the participant was happy to answer, they would simply speak their answer to the IVA and their speech would be recorded by a microphone placed behind the computer. Once participants had answered a question, they pressed the enter key on the keyboard to move onto the next one.

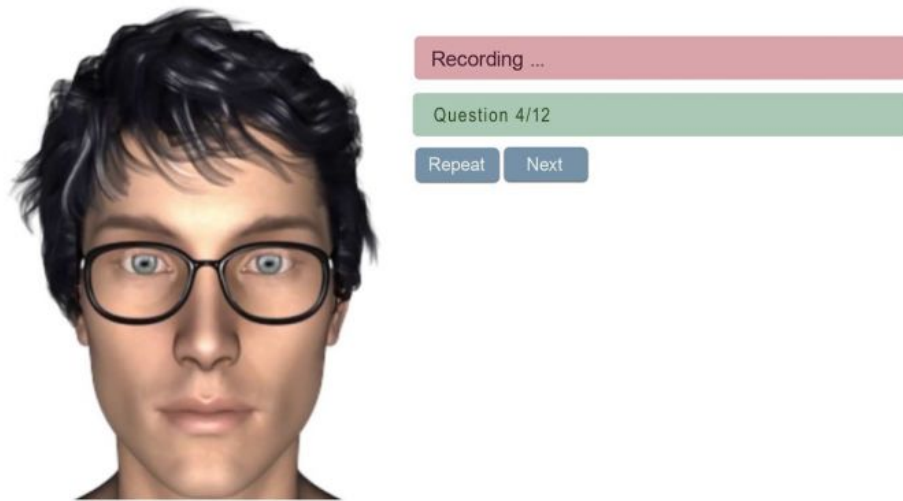


Figure 2.3: The Original IVA [Mirheidari, 2018, p.143].

The IVA system is a SVM-based classifier that consists of the components described above in 2.5.1.

There were three phases to these studies. The first phase consisted of an interview style conversation in which the IVA would ask participants questions designed to test their memory and uncover information about their symptoms. During this phase the participants are not constrained by time limits when they are answering the questions. The

second phase had the IVA instructing participants to complete SVF and PVF tests. Once participants started their answers to each of these tests, they were given 60 seconds until they had run out of time and were instructed to move on. The third phase instructed participants to complete a picture description task. Once again, there were no time restrictions imposed on participants when completing this task.

Slight adjustments were made to the questions the IVA played to participants over the course of the three years it was used to collect data. A full description of these changes can be found in Mirheidari [2018, p.145]. In total, recordings from 78 participants were collected. These participants belonged to one of four groups: Healthy Controls (HCs), people with Functional Memory Disorder (pwFMD), people with Mild Cognitive Impairment (pwMCI), or people with Neurodegenerative Dementia (pwND).

The newest version of the CognoSpeak IVA has had numerous upgrades. The key differences are that there is now a lot more variation in the appearance of the virtual agent. Users can select from a range of higher quality avatars based on their own personal preferences. These avatars are expressive and perform different gestures such as head tilts and nods to indicate to the participants that they are being “listened” to. There have also been changes made to the back-end infrastructure of the system, enabling the recruitment of higher numbers of participants via the web.

2.6 Conversation Analysis

Conversation Analysis (CA) is the study of talk in interaction, with close links to the field of ethnomethodology which investigates how social order is produced through interaction. CA emerged in the 1960s and 1970s as a method of investigating interaction whilst taking language as a main point of focus. Early investigations centred around the process of turn-taking within a conversation, and how those turns are organised and shaped by the discourse context in which they are found [Sacks et al., 1974]. Although CA emerged as a method of analysing the social interactions found within ordinary conversations, CA is now applied to a range of different interactional scenarios [Goodwin and Heritage, 1990]. Of particular interest for this thesis is the area of medical encounters, especially those

that revolve around memory concerns or complaints. This topic is discussed further in the background section of Chapter 5.

The remainder of this section focuses on key concepts within CA.

2.6.1 Turn-Taking

One of the most widely accepted principles of CA is that talk-in-interaction takes place in turns, and that this structures talk in such a way as to minimise the overlap of speech from different speakers. This organisation requires input from all members of the conversation and is an inherently cooperative process. Turns are key to the sequencing of conversations [Drew, 2013]. There are three main factors that shape the form of a turn; where in the sequence a turn is taking place, what is being done in that turn, and who the turn is addressed to [Drew, 2013, p.134]. The setting in which the conversation is taking place will also have an effect on the design of the turns in conversation. For example, in everyday conversation it is common to find turns being addressed to multiple people, with all members of the conversation having an input on who should be the next speaker. However, in an educational setting, the addressee of the turn is usually strictly dependent on decisions made by one speaker, in this case a teacher [Gardner, 2004, p.272].

2.6.1.1 Turn Constructional Units and Transition Relevance Place

Sacks et al. [1974] note that turns are incrementally built out of Turn Constructional Units (TCUs). These units could be words, phrases, clauses, or sentences, and are coherent and self-contained utterances [Clayman, 2013]. TCUs end with points of potential completion, named Transition Relevance Places (TRPs). These TRPs signal to the other conversational partners a potential point at which a turn could be finished, but they do not necessarily have to be. Rules of turn allocation become relevant at every TRP [Sacks et al., 1974].

2.6.2 Adjacency Pairs

Another important foundational principle of CA is that talk happens in a sequential and organised manner [Stivers, 2013]. This describes the way in which conversations set expectations around what will happen next. For example, when you greet someone they will probably greet you back; if they do not, there is a sense that something is missing or absent. Observing these back and forth sequences revealed what CA researchers term “adjacency pairs”- a pair of actions in conversation that are frequently found together [Gardner, 2004]. Other examples of adjacency pairs include question-answer pairs and farewell-farewell pairs. In a study of question-answer adjacency pairs in conversations with people with AD, Varela-Suárez [2018] found that the ability to complete a question-answer pair remains intact until very late stages of the disease, but the relevancy of the answers given by dementia patients decreases as the disease progresses.

2.6.3 Repairs and Trouble Sources

In CA, the term “repair” refers to the structure of how interactants deal with troubles in speaking, hearing, or understanding talk [Schegloff et al., 1977]. Repairs mark out something in the prior talk as a “trouble source”. Trouble sources can be a number of different things, such as a lack of attention being paid to the conversation, or a lack of knowledge surrounding a certain topic. They could also be instances of incorrect word usage, or speaker overlap. In people with aphasia, including those with some degree of CD, commonly observed trouble sources often take the form of false starts and disfluencies such as long pauses [Whitworth, 2003]. Trouble sources are often identified by their repairs, but there may be cases where a repair is unsuccessful. These are referred to as “failures”, although they are rare (p.363).

CA deals with repairs in two different ways; according to the initiation of the repair, or how the actual repair is carried out. Self-initiated self-repairs are the most common type of repair, and these involve the current speaker both initiating and producing the repair [Kitzinger, 2013]. An example of this kind of repair can be found in the recording from Participant 0252 in the IVA dataset. The participant (a healthy control) has been asked

if they remember what has been in the news recently:

0252 - HC - IVA

1 Pt: oh and the se- the home secretary has resigned hasn't she

In this case, “se-” is treated as a trouble source, as the beginning of an incorrect word. The repair sequence is “the home secretary”, which inserts a modifier of the original “se-[cretary]” that offers more information and addresses the potential trouble of lack of clarity or information.

Repairs can also be initiated by someone other than the speaker of the trouble source. Other-initiated repairs locate the trouble source in the prior turn, and indicate to the speaker of the trouble source that there has been some kind of misunderstanding. It is usually the speaker of the trouble source that completes the repair (eg., other-initiated self-repair). This frequently results in a repair sequence that is a repetition of the trouble source, but often adjusted in terms of the volume or rate of speech [Kendrick, 2015].

2.6.4 Preference Structure

Preference in CA refers to the fact that many conversational events engender alternative, but unequal, courses of action [Sacks, 1973]. These alternatives could be different utterance designs or different sequence choices (amongst others), but preference in this context does “not” refer to personal or subjective desires [Atkinson, 1985]. The concept of preference provides insights into how people navigate social interactions and co-construct conversational sequences, and preference structures are “culturally shared principles” which are “empirically evidenced by orderly ways of speaking that are produced in accord with those principles” [Pomerantz and Heritage, 2013, p.210]. Preference in conversation takes various different forms. For example, questions can be designed to show a preference for a particular answer. Pomerantz and Heritage [2013] present the following example (p.213):

1. “Do you belong to a church now?”
2. “You don't want that lamb chop do you?”

The authors argue that question one is an example of a question that prefers a “yes”

response, while question two prefers a “no”. They show that in response to yes/no questions, “where possible, speakers will avoid or minimise explicitly stated disconfirmations in favour of confirmations” (p.213). Therefore, the positive stance of question one is designed for agreement with “yes”, whilst the negative interrogative in question two is designed to be confirmed with a “no” (ie., “no, I don’t want it”).

Another social behaviour exemplified in the structure of preferred responses is that in conversation, people try to avoid or minimise rejections if they can [Atkinson and Drew, 1979]. This can also be observed in how people reject invitations. So-called dispreferred responses happen in a variety of different ways, such as including components in the response designed to minimise the rejection (“*I’d love to, but...*”), or to soften the rejection itself (“*I don’t think I can*” rather than an abrupt “no”). Additionally, delays and mitigations are commonly performed as part of a dispreferred rejection response, and these delays frequently take the form of filled pauses.

Other disfluencies can also be a marker of preference in conversation. In a study of a face-to-face medical consultation, Gill [2005] describes a situation in which the patient is about to praise the work of their previous doctor. As noted, when praise is directed at a person who is not present during the conversation it can often also act as a criticism of the present member of the conversation who is a member of the same category as the recipient of the compliment (p.466). The patient in this example produces numerous different disfluencies (filled and unfilled pauses, multiple repairs) before they formulate this compliment. Gill argues that the presence of these disfluencies demonstrates that the patient treats this as a sensitive matter and a dispreferred activity (p.267).

2.6.5 Analysis of Medical Dialogues

When researching methods of analysing medical dialogues, use of the Roter Interaction Analysis System (RIAS) is widely observed. This framework enables researchers studying doctor-patient interactions to categorise the communication components from both the doctor and the patient, capturing the fundamental elements of medical interactions [Roter and Larson, 2002]. RIAS organises all doctor utterances during a medical interview

into one of four groups; patient education and counselling, data gathering, relationship building, and activating and partnering. Each of these categories contains a group of communicative behaviours that have specific goals, such as biomedical information-giving in the patient education group (“*take this medication for seven days*”) or emotional talk in the relationship building group (“*I’m sure things are going to improve*”).

Although there are some similarities between the [RIAS](#) and [CA](#) methodologies, they are two distinct approaches that have their own benefits and drawbacks. The following table (2.3) briefly summarises the main differences between [RIAS](#) and [CA](#):

Comparison Points	RIAS	CA
Focus	General description of the interaction	Finely detailed descriptions of the interaction
Methodology Type	Quantitative	Qualitative
Study Design	Used in combination with statistical tests	Used alone or with other methods such as ethnography
Findings	General recommendations and suggestions for improving educational programmes	Detailed recommendations that are specific to the type of interaction e.g., strategies for ensuring a patient is understood

Table 2.3: Comparison of the [RIAS](#) and [CA](#) methodologies. Adapted from [Alsubaie et al. \[2022, p.21\]](#)

Although both methodologies can be applied to similar domains, the biggest difference between the two approaches is the resulting data. Using [RIAS](#) results in quantitative data, and is useful at investigating things such as what percentage of the medical dialogue was contributed to by the doctor vs the patient (for example, in a study using [RIAS](#) to analyse oncological consultations, [Ong et al. \[1998\]](#) found that oncologists contributed 60% to the consultation compared to the patients’ 40%). [CA](#) has the ability to describe in detail the actions being taken by both the doctor and the patient in such interviews, and probes what these actions are designed to achieve. Due to the amount of detailed analysis required for [CA](#), these studies typically include far fewer participants compared to studies that use [RIAS](#) [[Alsubaie et al., 2022](#)].

CA was chosen as a methodology for this thesis over *RIAS* for two main reasons. Firstly, the limited amount of data available for this analysis allows for a more in-depth approach. The second reason is how disfluencies are treated by each approach. In *CA*, attention is paid to disfluencies in speech in terms of what function they might serve in talk-in-interaction, along with what changes they prompt within a sequence. With *RIAS*, there is no specific attention paid to speech disfluencies. In terms of filled and unfilled pauses, *RIAS* lacks a method of coding them as events in a conversation which can result in their functions of maintaining the floor or signalling a shift in topic being overlooked [Sandvik et al., 2002, pp.237-238].

2.6.6 Conversation Analysis of Human-Computer Interactions

Traditionally, *CA* has focused on human-human communication. In recent years, however, there has been a growing body of research employing *CA* methodologies to examine human-computer interaction. Much of this work centres on leveraging *CA* to enhance the naturalness of virtual conversational systems, enabling them to sound and behave in a manner more akin to human interaction. Arend et al. [2017] conducted a case study using *CA* to analyse interactions between a human participant and a NAO robot. NAO robots are humanoid automatons that have been investigated for their potential as socially assistive robots in tasks such as human-robot interaction therapy for children with autism (see Shamsuddin et al. [2012]) or cooperative rehabilitation (see Assad-Uz-Zaman et al. [2019]). Arend et al. [2017] introduced the concept of “dis-balanced” communication, highlighting how the robot’s predetermined turn-taking rules frequently led to conversational challenges, such as unnatural pauses caused by the human participant waiting for the robot to signal its recognition of speech.

It is not uncommon to find conversational systems designed with *CA*-informed research. For instance, Stivers et al. [2009] discovered that, across the 10 languages investigated, confirmations in conversation are delivered 100–500ms faster on average than disconfirmations. Roddy and Harte [2020] incorporated these findings into their neural models,

enabling spoken dialogue systems to generate response times perceived as more realistic.

2.6.7 Disfluencies in Conversation Analysis

When researchers use CA to analyse speech they frequently encounter instances of disfluencies within that speech. Rather than glossing over these disfluencies, CA allows them to be treated as trouble sources; meaningful and interpretable parts of a speaker's turn which can be analysed sequentially to investigate what function they have in a conversation. Disfluencies may also serve as turn-holding devices, repair initiators, or markers of dispreferred responses [Schegloff et al., 1977]. There are examples of work in dementia-focused CA that have investigated the use of pauses in conversation. In a study of speech from people with dementia, Müller and Guendouzi [2005] found that long pauses in speech are used by the speakers in the same way that a minimal response would be; to signal to their conversational partner that the speaker is having some difficulty in completing their turn or formulating their response. Another CA study looking at conversations between people with dementia and their carers found that when the carers permit long pauses in the speech from the patients, the patients can be more successful in their communication as they are given time to process information and formulate responses [Perkins et al., 1998].

2.7 Summary

This chapter has presented the background knowledge required to understand the main concepts and methodologies used in this thesis. This started in Section 2.2 with an introduction to CD, paying particular attention to the levels of CD that are examined in the analysis chapters of this thesis; FMD, MCI, and ND. The focus of this section then turned to the effects that different levels of CD have on speech and language abilities, and what tests are currently used to investigate speech for the presence of CD (Section 2.3).

Section 2.4 presented a brief overview of the types of disfluencies investigated as part

of this thesis (pauses, repetitions, prolongations, speech errors, and repairs). We then discussed previously reported disfluency rates in healthy ageing adults, and compared this to disfluency statistics from people with different levels of CD (Section 2.4.2).

The next section presented the key background information about the automatic analysis of speech for the purposes of CD classification. This started with an overview of the main features of ACDC systems, beginning with ASR (Section 2.5.1.1) and ending with an overview of different classification algorithms (Section 2.5.1.5). We then presented the most common methods of evaluating ACDC systems in Section 2.5.2. This section ended with an introduction to the CognoSpeak system, which provided the human-computer memory clinic data analysed in this thesis.

The final section of this fundamentals chapter provided an introduction to CA (Section 2.6). We briefly introduced the history of CA and then presented the key principles of this methodology (turn-taking, sequence organisation, trouble sources and repairs, and preference structures). Another methodology commonly used for the analysis of speech in medical interviews, the RIAS method, was then presented and compared to CA in Section 2.6.5 alongside justifications as to why this thesis prefers the CA method.

The next chapter in this thesis (Chapter 3) presents our first manual disfluency study; comparing disfluency rates between the four different cognitive groups discussed above in Section 2.2.

Chapter 3

First Manual Disfluency Analysis

Contents

3.1	Introduction	60
3.2	Background	60
3.3	Methodology	68
3.3.1	The Disfluencies in Cognition Schema	69
3.3.2	Data	75
3.3.3	Disfluency Collection Process	82
3.3.4	Statistical Methods	83
3.4	Results	84
3.4.1	Total Disfluency Rates	84
3.4.2	Number of Fluent Words	84
3.4.3	Unfilled Pauses	85
3.4.4	Filled Pauses	87
3.4.5	Pause-to-Speech Ratio	89
3.4.6	Repetitions	90
3.4.7	Prolongations	90
3.4.8	Speech Errors	93

3.4.9	Summary	94
3.5	Discussion	95
3.5.1	Total Disfluency Rates	95
3.5.2	Unfilled Pauses	95
3.5.3	Filled Pauses	96
3.5.4	Pause-to-Speech Ratio	97
3.5.5	Repetitions	98
3.5.6	Prolongations	98
3.6	Conclusions	98

3.1 Introduction

This chapter presents the first manual disfluency analysis and addresses the following research questions:

1. How do the frequency and duration of speech disfluencies differ when participants engage in an interview-style task with a digital avatar in a simulated medical interview scenario compared to similar interviews conducted with human clinicians?
2. Can an analysis of speech disfluencies be used to differentiate between varying levels of cognitive decline?

To answer these questions, a manual disfluency study was devised to investigate the types of disfluency present in the speech of individuals interacting with a virtual agent during an interview-style task. This chapter introduces the development of the Disfluencies in Cognition (DisCo) schema, which is utilised to collect data on the frequency of various types of speech disfluencies. To situate this work, the background section of the chapter includes a discussion of different methods for defining and measuring disfluencies in speech.

Subsequently, the methodology of the analysis is presented, detailing the data sources, the process of capturing disfluency information from each recording, and the statistical techniques employed in this study. Following this, the results are presented and analysed, with a primary focus on the statistically significant findings. The chapter concludes with a discussion of these results and outlines directions for future work.

3.2 Background

As discussed in the Chapter 2, previous studies have focussed on the effects of Cognitive Decline (CD) on language and speech, although little work exists with a particular focus on speech disfluencies. Where such work does exist, the actual phenomena being investigated vary greatly between studies as there is no universally agreed disfluency taxonomy, making the task of directly comparing disfluency studies difficult. For example, a study

investigating the patterns of “pauses” in speech that actually only investigated unfilled pauses would draw completely different conclusions to a similar paper that also included an analysis of filled pauses. Yet, oftentimes such distinctions are not made.

This is also true of other disfluencies that are commonly mentioned in the context of CD, including repetitions and repairs. Whilst it is not uncommon to include such phenomena in disfluency studies, the definition of what constitutes a repetition or repair varies from study to study. For example, [Arslan and Göksun \[2022\]](#) define their repetitions with an example of a single word being repeated once in a sentence. There is no mention of how instances of multiple word repetitions or part word repetitions are handled, or if they are included in their analysis at all. The disfluency category in [Rohanian et al. \[2021\]](#) consists of “self-repairs” (which are not defined in the paper, making it unclear as to what exactly is being measured) and unfilled pauses, alongside “edit terms” which include discourse markers such as “*like*” and “*you know*”, which are rarely treated as disfluencies in disfluency studies.

[Gómez-Vilda et al. \[2015\]](#) include vowel prolongations in their “fillers” category alongside what many other studies refer to as filled pauses or hesitations. However, vowel prolongations in other disfluency studies are commonly treated as their own separate category of disfluency. For example, following a comprehensive description of the different ways of categorising disfluencies, [Eklund \[2004\]](#) notes that prolongations are disfluency events that are entirely separate to “fillers”, and constitute their own category of “hesitations without being silent” (p.208).

It is clear then that there are several disparities commonly seen when studying research into disfluencies. Although the problem of different terms being used to describe the same disfluency phenomena can make it difficult to directly compare studies, it rarely causes a problem that cannot be overcome with some additional reading or searching. However, the bigger issue occurs when it is not made explicitly clear in studies what the chosen terminology is supposed to be referring to. One of the main aims of this thesis is to present work that is unambiguous, ensuring that the procedures undertaken in the analysis chapters can be repeated on a wide range of speech data. Therefore, this

thesis clearly defines and describes each category of disfluency under investigation, and makes clear exactly how a non-fluent event should be categorised according to our schema (section 3.3.1).

This issue of a lack of clarity surrounding studies of disfluencies is not limited to investigations of CD. Discrepancies such as those described above occur throughout disfluency analysis work across a broad range of subjects including, but not limited to, psychology, speech production, teaching and education, forensic speech science, and discourse analysis. Of course, how disfluencies are studied will depend on the overall goal of the particular disfluency analysis and the level of detail it requires. For example, when Sigmund Freud was compiling his *Psychopathology of Everyday Life* in 1901 he described his theory that disfluencies represent hidden or suppressed feelings [Freud, 1989]. It did not matter to Freud if the disfluencies he observed met a certain duration threshold or number of occurrences. Rather, the importance was that the disfluency was present at all. In contrast, a 2015 study from Braun et al. that investigated the speaker-specific differences in hesitations (a category that would contain filled pauses and prolongations if included in this thesis) needed a narrow focus, and therefore puts forward a method of categorising hesitations according to very slight differences:

- Fillers of various kinds (vowel, vowel + nasal, nasal)
- Initial vowel lengthening
- Initial consonant lengthening
- Final vowel lengthening
- Final consonant lengthening

Due to the multidisciplinary nature of disfluencies it is unsurprising that there are numerous different taxonomies of disfluency. In some instances it may be sufficient to define a prolongation based on a subjective judgement of a phone that is longer than should be expected in normal-paced, fluent speech (such as in Eklund [2004]). However, other work may require a very granular definition of a prolongation, such as in the field of forensic phonetics.

This thesis pays particular attention to the Taxonomy of Fluency for Forensic Analysis (TOFFA) presented by McDougall and Duckworth [2017]. TOFFA was designed to allow for the analysis of disfluencies with a high level of phonetic detail in order to compare the interspeaker variation of how these disfluencies occur within speech. Their schema is presented in detail below:

1. Unfilled Pauses

In order to be classified as an unfilled pause according to TOFFA, a silence must last more than 200ms and occur within a single speaker's turn. These unfilled pauses can be further categorised into two different types:

1. Unfilled pauses at grammatical boundaries: *He came [unfilled pause] and I left shortly after*
2. Other unfilled pauses: *He came and I [unfilled pause] left shortly after.*

In addition, unfilled pauses found between the end of one speaker's turn and the beginning of another's are not counted in the classification. However, if the beginning of another speaker's turn starts with a filled pause and is then followed by an unfilled pause, that unfilled pause can be counted.

2. Filled Pauses

TOFFA defines a filled pause as a vowel that may or may not be followed by a nasal, and identifies three main groups of filled pauses:

1. Standalone centralised vowels
2. Centralised vowels accompanied by a nasal
3. An "other" category consisting of other vowels or even no vowels such as [m:].

There is no distinction between long or short forms of filled pauses according to TOFFA, although there is evidence from other studies (such as Kjellmer [2003]) that these may have different functions in speech.

3. Repetitions

There are three main types of repetitions identified in TOFFA:

1. Whole word repetitions: *This- this one is mine*
2. Part word repetitions: *Th- this one is mine*
3. Phrase repetitions: *This one- this one is mine*

TOFFA also includes a “multiple repetition” marker to annotate instances where a repeated segment is repeated more than once: *This- this- this one is mine*. The multiple repetition marker is used in addition to one of the three categories described above, resulting in annotations such as: *This- this- this [whole word repetition] [multiple repetition] one is mine*.

4. Prolongations

A prolongation according to **TOFFA** is one or more speech segments in a word that are prolonged. This does not include prolonged filled pauses. Prolongations are divided into three different subcategories depending on the kind of phone that is being prolonged:

1. Prolongations of a vowel, nasal, lateral, or approximant
2. Prolongations of a fricative
3. Prolongations of plosive not including aspiration

5. Interruptions

An interruption is defined as a change that the speaker makes by interrupting themselves as they are talking. There are two main types:

1. Interrupted phrases: *I was going to come- I was going to leave early*
2. Interrupted words: *ye- no*

TOFFA does not include “speech errors” such as phonetic additions, deletions, or substitutions as unique disfluency classifications. Rather, these would fall under the “interruptions” category.

This taxonomy is comprehensive and enables a very detailed disfluency analysis. However, this level of classification would be unnecessary for studies that do not require such levels of detail. For example, in the methodology of the present analysis as described below, it

is of no consequence to us whether a filled pause presents with a nasal or not. An analysis of disfluencies in fine phonetic detail is not appropriate in this case. Rather, the fact that a filled pause is present is the main concern.

An earlier, yet even more detailed, taxonomy of disfluencies was presented in Dell [1986] alongside a theory of retrieval processes in speech production. This taxonomy, which focussed primarily on “slips of the tongue”, consisted of an amalgamation of disfluency definitions from Stemberger [1982], Fromkin [1971], Garrett [1975], Shattuck-Hufnagel and Klatt [1979], and Dell and Reich [1981]. Dell’s taxonomy directly contrasts with TOFFA as it does not include classifications for phenomena such as pauses and prolongations, but does describe in high levels of detail the different kinds of speech errors. In fact, Dell identifies three main categories of errors (sound errors, morpheme errors, and word errors) which are further broken down into 35 different subcategories. As the present study does not examine speech errors in such detail it is not necessary to give an example of each of the 35 different categories. However, some examples of different speech errors are given below to highlight the specificity of Dell’s different speech error categories:

1. Anticipatory sound error: *reading list - leading list*
2. Noncontextual substitution sound error: *department - jepartment*
3. Anticipatory morpheme error: *my car towed - my tow towed*
4. Noncontextual substitution morpheme error: *conclusion - concludement*
5. Anticipatory word error: *sun is in the sky - sky is in the sky*
6. Noncontextual substitution word error: *pass the pepper - pass the salt*

In contrast to the granularity found in the taxonomies presented by McDougall and Duckworth [2017] and Dell [1986], Stasak et al. [2021] take a more high-level view of disfluencies. The purpose of Stasak et al.’s study was to investigate whether a group of healthy controls would produce fewer overall disfluencies compared to groups of participants that had scored highly on a depression rating scale. The authors designed their disfluency categories with a view to being able to add them as a features into their automatic system

for classifying suicidal behaviour. Therefore, the disfluencies had to be relatively simple to identify in order to accurately recognise them automatically. The decision to keep disfluency categories simple enough to identify automatically was carried through into the schema proposed in this thesis. They present two main categories of disfluency that encompass a broad range of phenomena:

1. Hesitations

Stasak et al. define this category as including any unnatural abrupt pauses, false starts, word or phrase repetitions, or abnormal prolongations. In order to classify the hesitations, the authors take into consideration the individual speech rate of each participant, noting that some participants spoke slower than others. However, there is no given threshold of duration for the classification of hesitations and hesitations are judged subjectively according to each participant's own idiosyncratic speech behaviours.

2. Speech Errors

Speech errors in this case are any deviations from the correct pronunciation of a given target word. This induces phonological additions, substitutions, and deletions alongside word and phrase repetitions and slips of the tongue. Although there is not much detail describing the conditions which must be met in order to classify something as a speech error, the authors do note that "incorrect" pronunciations that occur due to a regional accent or dialect were not counted as disfluencies.

One of the more comprehensive studies of disfluencies occurring specifically in patients with Alzheimer's Dementia (AD) was presented in 2018 by Cera et al.. This study into Portuguese speech classifies speech errors according to whether the error is phonetic or phonological in nature. Below is a description of each category with an example of each group taken from the study paper (in Portuguese):

1. Phonetic (motor) manifestations: this group includes occurrences of prolonged vowels and consonants, schwa insertion between syllables, schwa insertion between consonant clusters, and phone repetitions. For example: /'pedara/ instead of /'pedra/.
2. Phonological (linguistic) manifestations: includes substitutions of vowels, substitutions

of consonants, whole word repetitions, syllable repetitions, anticipatory speech errors, and phoneme substitutions. For example: /'riza/ instead of /'rizu/.

3. Phonetic or phonological manifestations: attempts at the correct pronunciation of a target word that do not cross the phoneme boundary but are produced with place, manner, or voice deviations. For example: /'kasi/ instead of /'klasi/. This group can include omissions (as in the example), additions and substitutions.

Whilst it is valuable to know whether certain disfluencies are of a phonetic or phonological nature, particularly when considering potential speech therapy treatments, the distinction between phonetic or phonological disfluencies is not made in the analysis presented in this chapter. The present study is interested in the frequency of disfluencies across different levels of CD, as opposed to investigating whether the disfluency is caused by phonetic-motor disorders or language disorders. It would certainly be interesting to investigate whether disfluencies produced at different levels of CD are predominantly phonetic or phonological in nature, but that was beyond the scope of this thesis.

A more recent study specifically investigating language in dementia was presented by Panesar and de Alba [2023]. Similarly to the work from Stasak et al. [2021], Panesar and de Alba investigated disfluencies with a view to using disfluency information in an Machine Learning (ML) model, although this time the model was designed to detect early stages of CD. Whilst Panesar and de Alba are not specifically interested in disfluencies and are instead more interested in language production parameters, their study does include an analysis of phenomena that could be considered disfluencies. Some examples of such phenomena are described below:

1. Repetition of words
2. Filled sound pauses
3. Timings of pauses (short, medium, or long)

Other measures that Panesar and de Alba investigate that are not related to disfluencies include measures of lexical richness, the frequency of pronoun use, and the completeness of sentences. Even though their study does not consider many different types of disflu-

ency, the authors do include measures of severity for each of the disfluencies they look at. For example, filled sound pauses can be classified as fluent, partial, or poor according to the length of each pause (with short pauses being given a “fluent” label, long pauses given a “poor” label, and medium pauses given a “partial” label). Each kind of disfluency is given a severity rating which correlates to a score, and by the end of the study the participants with a higher score were those participants with the fewest instances of disfluencies (alongside more lexically diverse speech, and more frequent pronoun use etc). Higher scores for language production also correlated with less severe levels of CD. Although the idea of a severity score is not used in the analysis presented in this chapter, it did inform some further work described in Chapter 4.

Given the different approaches to defining disfluencies as highlighted above, this thesis presents a new schema for categorising disfluencies for the purposes of investigating different levels of CD. The *DisCo* taxonomy as presented below combines the clearly defined categories and time-defined classifications of *TOFFA* with a less fine-grained approach to classifying speech errors as found in *Stasak et al.’s 2021* study but follows a similar approach to *Dell [1986]* in treating phenomena such as phonetic deletions, additions, and substitutions as their own individual classifications. Following the *DisCo* schema throughout this thesis ensures uniformity in the terminology used throughout the analysis chapters, and makes disfluency classification decisions as clear as possible.

3.3 Methodology

This section commences with an overview of the creation of the *DisCo* schema used throughout this thesis to perform manual disfluency analyses from a quantitative perspective. The data used in this first manual disfluency analysis is then presented, along with a description of the process of collecting disfluency information and a brief over of the statistical methods used as part of the analysis.

3.3.1 The Disfluencies in Cognition Schema

The aim of this first analysis was to investigate whether different kinds of disfluencies appear in speech at different levels of CD. This would serve as a precursory study to investigating whether disfluency information could help improve the accuracy of Automatic Cognitive Decline Classification (ACDC) systems. Therefore, a balance was struck between capturing a high enough level of detail whilst still being sufficiently simple to implement in the automatic system. The table below outlines the first version of the DisCo schema used in this analysis, which is based largely on the TOFFA framework [McDougall and Duckworth, 2017] discussed in Section 3.2. TOFFA was chosen as the starting point for our schema as it makes very clear what exactly constitutes each different disfluency category. However, as we hoped that we would eventually add this disfluency information into an automatic system, we decided to reduce the level of granularity from that seen in TOFFA. For example, whereas TOFFA differentiates between unfilled pauses at a grammatical boundary and unfilled pauses elsewhere, our schema does not. We did however keep the 200ms threshold presented in TOFFA for classifying pauses and prolongations. McDougall and Duckworth chose the 200ms threshold following from other studies such as Butterworth [1980]. Additionally, a large study of pause durations across five languages, Campione and Véronis [2002] found a trimodal distribution of pauses, with brief pauses being less than 200ms, medium pauses being between 200-1000ms, and long pauses being more than 1000ms. As we were not differentiating between medium and long pauses, we retained the 200ms boundary to differentiate between a short, fluent silence or a disfluent pause. This enabled us to make objective decisions when counting pauses and prolongations, rather than having to rely on subjective judgements which could differ from person to person.

Unlike TOFFA, we excluded most classifications that relied on small phonetic differences. For example, according to DisCo a filled pause is classed as a filled pause regardless of whether it contains a nasal or not. Some of our disfluency groups do however go into slightly higher levels of detail by containing subcategories, such as the repetitions group. This follows TOFFA by categorising repetitions in three distinct ways according

to whether the repeated segment is a whole word, whole phrase, or part of a single word. However, in instances where multiple repetitions occur concurrently **TOFFA** adds a [multiple repetition] marker in addition to the tag describing what kind of repetition has occurred, such as:

This- this- this- [whole word repetition] [multiple repetition] this one is mine

However, the **DisCo** schema makes use of the number of instances a repetition happened, so instead of tagging as a multiple repetition, each single repetition is marked in order to be counted:

This- [word repetition] this- [word repetition] this- [word repetition] this one is mine

Using the **DisCo** schema, we can tell that the sentence above contains three whole word repetitions, whereas **TOFFA** would only tell us that the repetition observed was a whole word repetition, and that there were multiple cases observed. There is no way of knowing exactly how many repetitions occurred in that particular sentence according to **TOFFA**.

As discussed in the previous section of this chapter, **TOFFA** does not include any categorisations for speech errors such as additions, deletions, or substitutions. This is stated to be due to a lack of such phenomena in the data used to create **TOFFA**. However, in **Stasak et al. [2021]** the number of speech errors (including those mentioned above) observed proved to be significantly different between their healthy control group and their depressed groups. Even though **Stasak et al.**'s study was not concerned with **CD**, no work could be found that had included such speech errors in a disfluency analysis of demented speech. Therefore, speech errors were included as a group of subcategories within this study to investigate whether they would appear more frequently than the numbers reported in **Stasak et al.**'s work.

It is important to note that the process of annotating the disfluencies in this experiment started as a somewhat iterative process. Occasionally, a disfluency was encountered that was not identifiable with the original iteration of **DisCo**. In such cases, the schema was

updated to reflect the new disfluency category, and recordings that had already been transcribed were revisited to make sure the transcriptions adhered to the updated version of the schema. For this reason, the malapropism category of disfluency was included to account for instances where an incorrect expression or word was used (such as in the recording from Participant 0277 in the phrase:

I look back on my school days with great fondness and, um [fp], [ufp] enjoy [mal]

In this instance, [mal] was chosen to categorise the use of “enjoy” in place of “enjoyment” or other suitable noun, as would be expected in normal, fluent speech. Table 3.1 demonstrates the DisCo schema used in this experiment. Following this is a more detailed description of each category, supported by examples taken from the data used in this study.

Disfluency	Description	Annotation
Unfilled pause	Silence >200ms	[ufp]
Filled pause	Typically a central vowel that may be followed by a nasal lasting >200ms E.g., [ə:m] or [(ə ^h)]	[fp]
Repetition	Part word repetition E.g., “there’s a ca- cat”	[pwrep]
	Whole word repetition E.g., “there’s a cat- cat”	[wrep]
	Phrase repetition E.g., “there’s a cat- there’s a cat”	[phrep]
Prolongation	Phone lengthened to >200ms	[pro]
Speech Error	Deletion - phone is deleted	[del]
	Substitution - phone is changed to something else	[sub]
	Addition - phone is added	[add]
	Malapropism - unrelated substitutions for a word	[mal]
Repair	Noticing an error has been made and going back to correct it	[repa]
Non-speech events	Lip smacking, coughs, laughs, sighs, deep breaths, etc	[nse]

Table 3.1: The DisCo Taxonomy.

Unfilled Pauses [ufp]

Unfilled pauses are any silences occurring within a participant’s utterance that lasts longer than 200ms. In the case of this analysis, an utterance is defined as a single turn by a speaker regardless of how many words or sentences comprise that turn. This usually results in each utterance being an answer to a question. The utterance ends when the participant decides that they have answered the relevant question in as much detail as they would like, and presses a button to move on to the next question. The initial silence that comes before a participant starts their response (even if the response is started with a filled pause instead of speech) is not counted as an unfilled pause, as this is not included in the utterance. The silences at the ends of an utterance (when the participant is deciding

whether or not they have finished talking) were counted as unfilled pauses (providing they met the 200ms threshold). Additionally, in instances of chains of filled and unfilled pauses (e.g., [fp] [ufp] [fp] or [ufp] [fp] [ufp]), each pause is counted providing it meets the 200ms threshold.

Filled Pauses [fp]

A filled pause is any silence-filling noise that is neither part of fluent speech nor a vocalisation related to coughs, throat clearing, laughter, et cetera. These can be vowels on their own, such as [ə:], or vowels followed by a consonant (typically a nasal such as [əm]).

Repetitions

DisCo classifies repetitions in three different ways, depending on what exactly is being repeated:

1. Part Word Repetitions ([pwrep])

Th- [pwrep] there's a cup on the table

2. Whole Word Repetitions ([wrep])

There's- [wrep] there's a cup on the table

3. Phrase Repetitions ([phrep])

There's a cup- [phrep] there's a cup on the table

As in TOFFA, contractions are counted as one word hence why the example in 2 above is classed as whole word repetition rather than a phrase repetition.

A phrase repetition according to DisCo is a repetition of something that is more than a sole word. This could be a [wrep] + [pwrep] as in:

I ha- [phrep] I have a cup

Or this could be a situation where multiple whole words are repeated sequentially:

I have a- [phrep] I have a cup

In cases of multiple repetitions, each repetition is marked separately. The final repetition is left unmarked as it is considered to be start of the fluent portion of the sentence being spoken:

I want- [wrep] want- [wrep] want- [wrep] want to go to the shop

It should be noted that in the case of part-word repetitions it is difficult to be sure that the repeated section is definitely a repetition and not a repair. For example, the “*Th-*” in the example above could theoretically belong to a number of different “*Th-*” words (e.g., “*this*”, or “*them*”). We make the assumption that if the incomplete segment matches the next fluent word then we have a part word repetition; if the incomplete segment varies from the next fluent word (i.e., *Thi- there’s a cup*) then we class it as a repair (see below).

Prolongations [pro]

A prolongation is any phone that is part of a word that is prolonged for more than 200ms (again, this threshold remains unchanged from [TOFFA](#)). This is most commonly vowels but could be any phone that meets the threshold.

Speech Errors

This is another broad category that includes a range of different disfluencies, each with their own annotation tag:

1. Deletions ([del]):

specific → *pecific* [del]

2. Additions ([add]):

optimal → *noptimal* [add]

3. Substitutions ([sub]):

specific → *sbecific* [sub]

4. Malapropisms ([mal]):

(in describing a picture of some dogs) *There’s five cats* [mal]

Repairs [repa]

The [repa] tag is added to signify a speech error that has been identified by the speaker which they then go on to correct:

I got in the car- no I got in the bus [repa]

Non-Speech Events [nse]

Whilst non-speech events are not classed as disfluencies as such, [DisCo](#) includes an [nse] tag which can be used to indicate vocalisations that are not speech but that might confuse automatic systems attempting to measure the durations of speech segments, such as laughs or coughs. By tagging such instances it becomes easier to remove them from calculations of the total speaker locution time, ensuring that only speech is being counted.

Table 3.2 illustrates how different speech error phenomena could be classified according to different disfluency taxonomies compared to the classification according to our schema.

3.3.2 Data

This experiment uses existing data that was collected as part of research testing the efficacy of an early version of what is now known as the [CognoSpeak](#) system. The [Intelligent Virtual Agent \(IVA\)](#) dataset was collected between 2016 and 2019 as part of a joint project between the University of Sheffield and the Royal Hallamshire Hospital [[Mirheidari, 2018](#)]. It includes recordings of 93 participants conversing and completing a number of different language tasks with a virtual agent. Each recording is structured as follows:

1. Interview: during the interview section of the recordings, participants are asked a set of predetermined questions by the virtual agent. Participants are free to respond for as long as they like, and simply press a button on a computer keyboard to move to the next question. The questions ask for a range of information and are based around the questions a human doctor would ask a patient in a memory clinic setting. Patients are asked a mixture of simple questions (such as *“tell me what problems you have had with your memory recently”*), and compound questions (such as *“who is most worried about your memory, you, or somebody else? And what did you do last weekend?”*).
2. Fluency Tests: after the interview portion of the recording, participants are asked to complete two fluency tests. The first is the [Semantic Verbal Fluency \(SVF\)](#) test in which participants are given 60 seconds to name as many animals as they can. The second is the [Phonemic Verbal Fluency \(PVF\)](#) test, which asks participants to name

Error	Type According to Dell [1986]	Type According to Cera et al. [2018]	Type According to DisCo
Sim swimmers sink	Phoneme anticipation	Substitution of vowel	Substitution
Some swummers sink	Phoneme perseveration	Preservation of phoneme	
Some swinkers sink	Cluster anticipation	Anticipation of phoneme	
Sim swummers sink	Phoneme exchange	Substitution of vowel + Substitution of vowel	Substitution + Substitution ([s λ m] \rightarrow [sim]) + ([swim α z] \rightarrow [sw α m α z])
Some simmers sink	Phoneme deletion	Omission	Deletion
Swum simmers sink	Phoneme shift	Transposition of phonemes	Addition + Deletion ([s λ m] \rightarrow [sw λ m]) + ([swim α z] \rightarrow [sim α z])
Some sinkers swim	Stem exchange	-	Malapropism
Some swimmers swim	Stem perseveration or word substitution	-	
Some swimmers drown	Word substitution	-	

Table 3.2: A comparison of errors according to different disfluency schema, adapted from Dell [1986].

as many words that begin with the letter “P” as they can in 60 seconds, excluding proper nouns.

3. Picture Description Task: the final task for participants is to describe a picture in as much detail as they can. Participants are shown the Cookie Theft picture, and given as much time as they need to describe what they can see.

For this initial experiment, only the interview section of each recording was used. As participants were given the freedom to be as concise or as descriptive as they wished during

the interview, the length of these recordings varied between participants. Spontaneous speech was the first choice for investigation as many studies have reported that read or planned speech often has a slower speech rate with fewer disfluencies (see [Howell and Kadi-Hanifi \[1991\]](#); [Pinto et al. \[2013\]](#); [Lickley \[2017\]](#)). Additionally, this analysis is more robust and relevant by using speech samples that are representative of speech that would be analysed in a real-world setting; in this case a memory clinic.

Whilst the virtual agent uses a predetermined set of questions for the interview portion of the task, our IVA subset contains some slight variations in the wording of some questions, depending on which iteration of the virtual agent was used during the data collection process. This could potentially influence the way participants respond to the questions. However, the analysis in this chapter is not concerned as much with the content of the responses given by the participants, but rather the disfluencies occurring in whatever they say. Therefore, the variations in the question formats were disregarded. [Appendix C](#) contains the sets of questions asked by the virtual agent to all participants, and highlights any small deviations between questions.

Each participant chosen for this study belonged to one of four cognitive groups; Healthy Controls (HCs), people with Functional Memory Disorder (pwFMD), people with Mild Cognitive Impairment (pwMCI), and people with Neurodegenerative Dementia (pwND). As this analysis was investigative in nature, four classes were chosen for analysis rather than the “Alzheimer’s Disease vs. Healthy Control” that is commonly seen in these kinds of studies. This allowed us to not only compare disfluencies at different levels of neurodegeneration, but also to compare to speech problems that are psychological in nature as in the case of FMD. There is little work investigating disfluencies in speech from people with FMD, so our study is an initial investigation into this. Some participants in the original set of 93 recordings had been interviewed twice. In these cases, only the initial interview was included in the data for this experiment. The resulting subset of data used for this experiment consisted of 55 recordings from different participants. [Table 3.3](#) shows the breakdown of participants according to their subject groups. The standard deviation (σ) is included to demonstrate the variance in each group [[Livingston, 2004](#)].

Subject Group	No. Participants	Male: Female	Age		No. Fluent Words in Recording		MMSE score	
			Mean	σ	Mean	σ	Mean	σ
HC	15	6:9	69.5	8	654.7	511.2	28.7	0.7
FMD	14	5:9	55.1	8.4	400.4	385.6	27.4	2.1
MCI	14	10:4	63.6	8.6	499.4	448.3	26.5	1.1
ND	12	10:2	68.1	7.9	490.7	500.8	23.6	4.7

Table 3.3: Participant Information for the first manual disfluency analysis, including number of participants, gender, age, number of fluent words per recordings, and participant Mini Mental State Examination scores. Statistics are given as group means and standard deviation.

One benefit of using this particular dataset is that there is already existing information surrounding the kinds of features that have been found to be useful in discriminating between the different cognitive groups (see [Mirheidari et al. \[2017\]](#); [Mirheidari et al. \[2019\]](#); [O’Malley et al. \[2020\]](#); [Beavis et al. \[2021\]](#); [Walker et al. \[2021\]](#)). These studies investigate a mixture of features, such as Conversation Analysis (CA) features (number of participant turns, average number of low frequency words, average number of topics discussed), lexical features (average number of prepositions, vowels, determiners, conjunctions), and acoustic features (average pitch, harmonics-to-noise ratio). It was therefore clear that disfluency information other than pauses had not been tested, and warranted an investigation.

As this data was collected continually over a number of years, many studies using the data have curated their own subsets to fulfill different criteria (such as being balanced for age, gender, or only containing participants diagnosed with Mild Cognitive Impairment (MCI), for example). This present work uses a subset called the IVA60, used previously in studies such as [O’Malley et al. \[2021\]](#). This subset contains 60 participants equally split into the four cognitive groups.

3.3.2.1 Data Considerations

Due to the difficulties surrounding the collection of sensitive data such as health data, this study used existing data that was already available to researchers at the University

of Sheffield. Using this data comes with a number of benefits. As mentioned above, previous research has used the IVA data for investigations, allowing the disfluency information found in this analysis to be added to the results from prior studies to create a more comprehensive overview of the data. Using existing data also benefits the many participants who kindly volunteered their time to contribute to this dataset. Whilst cognitive tests do not pose large risks to participants, it would also be unfair to ask participants to participate in more studies when we already have a good selection of existing data. The collection of this data started in 2016 as part of an earlier research project investigating the potential use of an intelligent virtual agent as a method of screening patients in a memory clinic [Mirheidari, 2018].

However, there are also a number of difficulties that come with using an existing dataset, and the IVA data is no different. What follows is a short description of some of the issues encountered when working with this particular dataset, and the steps taken to mitigate any negative effects.

Purpose of Collection

This data was collected in order to train an ML model to recognise different stages of CD through speech analysis. As part of a multidisciplinary project, lots of information was collected about different aspects of the participants' lives. For example, scores from a range of different memory tests are recorded, along with information from other clinical tests for cognitive decline such as Magnetic Resonance Imaging (MRI) scans and lumbar punctures. However, the amount of collected information varies between participants. Although there is usually a score from a cognitive test for each participant, the actual tests vary. This means that while some people may have recorded Montreal Cognitive Assessment (MoCA) scores, others have Mini Mental State Examination (MMSE) scores. All cognitive scores were converted to an MMSE score for ease of comparison.

Some participants have recorded information relating to any medication they may be taking, but some do not. This is potentially problematic as some medications may interfere with speech or cognitive function. Where possible, participants that were identified as having a comorbidity that is known to affect speech were removed from the subset of data.

However, information regarding comorbidities such as Parkinson’s disease or strokes was recorded inconsistently. This is potentially problematic as such comorbidities may also affect speech, making it difficult to discern whether disfluencies found in this study are a direct result of the effect of the CD each participant is experiencing, or a result of the comorbidity. For more information about how the aforementioned conditions affect speech see Proença et al. [2014], and Vidović et al. [2011].

Some participants have recorded PHQ-9 and GAD-7 scores (rating scales for severity of depression and anxiety, respectively). Those who scored highly on either test were removed from our sample. Again, this is due to the potential effects that depression and anxiety may have on speech (Cummins et al. [2015]; Pope et al. [1970]). However, it should be noted that participants that did not have this information recorded were still included in our sample, in an attempt to retain as many different participants as possible. Other participants that were removed from the sample were those observed to have speech impairments, such as a stammer. In one recording, the participant notes how they have recently been diagnosed with dysarthria. This participant was removed from the sample, as the aim of the present study is to try to identify patterns in disfluencies related to cognitive impairment, not disfluencies that may be associated with pathologies that may not related to CD.

For speech corpora that are used throughout different fields of linguistics such as sociolinguistics and phonetics, it is generally expected that researchers will collect comprehensive metadata concerning each participant’s linguistic experience [Niebuhr and Michaud, 2015]. This includes information regarding place of birth, socioeconomic status, and years of education. However, such detailed linguistic information is not available for participants in the IVA dataset. Broadly speaking, most participants in the original IVA dataset are native speakers of English and are from the greater Sheffield area.

Research has shown that disfluencies vary greatly depending on language proficiency (Temple [2000]; Rieger [2003]; Klapi et al. [2011]; Gurbuz [2017]). In cases where it was clear that a participant was not a native speaker of English (if, for example, they discussed in the recording their time spent growing up in a non-English speaking country)

they were removed from our subset. This allows us to compare results between people with the same language proficiency, reducing the variation in disfluencies that could be caused by factors other than CD.

Technical Difficulties

Although effort was made to keep the recording conditions as controlled as possible, some recordings include a large amount of background noise. As part of this analysis a number of gold-standard disfluency transcripts will be produced, which in turn could be used to train an ACDC system. Gold-star in this case would refer to transcripts that we can rely on for being highly accurate and checked by a human annotator, rather than transcripts that have been created automatically which may contain errors. In order to make the transcripts as accurate as possible it was imperative that the speech in all recordings was clear and easy to understand. Therefore, any instances of background noise that impacted the intelligibility of the speech were removed from the subset of data used for this task. There were also a few recordings in which the virtual agent appeared to malfunction, resulting in cases of the participants being asked the same questions repeatedly or the virtual agent interrupting participants mid-flow. In such instances these recordings were removed.

In total, five recordings were removed from the IVA60 subset due to the reasons above: two participants were found to be non-native speakers (one of whom also talked about having been diagnosed with a speech impediment) and the other three due to technical problems with the recordings. This brought the total number of recordings used in this analysis to 55.

Other Considerations

Before work on this experiment began, a decision was made to try to include as many different participants as possible in order to have as many gold-star disfluency transcripts for each subject group as possible. However, this came at the expense of ensuring that subject groups were balanced and representative of real-world data. As this study is investigating disfluencies according to level of CD and not according to gender or age, including as many possible examples of each CD group was prioritised over ensuring age

and gender were balanced in each group.

3.3.3 Disfluency Collection Process

Each recording was individually loaded into Praat [Boersma and Van Heuven, 2001], and the recordings were trimmed to include only the interview portion of each recording. Four text grid tiers were used to label the data as follows:

1. An orthographic transcription of the questions being posed by the virtual agent.
2. An orthographic transcription of the responses from the participant. All filled pauses were recorded here as “er”, and all other deviations from a fluent word were marked with a dash (for example, whole word repetitions would be recorded here “like- like so”).
3. Disfluency labels according to DisCo.
4. Speech of Accompanying Persons (APs) or researchers, or any other notes.

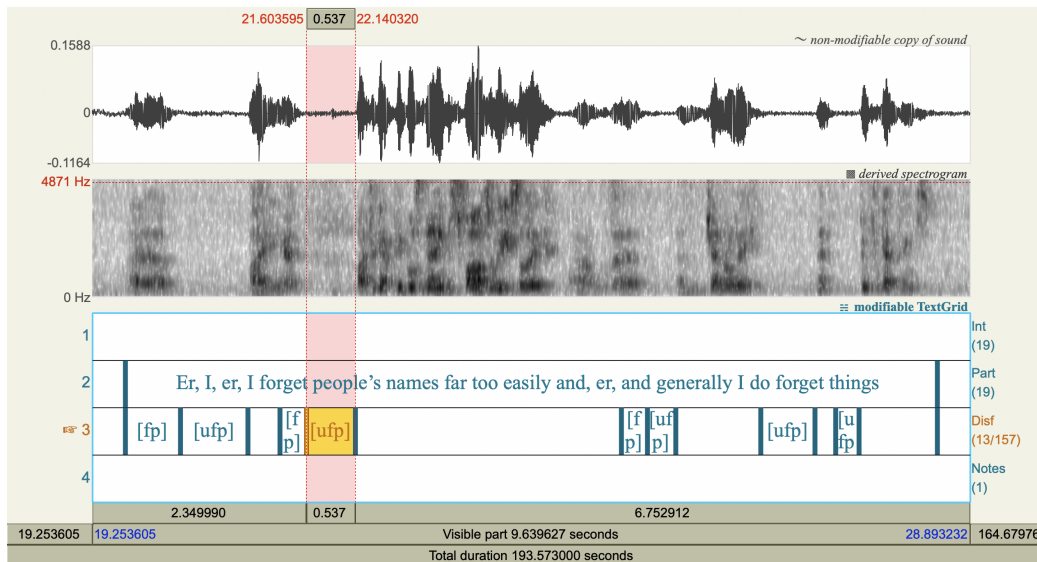


Figure 3.1: A view of the different Praat text grid tiers used in the manual disfluency analysis, with an unfilled pause highlighted.

Figure 3.1 illustrates the resulting text grids, and how each disfluency was tagged. By recording the disfluencies this way it was possible to quantify not only the number of each disfluency category present in each recording but also the durations of pauses and

prolongations. Each recording was listened to a minimum of four times. All transcriptions were completed by the researcher without the knowledge of which cognitive group the participant belonged to in an attempt to avoid any bias. Thorough notes were kept about each recording, and any questions or discrepancies were checked with other trained phoneticians. With help from collaborators in the department of computer science, a Python script was created to collect all the disfluency information from the text grid files.

3.3.4 Statistical Methods

Once all transcriptions had been completed and double checked, disfluency information was normalised by the number of fluent words in each recording to account for the range of different recording lengths. Fluent words were counted using a Python script which counted all words in the second text grid tier, minus “er” or any instances of a dash followed by a space to remove non-fluent words (e.g., “*the-* ”).

The first step of the statistical analysis was to find out whether the disfluency data were normally distributed. This was calculated in R using a Shapiro-Wilks test. The majority of the disfluencies were not normally distributed across the dataset, so the Kruskal-Wallis test for significance was used. In instances where the data were normally distributed, ANOVA was used to test for significance. Once significance had been established, a Dunn test with Bonferroni correction was applied to examine which were the significant groups, and to account for the multiple comparisons. All statistics (apart from averages) are performed on scores normalised by the number of fluent words spoken by each participant in their recording. Normalising in this way allows a more direct comparison of disfluency rates from person to person, independent of how long each participant spoke for. In early disfluency studies, it was common to assess disfluency rates per 100 words (see [Mahl \[1959\]](#), and [Faure \[1980\]](#)). The present study follows the method of [Fox Tree \[1995\]](#) in excluding disfluencies from this measure as it is difficult to tell the difference between a fluent and disfluent filled pause. Therefore, we normalise according to fluent words and report disfluency rates per 100 *fluent* words.

Participants in this study have a recorded cognitive test score. This is usually an **MMSE** score, but some patients have **MoCA** scores instead. In order to facilitate an easier comparison, all scores are converted to **MMSE** scores according to the conversion table from [Matías-Guiu et al. \[2018\]](#). This method of conversion has demonstrated high reliability when tested on a large scale ($n = 500$).

3.4 Results

This section details the findings from this first manual disfluency study which uses the **DisCo** schema to investigate the frequency and duration of disfluencies in speech. These are findings that are statistically significant or hold potential for being used as a diagnostic aid. Please see [Appendix A](#) for a table consisting of all findings.

3.4.1 Total Disfluency Rates

This analysis revealed high total disfluency rates across all cognitive levels investigated. Even our healthy controls exhibited much higher rates than have previously been reported (around six disfluencies per 100 words not including unfilled pauses). These rates can be found in [Table 3.4](#).

Cognitive Group	Total Disfluency Rate	Disfluency Rate Minus Unfilled Pauses
HC	27.8	13.7
FMD	34.3	13.5
MCI	43.5	20.8
ND	53.7	27.1

Table 3.4: Comparison of total disfluency rates and disfluency rates excluding unfilled pauses across the four different cognitive groups. Average rates are calculated per 100 fluent words.

3.4.2 Number of Fluent Words

Although no statistically significant difference was observed between the median number of fluent words produced by the different cognitive groups, we found that there were

participants in each of the three cognitively impaired groups who spoke a lot more than the other members of their respective cohorts (see Figure 3.2). For example, whilst the median number of words produced by the participants in the Neurodegenerative Dementia (ND) group was 334.5, one participant (Participant 221) produced 2003 words in their interview with the IVA.

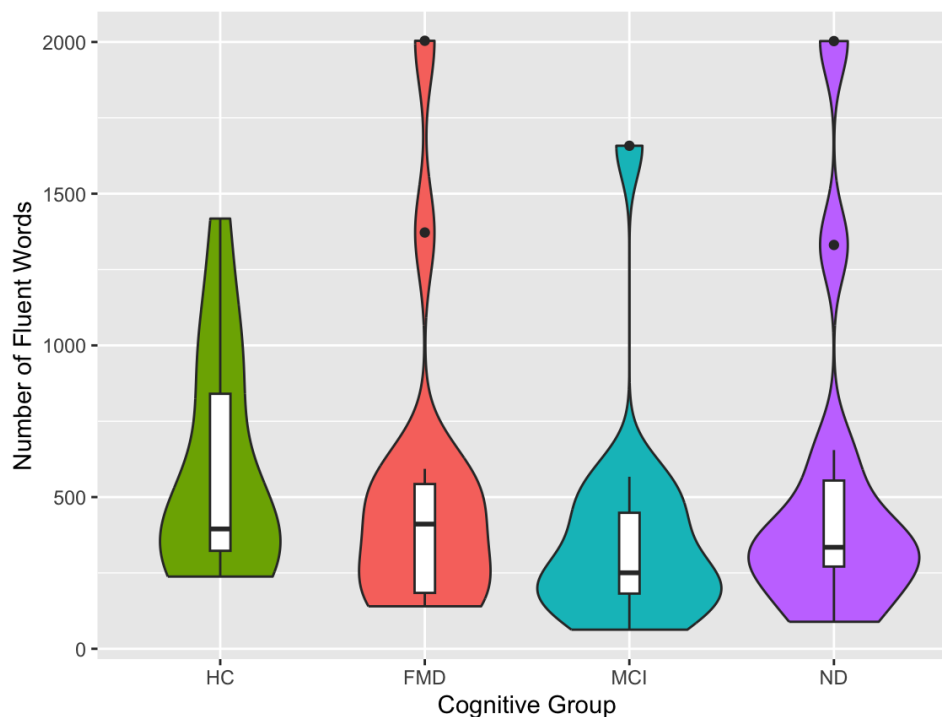


Figure 3.2: Violin plot showing the number of fluent words according to cognitive group.

3.4.3 Unfilled Pauses

3.4.3.1 Number of Unfilled Pauses

Figure 3.3 is a violin plot showing the number of unfilled pauses (per 100 fluent words) per cognitive group. Violin plots are used throughout this thesis as a method of visualising data rather than box plots, as violin plots allow us to visualise the distribution of the data within each cognitive group whilst box plots do not.

A Kruskal-Wallis test indicated that there was a significant difference in the number of unfilled pauses per 100 fluent words across the four different cognitive groups, $\chi^2(3, N = 55) = 19.77, p < .001$. The Kruskal-Wallis test was chosen as the data was found to be

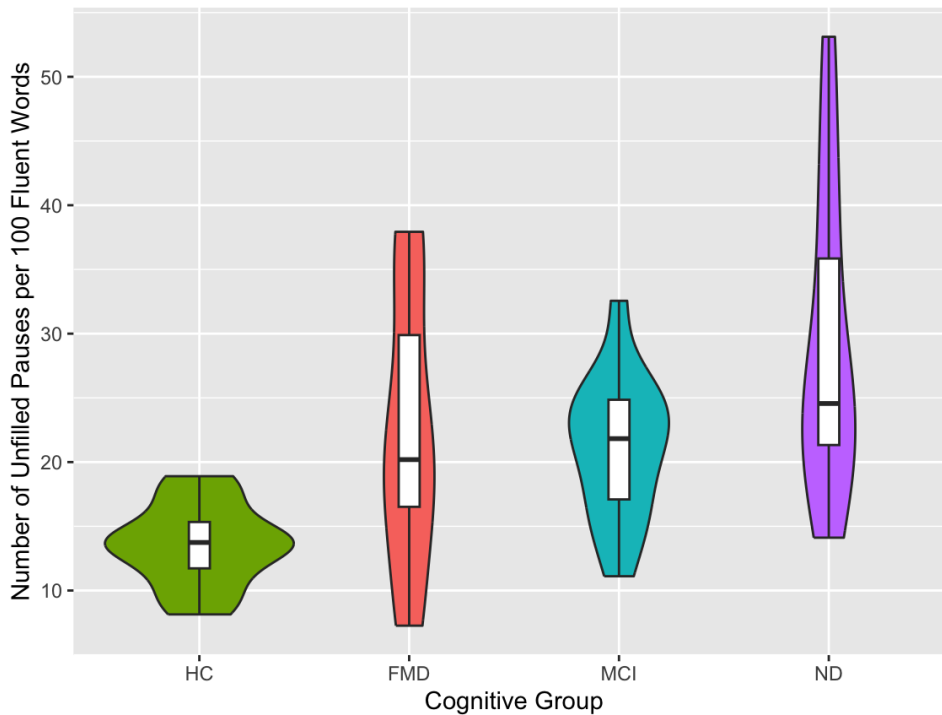


Figure 3.3: Violin plot showing the number of unfilled pauses per 100 fluent words according to cognitive group.

not normally distributed in a Shapiro-Wilkes test ($W = 0.91$, $p = <.001$). The median number of unfilled pauses per cognitive group was 13.74 for HCs, 20.19 for FMDs, 21.82 for MCIs, and 24.56 for NDs. Post-hoc comparisons using Dunn’s method with a Bonferroni correction for multiple tests indicated that the median number of unfilled pauses from people in the HC group was significantly smaller than the other three cognitive groups ($p = .03$ for FMD, $p = .01$ for MCI, and $p = <.001$ for ND). However, there were no significant differences in the number of unfilled pauses found between the FMD, MCI, and ND groups.

3.4.3.2 Average Length of Unfilled Pauses

Figure 3.4 shows the average length of unfilled pauses per cognitive group. Results from our analysis support findings from other studies (such as Singh et al. [2001]) that pause duration can be significantly different between control groups and cognitively impaired groups. A Kruskal-Wallis test indicated that there was a significant difference in the length of unfilled pauses across the four different cognitive groups, $\chi^2(3, N = 55) = 10.90$, p

= $<.01$). The Kruskal-Wallis test was chosen as the data was found to be not normally distributed in a Shapiro-Wilkes test ($W = 0.78$, $p = <.001$). The median average length of unfilled pauses (in seconds) per cognitive group was 0.55s for HCs, 0.89s for Functional Memory Disorders (FMDs), 0.82s for MCIs, and 0.98 for NDs. Post-hoc comparisons using Dunn's method with a Bonferroni correction for multiple tests indicated that the only significant difference was between the lengths of unfilled pauses in the HC group and the ND group, with the average length of unfilled pauses in the HC group being significantly shorter ($p = <.009$).

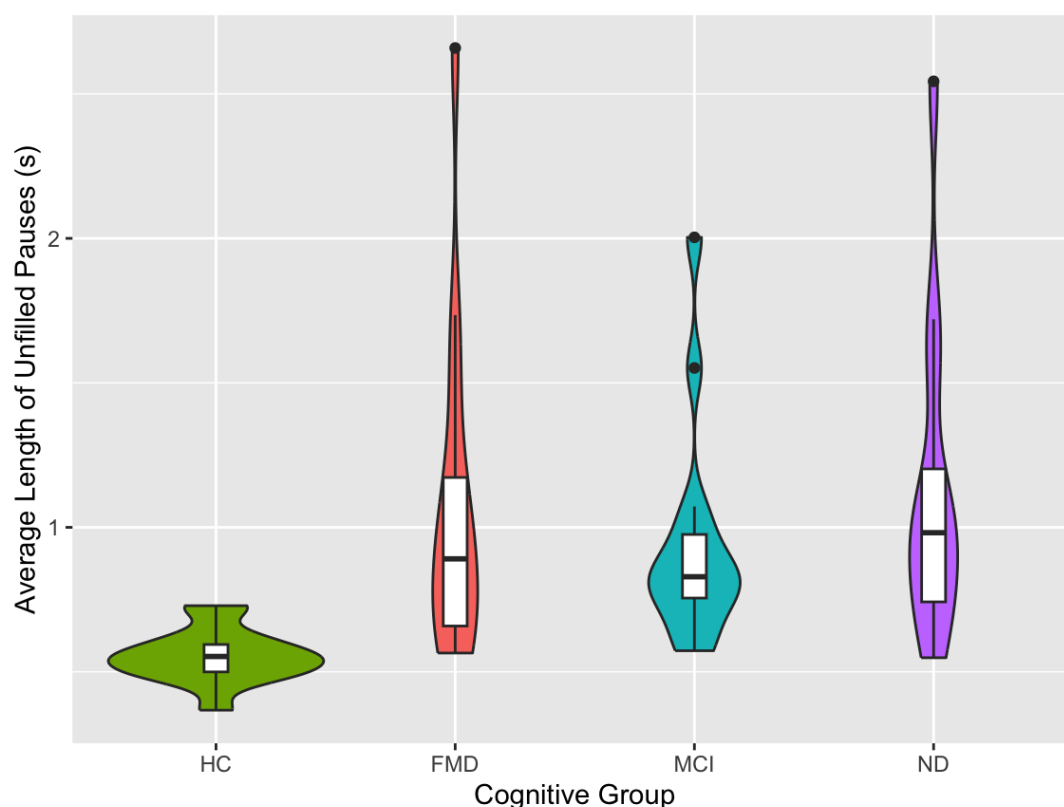


Figure 3.4: Violin plot showing the average lengths of unfilled pauses according to cognitive group.

3.4.4 Filled Pauses

3.4.4.1 Number of Filled Pauses

As demonstrated in Figure 3.5, our analysis did not find a significant difference in the number of filled pauses across the four cognitive groups ($p = .4$). However, there is a

slight trend of an increasing number of filled pauses as cognition worsens as shown in Table 3.5:

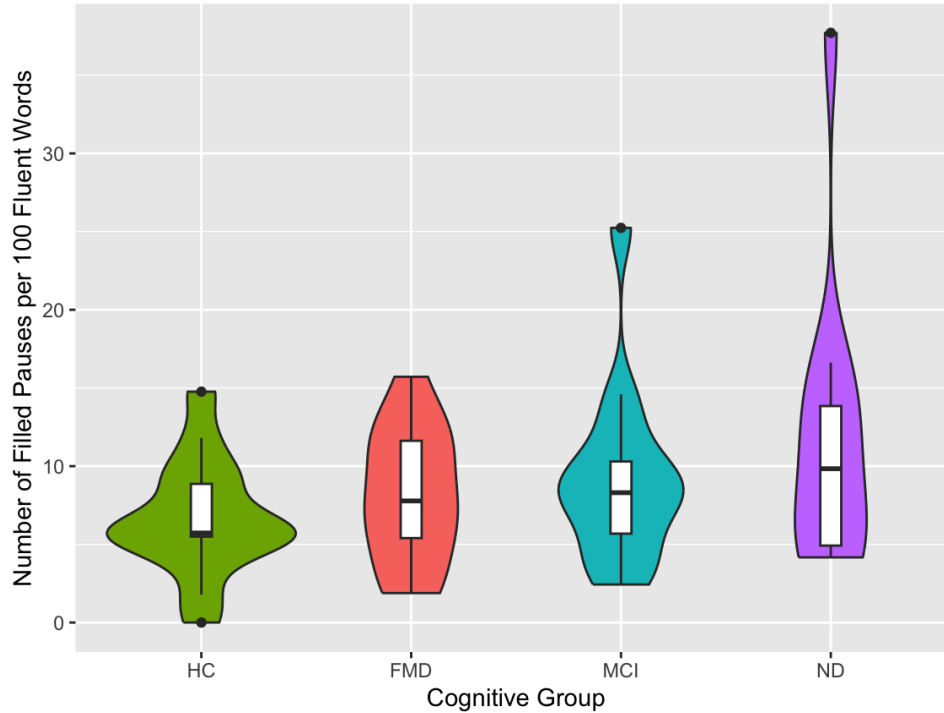


Figure 3.5: Violin plot showing the number of filled pauses per 100 fluent words according to cognitive group.

	HC	FMD	MCI	ND
Median Number of FP per 100 Fluent Words	5.7	7.7	8.3	9.8

Table 3.5: Comparison of the median number of filled pauses per 100 fluent words according to cognitive group.

3.4.4.2 Average Length of Filled Pauses

Although there were no significant differences in the number of filled pauses across cognitive groups, this study did find a significant difference in the average lengths of filled pauses. A Kruskal-Wallis test indicated that there was a significant difference in the length of filled pauses across the four different cognitive groups, $\chi^2(3, N = 55) = 11.32$, $p = .01$). The median average length of filled pauses (in seconds) per cognitive group

was 0.45s for HCs, 0.57s for FMDs, 0.54s for MCIs, and 0.55 for NDs. Post-hoc comparisons using Dunn’s method with a Bonferroni correction for multiple tests indicated that the only significant difference was between the lengths of filled pauses in the HC group and the FMD group, with the average length of filled pauses in the HC group being significantly shorter ($p = .01$) as demonstrated by Figure 3.6. Both the number and the average lengths of filled pauses were smaller across all cognitive groups when compared to the unfilled pauses.

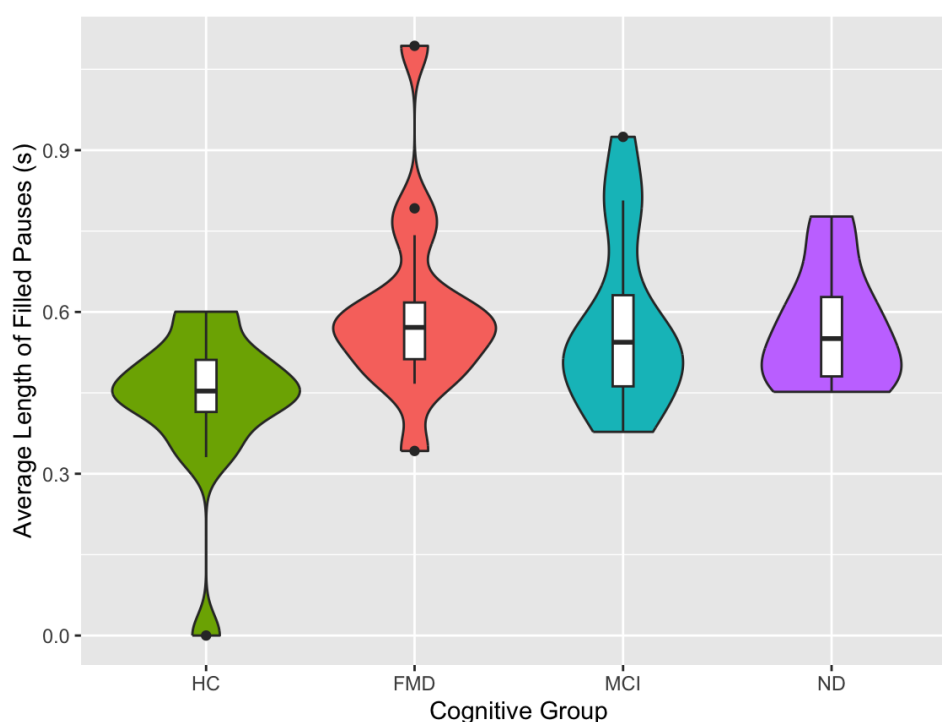


Figure 3.6: Violin plot showing the average lengths of filled pauses according to cognitive group.

3.4.5 Pause-to-Speech Ratio

Although pause to speech ratio is not technically part of the DisCo schema, it is easily calculated with the data resulting from a DisCo analysis. As pause-to-speech ratios are commonly investigated in studies into speech in CD, it is included as part of this analysis.

Despite FMD being the least “severe” of the cognitively impaired groups, participants in

this group exhibited the highest pause to speech ratio, as demonstrated in Table 3.6.

Cognitive Group	Rate of all Pauses Compared to Speech
HC	0.22
FMD	0.65
MCI	0.54
ND	0.59

Table 3.6: A comparison of the pause-to-speech ratios observed across cognitive groups.

3.4.6 Repetitions

A Kruskal-Wallis test indicated that there was a significant difference in the number of whole-word repetitions per 100 fluent words across the four different cognitive groups, $\chi^2(3, N = 55) = 12.35, p = .006$. The median number of whole word repetitions (per 100 words) per cognitive group was 0.38 for HCs, 0.09 for FMDs, 0.63 for MCIs, and 0.94 for ND. Post-hoc comparisons using Dunn’s method with a Bonferroni correction for multiple tests indicated that the median number of whole-word repetitions from people in the FMD group was significantly smaller than in the ND ($p = .004$) group (as shown in Figure 3.7).

In typical fluent speech, the average number of whole-word repetitions is around 1.5 per 100 fluent words according to Bortfeld et al. [2001]. As demonstrated in Table 3.7, we found the average number of whole-word repetitions to be smaller across the board in our study, with the ND group having the highest number of repetitions at 1.4 per 100 fluent words.

No other significant differences were found in the number of whole-word repetitions between groups ($p = .7$ for part-word repetitions and $p = .8$ for phrase repetitions).

3.4.7 Prolongations

3.4.7.1 Number of Prolongations

A Kruskal-Wallis test indicated that there was a significant difference in the number of prolongations per 100 fluent words across the four different cognitive groups, $\chi^2(3,$

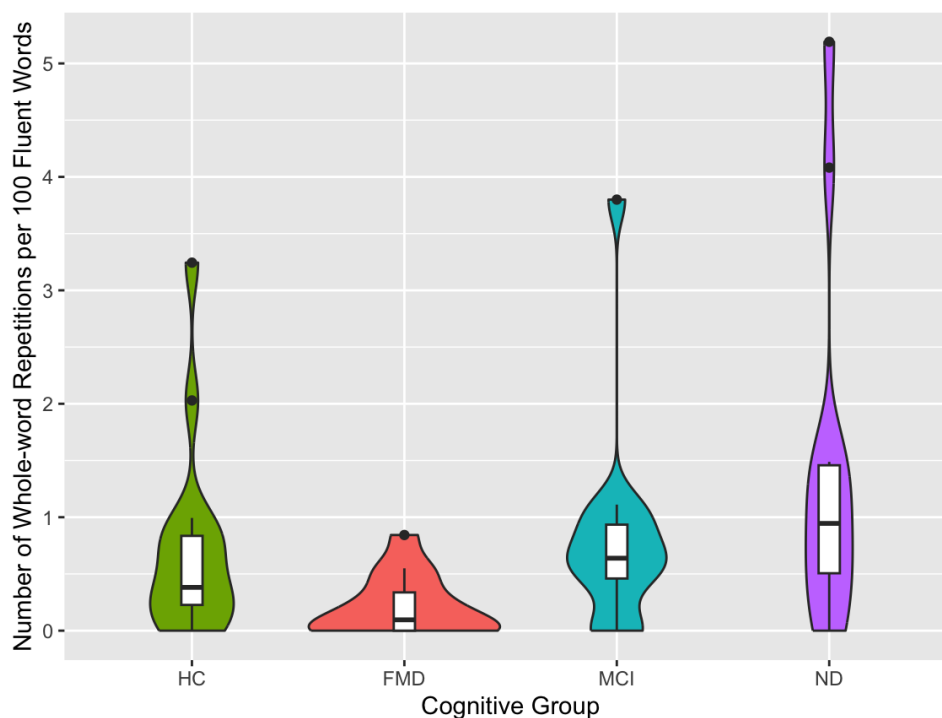


Figure 3.7: Violin plot showing the number of whole-word repetitions per 100 fluent words according to cognitive group

	HC	FMD	MCI	ND
Average Number of WREPs per 100 Fluent Words	0.7	0.2	0.8	1.4

Table 3.7: Comparison of the average number of whole-word repetitions per 100 fluent words according to cognitive group.

$N = 55$) = 12.92, $p = .004$. The median number of prolongations per 100 words per cognitive group was 3.89 for HCs, 6.07 for FMDs, 8.24 for MCIs, and 7.60 for NDs. Post-hoc comparisons using Dunn’s method with a Bonferroni correction for multiple tests indicated that the median number of prolongations from people in the HC group was significantly smaller than in the MCI ($p = .004$) group. No other significant differences were found between groups.

As can be seen in Figure 3.8, a participant in the MCI group and a participant in the ND group had far greater numbers of prolongations than the rest of their respective cohorts. Participant 87 (MCI group) had 50 prolongations per 100 fluent words, and

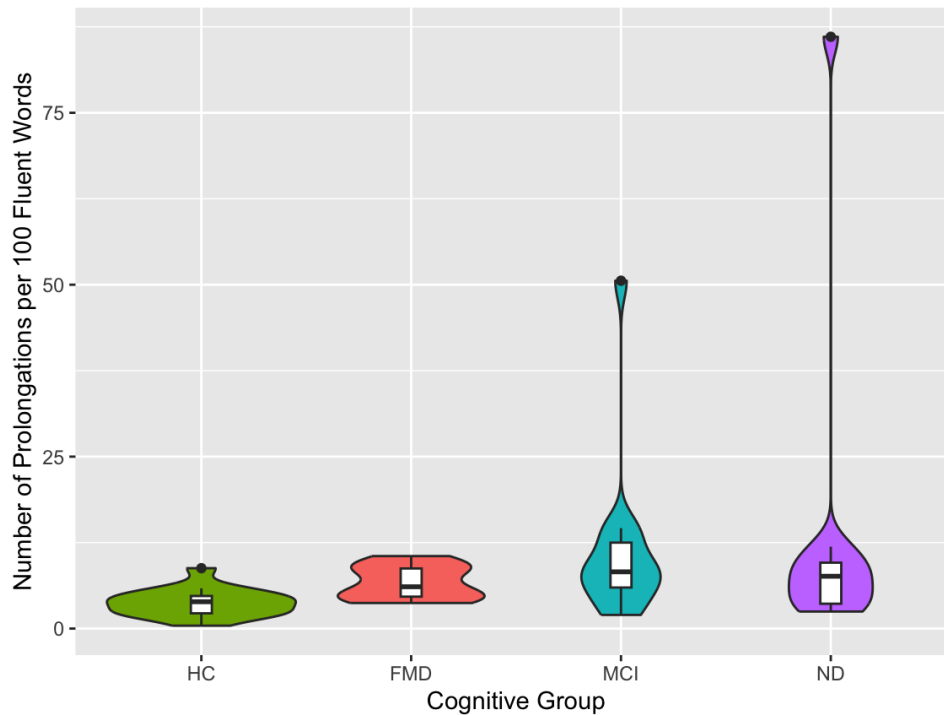


Figure 3.8: Violin plot showing the number of prolongations per 100 fluent words according to cognitive group

Participant 290 (in the ND group) had 86 prolongations per 100 fluent words. Upon reviewing these interviews, we found that these inflated numbers of prolongations could be due to the unusually slow speaking rates of each of these participants. This motivated the adjustment made to the method of recording prolongations in the second disfluency analysis (described in Section 4.3.2 in the next chapter).

3.4.7.2 Average Length of Prolongations

Figure 3.9 shows the average length of prolongations per cognitive group. A Kruskal-Wallis test indicated that there was a significant difference in the length of prolongations across the four different cognitive groups, $\chi^2(3, N = 55) = 8.60, p = .03$. The Kruskal-Wallis test was chosen as the data was found to be not normally distributed in a Shapiro-Wilkes test ($W = 0.92, p = 0.002$). The median average length of prolongations (in seconds) per cognitive group was 1.76s for HCs, 2.06s for FMDs, 2.16s for MCIs, and 2.13 for NDs. Post-hoc comparisons using Dunn's method with a Bonferroni correction for multiple tests indicated that the only significant difference was between the lengths

of prolongations in the **HC** group and the **MCI** group, with the average length in the **HC** group being significantly shorter ($p = .04$).

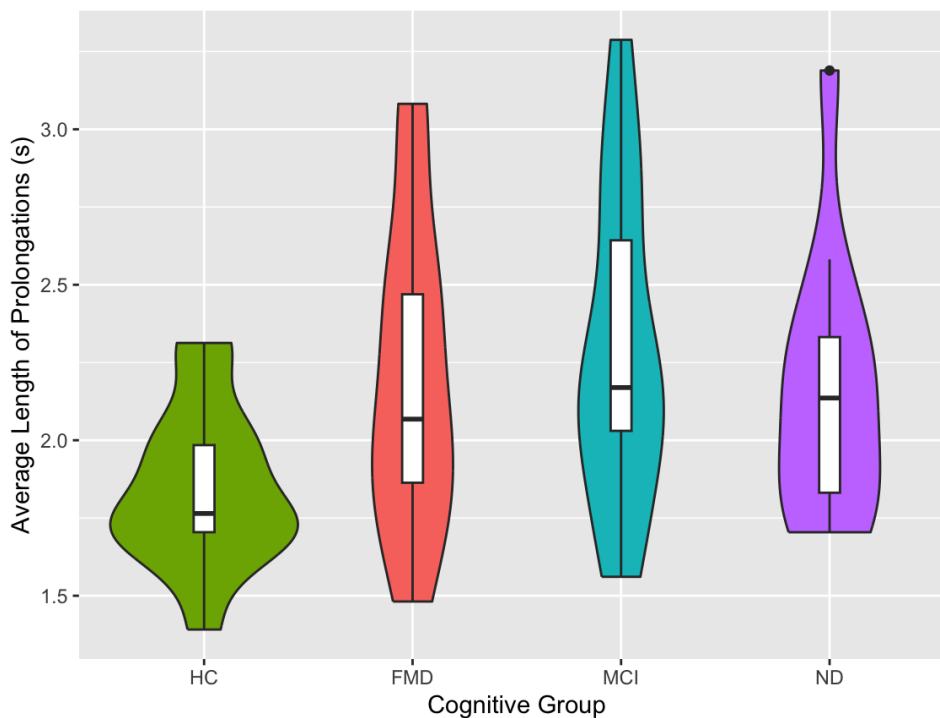


Figure 3.9: Violin plot showing the average lengths of prolongations according to cognitive group.

3.4.8 Speech Errors

There were not enough examples of any of the classes of speech error found in our data to facilitate a statistical analysis. Table 3.8 demonstrates how few examples of each type of error were found per group.

Cognitive Group	[de1]	[add]	[sub]	[mal]
HC	0	0	0	1
FMD	0	1	0	3
MCI	3	1	5	0
ND	4	5	2	6

Table 3.8: Comparison of the number of speech errors (deletions, additions, substitutions, and malapropisms) from across the different cognitive groups.

As we have seen across all the results, there is a slight trend of increasing amounts of errors as the severity of the cognitive decline increases. However, more instances of each error would be needed to assess whether this is statistically significant.

3.4.9 Summary

This small sub-section presents a summary of our findings. Section 3.4.9.1 presents a condensed version of the differences found between the HCs and patients in the cognitively impaired groups. Section 3.4.9.2 then explores the differences found among the cognitively impaired groups. After these short summaries, results for each category of disfluency are discussed in more detail.

3.4.9.1 Healthy Controls vs. Cognitively Impaired Participants

Statistically significant differences were observed between the HC group and the cognitively impaired groups across nearly all of the disfluencies examined in this study. Both the number and average duration of unfilled pauses were significantly smaller in the HC group. Although no difference was found in the number of filled pauses, the average duration of filled pauses in the HC group was significantly shorter than in the cognitively impaired groups, specifically the FMD group. Healthy controls exhibited the smallest pause-to-speech ratios, indicating that their rate of (all) pauses relative to the amount of speech produced was lower than that of the cognitively impaired groups. With regard to prolongations, HCs demonstrated significantly fewer and shorter prolongations compared to the participants in the cognitively impaired groups. Although there were insufficient instances of speech errors to conduct statistical testing, fewer instances of all types of speech errors were observed in the HC group than in the cognitively impaired groups.

3.4.9.2 Differences Between Cognitively Impaired Groups

The results indicated a statistically significant difference in the number of whole-word repetitions between the FMD and ND groups ($p = 0.004$). No significant differences were

observed in the number of part-word or phrase repetitions. This was the only significant difference found between the cognitively impaired groups.

3.5 Discussion

3.5.1 Total Disfluency Rates

Our study observed much higher disfluency rates than have previously been reported, even in the case of our HCs. We hypothesise that this is due to the unfamiliarity of the situation in which the data was recorded. Previous research has shown that disfluency rates directed at machines are lower than in spontaneous human-human conversation. However, these rates are taken from studies that involve participants interacting with automatic systems designed to carry out specific tasks. This results in interactions that are oriented to said tasks in which the human participants need only produce simple, short commands. Our results highlight that when asked more probing questions (to which long and detailed answers are expected) the high disfluency rates are reflective of an increased cognitive load in addition to a conversational style that the participants have not experienced before.

3.5.2 Unfilled Pauses

This analysis found statistically significant differences between the number of unfilled pauses in all cognitively impaired groups compared to the HC group. This aligns with findings from other studies, such as Yuan et al. [2021] which found that their group of people with AD had higher numbers of unfilled pauses of all durations (split into four bins ranging from <0.5s to >2s) when compared to the HC group. Sluis et al. [2020] analysed unfilled pause frequency and duration at three different cognitive levels; healthy controls, mild dementia, and moderate dementia. Although they did not find a statistically significant difference in the number of all pauses between groups, they did find that the HC group had a significantly lower number of long unfilled pauses (>2s) than the mild dementia group, which in turn also had significantly lower numbers of long

pauses than the moderate dementia group. Studies such as Rohanian et al. [2021] have already noted the salience of unfilled pauses as an identifier of CD, and have demonstrated that including unfilled pause information in a deep learning model for detecting dementia can help improve system accuracy.

In terms of the lengths of unfilled pauses, research supports the observation made in this study that there are significant differences in unfilled pause durations between cognitively impaired and cognitively healthy groups. Singh et al. [2001] found in their study of eight healthy controls and eight participants with AD that there were statistically significant differences in the mean duration of unfilled pauses between the two groups. It should be noted, however, that unfilled pause duration is not always found to be significantly different between control groups and cognitively impaired groups, even if the unfilled pause rate is significant (see Lofgren and Hinzen [2022]).

3.5.3 Filled Pauses

There is conflicting evidence surrounding whether or not the number and length of filled pauses is significantly different between HCs and cognitively impaired groups. One of the biggest issues here is in how statistics surrounding pauses are reported in this field. Often, it is not made clear whether researchers are including both filled and unfilled pauses in their studies. For example, Pastoriza-Dominguez et al. [2022] mention the distinction between filled and unfilled pauses, yet treat them both as one “pause” classification, making it difficult to separate how useful the filled pauses are alone at differentiating between healthy and cognitively impaired groups. Likewise, Pakhomov et al. [2011] treated both filled and unfilled pauses as one phenomenon. Other studies, such as Yunusova et al. [2016], do not explicitly describe how they define a pause with regards to timings or whether the pause is silent or filled. Zhu et al. [2022] differentiate between inter- and intra-sentential pauses, but again provide no information about whether the pauses they are measuring are filled or unfilled, or how exactly they differentiate between what they are classifying as “short” pauses and “long” pauses.

Filled pauses are typically thought of as a method of facilitating recall in spontaneous

speech. Research has shown that filled pauses can serve a “fluent communicative function”, such as aiding in turn-taking [Kosmala and Crible, 2022]. One possible explanation for the reduced number of filled pauses in our data (compared to the number of unfilled pauses observed) is that participants do not have a conversational partner with whom to participate in the usual turn-taking process observed in spontaneous conversation. In natural conversation, filled pauses are often used to give the speaker time to formulate their response whilst indicating to the listener that there is more information to follow [Cossavella and Cevasco, 2021]. It could be the case that our participants do not feel the need to fill dead air when talking to the avatar, so exhibit more unfilled pauses instead.

3.5.4 Pause-to-Speech Ratio

There is little information available surrounding the differences in pause-to-speech ratios between different levels of CD, despite the fact that pause-to-speech ratio has been reported as a significant predictor of more severe levels of cognitive impairment (see Pakhomov et al. [2010]). One study that used the original IVA dataset found a significant difference between HCs and ND/MCI groups when looking specifically at the pause-to-speech ratio in answers to only the remote memory questions [O’Malley et al., 2020]. The same study did not find any statistically significant differences for the FMD group. Results from the present study contrast with those from an earlier study based on a different subset of the IVA data. In Beavis et al. [2021], researchers found that the pause-to-speech ratio was generally lower for people in the MCI and FMD groups, compared to the HCs. As demonstrated in Table 3.6, our study found the lowest pause-to-speech ratio to be in the control group, and the highest in our FMD group. Our findings reflect the higher filled and unfilled pause rates observed in the three cognitively impaired groups, as presented above. One possible explanation for the differences in reported pause-to-speech ratios is that the present study uses a more fine-grained method of calculating these ratios. Our method involved calculating a total phonation time for each participant by summing the timings of only fluent segments of speech (so both filled and unfilled pauses, speech errors, and repetitions were removed). This is then compared to the total sum of the timings of both filled and unfilled pauses. In the Beavis et al. [2021] study, there is no clear definition

of what constitutes a pause and it is likely the case that filled pauses were not included in their “pauses” calculation. It is also important to note that while both studies use IVA data, two different subsets are used. As both studies use small amounts of data, any variation could result in quite a big difference in observations. For example, the [Beavis et al.](#) study may have included a couple of HCs with large numbers of pauses which simply were not included in the subset used in our experiment.

3.5.5 Repetitions

Little research exists examining the frequency and distribution of word repetitions at different levels of CD. Where such work does exist, the studies are typically conducted on either phonemic or semantic verbal fluency tests. Our analysis found that the number of word repetitions was significantly higher in the ND group when compared to the FMD group, which had the lowest number of word repetitions of all four cognitive groups. This rate was also significantly lower than word repetition rates you would expect to see in typical fluent speech [[Bortfeld et al., 2001](#)]. This could be due to the fact that our FMD group spoke the least, although we found no significant differences in the number of words spoken across all four groups.

3.5.6 Prolongations

Whilst we found a statistically significant difference between both the number and duration of prolongations between the HC and MCI groups, there is little work from other studies to support this finding. [Cera et al. \[2023\]](#) found in their study that the number of vowel prolongations was higher in patients with AD compared to the HCs. This study also investigated consonant prolongations, but found no statistical differences between cognitively healthy and cognitively impaired groups.

3.6 Conclusions

In summary, this analysis provides additional evidence in agreement with previous findings that significant differences in unfilled pause duration and frequency can be found between

control groups and participants with moderate to severe cognitive impairment. However, as shown above, this work did not find significant differences between cognitively impaired groups when analysing the number of filled pauses. This could be due to the fact that the human participants do not feel the need to signal to the avatar that they are formulating additional things to say, as they know that the avatar will be listening until they tell it not to. In terms of other disfluencies it was found that there were statistically significant differences in the number of word repetitions, prolongations, and additions between the HCs and the cognitively impaired groups. Whilst our findings come from a relatively small dataset, they do indicate that disfluencies may have the potential to hold useful information for diagnosing levels of CD

This study shows that differences in the number and length of disfluencies such as prolongations, unfilled pauses, and word repetitions can be a useful discriminator among different levels of CD. Due to conflicting information surrounding disfluencies and CD, this study remains cautiously optimistic about the usefulness of such features for aiding in the early detection of conditions such as AD, but keeps in mind that at this stage there is limited generalisability of the findings. The next experiment in this thesis uses a slightly amended version of the original disfluency taxonomy to investigate the incidences of disfluencies during a different cognitive task. This allows for a direct comparison between two different cognitive tasks administered by a virtual agent, and investigates the effect that the task difference has on the presence of disfluencies in speech.

Chapter 4

Second Manual Disfluency

Analysis

Contents

4.1	Introduction	102
4.2	Background	102
4.2.1	Picture Description Tasks	103
4.2.2	Task Differences	107
4.2.3	Automatic Detection of Disfluencies	108
4.3	Methodology of the Second Manual Disfluency Study	111
4.3.1	Data	111
4.3.2	Adjustments to Disfluency Schema	112
4.4	Results from the Second Manual Disfluency Study	117
4.4.1	Number of Fluent Words	117
4.4.2	Unfilled Pauses	118
4.4.3	Filled Pauses	120
4.4.4	Repetitions	121
4.4.5	Prolongations	122

4.4.6	Speech Errors	122
4.4.7	Discussion of Results from the Second Manual Disfluency Study	123
4.5	Comparison of Results from Manual Disfluency Studies One and Two	125
4.5.1	Total Disfluencies	125
4.5.2	Pauses	126
4.5.3	Repetitions	126
4.5.4	Prolongations	127
4.5.5	Speech Errors	127
4.5.6	Discussion	127
4.6	Automatic Cognitive Decline Classification	129
4.6.1	Background	129
4.6.2	Data	130
4.6.3	Methodology	131
4.6.4	Baseline System	132
4.6.5	Results	133
4.6.6	Discussion	133
4.7	Conclusion	134

4.1 Introduction

This chapter reuses the Disfluencies in Cognition ([DisCo](#)) schema described in the previous chapter and applies it to a different task from the Intelligent Virtual Agent ([IVA](#)) dataset in a second manual disfluency study. This chapter then presents a proof-of-concept study where disfluency information is added to an Automatic Cognitive Decline Classification ([ACDC](#)) system. This chapter addresses a further two research questions:

3. How do the patterns of disfluency vary from the interview task to a picture description task?
4. Can disfluency information improve the accuracy of an automatic cognitive decline classification system?

The background section of this chapter begins with a description of the variations in the presentation of disfluencies in spontaneous or semi-spontaneous speech that correlate with different tasks. This section also expands on the introduction to [ACDC](#) systems found in [Chapter 2](#) towards the beginning of this thesis, but with a particular focus on recent attempts at automatically deriving disfluency information.

The methodology section of this chapter then presents a new subset of the [IVA](#) dataset that is used for this second manual disfluency study, and discusses minor revisions made to the [DisCo](#) schema. The results section starts by describing the results from the second disfluency analysis in this thesis, along with a discussion of these results. This is followed by a description of the differences observed in the results from the two disfluency analyses presented in this thesis. This chapter ends with a description and the results of a proof-of-concept study in which disfluency information was incorporated into an [ACDC](#) system, the findings from which were published in [Thomas et al. \[2023\]](#).

4.2 Background

This chapter investigates the differences in the frequency and duration of disfluencies present in a picture description task across three different cognitive groups. The data

used in this experiment is taken from the IVA dataset. For this analysis, the chosen task is the Cookie Theft picture description task, a commonly used test for cognition. The following section describes this task, and its usage and efficacy of evaluating levels of cognition.

4.2.1 Picture Description Tasks

Picture description tasks provide a method of obtaining samples of spontaneous speech in order to assess cognition, often after a brain injury of some type (Eg., stroke). One of the most popular picture description tasks is the Cookie Theft task, which forms part of the Boston Diagnostic Aphasia Examination [Goodglass and Kaplan, 1983].



Figure 4.1: The Cookie Theft Picture Description Task from Goodglass and Kaplan [1983].

At first glance, the task seems simple. Participants are asked to describe the scene in the picture with as much detail as they can. Examiners are not allowed to ask for elaboration or to ask the patient to talk about details that they have not already mentioned. This makes a picture description task a clear choice for use with an IVA system, where the virtual assistant administering the test is not programmed to ask participants to expand

or clarify their original answers. During the task, seven main areas of a patient's language and cognition are being assessed, as detailed in Cummings [2019, pp.155-159]. Briefly, these main areas are:

1. The Salience of Information

The most salient information within this picture would be the activities of the three visible people. This is primarily what patients with Cognitive Decline (CD) focus on during their description of the scene. However, there is also a good deal of less salient information presented in the form of background details such as the plates stacked by the sink and the garden that is visible through the window. Typically, participants with no cognitive impairment will acknowledge the lower/lesser salience of this secondary information by describing it towards the end of their description, whereas participants with CD tend to neglect talking about this information.

2. Semantic Categories

In order to describe the picture fully, participants need to be able to use words from various different semantic categories. For example, inanimate entities such as *cup* and *stool*, animate entities such as *mother* and *boy*, concrete concepts like *falling* and more abstract concepts such as *daydreaming*. Typically, a person with advanced CD would struggle with including words from abstract concepts, and may show less variation in the number of different semantic categories they are able to produce words from.

3. Referential Cohesion

Referential cohesion is necessary for forming fully comprehensible descriptions and stories. As such, descriptions and sentences tend to be shorter when coming from people with cognitive impairment compared to Healthy Controls (HCs). Referential cohesion has been found to be difficult for patients with Alzheimer's Dementia (AD), with studies such as Ripich et al. [2000] noting that the use of all cohesive devices declines as severity of cognitive impairment increases.

4. Causal and Temporal Relations

The Cookie Theft picture is a static scene, but many events are happening as a result of something else. For example, the sink is overflowing *because* the woman has left

the tap on for too long. Again, in order to describe the picture to a full extent the participants must be able to describe the temporal order of events, which is something that people with Mild Cognitive Impairment (MCI) tend to have difficulties with.

5. Mental State Language

There are two kinds of mental states; cognitive mental states such as *knowledge* and *beliefs*, and affective mental states, also known as emotions. Theory of mind is the cognitive ability of attributing mental states to the minds of others and to oneself. If theory of mind skills are intact then a person would be able to describe things such as the tap overflowing because the lady *forgot* to turn the tap off. The lack of such language could be an indicator of CD.

6. Structural Knowledge

This concerns phonology, syntax, semantics, and speech motor skills, and is what clinicians can analyse for signs of anomia or aphasia; difficulties in lexical access and retrieval. Problems with forming structurally correct sentences could indicate CD.

7. General Cognition and Perception

This is investigated by looking at things such as how many aspects of the picture have been described more than once without the participant displaying awareness of the repetition. It also concerns whether and how the participant conveys information in a logical order.

When conducting a picture description task, the doctor or clinician administering the test should take into account all of the points mentioned above. However, there is a degree of subjectivity involved when clinicians are observing the results of a picture description task. Although picture description tasks are commonly used, there is currently no set guideline on how to grade the responses given by patients. Clinicians make a high-level judgement of the speech as a whole, which is taken into consideration along with the results of the other tests that patients will typically undertake in a memory clinic, such as verbal fluency tests. Whilst a general impression of a patient's fluency could contribute to an assessment of the structural knowledge of their speech, clinicians do not focus on disfluencies, and would not count or otherwise measure the number of disfluencies present

in a patient's speech.

A picture description task is, of course, not the only test from which the results will be analysed with subjectivity, as these tests do not produce concrete results (especially when compared to results from tests such as blood tests or lumbar punctures). As there is subjectivity to the assessment of speech in memory clinics, there is therefore potential for inter-rater differences. An audit of memory clinics in London in 2019 found “substantial” differences in the percentages of people aged 65 and over being diagnosed with dementia across different services [Cook et al., 2020]. This ranged from 22% to 100% as demonstrated in Figure 4.2.

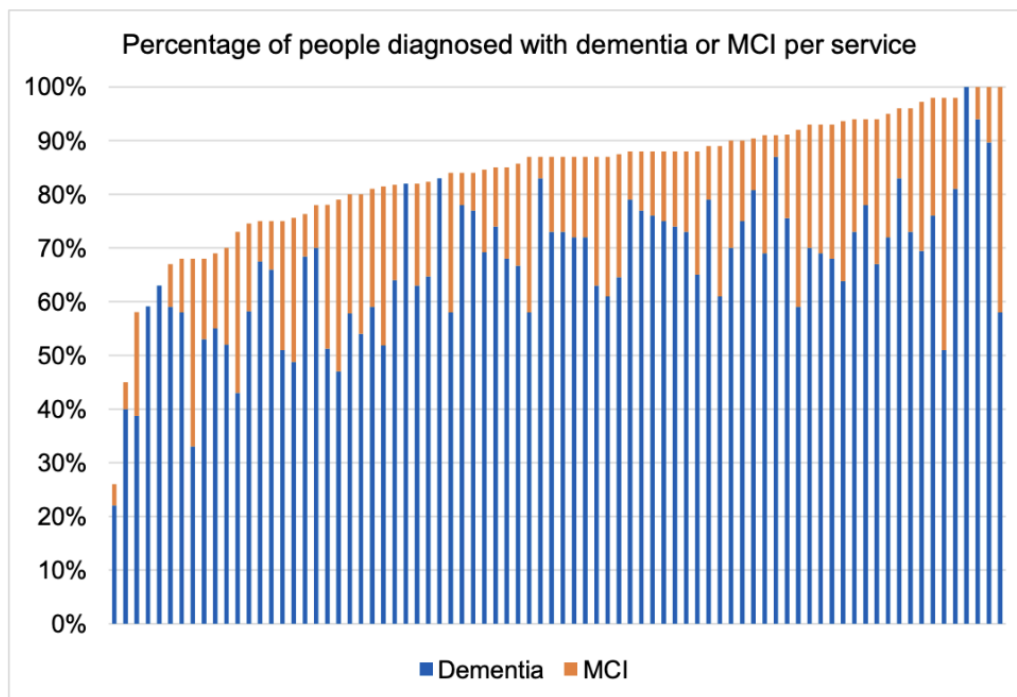


Figure 4.2: Percentage of people aged 65 and over diagnosed with dementia or mild cognitive impairment across different memory clinics in London [Cook et al., 2020].

The analysis presented in this chapter uses recordings of participants completing the Cookie Theft picture description task as a way of eliciting speech for a disfluency analysis. This counters some of the subjectivity associated with scoring picture description tasks because it will result in quantitative data, allowing for direct comparisons between participants and cognitive levels. This could potentially help reduce the amount of subjectivity involved in making diagnostic decisions, which in turn could help reduce the

variation seen between different services.

4.2.2 Task Differences

Linguistic studies have examined speech in many different modalities, from spontaneous conversational speech to scripted, read speech. It is important in our experiments to consider the kind of speech that clinicians (human or digital) are likely to encounter when conducting a spoken memory test. Although some cognitive tests may involve a reading task, the majority of tests will involve responding freely to questions. In the IVA dataset, the interview portion of the test allows participants to speak freely, although participants are answering a predetermined set of questions. This is similar to the picture description task, where all participants are given the same picture to describe, but decide themselves how much or how little they would like to talk. However, with the picture description task the possible responses that could be given by participants are much more constrained as they are all seeing the same picture (compared to the fact that they have probably all had different jobs, and done different things over the weekend). In the interview section of the data, it is not uncommon to see large jumps between topics whilst still falling under the umbrella of answering a question, for example the participant who ends up fondly recalling meeting the Princess of Monaco after being asked what school they attended. The scope for this kind of divergence from a question is much more limited in the case of a picture description task. For this reason, the speech resulting from a picture description task should fall somewhere between spontaneous speech and read speech. Scripted or read speech has been shown to contain far fewer disfluencies than spontaneous speech [Lickley, 2017]. As the speech resulting from the picture description task is not fully spontaneous but is also not scripted in any way, the resulting speech should contain “enough” disfluencies to allow for a thorough comparison to the fully spontaneous question responses.

Little work has focused on the levels of disfluency in healthy adults completing a picture description task. One such study [Duchin and Mysak, 1987] examined the performance of groups of healthy adult males of different ages on three different speech tasks; oral

reading, picture description, and conversation. Included in this study were two middle aged groups, consisting of 15 participants each, and two elderly groups, again consisting of 15 participants each. The first middle-aged group had a mean age of 49, the second had a mean age of 60. The mean ages of the elderly groups were 68 and 80. The disfluencies under investigation included part-word, whole-word, and phrase repetitions, interjections (repairs), and “dysrhythmic phonations” (prolongations). There were two main conclusions from this study; levels of disfluency across all three tasks did not differ between the middle-aged and elderly groups, and more disfluencies were present in the conversation task compared to the picture description. Given this information, we expect to observe fewer disfluencies in the speech elicited via the picture description task compared to the interview task that was analysed in Chapter 3).

4.2.3 Automatic Detection of Disfluencies

Although the field of [ACDC](#) is rapidly expanding, there is currently little work that includes an automatic analysis of speech disfluencies in dementia. The majority of automatic disfluency detection is used in fields such as second language learning, where it is important for learners of a language to receive feedback on the kind of mistakes they are making in their speech. Although these kinds of disfluency may vary from disfluencies that will be present in a speaker’s native language, many of the methods for detecting these disfluencies are transferable to the [ACDC](#) domain.

Aside from in the field of language learning, disfluencies are typically regarded as problems to discard from an automatically created transcript, or to be removed from speech that will be used to train a language model [[Mirheidari, 2018](#)]. Many applications of Automatic Speech Recognition ([ASR](#)) would not require the transcription of disfluencies such as filled pauses and repetitions. Usually, [ACDC](#) systems would be considered one of these application. For example, a study by [Tang et al. \[2023\]](#) tested three different commercially available [ASR](#) systems to create transcripts of speech to be used in an [ACDC](#) system. They found that of the three [ASR](#) systems tested, all of them produced transcripts which resulted in a system accuracy score equivalent to, if not better than, human transcribed

speech in a system that used linguistic and acoustic features. Whilst these results are promising for systems that use “traditional” acoustic and linguistic features, this approach will not work for an automatic analysis of disfluencies. This is due to the fact that the ASR transcriptions produced in the Tang et al. (and most automatically produced transcripts) do not include transcriptions of disfluencies. They are instead left out of the transcript, and if the transcript does not contain any disfluency information then the classification model has no disfluency features to analyse.

However, for applications that *do* require the assessment of disfluencies such as in the field of second language learning, there are a number of different methods of identifying disfluencies in speech, enabling them to be transcribed and analysed. Below is an overview of some of the most common methods of automatically detecting three kinds of speech disfluency: unfilled pauses, filled pauses, and repetitions.

4.2.3.1 Unfilled Pauses

Regardless of the application, the approach to identifying unfilled pauses remains relatively the same; examine the audio signal for regions of low energy or silence, and classify them as an unfilled pause according to some pre-decided duration threshold. This can be achieved via a process called Voice Activity Detection (VAD). Essentially, a VAD component is able to make a decision on whether a given window or segment of an audio file contains speech or not [Martin and Kolossa, 2012]. VAD can be used to detect silences within an audio file, as seen in Kaushik et al. [2010]. This provides an automatic way of detecting a pause within a sentence when used in combination with duration measures as a method of defining and labelling unfilled pauses. Jaiswal and Hines [2018] present an overview of different VAD algorithms, and assess their effectiveness for identifying silence in sound files.

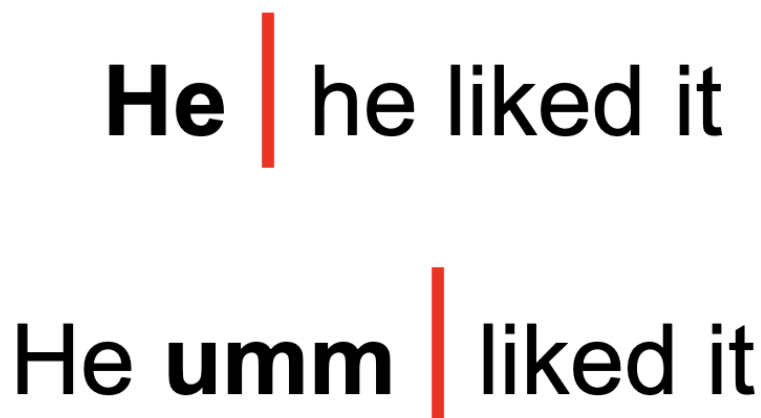
4.2.3.2 Filled Pauses

Filled pauses present a potentially more complicated task for automatic identification, as they often resemble phones in a similar way to speech. Early work from O’Shaughnessy and Gabrea [2000] employed a filled pause detection method including vowel identification

and duration measures. Fundamental Frequency (F0) stability was used to decide whether an identified phone could be classed as a vowel. A duration threshold of 120ms was chosen, and any vowels found to last longer than the threshold were classified as a filled pause. Whilst this is a simple process, it is somewhat flawed. The authors note that this method frequently misclassified intended speech sounds such as word-initial or word-final schwas as filled pauses. One potential method of mitigating this would be to raise the duration threshold for filled pause classification. However, this would not work for differentiating between filled pauses and prolongations. [Barczewska and Igras \[2013\]](#) present a method based on the tracking of the first and second formants, as well as segment duration. They used the first and second formants to track a central vowel, which they classed as a filled pause. Other vowels, or cases where there was movement in the formants, resulted in the segment being classed as a prolongation. Their study reports a classification accuracy of around 68%.

4.2.3.3 Repetitions

[Shriberg et al. \[1997\]](#) developed a method of automatic disfluency detection that uses decision tree classifiers and prosodic features. This included detection of filled pauses, repairs, and repetitions. The basic goal of their model was to identify the boundary in speech at which a disfluency occurs, as demonstrated in [Figure 4.3](#).



He | he liked it

He umm | liked it

Figure 4.3: Examples of the boundary (red) between a disfluency (bold) and fluent speech. Adapted from [Shriberg et al. \[1997\]](#).

Their aim was to improve natural language understanding and speech language models in general, and they hypothesised that because their model is based only on prosodic information that the model would be applicable to situations when word recognition is unreliable. Each disfluency boundary in the database was represented by a feature vector which also contained acoustic and gender information. A decision tree is used to output posterior probability estimates of disfluency events. The authors also examined a classifier-based n -gram Language Model (LM) for this purpose, and the results of both were compared. They tried their models on the different disfluency types, and found that overall the decision tree performed well, achieving an accuracy of 77.5% when detecting word repetitions. The tree revealed that the most important features for classification in this case were duration and the distance from pause.

4.3 Methodology of the Second Manual Disfluency Study

This section introduces and then presents the second manual disfluency analysis undertaken as part of the work investigating how well disfluencies can differentiate between different levels of CD.

4.3.1 Data

This experiment again uses a subset of the IVA data, however some changes were made which resulted in a slightly different subset to that used in the first disfluency analysis.

As the first study did not identify any disfluencies capable of distinguishing between Functional Memory Disorder (FMD) and other levels of CD, and given the limited research on the effects of FMD on speech, this second study does not include FMD as a cognitive group for investigation. Instead, we included additional participants from the two other groups (MCI and Neurodegenerative Dementia (ND)), sourced from the original IVA dataset (N = 93). Specifically, we included all available participants in the MCI and

ND groups who had recorded picture description tasks and met the inclusion criteria outlined in Section 3.3.2.1. This required good technical recording quality and excluded participants who were non-native speakers of English, had diagnosed pathological speech conditions, or had comorbidities known to affect speech. Of the original 15 HCs used in the previous manual disfluency analysis, two participants did not complete picture description tasks. This adjustment brought the total number of participants included in this second study to 48.

Some participant information was missing; the table below shows the participant information that had been recorded at the time the data was collected.

Subject Group	No. Participants	Male: Female	Age		No. Fluent Words in Recording		Mean MMSE Score
			Mean	sd	Mean	sd	
HC	13	4:10	69.4	8.3	165.5	88.3	28.7
MCI	17	11:6	62.3	8.1	97.9	86.1	26.7
ND	18	7:9*	69.4	7.0	90.2	54.1	23.1

Table 4.1: Participant information for the second manual disfluency study including number of participants, gender, age, number of fluent words per recordings, and participant Mini Mental State Examination scores. Statistics are given as group means and standard deviation. *Gender information missing for two participants.

4.3.2 Adjustments to Disfluency Schema

The original disfluency schema was updated for this experiment in order to mitigate certain issues that arose with the original version. Primarily this surrounded the classification of prolongations. An investigation of speech rates of our participants from the first experiment showed that the majority of them spoke much slower than would typically be expected in interview-style speech (approximately 190 words per minute; Tauroza and Allison [1990]). This is probably due to the ages of our participants. Our participants' speech rates ranged from 60 words per minute (ND group) to 197 words per minute (the HC group). The median for the entire subset was 119 words per minute, although this will have been impacted by the very low rate in the ND group. It seemed unfair to mark prolongations (a prolonged segment lasting >200ms) as a disfluency when analysing slower

speech, as the rate of prolongations would be unnaturally high. In an attempt to combat this, we introduced a rating scale to each prolongation notation. Two different rating scales were included for each prolongation notation. Both of these ranged on a scale from one to three. The first was a scale of perceived intentionality. This was included because in the first round of manual disfluency annotations there were multiple instances of prolongations that appeared to have occurred for specific reasons such as emphasis. As this kind of classification relies on some subjectivity from the annotator, the inclusion of an “intentional \longleftrightarrow disfluent” scale allowed prolongations to be double checked for an assessment as to whether the prolongation was in fact a disfluency. This builds on the severity ratings of disfluencies proposed by Panesar and de Alba [2023], as described in Chapter 3.

The second scale was a rating of annotator certainty. This was useful in instances where segments met the prolongation length threshold of $>200\text{ms}$ but did not appear to be out of place, such in the case of speakers with a slower speech rate. This resulted in annotations such as the following:

[pro 1-1]: a prolongation that the annotator is certain was for emphasis or some other effect.

[pro 3-1]: a prolongation that the annotator is certain can be classed as disfluent.

[pro 1-3]: a prolongation that the annotator feels is intentional, but cannot be sure of as it meets the criteria for being classified as a “true” prolongation.

Other changes to the original disfluency schema include the addition of three more speech error classifications; blends, lexical selection errors, and circumlocution errors. Blends occur when two words that would both be correct given the context are muddled together, such as when one participant was describing the stool which is about to fall over:

Participant 0249: [tɒp^əlɪŋ] + [tɪpɪŋ] = [tɪp^əlɪŋ]

The lexical selection errors constitute a group of errors resulting as a modification of the original malapropism group. The definition was changed to better reflect the data that had already been observed in the first experiment. The key difference between a lexical

selection error and a malapropism, in this case, is that words used in a lexical selection error do not have to sound similar to the intended word, although in some cases they may. Malapropisms could be thought of as a subset of lexical selection errors, but as there were so few examples in the first study that it felt superfluous to include it as a category of its own in this second study. Some examples of lexical selection errors in the dataset include:

Participant 0263: “there’s tie-backs on the *carton*” rather than *curtain*

Participant 2111: “the mother is doing the *shop*” rather than *pots*

The third new category of speech errors is not technically an error or disfluency, as this phenomenon makes up part of natural spontaneous speech. However, DisCo2.0 provides a label for circumlocutions ([*cir*]), when words such as “thing” or “whatsit” are used in place of the intended word. Whilst not strictly a disfluency, this was observed to be a very common phenomenon in the first study so adding this category to DisCo2.0 allowed us to investigate whether the prevalence of this phenomenon was higher at more severe levels of cognitive impairment compared to the HCs (research such as Hier et al. [1985] suggests that the amount of circumlocution in speech increases as cognitive ability decreases). The updated disfluency schema can be found in Table 4.2.

The other main methodological change is the structure of the Praat text grids. Unlike in the first study, this second study did not have a text grid tier containing a transcription of the participant’s speech. This speech had already been manually transcribed and checked by annotators, so it was not necessary to re-transcribe the speech orthographically for inclusion in the text grids for this analysis. In this task there is no interference or any additional questions from the virtual agent, so that did not need to be included in a tier. Instead, the first text grid tier contains the disfluency annotations. As the manual part of this experiment was undertaken with a specific view to automating the process, additional tiers were included to contain syllable information in the hopes that this might be helpful when trying to automate some of the disfluency analysis. A tier including information marking phrase repetition boundaries was also added, resulting in the following overall tier structure:

1. All disfluency annotations.
2. Syllable boundaries that correspond to those disfluencies.
3. Phrases that are repeated in a [phrep] ([phr1], [phr2], etc).
4. Final syllable of repeated phrase if it's not a complete word.

This resulted in text grids with the format as depicted in Figure 4.4.

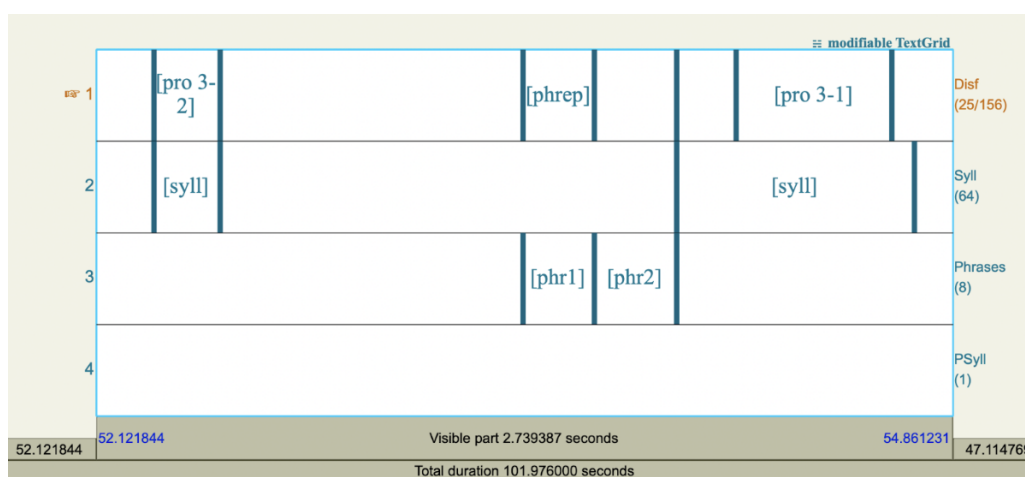


Figure 4.4: A view of the different Praat text grid tiers used in the second manual disfluency analysis.

The process of transcribing the data was largely the same as the process in the first manual disfluency analysis (Chapter 3). The annotator was not aware of the cognitive group that each participant belonged to in an attempt to mitigate bias. Each recording was listened to at least four times. Comprehensive notes were kept alongside the transcripts to serve as an aide-mémoire and to record any questions or queries that needed to be double checked by a separate phonetician. Once the annotation was complete, statistics were calculated and analysed in the same way as the previous manual disfluency analysis.

Disfluency	Description	Annotation
Repetition	Silence >200ms	[ufp]
	Does not count if at beginning of speaker's turn	
Filled pause	Lasts >200ms	[fp]
	Usually a vowel	
	May be followed by a nasal	
Repetition	Part word repetition	[pwrep]
	Whole word repetition	[wrep]
	Phrase repetition	[phrep]
Prolongation	Segment lasting >200ms	[pro x-x]
	1st Scale = Prosodic - Unintentional	
	2nd Scale = Certain - Uncertain	
	(e.g. [pro 3-1])	
Speech Error	Deletion (phone is deleted)	[dele]
	<i>Specific - pecific</i>	
	Substitution (phone changed to something else)	[sub]
	<i>Flap - flip</i>	
	Lexical selection error (usually based on semantic relations)	[lex]
	<i>Washing - drying</i>	
	Circumlocution (using "thing" etc in place of correct word)	[cir]
	<i>Out of the thing - out of the jar</i>	
Addition	Addition (phone added)	[add]
	<i>Favourite - fravourite</i>	
	Blend (two words blended together unintentionally)	[ble]
	<i>Toppling + tipping = tippling</i> (either word would be appropriate in the context)	
Repair	Noticing the error and then correcting it (false start)	[repa]
	Could be saying the word incorrectly so starting again, or starting a sentence and then going back and changing it	
	Edit to add: could also be the beginning of a word/phrase that is left incomplete but not repeated	
Non-Speech Event	Not silence, but not a filled pause. Could be laughs, coughs, sighs, etc	[nse1]
		[nsec]
		[nses]
Ignore	Indicates that this segment of the recording should be ignored, could fall outside of the turn	[ig]

Table 4.2: The DisCo2.0 Taxonomy.

4.4 Results from the Second Manual Disfluency Study

As with the previous manual disfluency analysis, this study used statistical tests to assess the significance of differences in disfluencies between the different cognitive groups. Kruskal-Wallis tests were performed for each of the different disfluency classes as none of the data were normally distributed. Dunn's test with Bonferroni corrections were then used to investigate where the significant differences were between groups; these results are described below.

4.4.1 Number of Fluent Words

Although the picture description task is a somewhat closed task with a limited number of points to talk about, there was some variation in the number of fluent words produced by each cognitive group.

A Kruskal-Wallis test indicated that there was a significant difference in the number of fluent words produced across the three different cognitive groups, $\chi^2(2, N = 48) = 10.93, p = .004$. The Kruskal-Wallis test was chosen as the data was found to be not normally distributed in a Shapiro-Wilkes test ($W = 0.83, p = <.001$). The median number of fluent words per cognitive group was 132 for HCs, 72 for MCIs, and 73.5 for NDs. Post-hoc comparisons using Dunn's method with a Bonferroni correction for multiple tests indicated that the median number of fluent words produced by people in the HC group was significantly larger than the other two cognitive groups ($p = .01$ for MCI and $p = .009$ for ND). There was no significant difference in the number of words produced between the MCI and ND groups.

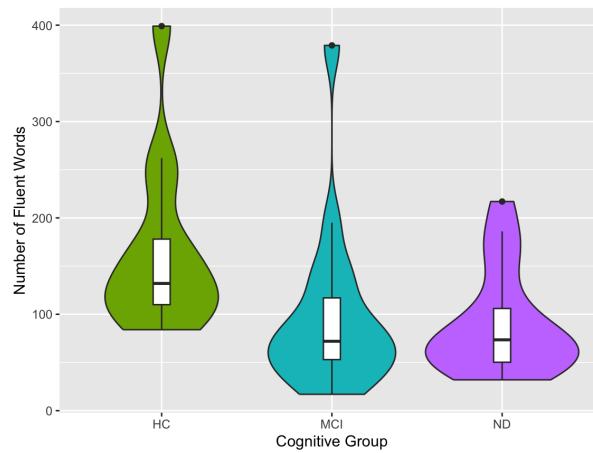


Figure 4.5: Violin plot showing the number of fluent words according to cognitive group.

4.4.2 Unfilled Pauses

4.4.2.1 Number of Unfilled Pauses

Figure 4.6 is a violin plot showing the number of unfilled pauses (per 100 fluent words) per cognitive group. A Kruskal-Wallis test showed that there was a significant difference in the number unfilled pauses per 100 fluent words produced across the three different cognitive groups, $\chi^2(2, N = 48) = 10.61, p = .004$. The Kruskal-Wallis test was chosen as the data was found to be not normally distributed in a Shapiro-Wilkes test ($W = 0.96, p = .2$). The median number of unfilled pauses (per 100 fluent words) per cognitive group was 13.5 for HCs, 17.1 for MCIs, and 19.4 for NDs. Post-hoc comparisons using Dunn's method with a Bonferroni correction for multiple tests indicated that significantly fewer unfilled pauses were produced by the HCs than the NDs ($p = .004$). There was no significant difference in the number of unfilled pauses produced between the HC-MCI groups or the MCI-ND groups.

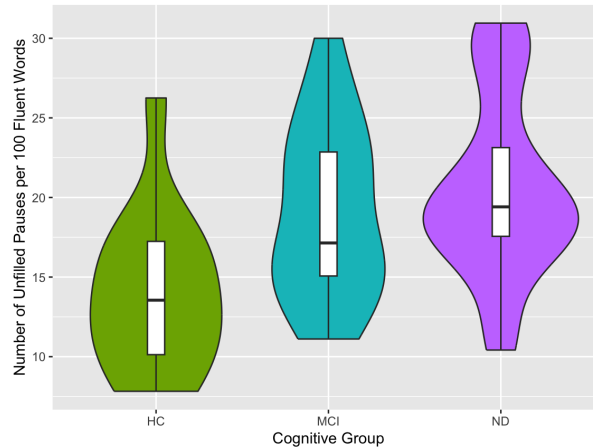


Figure 4.6: Violin plot showing the number of unfilled pauses per 100 fluent words according to cognitive group.

4.4.2.2 Average Length of Unfilled Pauses

As demonstrated in figure 4.7, the HCs had shorter average unfilled pauses than both the MCI and ND groups. A Kruskal-Wallis test showed that there was a significant difference in the number unfilled pauses per 100 fluent words produced across the three different cognitive groups, $\chi^2(2, N = 48) = 19.70, p = <.001$. The Kruskal-Wallis test was chosen as the data was found to be not normally distributed in a Shapiro-Wilkes test ($W = 0.84, p = <.001$). The median average length of unfilled pauses per cognitive group was 0.6s for HCs, 1.08s for MCIs, and 1.08s for NDs. Post-hoc comparisons using Dunn's method with a Bonferroni correction for multiple tests indicated that unfilled pauses from the HC group were significantly shorter than both the MCI group ($p = .002$) and the ND group ($p = <.001$). There was no difference in the average unfilled pause length between the two cognitively impaired groups.

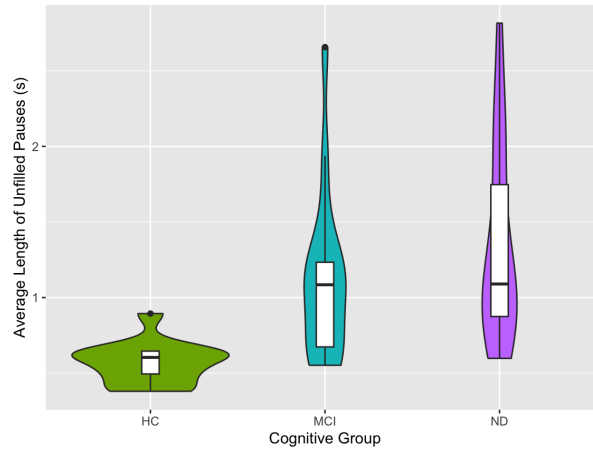


Figure 4.7: Violin plot showing the average length of unfilled pauses (s) according to cognitive group.

4.4.3 Filled Pauses

4.4.3.1 Number of Filled Pauses

Figure 4.8 shows the number of filled pauses produced during the picture description task per 100 fluent words across the three groups. No statistically significant difference was found across the groups ($p = 0.64$).

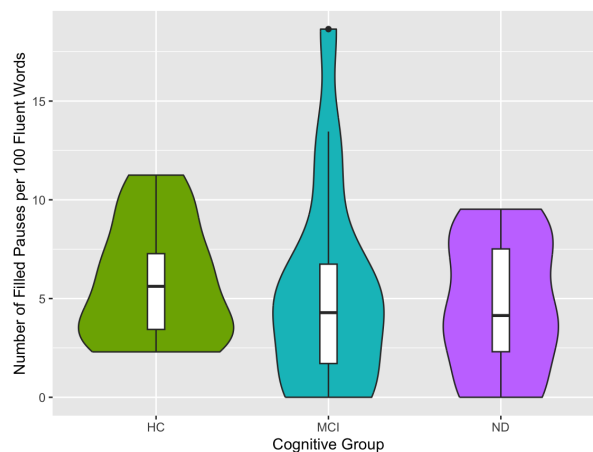


Figure 4.8: Violin plot showing the number of filled pauses per 100 fluent words according to cognitive group.

4.4.3.2 Average Length of Filled Pauses

Figure 4.9 shows the average lengths of filled pauses (s) in the picture description task across the three cognitive groups. No statistically significant difference was found between groups ($p = .53$).

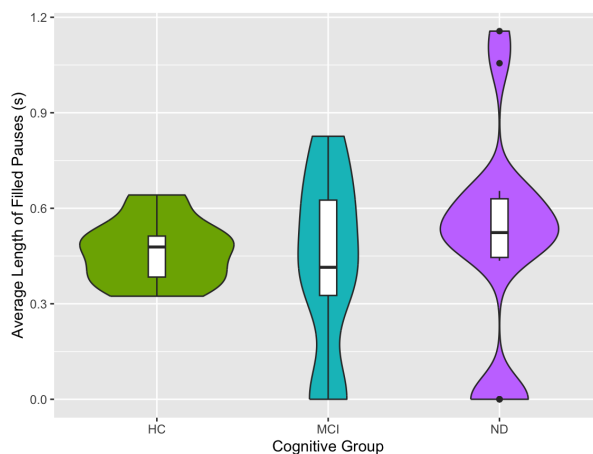


Figure 4.9: Violin plot showing the average length of filled pauses (s) according to cognitive group.

4.4.4 Repetitions

There were not enough instances of any of the different kinds of repetition found in this data to perform statistical testing. Table 4.3 shows the total numbers of repetitions produced per cognitive group. Although the number of word repetitions seems elevated in the HC and MCI groups when compared to the ND group, the majority of these repetitions came from a singular participant in each group. These participants accounted for eight of the word repetitions in the MCI group, and nine of the instances in the HC group.

Cognitive Group	Part Word Repetitions	Whole Word Repetitions	Phrase Repetitions	Total
HC	1	12	8	21
MCI	3	15	5	23
ND	8	9	7	24

Table 4.3: Number of part-word, whole-word, and phrase repetitions produced in the picture description task according to cognitive group.

4.4.5 Prolongations

Figures 4.10 and 4.11 demonstrate that this second analysis found no statistically significant differences between the number and average length of prolongations across the levels of CD.

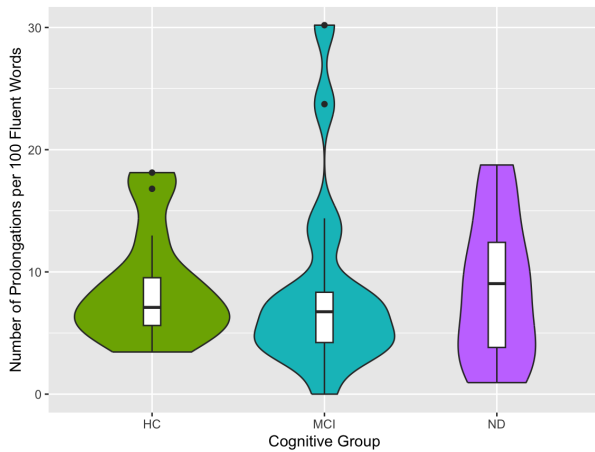


Figure 4.10: Number of Prolongations

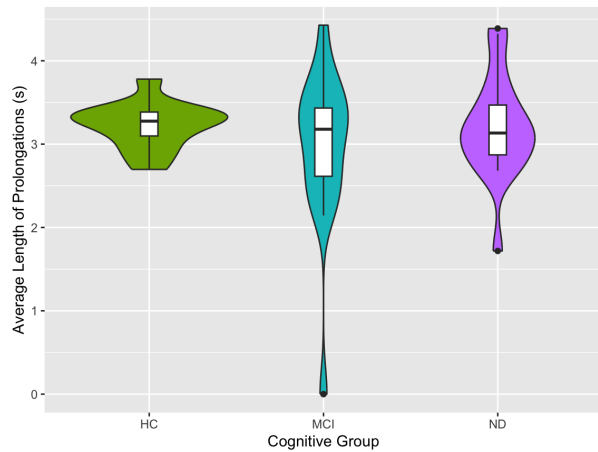


Figure 4.11: Length of Prolongations

4.4.6 Speech Errors

Despite the adjustments made to the disfluency schema, little information was gained from investigating the different kinds of speech errors in this experiment. Only very small numbers of each speech error were present in the data. For example, only one instance of a circumlocution error was found, and only two instances of deletions, as shown in Table 4.4.

Cognitive Group	[add]	[sub]	[dele]	[cir]	[ble]	[lex]	Total
HC	0	1	0	0	1	1	3
MCI	0	0	1	0	1	0	2
ND	0	0	1	1	0	3	5

Table 4.4: The number of additions, substitutions, deletions, circumlocutions, blends, and lexical retrieval errors produced in the picture description task according to cognitive group.

4.4.7 Discussion of Results from the Second Manual Disfluency Study

This analysis found significant differences between the number of fluent words produced by people at different levels of CD when completing a picture description task. One possible explanation for this is that participants in our control group are aware of the fact that they are cognitively healthy. This could perhaps lead to participants feeling as though they need to prove that they are healthy by really trying to inflate their vocabulary and the number of features of the picture they decide to talk about. A similar conclusion was drawn in a recent study analysing the same IVA data, which concluded that healthy controls are both willing and able to “show off” their memory skills, whereas patients with AD are generally unable to do the same [Walker et al., 2023]. However, it could also be the case that the participants in the cognitively impaired groups lack the capacity to pick up as many details to talk about as the HCs. In likelihood, it is some combination of the two factors.

Below are examples of the responses given in the picture description task; one from a participant in the control group and one from a participant in the ND group.

Participant 0251 - Healthy Control:

“Well a little boy is climbing on a stool, he looks as though he’s trying to get some biscuits out of a jar. He’s got the lid off but it’s... Got a biscuit. He looks as though he might be handing it to presumably his sister but the stool is overbalancing. He’s going to go with an almighty crack and presumably... The children’s mother has got her back to them so

she's not really noticing so it's going to be a bit of a shock for her. They're all in the kitchen. She's busy drying a plate but she's obviously distracted because she's, um, the tap isn't turned off and there's a flood coming out of the sink cascading onto the floor. I should think her feet would be wet by now. But whether she's daydreaming there doesn't appear to be anything happening out of the window for her to look at so I'm not quite sure what's going on there, but it looks as though there's a potentially rather messy and quite dramatic scene about to happen."

Participant 2173 - Neurodegenerative Dementia:

"Right okay, um, well this- the children seem to be wanting something at the top of the cupboard and they're trying to get to it but the stool's tipping over, er, and mum's there and not even looking really, she's drying the dishes."

From the two examples above, it is easy to observe the differences in the detail of the responses. Participant 0251 captures most of what is happening in the scene. Their response follows an order of salience, with the main points of the picture being mentioned first before moving onto the less salient features (such as the fact that there is nothing happening outside the window). This participant talks about abstract concepts such as the mother daydreaming, and also references causal relationships such as the sink overflowing *because* the lady in the picture has left the tap on. This participant also seems to have no difficulty with referential cohesion. In contrast, Participant 2173 relays much less detail about the picture, only describing the most prominent aspects of the scene.

Our data showed that the number of fluent words produced by participants decreased as the level of CD increased. This contrasts with findings from Ripich and Terrell [1988] that showed that people with AD produced twice as many words as their HC group. However, this study involved people with dementia conversing with a human interviewer, which could be the cause of the inflated number of words that Ripich and Terrell observed.

In terms of disfluencies, our analysis found significant differences in the rate and average length of unfilled pauses between the HCs and the cognitively impaired groups. There

were no significant differences observed for the rate and average length of filled pauses. There were not enough instances of repetitions or speech errors present in the Cookie Theft data to perform a statistical analysis. This analysis also did not reveal any statistically significant differences in the rate and length of prolongations in the Cookie Theft data.

4.5 Comparison of Results from Manual Disfluency Studies One and Two

The first disfluency analysis presented in this thesis looked at disfluencies produced in natural, spontaneous speech. The present analysis investigated disfluencies from speech that is more constrained in terms of the topic, but still produced spontaneously. Numerous differences were found in terms of the presentation of disfluencies between the two groups, as discussed below.

4.5.1 Total Disfluencies

Our results found that, regardless of the task, disfluency rates exhibited in speech increase along with the severity of CD. The increase is particularly noticeable in the interview task, where HCs had a disfluency rate of 13.7 compared to the rate from the ND group of 27.1, shown in Table 4.5.

	Interview Task		Picture Description Task	
	Total Disfluency Rate	Disfluency Rate Excluding UFP	Total Disfluency Rate	Disfluency Rate Excluding UFP
HC	27.8	13.7	25.6	12.5
MCI	43.5	20.8	33.1	16.1
ND	53.7	27.1	36.9	16.6

Table 4.5: Comparison of total disfluency rates and disfluency rates excluding unfilled pauses between the two different virtual agent tasks. Average rates are calculated per 100 fluent words.

4.5.2 Pauses

Table 4.6 shows the general trend of higher levels of impairment correlating with higher rates and longer lengths of pauses across both tasks. One of the biggest differences found was in the average rate of filled pauses for the ND group between the two tasks. In the interview task, the rate of filled pauses was more than twice as much as the rate of filled pauses found in the picture description task.

	Interview Task				Picture Description Task			
	Average UFP Rate	Average UFP Length	Average FP Rate	Average FP Length	Average UFP Rate	Average UFP Length	Average FP Rate	Average FP Length
HC	14.17	0.54	7.48	0.43	13.15	0.57	5.39	0.48
MCI	22.67	0.87	8.25	0.55	16.99	0.95	7.20	0.48
ND	26.57	0.92	10.70	0.53	20.01	1.33	4.86	0.61

Table 4.6: Comparison of pause rates and lengths between the two different virtual agent tasks. Average rates are calculated per 100 fluent words.

4.5.3 Repetitions

It is not as easy to see patterns in these results as it was for the pause results discussed above. However, this data does show that generally there are lower repetition rates in the picture description task when compared to the interview task. It is also generally the case that the rate of repetitions increases with the severity of the CD (Table 4.7).

	Interview Task			Picture Description Task		
	Average PWREP Rate	Average WREP Rate	Average PHREP Rate	Average PWREP Rate	Average WREP Rate	Average PHREP Rate
HC	0.32	1.08	0.29	0.04	0.55	0.37
MCI	0.58	1.60	0.31	0.18	0.84	0.30
ND	1.05	1.68	0.48	0.55	0.55	0.43

Table 4.7: Comparison of part-word, whole-word, and phrase repetition rates between the two different virtual agent tasks. Average rates are calculated per 100 fluent words.

Task	Interview Task			Picture Description Task		
	HC	MCI	ND	HC	MCI	ND
Cognitive Group						
Min # [pro]	0.42	1.98	2.48	3.45	0	0.94
Max # [pro]	8.78	50.6	86.1	18.1	30.2	18.8
Standard Deviation	2.05	12.0	23.1	4.58	7.76	5.86
Mean	6.68	11.2	13.4	8.79	8.73	8.54

Table 4.8: Comparison of prolongations per 100 fluent words between both tasks.

4.5.4 Prolongations

The control group had a higher number of prolongations per 100 fluent words in the picture description task compared to the interview task. The number of prolongations produced by participants in the **MCI** and **ND** groups remained similar, although we observed a wider range of prolongation rates for the **ND** group in the picture description task (see Table 4.8).

4.5.5 Speech Errors

The presence of all speech errors was reduced in the Cookie Theft data. Only 12 instances were found across all 48 recordings. Although the number of speech errors was still small in the interview data, 32 speech errors were found in total. This suggests that speech errors are more common in spontaneous, interview-style speech compared to picture descriptions. Although when comparing total speech error rates across the two tasks we observe a slight increase for the healthy controls in the picture description task, in practice this relates to only one more error than was found in the interview task (Table 4.9).

4.5.6 Discussion

The incidences of speech disfluencies were much higher in the interview task when compared to the picture description task. The most commonly observed disfluency across

	Speech Error Rate Interview Task	Speech Error Rate Picture Description Task
HC	0.01	0.09
MCI	0.17	0.06
ND	0.26	0.12

Table 4.9: Comparison of total speech error rates per group between the two different virtual agent tasks. Average rates are calculated per 100 fluent words.

both tasks were the unfilled pauses, but even when these are excluded from the total disfluency rates we still observe higher rates in the interview task. The disfluency rates observed here are higher than commonly reported rates for normal, spontaneous speech in healthy adults (around 6/100 words). There are two factors at play here. Firstly, the majority of our subjects are not healthy controls. Our research supports the general finding that the severity of CD can result in more disfluencies. Although previous research has found that this is mainly observed in higher levels of unfilled and filled pauses, our research demonstrates that the frequency of all disfluencies increases with the severity of CD. Secondly, our studies involve participants conversing with a digital avatar, rather than another human being. As discussed in Section 2.4.4, research so far has suggested that disfluency rates are typically lower in human-computer interactions. However, this view was predominantly based on human responses that are much shorter (responses of around five words long) than those in our datasets, where both tasks result in participants forming long and complex sentences.

Turning to individual disfluency groups, our study found that participants exhibit more of each category in the interview task compared to the picture description task (see Appendix D for a table of all findings from the picture description task). The ND group exhibited more than twice as many filled pauses in the interview task compared to the picture description task. It is possible that the cognitive load of the first task is contributing to this difference. Fraundorf and Watson [2011] note that filled pauses (referred to as fillers in their study) aid in recall. Recall is important in the interview task as patients are asked a mixture of both long term and short term memory questions, whereas recall is not necessary in the picture description task. Research has also shown that filled pauses

can be used as a method of saving face when trying to remember the answer to a question [Smith and Clark, 1993]. This opens up another discussion; whether people feel the need to “save face” when talking with a digital avatar. To answer such a question is beyond the scope of this thesis, but interesting discussions on this topic can be found in Baron [2015] and Guzman [2018].

4.6 Automatic Cognitive Decline Classification

In order to investigate whether disfluency information could help improve the accuracy of an ACDC system, a proof-of-concept style experiment was designed to test a baseline system with different feature combinations (including manual and automatically extracted disfluency information) to see which combination provided the best classification accuracy. This work was presented at the International Congress of Phonetic Sciences in 2023, and the sections below describe this study and its results [Thomas et al., 2023]. The work on automatically extracted features was completed by a colleague from the Department of Computer Science at the University of Sheffield, whilst the disfluency analysis presented in the first half of this chapter provided the transcripts and disfluency feature set.

4.6.1 Background

This proof-of-concept experiment focussed specifically on using interpretable features for ACDC. In the fields of Machine Learning (ML) and Artificial Intelligence (AI), interpretability refers to how much of an algorithm, feature set, or system as a whole can be understood by humans. The more interpretable an ML model is, the easier it is for humans to understand the decisions that are made by the system (a gentle but thorough discussion surrounding interpretability in ML can be found in Molnar [2022]).

Interpretability is especially important when working within a healthcare field as numerous studies have shown that both doctors and patients lack trust in ML models, which is exacerbated by a lack of understanding in how these models work (Feldman et al. [2019]; Juravle et al. [2020]; Hallowell et al. [2022]). The interpretable features used for this study were traditional linguistic features and newly-proposed disfluency features based on the

manual disfluency analysis presented above. **ACDC** systems typically do not make use of disfluency information other than those that can be extracted acoustically, such as un-filled pauses. The use of transcript-derived disfluency information relies on the existence of transcripts that capture disfluency information. As discussed in the Chapter 2, **ASR** transcripts usually do not contain such information.

We also included some acoustic features in this study. These features are interpretable in the sense that they are derived from algorithms that have been hand-coded. However, the lack of interpretability for acoustic features comes from the abstractness of the features (for example, what does a high 13th value of an **MFCC** actually mean?), along with the fact that there are vast numbers of these features used, making it difficult to detangle exactly what is working for the classification system. However, acoustic features are very commonly used in **ACDC** systems, so we included them in order to compare our disfluency features with a “typical” **ACDC** pipeline.

The aim of this study was to test whether disfluency features in particular could improve the accuracy of **ACDC** systems whilst also providing interpretable information. This was assessed in two ways; training the model on the manually produced disfluency transcripts, and training the model on automatically produced disfluency analytics.

4.6.2 Data

This study used the same subset of participants as used in the manual disfluency analysis as described above (48 Cookie Theft picture description tasks). Participants belonged to one of three groups; **HC**, **MCI**, or **ND**. For the classification task we used a binary classification of either healthy control or cognitively impaired, combining the **MCI** and **ND** groups together, as we were working with a small dataset. Each recording in our experiment came from a different participant to ensure there was no overlap between our training and testing sets.

4.6.3 Methodology

4.6.3.1 Manual Disfluency Features

The manual disfluency transcripts were made according to [DisCo2.0](#) (see above). As with previous analyses in this thesis, any diagnosis labels were hidden during annotation to mitigate potential bias.

4.6.3.2 Automatic Disfluency Features

[VAD](#) was used to identify portions of the recordings that contained speech, and a Common Voice [[Ardila et al., 2019](#)] based [ASR](#) system was used alongside a phoneme recogniser to transcribe the speech and disfluencies present in the recordings.

4.6.3.3 Automatic Linguistic Features

Linguistic features used for this task included content density, speech rate, syntactic complexity, and utterance length. These were extracted using a combination of part-of-speech taggers (POS), tokenisers (Tok), semantic taggers (Sem_Tag), syntactic tree parsers (Tree), word lists (W_List), co-reference taggers (CoRef), and part-of-speech pattern matchers (POS_Pat). [Table 4.10](#) details these features.

Feature(s)	n=#	Automated Components
Content Density	1	Tok, POS, W_List
Part-of-Speech Rate	45	Tok, POS
Reference Rate to Reality	1	Tok, POS
Personal, Spatial and Temporal Deixis Rate	3	Tok, POS, W_List, CoRef
Relative Pronouns and Negative Adverbs Rate	2	Tok, POS, W_List
Lexical Richness	3	Tok
Action Verbs Rate	1	Tok, POS, Sem_Tag
Frequency-of-Use Tagging	1	Tok, W_List
Propositional Idea Density	1	Tok, POS, POS_Pat
Mean Number of Words in Utterance	1	Tok
Number of Dependent Elements Linked to the Noun	2	Tok, POS, Tree
Global Dependency Distance	2	Tok, POS, Tree
Syntactic Complexity	1	Tok, POS, Tree
Syntactic Embeddedness	2	Tok, POS, Tree
Utterance Length	2	Tok

Table 4.10: Linguistic features used within this study (as used by [Fraser et al. \[2016\]](#)), the number of features (n) within each category, and the automated elements needed for each feature to function.

4.6.3.4 Automatic Acoustic Features

Acoustic features were extracted using the openSMILE toolkit in Python [[Eyben et al., 2010](#)]. Commonly used feature sets (eGeMAPS, emobase, ComParE) were chosen as these have been well established for use in the field (as described in Section 2.5.1.2).

4.6.4 Baseline System

Support Vector Machines (SVMs) were trained using the acoustic feature sets mentioned above and were developed using 5-fold cross validation. For a binary (HC-MCI/ND) classification task, the baseline performance was 78.4%.

4.6.5 Results

An SVM trained only on the manual disfluency features was found to be the most accurate in this classification task, with an accuracy of 88.8% (Table 4.11). Whilst the model trained on only automatic disfluency features demonstrated only a slight improvement to the baseline model, a substantial improvement was seen when automatic disfluency features were combined with the traditional acoustic and linguistic features (83.2%).

Model	Accuracy
Acoustic Only	77.6%
Linguistic Only	57.6%
Acoustic + Linguistic	68.8%
Manual Disfluency	88.8%
Automatic Disfluency	78.4%
Acoustic + Linguistic + Automatic Disfluency	83.2%

Table 4.11: Binary SVM Accuracy Results (5-fold cross validation)

4.6.6 Discussion

Our results demonstrate that disfluency features have the potential to enhance the performance of ACDC systems. Although manually produced disfluency features performed the best, there was still a large improvement in accuracy when automatically produced disfluency features were used in combination with acoustic and linguistic features.

Our results also found that traditional linguistic features performed the worst on our data in the binary classification task (with an accuracy of 57.6%). This contrasts with the performance of these features on the Pitt Corpus, which we found produced an accuracy of 78.8%. The Pitt Corpus is part of the DementiaBank dataset [Becker et al., 1994], and is arguably the most widely used corpus of dementia affected speech. This corpus is used in the popular Interspeech ADReSS challenges, designed to encourage researchers to actively participate in dementia recognition tasks (see Luz et al. [2020]). However, this results in a situation where almost all research into ACDC systems use the Pitt corpus data, which in turn results in feature sets becoming fine-tuned to that data. This means

that feature sets that seem to result in high system accuracy only achieve those results on specific data, which is a problem when trying to use the feature sets on any other data. This poses significant challenges for the generalisability of feature sets, and suggests that the high accuracy commonly reported for [ACDC](#) systems is not a realistic representation of the current capabilities of these systems for use on real-world data.

4.7 Conclusion

This chapter presented a second manual disfluency analysis. The [DisCo](#) schema was updated to provide more accurate classification of prolongations and speech errors, and was used to facilitate the creation of disfluency transcripts of picture description tasks from people at three different levels of cognitive ability ([HCs](#), people with [MCI](#), and people with [ND](#)). Differences were found in the presence and duration of disfluencies between the interview task presented in chapter two and the picture description task presented here. Whilst there were fewer statistically significant differences in the disfluencies from the picture description task, a proof-of-concept study which used disfluency information in an [ACDC](#) system found that disfluency information improved system performance whether this information was produced manually or automatically.

Our results also demonstrate that even if statistical analysis does not find significant differences in disfluency information, the information is still useful in an [ACDC](#) pipeline. This is because significance tests look at the discriminatory power of one feature at a time (for example the length of filled pauses). However, the [SVM](#) is able to exploit the relationship between multiple different features and the experiment demonstrates that when the disfluency features are used in combination with each other they have the potential to greatly improve the accuracy of cognitive decline classification.

Chapter 5

A Conversation Analysis of Human-Avatar Data

Contents

5.1	Introduction	137
5.2	Conversation Analysis in Medical Interactions	137
5.2.1	Problem Presentation Phase	138
5.2.2	Communicating with People with Neurodegenerative Dementia	145
5.2.3	Using Conversation Analysis to Study Human-Computer Interactions	146
5.2.4	Conversation Analysis as a Diagnostic Tool	148
5.3	Methodology	150
5.3.1	Participants	151
5.3.2	Procedure	152
5.4	Analysis	153
5.4.1	Comparison of Problem Presentation Phases Initiated by a Human Doctor vs an Intelligent Virtual Agent	154
5.4.2	Case Study: Participant 096/0221's Problem Presentations	177

5.5 Conclusion	184
-----------------------	------------

5.1 Introduction

This chapter addresses our final research question:

5. How do patients construct their problem presentation phases in a medical interview with a human doctor versus a digital avatar?

This chapter investigates how patients construct their problem presentation phases when addressing a digital avatar. We begin with a continuation of the background to Conversation Analysis (CA) presented in Chapter 2, but with a specific focus on how CA is used in the field of dementia and health studies. Particular attention is paid to how disfluencies are treated in CA, and what this can tell us that extends beyond a purely quantitative approach to investigating speech. The analysis begins with an investigation into how patients are formulating their responses to the opening questions of a problem presentation phase in a medical consultation. The data used in this chapter primarily consists of the Intelligent Virtual Agent (IVA) data used in previous chapters, along with some control data where a human doctor is leading the consultation instead of an avatar. We then narrow our focus to a single participant who had completed both the human doctor and the virtual agent studies, and present a microanalysis of the similarities and differences observed between the two recordings. This is followed by a discussion of the findings from this investigation.

5.2 Conversation Analysis in Medical Interactions

As discussed in Chapter 2, CA is the study of talk in interaction. CA has been used to investigate medical interviews and doctor-patient interactions almost since the emergence of CA as a methodology.

One of the earliest examples of an investigation into doctor-patient conversations was presented in Byrne and Long's study of primary care encounters in 1984. This study involved an analysis of 2,500 doctor-patient interactions, and focussed primarily on the

behaviours of the doctors during these interactions. Building on this work, [Heritage and Maynard](#) published their collection of papers in 2006 which again investigated doctor-patient interactions. However, unlike [Byrne and Long](#)'s study, this work was centred around a co-constructive approach examining the conduct of both the doctor and the patient, and the relationships between the two. [Heritage and Maynard](#) posit that a set structure can be observed across a range of different acute primary care visits, as shown in Figure 5.1.

- I Opening: doctor and patient establish an interactional relationship
- II Presenting Complaint: the patient presents the problem/reason for the visit
- III Examination: the doctor conducts a verbal or physical examination or both
- IV Diagnosis: the doctor evaluates the patient's condition
- V Treatment: the doctor (in consultation with the patient) details treatment or further investigation
- VI Closing: the consultation is terminated

Figure 5.1: The Overall Structure of Primary Care Visits according to [Heritage and Maynard](#) [2006, p.14].

The analysis presented in this chapter focuses specifically on section two of the above structure; the presenting complaint or the *problem presentation phase*. As noted above, this phase starts after the expected introductions or re-familiarisations have been made between the doctor and the patient. The problem presentation phase can be analysed in two distinct ways; in terms of how the doctor is asking the patient about their concerns, or the manner by which the patient chooses to present their concerns to the doctor.

5.2.1 Problem Presentation Phase

The problem presentation phase of a medical interview is the only phase in which patients are able to describe their illnesses on their own terms [[Heritage and Robinson, 2006](#), p.89]. It is common to observe patients justifying their need for a visit to a doctor [[Teas-Gill and Roberts, 2012](#)]. This often manifests as patients demonstrating that they have tried to deal with the problem themselves first, or that they have waited some time before visiting a doctor to see if things would improve on their own [[Halkowski, 2006](#)].

Problem presentation phases will also differ according to whether or not the patient is accompanied during their visit. In a study investigating how parents present their children's health concerns to a paediatrician, [Stivers \[2002\]](#) found two different methods commonly employed. Firstly, a “symptoms only” problem presentation involves parents simply describing the symptoms that their child is experiencing. The second category, “candidate diagnosis”, includes the addition of a suggested diagnosis alongside the description of the symptoms. [Ijäs-Kallio et al. \[2010\]](#) built on this work and proposed a further two categories in addition to the two by [Stivers](#). These are the “diagnosis implicative symptom description” category, in which there is an implied diagnosis of the problem by the patient, or the “candidate diagnosis as background information” category in which patients provide information about previously diagnosed conditions that they feel may be relevant to their current symptoms.

[Lee and Kim \[2015\]](#) focussed more on the differences between how a patient describes their own symptoms compared to how an Accompanying Person (AP) describes the patient's symptoms on their behalf. Their study investigated problem presentation phases taking place in an emergency department of a hospital. They found that when patients describe their own symptoms they are typically brief and include descriptions of the sensations they have experiencing such as pain or discomfort. APs however tended to describe symptoms from an observational point of view, and include much more patient history in their description of what is wrong. The brief responses from patients could be linked to the fact that they know that they are in an emergency department, and are aware that responses should be brief given the high stress environment they are in. This would contrast with a problem presentation phase taking part in a General Practitioner (GP)'s office, where patients know that they have an allotted time and that there is no particular need to rush their problem presentations.

There are two main decisions that conversational partners must make in the context of a problem presentation phase that are of particular importance to the analysis presented in this chapter. Firstly, the doctor must decide how much of their knowledge of the patient's condition should be revealed to the patient. Secondly, a patient needs to decide

how much repetition would be acceptable in their response given what the doctor has already revealed they know.

5.2.1.1 Question Design

The initiation of the problem presentation phase can have a significant impact on how patients respond. In a study of 182 GP-patient interactions, [Robinson \[2006\]](#) notes that even very slight differences in the phrasing of questions by the doctor can result in changes to the action that the question is treated as performing, as shown by the type of response provided. The way in which a doctor phrases these questions displays their perceived reason for the visit. [Robinson](#) identifies three different types of visit that will influence the approach the doctor takes during the problem presentation phase.

Firstly, there are patients who present with a problem that has not been discussed previously with their doctor. This builds on prior work from [Heath \[1981\]](#) in which the author was able to categorise the type of questions used by the doctor in opening the problem presentation phase according to whether the appointment was instigated by the patient (a new appointment) or instigated by the doctor (a return appointment). [Heath](#) found that the new appointment interviews started with open questions such as “*How can I help you?*” or “*What can I do for you?*”. Return appointments were designed with more specificity and frequently referred to known symptoms, such as “*How’s your arm?*” or “*Ah, it’s your foot isn’t it?*” (p.76). A similar investigation into doctor-patient consultations from [Gafaranga and Britten \[2003\]](#) found that the person who initiated the appointment (doctor or patient) was of little significance to their findings, and differences arose depending on whether the visit was new or a follow-up. The key findings from this study were that there are certain “rules” that doctors follow in doctor-patient interactions (such as the rules identified above, like asking “*How are you?*” for a follow-up consultation but starting consultations with a new patient with “*What can I do for you?*”). If these rules are broken, there is scope to repair the sequence. The authors demonstrate this by presenting an example of a doctor opening a problem presentation with “*What can I do for you?*” despite having previously met with the patient about an ongoing condition. This prompts the patient to remind the doctor that they have met before,

resulting in the doctor repairing the sequence by acknowledging his mistake. However, the authors demonstrate that if in instances of “deviations” the sequence is not repaired, there could be resulting discrepancies in the consultation. These are important observations to note when considering data from the IVA dataset. The virtual agent is unable to ask for clarification on any points raised by the patient during their problem presentation phase, and if there are any deviations the virtual agent cannot initiate a repair.

The second type of visit identified by Robinson [2006] are those in which the consultation revolves around an issue that was already discussed at an earlier point. These follow-up visits are centred around the patient’s recovery and how effective the treatment decided on in the previous consultation has been at addressing the patient’s concerns.

The final type of visit concerns routine concerns. These arise from long-term conditions that are under control but require some monitoring, such as a patient dealing with high blood pressure or type one diabetes. These visits are somewhat frequent, and are typically initiated by the doctor formulating their questions around whether the patient is experiencing any new concerns related to their ongoing condition (Robinson focuses on “*What’s new?*” type questions for routine visits).

For both the Hallamshire and the IVA data sets used in this analysis, all visits are treated as *new concern* visits. Participants in the Hallamshire dataset may have spoken to their GP about their memory concerns, but this is the first time the participants are seeing a specialist in a memory clinic. Those in the IVA dataset may have seen a specialist clinician before, but this is their first time interacting with a virtual agent that is taking on the role of a clinician.

In addition to displaying the type of overall interaction that they expect to have with a patient, doctors must also design the talk that they will use to open the problem presentation phase of medical interviews. Heritage and Robinson [2006] identified five main types of questions that doctors use when inviting patients to describe their symptoms:

I. General Inquiry Questions

These questions can vary in terms of how much preexisting knowledge the doctor presents

themselves as having, but do not constrain the content of what the patient can respond with. Some examples of general inquiry questions include:

“How can I help you today?”

“What’s going on?”

II. Gloss for Confirmation Questions

These questions are formatted as yes or no questions, but invite further explanation from the patient. They mention symptoms but in a general manner, such as:

“So you’ve been feeling sick?”

“Sounds like you’re uncomfortable?”

III. Symptoms for Confirmation Questions

This type of question is a request for confirmation of concrete symptoms. They constrain patients from repeating information that the doctor has already signalled they’re aware of, and discourage elaboration.

“So you’ve got a headache and a sore throat?”

“Your knee has been hurting for three weeks?”

IV. How Are You Questions

These questions are formatted to elicit general information, rather than an immediate invitation to a problem presentation. This type of question can come with some ambiguity, probably due to the fact that these questions are commonly observed in ordinary conversation openings and typically invite a general and lacklustre response [Jefferson, 1980]. Examples of these questions are:

“How are you feeling?”

“How are you doing?”

V. History-Taking Questions

These are close-ended questions that require highly constrained responses. Although

not a common occurrence according to Heritage and Robinson's study, it is possible for patients to work around the constraints presented by this type of question and respond with a problem presentation. An example of this kind of question would be:

"Have you had any aches and pains?"

As part of this work, Heritage and Robinson investigated the frequency of the above question types in different medical consultations, and found general inquiry questions to be the most common (62%). They also found that patients responded with significantly longer problem presentations and included more information about current symptoms when faced with a general inquiry question rather than a confirmatory question.

5.2.1.2 Patient Response

As discussed above, the way a question is phrased by a doctor will have an effect on how the response is phrased by the patient. Another factor that will inform a patient's problem presentation phase is their own knowledge of the symptoms they have been experiencing. Heritage and Robinson [2006] define two broad categories; *known* and *unknown* problems. Known problems can be "routine"; things that numerous patients will probably encounter at some point in their lives, such as the flu or a sore throat. They could also be "recurrences", where a patient is experiencing symptoms that have previously been the object of medical diagnosis and treatment (p.50). Unknown problems on the other hand are those that a patient has not experienced before. Patients experiencing unknown problems may have difficulty describing exactly what the problem is to their doctor. They may also include information about what they thought the problem was, until something happened to make them question their original stance and thus prompting them to seek medical advice. Halkowski [2006] names this device a "*At first I thought X*", and argues that this is a method of displaying oneself as reasonable and not dramatic by investigating mundane causes first, only seeking help when turning up empty-handed. This serves as a way of justifying the patient's visit.

Elsley et al. [2015] show how the details of patients' problem presentation phases provide diagnostic value to clinicians. Their study included 25 participants who were diagnosed

with either Neurodegenerative Dementia (ND) or Functional Memory Disorder (FMD), and their analysis revealed that each condition had a distinct conversational profile. Patients with FMD were able to respond in detail to questions and often included additional details that were unprompted. These patients were able to respond to compound questions, and to display knowledge of when they were repeating themselves. Conversely, patients with ND could offer little detail in response to questions, often failed at addressing both parts of a compound question, and frequently repeated themselves without displaying any awareness of doing so. This study provided the basis for the collection of data which formed the Hallamshire dataset (described in more detail in Section 5.3.1).

5.2.1.3 Ending the Problem Presentation Phase

An additional area of interest is how the problem presentation phase is completed, and how the consultation moves on afterwards. This is an area in which there are sometimes mismatches in the agendas of the doctor and the patient. The doctor is acutely aware that they have other appointments scheduled, and that there is specific information they require from the patient before they can move on. The doctor does not necessarily need a full medical history from the patient. However, for the patient, they are not aware of the level of information required by the doctor, as they do not share the same level of medical knowledge. This can often result in patients taking the floor for longer than the doctor has time for.

It is not uncommon to observe doctors interrupting patients before they are finished recounting their symptoms. Beckman and Frankel [1984] found that in 69% of the medical interviews they examined, the doctor interrupted a patient's problem presentation phase by directing the conversation towards a specific symptom, only letting the patients talk for an average of 18 seconds before taking the floor and moving the conversation forward themselves. In only one case of the interviews that had been interrupted was a patient able to go on to successfully complete their problem presentation. In a follow up to this work 15 years later, Marvel et al. [1999] found that of the 264 doctor-patient interviews they analysed only 28% of the patient's initial problem presentation concluded without being interrupted or redirected by the attending doctor. These findings raise concerns

that symptoms could go undiscussed and that valuable information could be missed. This points to a potential benefit of using an IVA to conduct medical interviews. The virtual agent that is used as part of the CognoSpeak system does not have the ability to interrupt a patient as they are describing their problems. Therefore, patients have room to describe their symptoms as much as they wish, and the responsibility of moving the conversation forward to the next stage of the medical interview rests solely on the patient.

5.2.2 Communicating with People with Neurodegenerative Dementia

Although problem presentation phases have been studied regularly, and there are accepted stages of medical interviews that both the doctor and the patient will both adhere to, there are some conversational barriers that are specific to conversing with people experiencing cognitive impairment. For example, people with Alzheimer's Dementia (pwAD) often exhibit difficulties in participating in communication, which manifests as problems with events such as greeting behaviours or engagement with the interaction [Rousseaux et al., 2010]. This could result in shorter or more disjointed problem presentation phases. However, various studies have found that whilst Cognitive Decline (CD) can affect how people with dementia communicate, certain structural aspects of conversation are retained. Hamilton [1994] conducted a longitudinal study of herself conversing with a dementia patient over a period of four years. Hamilton found that while the patient's ability to formulate appropriate responses decreased over time in line with the effects of their dementia, the patient retained the ability to participate appropriately in turn-taking structures. This is highlighted when Hamilton notes that at the start of the study the patient was able to issue requests, ask for clarification, and express wishes (amongst other things), but by the end of the study the patient was only able to respond to utterances and had seemingly lost the ability to initiate a verbal exchange. A similar conclusion was reached by Müller and Guendouzi [2005] who observed that conversation skills were maintained in their study participants despite the memory problems the participants were experiencing. Even when memory disruptions were severe, participants were able to con-

tribute to the continuation of the conversation (p.400).

These studies demonstrate that whilst people with Neurodegenerative Dementia (pwND) can still participate in conversations they often require prompts to do so. They may also have difficulty beginning new topics, or even maintaining a conversation beyond a minimum exchange. This suggests that for effective communication with pwND, the conversational partner may need to adjust their usual communicative practices.

5.2.3 Using Conversation Analysis to Study Human-Computer Interactions

Although CA developed as a way of understanding human-human communication, a more recent body of research has seen CA employed to analyse human-robot interaction. Broadly speaking, there are three main types of verbal human-computer interactions.

Firstly are the interactions in which a computer is required to respond to a request from a user. These systems are becoming increasingly popular thanks to the rise of smart home technologies such as Amazon's Alexa, or on-device intelligent assistants such as Apple's Siri. Generally speaking, these assistants are designed to answer basic questions (such as "what will the weather be like tomorrow?") or perform simple tasks ("set a five minute timer"). These systems are not (yet) designed to converse with users and as such are unable to hold back-and-forth conversations with users [Skantze, 2021].

The second type of interactions are those with social robots. These systems are designed to perform a broad range of tasks, but the differentiating factor is that these robots give a sense of real "communication" with the user. They respond to questions, hold conversations, and some can even simulate human emotions. Such systems hold promise for being companions of elderly people and could help to alleviate loneliness amongst dementia patients. Gasteiger et al. [2021] provide a comprehensive review of computer agents that are designed to combat loneliness in the ageing population.

The third type of interactions are an inverse of the first; the system is designed to ask questions to users in order to find out some specific information. The CognoSpeak system (Section 2.5.3) belongs to this group. Such systems are not yet as commonplace as the

first type, but companies such as [Therapy Box Ltd \[2024\]](#) (the company working with researchers at the University of Sheffield on the CognoSpeak system) and [Thymia Ltd \[2024\]](#) (who are working towards systems to objectively measure mental health conditions through speech analysis) are working to change this.

Most of the work surrounding verbal human-robot interaction is concerned with how the robot in question can be made to appear more natural; there is little existing work that focuses on how humans are adjusting their own behaviour and language towards the robot they are interacting with. One study from [Pelikan and Broth \[2016\]](#) used CA to investigate the changes humans make when interacting with a humanoid robot called Nao. Participants were sat in front of a Nao robot and told that the robot would initiate the interaction and participants should follow the robot's lead. This study found that many key constructs of human interaction were maintained throughout the human-robot interaction. For example, most participants would follow normal conversational rules and introduce themselves to the robot after it initiated a conversation (typically with a "hello" statement). The human participants also adjusted their behaviour depending on what the robot was doing. They quickly learned that when the robot asked a yes or no question, responses such as "sure" or "of course" should be abandoned as they were not recognised as acceptable answers by Nao. Participants were also observed simplifying their language as the interaction continued, eventually using primarily single words in the conversation. When it appeared to participants that Nao was having a hard time "hearing" what they were saying, they were quick to adjust the volume of their speech or alter their pitch to emphasise specific pieces of information (p.4929). This study emphasises the fact that cognitively healthy adults can adjust quickly to their conversational partner, even if that partner is a machine.

One of the more prominent researchers investigating how humans converse with machines is Clifford Nass. The bulk of Nass' work surrounds his claim that humans behave the same way with robots as they do with other humans. His research suggests that not only do humans attribute human-like qualities such as gender to machines [[Nass and Brave, 2005](#)], we also tend to treat machines with politeness despite *knowing* that we are not interacting

with a sentient being [Nass, 2004]. A key feature of Nass' theory is that this process is unconscious ("mindless"), and is a result of social evolution rather than a conscious choice [Nass, 2004]. However, this view is contested by Kerstin Fischer, another prominent researcher in the field. Fischer's work argues that the fact that there is variation in how people talk to robots [Fischer, 2006] along with the fact that some users do not treat machines like humans [Fischer, 2011b] challenges Nass' hypothesis. Fischer instead suggests that the biggest influence on how a human will communicate with a machine is dependent on what each person believes about the capabilities of the machine, and this will change throughout the course of an interaction as the human learns more about the machine [Fischer, 2011a].

5.2.4 Conversation Analysis as a Diagnostic Tool

Conversation analytic methodology holds that talk-in-interaction contains claims and displays of understanding, but does not inherently reflect inner psychological states [Antaki and Wilkinson, 2013]. This does not, however, preclude the use of CA as a diagnostic aid; the ways that talk is used may reflect different underlying cognitive states. Plug et al. [2009] used CA to identify differences in how people described the experience of having an epileptic seizure vs. a non-epileptic seizure. They found numerous different linguistic, topical, and interactional diagnostic features. Those with non-epileptic seizures tended to avoid describing their own subjective seizure experiences, whereas those with epilepsy would volunteer such information and could discuss their symptoms in detail. Epileptics also describe an active effort to fight against a seizure, whereas non-epileptics give no descriptions of actively struggling against a seizure. Findings from this study informed a further investigation by Ekberg and Reuber [2015]. This second study investigated the differences between how neurologists approached the task of history-taking in routine clinical interviews compared to the approach taken in the research interviews analysed by Plug et al. [2009]. In the research interviews, the clinician generally started with open-ended questions, allowing patients to decide for themselves what they wanted the focus of the conversation to be. In contrast, the opening questions of the routine medical interviews were much more constrained and focussed the topic immediately on seizures. As

Ekberg and Reuber note, this is a time-effective method of gathering information from the patients but it does limit how much the patient can contribute to the conversation. This in turn limits the potential for identifying some of the linguistic differential diagnostic features identified in the earlier research interviews. Ekberg and Reuber therefore suggest that doctors should take a more passive approach to these medical interviews, allowing the patients to produce uninterrupted speech which can then be used for a diagnostically relevant linguistic and conversation analysis.

Abbas et al. [2022] investigated the conversational differences between neurotypical children and children with autism. They found that, compared to the neurotypical children, the children with autism take fewer turns in a conversation, produce speech that is abundant with pauses, and may repeat the same words or phrases. The authors claim that an investigation of these differences can help clinicians diagnose autism.

Else et al. [2015] identified numerous differences in conversational skills between people with Functional Memory Disorder (pwFMD) compared to pwND. This study resulted in the development of different conversational profiles which can be used in differentiating between ND and FMD. These profiles are dependent on two key differences; who attended the memory clinic and how the patients responded to the neurologist's questions during the visit. In terms of APs, it was found that 91% of their participants with ND were accompanied, whereas only 40% of patients with FMD were accompanied. In the ND group the APs acted as spokespersons for the patients, particularly during the patients' problem presentation phases where the APs would offer up more information about symptoms or more details about the history of the participants. In terms of how the patients were responding to the neurologist's questions, one question in particular resulted in a clear distinction between pwND and pwFMD. In most of the interviews Else et al. investigated, the clinician would inevitably ask some variation of the question "*Who is the most concerned about your memory?*". Upon analysing the patients' responses they observed that all of the patients with FMD said that they themselves were the most concerned about their memory problems. In four out of five of the participants with ND, it was found that the AP was the most concerned about the memory issues, and frequently the ND patients themselves were not aware of any problems or failed to respond to the

question.

Mirheidari et al. [2017] built on the work from Elsey et al. [2015] and incorporated CA derived information into an automatic system designed to differentiate between pwND and pwFMD. Leveraging CA information in their automatic system yielded a system accuracy of 97%, demonstrating the diagnostic potential of systems that can automatically analyse conversational data.

5.3 Methodology

The analysis presented in this chapter examines the differences in how people interact with a real-life clinician versus a virtual agent asking them questions about their memory complaints. We make use of two existing datasets from the university of Sheffield. The bulk of the data used in this analysis is the same data that was used in the first manual disfluency analysis (Chapter 3). We include responses from people with FMD, Mild Cognitive Impairment (MCI), and ND. In addition to the IVA data, we make use of nine responses taken from participants in the Hallamshire dataset [Elsey et al., 2015]. This dataset consists of recordings of people interacting with a human clinician who is asking them questions in order to investigate memory complaints. A total of 99 patients were recruited as part of this study, and their neurology consultations in a memory clinic service at the Royal Hallamshire Hospital in Sheffield were recorded. A number of different clinicians conducted the consultations in this dataset, but the present analysis focuses on recordings from just one clinician (allowing the analysis of nine responses). This decision was made in an attempt to control for the variation that will be exhibited by different clinicians during a medical interview. Before patients attended the memory clinic they were told that they could bring someone with them. Patients completed a neuropsychological battery of tests including the Mini Mental State Examination (MMSE), short and long term memory tests, and tests for executive functioning. The collection of this data was funded by the National Institute of Health Research, and full consent was obtained from all participants.

In both datasets, the first question asked to participants (after necessary introductions are

made) is designed to discover what they have been experiencing lately in terms of their memory problems, and introduces the problem presentation phase as discussed above. In the IVA dataset, the question is “*Tell me what problems you’ve noticed with your memory recently*”. In the Hallamshire dataset there is more variation in how this question is phrased, although these questions do all belong to the general inquiry type of questions identified by Heritage and Robinson [2006], as discussed in Section 5.2.1.1.

The main difference in the interview sections of the recordings from both datasets is that the IVA data, the participant takes as much (or as little) time answering the questions as they like, and then presses a button on a computer to move on. In the Hallamshire data, the human clinician is able to ask for clarification if necessary, or ask for more detail from the participant. Of course, it is also possible that the clinician could choose to interrupt the participant, or redirect them if it seemed they were veering off-topic. Our investigation is particularly interested in the differences between the responses to the initial question about participants’ symptoms, therefore the entire response to this question is considered as part of our analysis. In the IVA data, we treat the problem presentation phase as finished when the participant makes the conscious decision to move on from their own problem presentation, something that would be unlikely to happen in a human-human scenario where the decision to move the conversation onto the next phase would be more co-constructive and involve both the patient and the doctor.

5.3.1 Participants

A total of 56 problem presentation phases from 55 different participants were analysed as part of this work. 47 of these participants come from the IVA data, and the remaining nine are from the Hallamshire dataset. These nine responses form our control group of examples of people with ND attending a memory clinic in a real-world scenario. Additionally, we did not include any responses from healthy controls in this study. This is for two main reasons; first, the healthy controls in the IVA data *know* that they are healthy controls. This means that when they are asked the question about their memory concerns by the virtual agent, their responses are primarily something akin to “*There’s*

nothing wrong with my memory” or *“I’m just here to help with the study”*. Second, the Hallamshire dataset does not contain responses from any healthy controls for much of the same reason; it would be unusual that a cognitively healthy person would find themselves the subject of a memory-specific medical consultation in a specialised memory clinic if they had never made any complaints about their memory, and therefore there is no rational basis for collecting such data. We chose to have a mixture of participants across different levels of **CD** and across different datasets to allow for two main comparisons. The first investigates the differences between people with **ND** conversing with a human versus a virtual agent. The second investigates the differences that the level of impairment (**FMD** vs **MCI** vs **ND**) have on conversations with the digital avatar.

Some information on participant age and **MMSE** scores was unavailable. Table 5.1 shows the number of participants according to **CD** group and dataset.

After an initial inspection of the recordings from each dataset, it was discovered that Participant 096 in the Hallamshire dataset seemed to be the same person as Participant 0221 from the **IVA** dataset. This was confirmed by verifying the participant’s date of birth and conducting an in-depth comparison of the two files, something that has not been noted in previous studies that. For the purposes of the present study, the inclusion of both the responses to the digital and human doctor from the same person enabled us to investigate the changes that occur as a direct result of who is opening the problem presentation phase. We do not foresee this having any detrimental impact on the analysis presented below, as even though the participant had prior experience of completing memory tests by the time the **IVA** data was recorded, this was still his first time completing them with a digital avatar (as is the case for the other participants in the **IVA** dataset).

5.3.2 Procedure

All transcriptions were created based on the Jefferson transcription conventions (described fully in **Bolden and Hepburn [2018]**). The full transcription conventions can be found in Appendix E, but the most commonly used symbols are described in Table 5.2. The author

Diagnostic Category	IVA		Hallamshire		Total
	Male	Female	Male	Female	
FMD	6	8			14
MCI	10	6			16
ND	11	6	5	4	26 (17+9)
Total					56

Table 5.1: Participant Information for the Conversation Analysis.

of this thesis was the transcriber and was unaware of the participants' diagnoses in an attempt to mitigate any bias.

Dr / VA / Pt / AP	Denotes who is talking; Dr = human doctor, VA = virtual agent, Pt = patient, AP = accompanying person
(.)	Identifies a silent pause of less than 20ms
⋮⋮⋮	Identifies the lengthening of a segment, each colon representing 10ms
...	Identifies a break in the transcript
(1.2)	Silent pauses in seconds

Table 5.2: Frequently Used Transcription Symbols.

In the *IVA* transcripts the final time stamp is the amount of time that passes before the participant moves on to the next question from the virtual agent. For the Hallamshire dataset, the problem presentation phase was everything from the point of the doctor first asking about symptoms to the doctor moving the conversation on to the next phase of the interview. Hallamshire transcripts do not end in a time stamp and instead end with the last thing the patient or *AP* says before the doctor moves the conversation to the examination phase of the consultation.

5.4 Analysis

This section details the findings from this analysis. First, We compare the design of the human doctor's questions to those of the virtual agent, before turning to the responses

produced by the participants. We then proceed to the case study of the participant who had completed both the digital and human doctor interviews.

5.4.1 Comparison of Problem Presentation Phases Initiated by a Human Doctor vs an Intelligent Virtual Agent

This section presents the differences between the problem presentation phases of people talking to a human doctor, versus talking to a virtual agent. We identified two distinct types of problem presentation phase, based on whether or not the patient accepts their condition.

5.4.1.1 Design of the Initiation of the Problem Presentation Phase

In the IVA dataset, the problem presentation phases are opened with the question *“Tell me what problems you’ve noticed with your memory recently”*. This question falls somewhere between a Type I (general inquiry) and a Type III (symptoms for confirmation) question according to the classification from Heritage and Robinson [2006] described above. This question is vague in the sense that it does not mention memory-specific symptoms as we would see in a true Type III question (perhaps something like *“So you’ve been forgetting people’s names?”*), but at the same time is not so vague that the patient could start talking about non-memory related symptoms, such as a sore throat.

In the Hallamshire data, we observed greater variation in the question design. Whilst the same doctor is present in all of the Hallamshire recordings used as part of this analysis, we found two different types of questions used when inviting a participant to start their problem presentation: open-ended and constrained.

1. Open-Ended Questions

Open-ended questions would fit into the category of Type I questions from Heritage and Robinson [2006] as described above. These questions do not limit the patient in terms of what they can respond with. However, the memory clinic setting of the interviews does in itself reinforce to the patients that they are supposed to be talking about their

memory problems, even though the doctor has not explicitly mentioned them in his question.

Examples from our data of these kinds of questions are:

033

1 Dr: what's been the problem

017

1 Dr: can you tell me what kind of problems you've been having

056

1 Dr: so how can I help you

2. Constrained Questions

The second type of question found in the Hallamshire data is similar in form to the question asked by the virtual agent (*"Tell me what problems you've noticed with your memory recently"*). These questions still allow the patient to talk about a wide range of symptoms, but this design makes it absolutely clear that the aim of this consultation is to discuss symptoms that are specifically related to memory concerns. For example:

083

1 Dr: can you just tell me what problems you've been having with
2 your memory and when you first noticed them

096

1 Dr: and from your perspective have you noticed any difficulties or
2 problems with your memory

One important difference in the way questions posed by the human doctor are designed is that some of them are compound questions, and some are not. For example, the question from recording 083 above is a compound question. This form of question requires the patient to respond to two different requests in their singular response. Studies such

as Elsey et al. [2015] have found that pwND often struggle to answer both parts of a compound question, and therefore this could potentially be used as an indicator that a person is experiencing dementia or dementia-like symptoms. Results from our analysis support this claim, with patients needing reminding that there is still part of the question unanswered, such as in the following example from the Hallamshire dataset:

083
1 Dr: can you just tell me what problems you've been having with
2 your memory and when you first noticed them
3 Pt: when i first noticed them
4 Dr: hm
5 Pt: be back a while now (2.7) can't remember exactly when
6 Dr: okay
7 AP: yeah i would say it's approximately, what, two year i think
8 Pt: roughly that yeah
9 (2.1)
10 Dr: and can you give me an example of how your memory has let you
11 down

In the example above, the doctor restates the first half of the question as the patient has only responded to the latter (see lines 5-8). In this instance, the patient makes no reference to the fact that they have forgotten to answer something and instead appears unaware that they had left part of the question unanswered.

We also observed examples of participants who did not respond to both parts of a compound question but *did* display awareness that they had forgotten something:

089
1 Dr: so can you tell m- i've read the letter from your general
2 practitioner .hhhh can you tell me what (0.3) memory problems
3 you've noticed and what your expectations are from (the)
4 clinic today
5 Pt: err::: i've noticed that me memory's not very good people tell
6 me things and i don't remember what they've, they've, you know
7 (.) when i'm trying to recall it (0.5) i don't remember it
8 (0.6) um:: events i'm okay with (.) it's (.) it's more or less
9 conversations that i dont (.) i dont remember (.) um:::
10 (1.3)
11 Pt: what else did you say ((laughs))

The difference between the question types from the Hallamshire dataset analysed above

and those identified by [Heritage and Robinson](#) is that the latter are taken from an analysis of general medical interviews. In both of the datasets used as part of this analysis it is already known by both the patient and the clinician that the ensuing consultation will be about memory complaints. Given this and the fact that the interviews from both datasets take place in a memory clinic, it makes sense that the questions asked by the doctor are not entirely general, but they are general within the constraint of discussing memory problems.

5.4.1.2 Patients' Responses

Patients' responses can be broadly categorised into two main groups; those who display acceptance of their condition and those who deny having memory problems. These groups are described in more detail below, with each of the groups containing a mixture of participants from both datasets. We found that rather than the interlocutor influencing how the patient responded, this seemed to be more affected by whether or not the patient accepted the reality of their condition. In turn, this was evenly spread across the diagnostic groups, with no correlation found between the degree of [CD](#) and the type of response. [Figure 5.2](#) shows the percentages of participants that produced each response type. [Table 5.3](#) breaks down the response types according to dataset, and [Table 5.4](#) breaks down the response types according to diagnostic group.

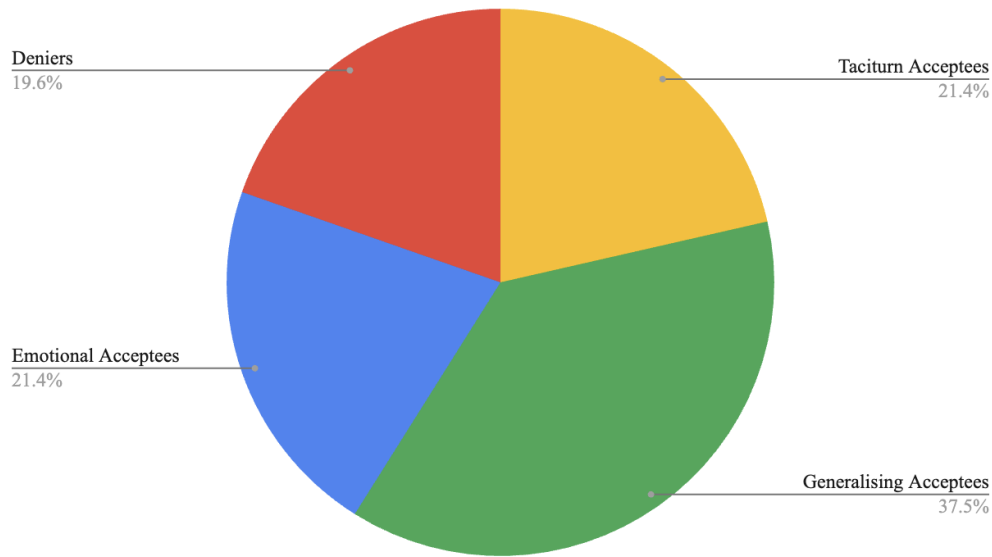


Figure 5.2: Categorisation of Patients' Responses

Category of Response	No. Participants from Hallamshire Dataset	No. Participants from IVA Dataset	Total No. Participants per Response Category
Denier	1	10	11
Emotional Acceptee	2	10	12
Generalising Acceptee	4	17	21
Taciturn Acceptees	2	10	12
Total No. Participants per Dataset	9	47	56

Table 5.3: Breakdown of Response Types According to Dataset

Deniers

This group comprises three types of response: those that fully deny the existence of memory concerns, those that attribute potential symptoms of memory problems to other things, and those that downplay the extent to which their symptoms are affecting their lives. 11 of the 56 conversations included in this analysis belong to this group, and the group consists primarily of a mixture of patients diagnosed with *MCI* and *ND*. Exam-

Category of Response	No. Participants w. FMD	No. Participants w. MCI	No. Participants w. ND
Denier	1	5	5
Emotional Acceptee	3	3	6
Generalising Acceptee	8	4	9
Taciturn Acceptees	2	4	6
Total No. Participants per Cognitive Group	14	16	26

Table 5.4: Breakdown of Response Types According to Diagnostic Group

ples 1-3 demonstrate how participants downplay any memory problems they may describe.

Example 1

2106 - ND - IVA

1 VA: tell me what problems you've noticed with your memory recently
2 (1.3)

3 Pt: er::m (.) i do not pick things up
4 (1)

5 very good with- (.) 'cause of me hearing (0.2) .hhh (.) so i
6 get things wrong (0.6) and sometime me memory (0.9) is at
7 fault (0.8) but normally (0.2) .hhh i'm pretty good (0.4)
8 round the house::: (0.2) doing things (.) .hhh (0.4) manually
9 i'm fine ↓ (0.3) .hh (0.2) >it's just if somebody asks me a
10 question or i- i want-< .hhh i know the person .hhh but i
11 can't put a- a- er::: a name to him (.) .hh (.) and then (.)
12 FIVE minutes later >it comes into me head and i-< (0.3) i'm
13 away (3.05)

The above example, from a participant diagnosed with **ND**, includes numerous instances of minimising and downplaying symptoms. The participant starts out describing that they do not pick things up very well, but is quick to add that this is because of their hearing and not anything else. In lines 6-7 the participant relents that their memory is sometimes at fault but again is quick to point out that normally they are fine, and this is

backed up by examples of things that the participant has no trouble with such as doing things around the house. This pattern is then repeated from line 9-onwards where the participant describes sometimes having difficulties recalling names but again reassures the clinician (in this case the virtual agent) that they go on to remember what it was that they had forgotten, and that everything is fine a few minutes later.

Example 2

0212 - MCI - IVA

1 VA: tell me what problems you've noticed with your memory recently
2 (1.8)

3 Pt: .hhh (0.2) er::::: (0.8) obviously i'm sixty nine year old
4 and i er (0.2) i haven't got the memory that i used to (0.9)
5 ((coughs)) (0.2) but i can er:::
6 (1.5)
7 i can more or less (0.3) carry on life as normal (0.2) (well)
8 i just have a or- really bad memory (0.8) ((coughs)) (0.7)
9 er::: maybe the dates or that kind of thing
10 (1.0)
11 so::::
12 (1.8)
13 i don't think i can give you anything else i'm sorry
14 (0.4)

This next example, from a participant with MCI, provides another example of downplaying symptoms by identifying that there might be some small issues, but that in other areas everything is fine (*"I can more or less carry on life as normal"*). This participant also seems to contradict himself. He states that he has a *"really bad memory"* but goes on to give the example of *"maybe the dates or that kind of thing"*. This problem presentation also demonstrates a clear example of a patient being polite to the virtual agent by apologising, despite knowing that it is a machine (line 13). This example supports Nass' hypothesis that humans have evolved to be polite, and that this remains intact even when talking to something that we know cannot respond to us.

Example 3

0229 - MCI - IVA

1 VA: tell me what problems you've noticed with your memory recently
2 (2.7)

3 Pt: i have struggled with

4 (1.6)
 5 er::: some memory loss
 6 (2.0)
 7 w- which (0.8) i think has seem to improved (0.9) i've also
 8 struggled with my (.) mood (0.9) which (.) my wife confirmed
 9 today that i seem to be m- (.) much more stable
 10 (6.5)

The extract from this participant follows the two previous examples by downplaying the symptoms that they mention in their problem presentation phase. However, this participant mentions the opinion of his wife, who presumably does not have any memory problems. This serves to reinforce the participant's view that he is doing better now (line 7). We also observe some emotive language (expressing the participant's feelings rather than being a neutral description) being used in this extract (line 3- "*struggled*") but the potency of this word is reduced by the confirmation that the participant is "*much more stable*" according to an outside observer. This example (and Example 1 above) both end with very long, unfilled pauses before the patient presses the key to move forward.

The three examples above were all taken from the IVA data. However, it is not just participants conversing with the virtual agent that deny (or downplay) their diagnosis or symptoms. The extract below is from the Hallamshire dataset:

Example 4

029 - ND - Hallamshire
 1 Dr: from (.) your own perspective >what would you say have been
 2 the main things< you've been (0.5) er struggling with and- and
 3 why you've come
 4 (1.3)
 5 Pt: (hand) (0.7) um (0.2) er
 6 (2.0)
 7 Dr: struggling with a hand there (0.8)
 8 Pt: hm (0.4)
 9 Dr: did you [(per chance)]
 10 Pt: [(that one)]
 11 (0.3)
 12 ...

This example is particularly interesting. It seems that despite the fact that this consultation is taking place in a memory clinic, this patient may not be completely aware

of their surroundings or may not have inferred from those surroundings that they should be talking about memory problems. Instead, the patient seems preoccupied by the issue with his hand. At this point in the consultation, the AP steps in to give the doctor more information about the hand problems, and to inform him that they are dealing with the hand and will be attending an appointment to address the issue at a later date:

Example 4 (cont.)

13 AP: we've got an appointment at erm::::: (0.5) ((tuts)) Mexborough
14 this afternoon for the [finger]
15 Dr: [oh to get that looked at]
16 AP: that's- that's you're hand [in't it]
17 Dr: [i see]
18 AP: you're wanting to [get that done] but
19 Dr: [oh yeah]
20 (0.8)
21 ...

The doctor then attempts to re-start the problem presentation by asking the patient more specifically if he has had any issues with his memory:

Example 4 (cont.)

22 Dr: (has) there been a- a concern about things like your memory
23 and- and managing day to day tasks
24 (0.7)
25 Pt: not really
26 (0.5)
27 ...

The above highlights the first response from the patient that directly challenges the assumption that he has been experiencing memory concerns. The doctor attempts to gather more information from the patient:

Example 4 (cont.)

28 Dr: mhm (0.8) and (0.2) do you think (0.4) if- for yourself kind
29 of managing day to day do you struggle with things or do you
30 manage things fairly well
31 (.)
32 as far as you're concerned [at home]
33 Pt: [mostly]
34 ...

The patient's response of "mostly" on line 33 is delayed; it does not come within the Transition Relevance Place (TRP) located at line 31. This late placement, along with his use of minimising language, may indicate that the patient will ultimately admit to memory problems. The doctor continues to attempt to find out more information from the patient, by repeating his "mostly" in a turn-initial position and probing for something "in particular":

Example 4 (cont.)

35 Dr: mostly (0.4) is there anything in particular that you struggle
 36 with (0.8)
 37 Pt: well (hhenhh) ((gestures with his bad hand))
 38 Dr: apart from the hand obviously [(hhenhh)]
 39 Pt: [(hhenhh)] yeah
 40 ...

In attempting to elicit more information from the patient, the doctor asks another question that does not explicitly signal to the patient that he is interested in memory concerns. This results in the patient referring back to the problem with his hand. The doctor notices this quickly and acknowledges the hand problem, but proceeds to clarify to the patient that they should be talking about cognitive issues:

Example 4 (cont.)

41 Dr: things like kind of memory and orientation and um:::: .hhhhh
 42 things to do around the house (0.2)
 43 Pt: yeah (.) ((clears throat)) (0.3) (i can) get round (.) (our)
 44 house yeah (0.3) [yeah]
 45 Dr: [good] okay

This final attempt by the doctor to encourage the patient to talk about their memory problems is met with another denial type of response, where the patient uses the doctor's example of "doing things around the house" to exemplify that he is fine.

As discussed earlier in Section 5.2.4, research has found that the presence of an AP in clinical interviews can be a reliable indicator of dementia. Larner [2005] reported that if attending a memory clinic with an AP were a diagnostic test it would have a sensitivity of 100%. Elsey et al. [2015] found that the AP often acts as a spokesperson for the patient

and aids them in describing their symptoms. In this example (and the next), an AP is present. However, they behave rather differently. In Example 4, the AP does not talk about memory symptoms at all and instead only speaks to clarify the problems with the participant's hand. Whilst the AP in this instance is not denying that her husband has been experiencing memory concerns, she also does not offer up any additional information that could help the doctor discern what the patient has been experiencing. In Example 5, we see both the participant and their AP denying memory issues. This example comes from the IVA data:

Example 5

```

0276 - ND - IVA
1  VA: tell me what problems you've noticed with your memory recently
2      (2.9)
3  Pt: when
4      (1.9)
5  AP: °you-° i don't think you've had any problems with your memory
6  Pt: yeah
7      (0.7)
8  AP: ((after pressing button)) as such

```

Neither the patient nor the AP presents any memory issues in response to the virtual agent's question. However, the increment "*as such*" by the accompanying person (line 8) suggests that there may be complaints related to cognitive issues in some other way. If this were to happen in a setting with a real human doctor it is likely that the doctor would then prompt for further information (as we saw in Example 3 from Participant 029 above, where the patient's response on line 33 prompts the doctor to continue to dig for information). However, because this patient is presenting to the virtual agent, there is no further questioning and therefore there is the possibility that some symptoms have gone unreported in this instance.

Out of the 14 responses from patients with FMD, only one falls into the *denier* category. This is unsurprising; given that one of the hallmarks of FMD is *subjective* memory problems that often do not show up on clinical tests, it would unlikely for a patient to be referred to a memory clinic if they did not perceive themselves as having any symptoms. The participant whose response does fit this group demonstrates in their problem presen-

tation phase that they no longer believe they have memory problems at all, attributing the difficulties they experienced instead to a lack of confidence:

Example 6

0217 - FMD - IVA

- 1 VA: tell me what problems you've noticed with your memory recently
 2 (2.2)
 3 Pt: my problems weren't actually to do with me::mory (0.3) it was
 4 my confidence and (.) i was (0.3) making assumptions when i
 5 was right that in fact that i had made mistakes (0.3) i had
 6 forgotten to do something (0.4) er::: so i::: was
 7 misidentifying it as memory
 8 (1.5)

Acceptees

Unlike the responses in the examples above, which are characterised by denial or varying degrees of minimisation, the participants whose responses are included in the *acceptee* group consistently demonstrate some acknowledgment of having a memory problem. This analysis revealed three distinct groups of acceptees that span the different datasets and levels of cognition, these are detailed below.

1. Emotional Acceptees

Participants belonging to this group describe the symptoms they have been experiencing and how these symptoms are having a negative effect on their day to day functioning or emotional wellbeing.

Example 7

2105 - MCI - IVA

- 1 VA: tell me what problems you've noticed with your memory recently
 2 (1.6)
 3 Pt: i can't remember (.) my mother and father properly (0.3) which
 4 is very upsetting (.) .hhhhh (.) i'm having to use
 5 photographs:::: (.) and a:::nd things like that to remind me
 6 and (0.2) what me children tell me (0.3) that they did with
 7 their grandma and grandad because i can't remember .hhhhhh
 8 (0.4) a:::nd (0.2) also i'm coming out with (0.7) weird words
 9 (0.3) that i wouldn't (.) normally (0.2) use
 10 (1.6)

Participant 2105 describes their specific memory related symptoms and how terrible this is making them feel (line 4- “*very upsetting*”). There is no language that downplays what the participant has been experiencing and there is no attempt by the participant to attribute these symptoms to something less serious (cf. Example 1 from participant 2106 and their claim of hearing issues).

Example 8

0219 - MCI - IVA
 1 VA: tell me what problems you’ve noticed with your memory recently
 2 (1.3)
 3 Pt: um::::: (.) .hhhh i’m unable to remember (0.3) um::::: (.)
 4 dates times (0.7) um::: things that have happened hh. a::::nd
 5 um::: it’s very distressing
 6 (1.8)

Some participants belonging to this group are less descriptive than others in their problem presentations. However, we still observe emotive language in the short problem presentation phases from these participants, such as line 6- “*very distressing*” above from Participant 0219.

Example 9

043 - ND - Hallamshire
 1 Dr: what about recently what sort of things do you found that
 2 you’re struggling with
 3 (3.3)
 4 ...
 36 Pt: (i’ve had) to give up um
 37 (1.6)
 38 rainbows and guides (0.8)
 39 Dr: oh i see (so) you used to volunteer with that was it
 40 Pt: i used to be a leader
 41 Dr: used to be a leader okay (0.5) and again do you find it was
 42 because of [memory difficulties is it you stopped]
 43 Pt: [i couldn’t cope with it yeah] i couldn’t remember
 44 their names i couldn’t remember what (i) was supposed to be
 45 doing
 46 Dr: i see [okay]
 47 Pt: [just got too difficult]

The extract above from Participant 043 is an example of an emotional acceptee talking

with the human clinician. After the clinician's initial invitation to start the problem presentation phase there is some back and forth with the AP in which they describe some of the symptoms the participant has been experiencing, such as asking the same question multiple times or forgetting both old and new memories (lines 4-35). There is no emotional language used by either the patient or the AP during this back and forth. It is not until the participant starts talking about a scenario that's specific to their own experience (giving up volunteering, from line 36) that they use more emotive language (*"I've had to give up"* rather than *"I no longer"* or *"I've stopped"*).

2. Generalising Acceptees

This was the largest group of responses we found, with 37.5% of responses belonging to this category from across the different interview modalities and cognitive groups. This group consists of people who are accepting of the fact that they have memory problems, but who tend to speak in very general terms. Some participants give examples, but these examples are not specific to them. Rather than giving details, they instead use vague language (such as *"forgetting things"* or *"that sort of thing"*). The use of such circumlocution in this group indicates that participants are aware of their symptoms (because they are able to name them) but simultaneously they claim that the symptoms are not necessarily severe enough to warrant investigation.

Example 10

```

0218 - MCI - IVA
1  VA: tell me what problems have you had with your memory
2      (1.7)
3  Pt: err:::: forgetting dates (0.5) err:::: forgetting times of
4      things (0.4) erm:::::
5      (1.7)
6      just (0.3) o- odd (.) common or garden things
7      (2.5)

```

In the example above, Participant 0218 gives some examples of the difficulties they have been experiencing, but these are general examples: *"forgetting dates...forgetting times"*. Furthermore, these kinds of memory lapses are things that even cognitively healthy ageing adults may experience from time to time [Kirk, 2023]. The problem presentation phase

here is ended with a frequently used common phrase (line 6- “*just odd common or garden things*”). By using this phrase, the participant claims that they are not particularly concerned with the memory issues they have just described, and implies that his concerns are not particularly special or require attention.

Example 11

0238 - ND - IVA

1 VA: tell me what problems have you had with your memory recently
 2 (2.0)

3 Pt: .hhhhhh
 4 (1.1)
 5 er:::::
 6 (2.2)
 7 things like
 8 (1.0)
 9 walking into a (.) room and (.) forgetting (.) what (0.8) i've
 10 gone for (0.2) er:::::
 11 (1.5)
 12 that sort of thing
 13 (2.4)
 14 walking upstairs (0.2) and
 15 (1.2)
 16 forgetting
 17 (1.6)
 18 why i've (0.2) gone
 19 (1.7)
 20 that sort of thing
 21 (1.2)

We see a similar thing happening in the example above from Participant 0238. The virtual agent is given some examples of general issues that the participant has experienced, and then the participant ends the problem presentation phase with a general statement (line 20- “*that sort of thing*”). There is a notable absence of any language that would suggest the participant is particularly distressed or worried about their condition. Participant 0238 also exhibits multiple long, unfilled pauses throughout their problem presentation. Often these appear at TRPs (lines 6, 12, and 19). The fact that these unfilled pauses are so long could be an artefact from the human-human conversations the participant is used to having. For example, these TRPs indicate places where a human doctor could ask for

elaboration from the patient.

In another example, this time from a patient with [FMD](#), we observe the recounting of a general memory complaint (difficulty finding words). Notably, the participant describes this issue in a very general manner, without offering any insight into how the experience made them feel at the time, nor providing specific details about where they were or what they were doing when it occurred:

Example 12

```

0259 - FMD - IVA
1  VA: tell me what problems have you had with your memory recently
2      (1.8)
3  Pt: um:: (0.4) i've had problems finding words
4      (0.4) er::: remembering things (.) doing things
5      and getting distracted (0.8) and then f:::::orgetting
6      (1.7)
7  Pt: yeah (.) just forgetting (.) and getting muddled (1.2)

```

3. Taciturn Acceptees

Patient responses that belong to this group provide very little information about a patient's symptoms. Symptoms may be mentioned briefly but participants do not expand on them, and instead keep their answers very brief.

Example 13

```

0215 - MCI - IVA
1  VA: tell me what problems have you had with your memory recently
2      (1.0)
3  Pt: um:::::::::: (0.4) forgetting things
4      (1.8)

```

The participant in Example 13 above does not deny the presence of symptoms, but makes no attempt at any kind of description of those symptoms. There are also no examples of the kinds of things they might be forgetting. The participant exhibits quite a long hesitation followed by a pause, before verbalising their short answer (line 3). Other participants from the [IVA](#) dataset responded in a very similar manner, but without the use of filled pauses:

Example 14

0235 - MCI - IVA

- 1 VA: tell me what problems have you had with your memory recently
 2 (0.6)
 3 Pt: short term memory (.) hopeless
 4 (1.2)

Again, this participant discloses very little information and decides to move the conversation forward instead of offering any more detail. This response only tells us that the participant has noticed issues with their short term memory, but we have no indication of what kind of issues these might be. Note also that this participant does not use a complete sentence. As this participant is diagnosed with **MCI**, his linguistic abilities should be preserved enough that complete sentences are how he usually communicates. This incomplete sentence could therefore be the result of the fact the participant is conversing with the virtual agent, as this short and choppy sentence is similar to the types of commands that people direct to in-home smart devices (which the participant may be familiar with).

The taciturn responses from patients with **FMD** are no different, containing very little information about what they have experienced and no sense of how this might be affecting them emotionally:

Example 15

0209 - FMC - IVA

- 1 VA: tell me what problems have you had with your memory recently
 2 (1.2)
 3 Pt: err::: (1.4) sometimes i forget (.) um::: (1.4) things
 4 (0.2)

Whilst the examples above come from the **IVA** data, we also have examples of taciturn acceptees in the human-human Hallamshire data interviews:

Example 16

033 - ND - Hallamshire

- 1 Dr: could you describe what's- what you understand about why
 2 you're here (0.4)
 3 Pt: u:::m (0.5) not entirely
 4 (2.2)
 5 Dr: what's been the problem (0.4)

6 Pt: um:::::::::: (0.5) memory shortage (0.2)
7 ...

Within the first four turns of the extract from this participant, we can see that whilst the patient knows that there is a problem (line 6) he has already signposted to the doctor that they might have difficulties talking in detail about this, as he is not entirely sure why he is talking to the doctor in the first place (line 3). When the doctor confirms that the patient has experienced memory shortage, he prompts the patient for more information:

Example 16 (cont.)

8 Dr: okay
9 (3.8)
10 so you feel you've had memory shortage
11 Pt: yes
12 (4.6)
13 Dr: and (.) could you::: give me an example of last time your
14 memory (0.3) let you down↑
15 (1.7)
16 Pt: u::::::::::m
17 (2.4)
18 ...

The patient again confirms that he has been having memory shortage. However, he has difficulty forming an example of his memory letting him down, as evidenced by the long pause in line 17. At this point, the AP steps in to try to answer the doctors questions:

Example 16 (cont.)

19 AP: you can't- you've lost your sense of direction (0.4) does that
20 count
21 Pt: right
22 AP: (hhhhh)
23 Pt: ((chuckles)) (0.3)
24 AP: .hhhhh never needed a navigator before (hahaha (0.8) (hhh)
25 (2.4)
26 u:::::m (0.6) can't remember how to:: (0.3) maintain the
27 computers
28 (1.4)
29 Pt: a:::::greed
30 (1.4)
31 AP: u:::::m you can't remember how to do your job any more
32 (8.6)

33 ...

We can see from the above in lines 21 and 29 that the participant is accepting of the points being raised by their AP. However, the participant is not offering any additional information. There is an unusually long pause (line 32) before the doctor attempts to get the participant to give their own account of their memory problems, rather than directing additional questions to the AP:

Example 16 (cont.)

34 Dr: .hhhhh and could you- a- a- again i appreciate (0.3) y- (.)
 35 you might have difficulties er::: doing this but could you::
 36 give me some examples of (0.7) of problems you've had with
 37 your memory(0.3)
 38 Pt: u::::m hhh.
 39 (5.1)
 40 i
 41 (1.8)
 42 ((patient looks at AP))
 43 AP: no (dad) this is down to you well is- what do y- your
 44 perception not mine
 45 ...

Once the doctor has asked the patient to give some examples of the memory problems he's been having, he hesitates and then pauses for a long time (line 39). This is an indicator that the patient is having difficulty formulating his response. He looks to his AP for help, but she reinforces to him that the doctor wants the patient to give his own examples. After a pause, the patient continues:

Example 16 (cont.)

46 Pt: (1.4)
 47 i::: (.) can't remember how to do my job
 48 (2.1)
 49 Dr: you can't remember how to do your job
 50 Pt: yeah
 51 (6.2)
 52 ...

This is the first time in the consultation that the patient has been able to give a specific example of the kind of memory problems he has been experiencing. However, the infor-

mation he offers up is not new and is in fact a repetition of what his AP said back in line 31. The doctor confirms what the patient has just said by way of repetition (line 49), and then we see another long pause in the conversation. The doctor has left enough time for the patient to expand a bit more, but after he fails to do so the doctor reengages the conversation by inviting the patient to resume talking by way of describing what their job involves:

Example 16 (cont.)

```
53   Dr:  and (0.6) what does that involve
54         (1.3)
55   Pt:  er::: (.) joinery
56         (7.7)
57   ...
```

Once again, there is a large pause in the conversation at this point. The doctor prompts the patient two more times for a description of specific problems related to his memory, but the patient does not respond (data not shown), and the doctor then ends the problem presentation phase.

5.4.1.3 Summary of Patients' Responses

The analysis presented above has demonstrated the different kinds of problem presentation phases observed in our data. There are two main kinds of problem presentation; one in which participants deny their symptoms, and one where participants accept them. In terms of the acceptees, we found three main approaches in how people accept their conditions. Some participants talk about the emotional toll their symptoms have been taking whilst describing specific instances where their memory has let them down. Some participants talk only in very general language and do not offer any specific examples. The remaining participants do not talk much at all. This is more obvious in the recordings from the IVA data, where participants can say something as short as “forgetting things” for their problem presentation phase and then manually move the conversation forward. Taciturn acceptees from the Hallamshire data are questioned more by the human doctor, who frequently presses these participants to expand on what little information they have already given. However, we found that this was usually unsuccessful and would result

in either the AP stepping in and acting as a spokesperson for the patient or the patient discussing memory issues that have already been mentioned earlier on in the consultation.

Both the accept and deny groups contain a mixture of participants from both the human-human and human-avatar datasets, as well as a mixture of people diagnosed with all three different levels of cognitive decline investigated in this chapter. Whilst this analysis was not able to uncover diagnostically valuable conversational profiles, our analysis does suggest that as far as patients are concerned there may not be much of a difference between presenting their problems to an avatar instead of a human being. The next section highlights the relationship between how the doctor's questions are phrased, and how the patients respond.

5.4.1.4 The Relationship Between the Phrasing of the Doctor's Questions and the Patients' Responses

The results above demonstrate that even when participants are asked the exact same question from the virtual agent, the way they formulate their responses will vary from person to person. Recordings from the IVA dataset show examples of all the different kinds of responses identified by our analysis, so controlling for how the beginning of the problem presentation phase is worded by the doctor seems to have little effect on what kind of response the participant will offer.

This assertion is exemplified by two particular recordings from the Hallamshire dataset. In the examples from Participants 029 and 043, the doctor uses the word "*struggle*" when asking about the problems the participants have been experiencing. This could be seen to project some emotion onto the response; before the patient has even started describing what is wrong with them they are being told that they are struggling with something. Indeed, Participant 043 does respond with an emotional description of her symptoms, detailing the negative impacts of her memory problems. Participant 029 however never uses emotional language in his problem presentation phase, even though the doctor uses the word "*struggle*" four times throughout his questioning.

5.4.1.5 Discussion of Results from the Comparative Study

It is perhaps not entirely surprising that some of the participants in the IVA dataset give such short responses to the virtual agent. There is no prompting or fishing for more information from the avatar, so responses can be short and that won't be questioned by the digital system. However, we also saw examples of taciturn responses from the Hallamshire data. Despite the virtual agent's inability to ask for more information, several participants gave a detailed account of their memory complaints. The questions asked by the human doctor were consistently general in nature but with a specific focus on memory complaints as the interviews are taking place in a memory clinic. This was the same for the question played by the avatar in the IVA data.

In terms of participants' responses, two main groups were identified (deniers or acceptees). Both of these groups contain participants from both datasets, and participants at the three different levels of CD.

The first group of responses deny the existence of any memory concerns, or downplay their symptoms by noting that they have seen some improvement in their condition recently. Participants in this group may also attribute their symptoms to some other cause, such as hearing loss. We found examples of this type of problem presentation in both the human-human and human-avatar interviews. The concept of denial from a patient in a clinical setting is not new, nor is it constrained to only diagnoses of neurodegenerative disorders. In fact, some amount of denial is considered healthy to some extent as it can help to ward off excessive worry, depression, or fear (see [Ness and Ende \[1994\]](#) for more discussion on this topic). However, some of the patients in this group could be dealing with a level of anosognosia (being unaware of their condition due to their cognitive deficits). Although anosognosia may be caused by a broad range of different conditions, it is particularly pervasive in dementia. According to a study from [Wilson et al. \[2016\]](#), anosognosia may eventually be a symptom in almost all patients with dementia. Research has also found that anosognosia can be experienced by people with MCI, although this is less common [[Morris and Mograbi, 2013](#)].

The reason for the patients' denial and lack of detail regarding symptoms may not be of

much consequence to a digital system where the content of the speech is not as important to the classification process as the acoustic or linguistic features that the system measures. Although these systems could theoretically be used to make the diagnosis alone, given larger training sets and computational improvements, realistically the output of a system such as CognoSpeak would never be the sole determinant of a diagnosis and rather would be used in conjunction with a doctor or clinician's assessment. There are two main reasons for this; the issue of explainability and trust, and the issue of responsibility in the case of something going wrong. [Heinrichs and Eickhoff \[2020\]](#) discuss these issues in detail. They also suggest that Artificial Intelligence (AI) systems that can provide diagnostic analytics about a patient's condition would be best used if they could demonstrate some kind of report that clinicians could use to interpret the model's decision. This is where it is particularly important to have more detail about symptoms, and where patients belonging to the denier group would need to be pressed more for information. This suggests that an Automatic Cognitive Decline Classification (ACDC) system could be particularly useful in cases where patients reveal little information about their symptoms. However, our research has also demonstrated that it may still be difficult to elicit more information from deniers or taciturn patients even when they are being interviewed by a human doctor and repeatedly being asked for more information, as in the example from Hallamshire Participant 029.

The second group of participants generally accept that they are experiencing memory loss, but we found distinct differences in how these participants addressed their concerns. Firstly are those participants who talk not only about the physical symptoms they have been experiencing, but also how this is affecting them emotionally. They describe, often in detail and using emotive language, specific incidents where their memory has let them down which is then reinforced by descriptions of the negative effect on their daily lives or emotional wellbeing. Our second group of acceptees admit to having memory issues but only talk about them in a very general manner, often failing to give specific examples and instead relying on broad statements such as forgetting people's names or what the current date is. Even when pressed by the human doctor for more information the participants belonging to this group struggle to respond accordingly.

The final group of participants who accept the fact that they have been experiencing memory issues are those who seem to have the most difficulty with their problem presentation phases, and this is often demonstrated by short, non-specific answers and more frequent hesitations and pauses. When these participants are presenting to the digital avatar and are in charge of moving the conversation along themselves they frequently respond with very short, vague answers. This was also observed in the human-human group, where we found examples of the human doctor trying his best to get a taciturn patient to talk more about their symptoms, but to little effect. One particularly interesting finding from these results is that taciturn patients do not necessarily divulge more information when conversing with a human compared to a digital avatar. Our examples of taciturn patients in the Hallamshire dataset show that even when being repeatedly asked for more information, they provide little or none. This supports the idea that while the virtual agent does not ask follow up questions, it does not necessarily need to and there is no guarantee that this would result in patients divulging more information.

5.4.2 Case Study: Participant 096/0221's Problem Presentations

During the process of analysing the Hallamshire and the IVA datasets it was discovered that one of the participants had completed both studies. This participant had attended the Hallamshire human-human interview in June 2014, and then went on to complete the IVA human-avatar interview in September 2016. Their diagnosis of ND remained the same for both tests, although at the time of the Hallamshire interview they had an MMSE score of 24 which had decreased to a 23 by the time of the IVA interview. However, both of these scores are firmly within the mild ND category and the difference does not indicate a substantial decline in the participant's cognition during the time that passed between the two recordings.

This analysis compares the patient's problem presentations from each dataset. For clarity, examples of the transcriptions are presented side-by-side in a table, with the transcripts on the left coming from the human-human Hallamshire data and those on the right coming

from the human-avatar IVA data. Despite the large differences in how the doctor and the virtual agent’s initial questions are phrased, we found numerous similarities in the content of the responses from the participant. We also observed a more frequent use of hesitations and pauses in the patient’s recording from the IVA dataset.

5.4.2.1 The Doctor’s Opening Question

Table 5.5: Extract 1 - Doctor’s Questions

Dr: have you noticed any problems with your memory and if so what and what are your expectations from this clinic ...	VA: what problems have you had with your memory ...
--	--

The virtual agent plays a straightforward question to initiate the patient’s problem presentation phase. As seen above, this question does not ask for specifics from the patient, but does inform him that his response should focus on memory problems. The question from the human doctor is another story entirely and is in fact three different questions compounded into one; *“Have you noticed any problems with your memory?”*, *“If you have, what have you noticed?”*, and *“What are your expectations from this clinic?”*. As discussed earlier on (Section 5.2.1.2), research has shown that people with dementia may have trouble answering all parts of a compound question. Having three parts to this compound question makes it particularly difficult for the participant to follow, as evidenced below.

5.4.2.2 The Patient’s Responses

After the doctor has initiated the problem presentation phase with the questions above, the patient begins his response. With the human doctor, the patient starts by addressing the first part of the compound question that was asked to him (*“Have you noticed any problems with your memory?”*). This is a very similar question to the one posed by the virtual agent in the sense that both questions are asking about memory prob-

lems in particular. This results in similar opening statements from the patient in both recordings:

Table 5.6: Extract 1 - Patient's Responses

Pt: well basically (0.4) i've noticed a deterioration (0.5) gradually (0.3) with me memory (0.5) ...	Pt: memory (0.3) i:::: have to keep (0.9) repeating myself a lot (0.6) i::::: get confused (0.7) ...
---	---

At first glance, these responses seem to be functioning in the same way; confirming that the patient has experienced memory problems. One notable difference between the two is the structure. The response from the Hallamshire data mirrors the wording of the doctor's question (*"Have you noticed-"*, - *"I've noticed"*). It then quantifies the progression of the decline, letting the doctor know that whilst there have been changes with the patient's memory these have happened slowly.

The response from the IVA data is less syntactically complex, and does not make use of any conjunctions. However, the participant is quick to give the virtual agent examples of how his memory has been letting him down. These examples are interspersed with hesitations that are longer than 0.5 seconds in duration, something we see no example of in the recording from the Hallamshire dataset.

In the Hallamshire recording, the patient's problem presentation phase then goes on to address the final part of the doctor's opening question (*"What are your expectations from this clinic?"*):

Extract 2 - Patient's Responses

1 Pt: i'm hoping that (2.0) there'll be something (0.4) fulfilling
2 that's gonna come from (0.9) yourself and other colleagues
3 (1.1)
4 that will result in
5 (1.1)
6 a magic pill (as you might say) that (0.8) after all the tests
7 that will (.) eradicate that (0.5)
8 ...

This segment of the problem presentation phase is not a direct reflection of what the

patient is *expecting*. Rather, the patient focuses on what he *hopes* might be the outcome. This slight change suggests that the patient does not really believe that the doctors will indeed find a magic pill to stop his memory deteriorating further (he does not expect that this will happen), but this wording makes it clear that this would be an example of a best case scenario for the patient.

We do not see any mention of a “*magic pill*” in the recording from the IVA data, as the virtual agent did not ask what the patient was expecting to happen. However, from this point onwards, both problem presentations seem to converge. In both examples, the patient goes on to talk about things that he has already tried in an attempt to address his memory issues:

Table 5.7: Extract 3 - Patient’s Responses

<p>Pt: err::: (0.5) i know i've had that many tests and scans (1.0) but nothing's been err::: (1.2) really (0.5) err::::: (0.2) fulfilled as far as saying that (0.4) ...</p>	<p>Pt: i've (0.5) took medication .hhhhh (0.4) which er didn't work (0.6) ...</p>
---	--

The extract from the Hallamshire data is more descriptive than its IVA counterpart. However, both of these segments are doing the same thing; they are indicating to the doctor that the patient has already sought help for his condition but has been unsuccessful. In the Hallamshire data, this is achieved by telling the doctor about the tests and scans that the patient has undergone which did not result in any clarity as to what is causing the memory issues. In the IVA data, the patient talks about the medication that he has been taking, which does not seem to have helped him.

In both recordings the participant now steers the conversation towards another concern he has been having involving mobility issues. It is at this point in the problem presentation that the patient really starts to talk in an emotive way, making it clear to the doctor that his mobility issues are having a very detrimental effect not only to his physical health, but also to his mental wellbeing:

Table 5.8: Extract 4 - Patient's Responses

Pt: er::::: (0.6) and obviously (.) with this falls i have a lot of falls	Pt: but also (0.6) i::::: cannot walk i have a:::: mobility scooter (0.5) which is very (0.7) helpful but frustrating ...
Dr: mhm (0.6)	
Pt: and i am (1.1) most concerned about that because (1.0) ...	

In the Hallamshire data, this is the first time that the patient gives a real sense of his worries. Up until this point he has talked about “*gradual deterioration*” with his memory and his hopes for a “*magic pill*”, and now he is really getting to the crux of his issue, stating that he is most concerned about the multiple falls he has been experiencing. This is also the only point during the Hallamshire problem presentation that we have any kind of response from the doctor. His “*mhm*” in this example is signalling to the participant that the doctor has understood, and is giving him permission to continue steering the conversation towards this slightly different topic.

It seems that in the two years since the Hallamshire recording the patient's mobility issues have worsened. He is now unable to walk and has to use a mobility scooter. The patient goes on to describe why this is so distressing:

Table 5.9: Extract 5 - Patient's Responses

Pt: i've always been (1.4) you don't know >this is probably not relevant to the situation but i've always been< (0.9) keen eh sportsman ...	Pt: because i used to walk (0.7) er::::: (0.2) very (0.3) very long way my age is:::: er (1.0) seventy six (0.6) i used to be extremely (0.7) fit ...
---	---

The two extracts above highlight a difference in how the patient is framing his answer depending on which doctor he is talking to. In the IVA data, he is continuing the topic

of his mobility issues and is using examples of how he used to be very fit to demonstrate how these issues are particularly difficult for him to deal with. In the Hallamshire dataset the patient is doing the same thing, he has already brought up the subject of his mobility issues and now he is justifying why these symptoms in particular are the most upsetting to him. However, the patient also includes the statement “*this is probably not relevant to the situation but*”. The phrasing of this sentence holds a lot of information for the doctor. It signals to him that what the patient will talk about next is going to veer away from the memory issues which have been the focus of the conversation so far. It also tells the doctor that if he does not think the coming information is relevant it is okay to interrupt or ignore because it is probably not relevant anyway. We do not see this kind of hedging when the patient is talking to the virtual agent, despite the fact that he is making the same shift in topic from memory to mobility issues. However, we do find more pauses and hesitations.

The final parts of both problem presentation phases include justifications as to why the mobility issues are so distressing for this patient in particular:

Table 5.10: Extract 6 - Patient’s Responses

<p>Pt: and then (0.6) when something like this happens it makes it- well it’s bad enough for everybody (0.8) er:::: but it makes it worse (.) when you’ve been (.) used to being fit and then you’re not (1.1) and er:: when y- you can’t get to the bottom of what’s causing it (0.8) er:::: (.) it is frustrating</p>	<p>Pt: i::::: (0.3) held the yorkshire record for the two hundred metre- (0.4) two hundred metres .hhhhh (.) i walked the pennine wa:::y (.) when i was fifty .hhhh (0.3) so all this that’s happened (0.4) is really a big body blow to me (1.5)</p>
---	---

In both of the extracts provided above we see more examples of emotive language being used by the participant to describe his feelings about struggling with something he used to be so good at. In the Hallamshire recording the participant is paying particular attention

to the fact that he knows mobility issues are unpleasant for anyone who is experiencing them, but is using his past as an athletic person to further justify why the experience is worse for him. In the IVA recording the participant is continuing the list of his sporting achievements. This again is pointing out the fact that he was not an ordinary person, he held records and completed challenging walks even in his fifties, so it is no wonder he is having a particularly tough time dealing with the fact that he now has to use a mobility scooter.

5.4.2.3 Discussion of Results from the Case Study

Considering the time that passed between the recording of the Hallamshire and the IVA interviews from this participant, both problem presentation phases share numerous similarities. In the Hallamshire interview the doctor opens the problem presentation phase with a compound question which the participant does a good job of following and responding to, although he does not give specific examples of memory issues to answer the second part of the compound question (what those memory problems might be). In the recording from the IVA dataset the virtual agent's question is much easier to follow and does indeed result in examples of memory problems from the participant. In both recordings the participant shifts the main focus of the consultation from memory-specific problems to mobility problems. In the human-human recording the participant exhibits some hedging behaviour, which is not observed in the human-avatar recording. In both interviews the patient is descriptive and uses emotive language when talking about his mobility issues. He also gives descriptions of these mobility issues in both recordings, although more hesitations are observed in the IVA data.

The results of this analysis suggest that patients might not need much input from a doctor during their problem presentations to talk about their concerns, even if they think some of the issues raised might not be directly relevant to the original topic of the consultation. In the recording from the IVA dataset there is no interjection from the virtual agent to signal to the patient that he is being listened to and is justified talking about his mobility issues. However, the patient continues all the same.

One difference that was observed between the two interviews was the hesitation markers present in the participant's problem presentation phases. In the human-avatar data the participant exhibits more hesitation behaviours than in the human-human data, although they are present in both recordings.

5.5 Conclusion

This chapter has compared the problem presentation phases in different clinical interview settings using the methodology of CA to examine variations in how patients present their concerns when engaging with a digital avatar compared to a human clinician. Distinct differences were identified in the ways patients presented their concerns, though these varied across the three levels of CD investigated and between the two interviewer modalities. The broader results of our study suggest that adopting a more open-ended approach to medical interviews (such as those facilitated by the digital avatar) may not necessarily elicit more detailed problem presentations from patients.

The case study of participant 096/0221 revealed numerous similarities in his problem presentations across both settings. However, notable differences were observed in the frequency and duration of hesitations in his speech when interacting with the virtual agent. Early research had posited that speech disfluencies are less frequent when humans communicate with machines [Oviatt, 1995], but our analysis contradicts this assertion and demonstrates that this is not always the case.

Chapter 6

Conclusions and Further Work

Contents

6.1	Conclusions	186
6.1.1	Assessment of Contributions	188
6.2	Limitations and Further Work	192
6.2.1	Disfluency Analysis	193
6.2.2	Automatic Cognitive Decline Classification	194
6.2.3	Conversation Analysis	195
6.3	Concluding Remarks	196

6.1 Conclusions

Cognitive Decline (CD) can happen as a result of regular ageing but may also have pathological causes. *Dementia* is a frequently used term in research and conversations surrounding pathological CD that refers to a collection of syndromes and diseases characterised by a loss of cognitive abilities severe enough to impact day-to-day life. The severity of dementia symptoms can range from mild (which includes things such as forgetting names, word-finding difficulties) to severe (forgetting important personal events in one's past, experiencing delusions and hallucinations, or exhibiting personality and behavioural changes). The most common type of dementia, Alzheimer's Dementia (AD), accounts for around 60% of all cases of dementia in the United Kingdom. AD has devastating economic and emotional effects on those diagnosed, in addition to their caregivers and loved ones. It is estimated that AD and AD-related care costs £34.7 billion annually in the U.K alone.

There are no cures for AD and other dementias, and symptoms will get progressively worse over time. Due to how similar symptoms of different types of dementia can be, receiving a diagnosis of dementia can take a long time. Those aged 65 and over typically wait 2.8 years, and younger people face an average waiting time of 4.4 years before receiving a diagnosis. The most common tests for dementia are expensive and require specialised equipment and trained doctors or neurologists. These are often used in combination with cognitive tests designed to assess different areas of cognitive functioning, including speech and language abilities. Difficulties with speech production are often reported as one of the earliest signs of dementia and CD. This thesis investigated speech from people who have been diagnosed with different levels of CD, and presented results of Conversation Analysis (CA), analysis of the frequency and duration of speech disfluencies, and a proof-of-concept study that suggests using disfluency information can enhance the accuracy of Automatic Cognitive Decline Classification (ACDC) systems.

Numerous different approaches to analysing speech have been used to investigate what else we can learn about different levels of CD. A popular qualitative method for analysing speech, CA, has been used to investigate how people navigate medical consultations. Some

studies have even found distinct conversational profiles that can help provide diagnostic information to doctors and clinicians working with patients with CD. A qualitative approach to speech analysis provides rich, in-depth information about the complexities of speech and how this varies across different situations.

Research has shown that the earlier CD can be diagnosed, the better the outcomes for the patient. Therefore, there is an urgent need for cheaper, quicker, and more accurate tests for CD. With recent advancements in Machine Learning (ML) and Artificial Intelligence (AI), a growing body of research has investigated the use of speech as a non-invasive biomarker of CD. Studies using ACDC systems have indicated that ML and AI methods can be used to accurately identify speech differences between healthy controls and people diagnosed with CD. These systems could be used to help accelerate the process of diagnosing cognitive impairment, as these systems only require recordings of speech which is quick and easy to collect, and is not as expensive as traditional tests.

Previous studies have suggested that differences in pause length and duration can help to discriminate between different levels of CD. Despite this, little work exists investigating the usefulness of other disfluencies for this task. Due to the fact that there is such a well-documented link between CD and difficulties with speech production, speech disfluencies could provide valuable information that could help improve the accuracy of ACDC systems. However, quantitative analysis of speech can only give us so much information. Combining the knowledge gained through disfluency analysis with that revealed by the qualitative CA paints us a broader picture of the intricate relationship between cognitive decline and speech.

The findings related to the five research questions addressed in this thesis are summarised below.

6.1.1 Assessment of Contributions

6.1.1.1 Research Question One

How do the frequency and duration of speech disfluencies differ when participants engage in an interview-style task with a digital avatar in a simulated medical interview scenario, compared to similar interviews conducted with human clinicians?

To complete the first manual disfluency analysis we first devised a new schema, the Disfluencies in Cognition (DisCo) schema, as a method of accurately identifying and measuring disfluencies in speech. The schema was developed in order to balance the pressures of capturing a high enough level of detail whilst still being sufficiently simple to implement in the automatic system. Given that previous research into disfluencies has suggested an average rate of six disfluencies per 100 words in spontaneous, every-day speech from healthy individuals, our research suggests that talking to a digital avatar rather than a human resulted in higher disfluencies rates (13.7 disfluencies per 100 words for our healthy controls). This is particularly interesting given that previous research such as Oviatt [1995] found lower disfluency rates when humans were interacting with machines.

Bortfeld et al. [2001] suggested three possible explanations for why previous research has found lower disfluency rates in human-computer speech. Firstly, they suggest that people may be more careful when speaking with machines. Whilst our results cannot comment on how careful participants may or may not have been when communicating with the Intelligent Virtual Agent (IVA), our research does suggest that the unfamiliarity of the conversational situation (conversing with a virtual agent which is taking the role of a clinician in a memory clinic) is having an effect on the presence of disfluencies in speech. This is demonstrated by the high disfluency rates exhibited by the healthy participants in their responses to both IVA tasks.

Secondly, they suggest that disfluencies are related to coordination processes that are different with machine partners than with human partners. The datasets used for the disfluency analyses in this thesis did not provide any video data, so an analysis of the

link between gestures and disfluency was not possible in this case. Results from our conversation analysis in Chapter 5 suggest that processes of communication are actually rather similar between human-human and human-machine interaction. This is exemplified by the similarities between the problem presentation phases from Participant 096/0221, despite the changes in their conversational partner.

Thirdly, both Bortfeld et al. and Oviatt [1995] suggest that the structured nature of human-computer interactions require less planning from human participants, resulting in fewer disfluencies. Whilst this may certainly be true of spoken human-computer interactions that follow a strict pattern (such as automated reservation or banking systems), our research suggests that spontaneous and less-constrained “conversation” between humans and computers results in the production of more disfluencies. This is potentially due to the unfamiliarity of having those kinds of interactions with a virtual agent instead of a human doctor. This finding has implications for various different fields. For example, as the use of systems such as CognoSpeak increases, researchers should be aware that the speech these systems elicit are likely to contain higher rates of disfluencies. Not only will this impact researchers wishing to use such data, but could also affect how well these systems work as disfluencies have been shown to have negative effects on Automatic Speech Recognition (ASR) systems.

6.1.1.2 Research Question Two

Can an analysis of speech disfluencies be used to discriminate between different levels of cognitive decline?

Our first manual disfluency analysis in Chapter 3 revealed that the number and length of unfilled pauses was significantly different between our Healthy Controls (HCs) and our participants with Neurodegenerative Dementia (ND) ($p = <0.001$ and ($p = <.009$) respectively). In addition, the number of unfilled pauses were significantly higher in the Mild Cognitive Impairment (MCI) and Functional Memory Disorder (FMD) groups when compared to the control group ($p = 0.01$ and $p = 0.03$). This supports previous findings (for example from Vincze et al. [2021]; or Yuan et al. [2021]) that suggest the rate and

duration of unfilled pauses increases with severity of CD.

In terms of other disfluencies, we found that the number of word repetitions were significantly different between the FMD and ND groups ($p = 0.004$), with the FMD group exhibiting a rate that was smaller than previously reported word repetition rates for healthy individuals [Bortfeld et al., 2001]. The MCI group exhibited more frequent prolongations than all other groups. This difference was statistically significant when compared to the HCs ($p = 0.004$). There was also a statistically significant difference when considering the length of prolongations between the HCs and the MCI group, with the prolongations in the HC group being significantly shorter than those of the MCI group ($p = 0.04$).

Our results demonstrate that an analysis of disfluencies can yield statistically significant information that could help to differentiate between different levels of CD. We also found that the amount of speech elicited by the digital avatar was sufficient to allow for this disfluency analysis to be conducted.

6.1.1.3 Research Question Three

How do the patterns of disfluency vary from the interview task to a picture description task?

Our second manual disfluency analysis, presented in Chapter 4, found that disfluency rates were higher in the interview task when compared to the picture description task. We hypothesise that this is due to the more spontaneous nature of the interview task, in line with previous research such as Fraundorf and Watson [2011] which found that rates of disfluencies such as filled pauses increase during recall tasks. The questions asked of the participants during the interview task include a mixture of long and short-term memory questions (such as “Tell me what you did when you left school?” compared to “What has been in the news recently?”). The picture description task does not require such recall, as patients can rely on the picture in front of them to provide all the information they need to produce a description. Although the disfluency rates are lower in the picture description task, we again observe higher rates than have previously been reported for healthy adults completing a similar task. For example, our healthy controls had a word

repetition rate of 0.55 per 100 words. Duchin and Mysak [1987] report a word repetition rate of 0.31 for their group of healthy adults aged between 65-74, comparable to the ages of the healthy participants used in our study. Whilst this is a small increase, the rates of disfluencies produced by the healthy participants in our analysis are consistently higher than those reported by Duchin and Mysak. This could suggest that the presence of a digital avatar rather than a human interviewer is having some effect on the frequency of speech disfluencies produced during this task.

In addition, we found a general trend of the frequency of all disfluencies increasing along with the CD. This supports previous findings in terms of unfilled pauses but also suggests that (in line with our results from Chapter 3) other kinds of speech disfluency may also be reflective of levels of CD.

6.1.1.4 Research Question Four

Can disfluency information improve the accuracy of an automatic cognitive decline classification system?

Our proof-of-concept study presented in Chapter 4 demonstrates that including disfluency information in an ACDC system can enhance the classification accuracy. Other work in this thesis has found that disfluency rates are higher when people are speaking with the digital avatar. This implies that fully automatic data collection and classification systems such as CognoSpeak are well positioned to leverage disfluency information to improve classification accuracy. This is especially important when considering the need for interpretable features in medical machine learning systems, where transparency is key to helping doctors and clinicians trust the decisions made by these systems.

6.1.1.5 Research Question Five

How do patients construct their problem presentation phases in a medical interview with a human doctor versus a digital avatar?

Chapter 5 revealed that patients' problem presentation phases can be classified depending on whether or not they are accepting of their memory issues (classified as "acceptees"

and “deniers”). Acceptee responses could be categorised as one of three types: emotional, generalising, and taciturn. We found that there was no correlation between category of response and level of *CD*. We also found that there was no correlation between category of response and type of interviewer (human clinician vs digital avatar). We found no evidence that taciturn patients divulge more information to the human doctor when compared to the digital avatar. This suggests that whilst many patients are receptive to further inquiries, there are also patients who are not. This has implications for the design of systems such as *CognoSpeak*, and suggests that having a system that is able to interject with additional questions may not always elicit more information from participants.

The microanalysis of Participant 096/0221, who answered questions from both the human and digital doctor, revealed numerous similarities in the way he presented his symptoms on both occasions. This suggests that participants treat the digital avatar much in the same way as they do the human doctor. We found that the main observable difference between the two problem presentation phases was that more frequent and longer unfilled pauses at transition relevance places were observed in the computer-directed speech. As this patient’s Mini Mental State Examination (*MMSE*) score had barely changed in the time between the two recordings, we assert that this increase in disfluencies is largely due to the unfamiliarity of the situation from which the data was collected and the fact that the avatar does not have the ability to “chime in” during the patients’ descriptions, and therefore the length of these pauses depends entirely on the patient instead of being a co-constructive act which can also be influenced by the conversational partner.

6.2 Limitations and Further Work

The following section discusses the limitations of the work completed in this thesis alongside suggestions for further work. These are addressed according to the three different methodologies used in each of the analysis chapters.

6.2.1 Disfluency Analysis

The disfluency analyses conducted for this thesis were time consuming and required a trained phonetician to complete. This makes such analyses an unlikely choice for use in a clinical setting. Although our analyses included a fair amount of participants (55 for Analysis I and 48 for Analysis II), more data would have provided a more representative view of disfluencies. However, some researchers have argued that as little as five or six participants per group allows for statistically sound generalisations to be made about the collected data [Schleef and Meyerhoff, 2010]. Whilst there were participants in the larger datasets that we did not include in our subsets, we have given justifications for this where relevant (including low recording quality, too much overlapping speech, etc). Whilst we have suggested that the disfluency analysis would remain largely the same for a variety of accents and dialects in the UK, we do accept that some adjustments may need to be made to the [DisCo](#) schema in order to investigate disfluencies across different languages.

Work also needs to be done to investigate conditions that are frequently found to be comorbid with dementia, such as depression. Recent studies (such as [Koops et al. \[2023\]](#)) have shown that depression can affect speech in different ways, including reducing the speech rate and resulting in more self-references. A thorough investigation into such conditions is required to ensure that the disfluencies found are actually a result of the [CD](#) under investigation and not anything else.

As part of the work in this thesis, numerous discussions were held with doctors and clinicians about the feasibility of using both disfluency analysis and [ACDC](#) systems for healthcare purposes. One of the major points raised is how to decide what constitutes a piece of information that would actually affect a clinical decision. For example, where is the point at which the number of repetitions observed in a person's speech becomes clinically significant? Recent work from [Kothare et al. \[2022\]](#) introduced the concept of a Minimal Clinically Important Difference ([MCID](#)) used in conjunction with an automatic system intended to investigate speech and facial metrics for Parkinson's Disease assessment. However, in order to define an [MCID](#), gold-standard diagnostic labels are required. These are difficult to obtain in cases of diseases such as dementia due to the difficulties

doctors and clinicians face in correctly diagnosing these conditions. This is an important concept that should be considered when continuing work on automatic machine learning systems for clinical purposes.

6.2.2 Automatic Cognitive Decline Classification

Whilst we found an improvement in classification accuracy when testing a simple [ACDC](#) system on our data, there needs to be an investigation into how well disfluency features perform when being tested on other datasets. It would also be interesting to find out how well the disfluency features perform when used alongside more complex classification algorithms.

Further work should also examine how well the disfluency features from both the picture description task and the interview task work when used together. We hypothesise that this would lead to greater advances in classification accuracy. It would also be valuable to investigate how well disfluency features perform on languages other than English.

Our results indicate a positive step towards improving [ACDC](#) systems, but there are numerous considerations that need to be addressed when deliberating the use of such systems in a healthcare environment:

Firstly, [ACDC](#) systems rely heavily on technology, including hardware and software components. Any failures or glitches could disrupt how well the system is working, leading to inaccurate classifications or even system downtime.

Secondly, [ACDC](#) systems typically process large volumes of sensitive data, which in this instance would include patient health information. This raises concerns about data privacy and security. If not properly secured, the data stored by [ACDC](#) systems could be vulnerable to unauthorised access, data breaches, or misuse, potentially compromising patient confidentiality and trust.

Thirdly, like all machine learning systems, [ACDC](#) systems are susceptible to bias, which may result in unfair or discriminatory outcomes, particularly if the training data used to develop the system is unrepresentative of the target population. Biases in the system's

algorithms or decision-making processes could lead to disparities in healthcare delivery or exacerbate existing inequalities.

Finally are the issues of interpretability and trust, as discussed in Section 4.6.1. Despite efforts to improve interpretability, many ACDC systems are still considered to be “black boxes”, making it challenging for both clinicians and patients to understand how decisions are made. A lack of transparency and interpretability can undermine trust in the system, especially in healthcare contexts where decisions have significant consequences for patient care. In particular, patients may feel uneasy or distrustful if they perceive that their healthcare decisions are being influenced or overridden by automated systems, potentially undermining patient autonomy and the trust that patients hold in their doctors.

Overall, while ACDC systems offer promising opportunities to enhance clinical decision-making and healthcare delivery, it is essential to carefully address the potential downsides and challenges associated with their use to ensure that they are deployed ethically, responsibly, and effectively for healthcare settings.

6.2.3 Conversation Analysis

While the results of our conversation analysis suggest little correlation between level of CD, interlocutor type (human-human vs. human-avatar) and type of response, we acknowledge that further investigation on this data could still potentially reveal diagnostic profiles (following from the work presented by [Elsey et al. \[2015\]](#)).

Our case study revealed that (at least in this case) participants treat the digital avatar in much the same way as the human clinician. It would be interesting to investigate whether the same thing can be observed across a different range of participants, including younger participants who may be more familiar with directing speech at computers and therefore having more preconceived notions of how they should adjust their own speech in order to be better “understood” by the system. Expanding on this point, it would be beneficial to include a more diverse range of participants. The participants involved in our study represent a somewhat homogeneous group. It would be interesting to investigate whether the types of problem presentation phases produced by these participants gener-

alise to other groups. Exploring how different demographic groups interact with digital avatars and human clinicians could provide insights into the homogeneity or variability of communication strategies in healthcare scenarios. Additionally, such investigations could inform the design and implementation of future ACDC systems, helping to further tailor automatic systems to better meet the diverse needs of users across different age groups and backgrounds.

With the ongoing improvements that are being made to the CognoSpeak system, it would be an interesting future direction to compare the IVA data used here to the more recently collected CognoSpeak data. This would allow us to compare any differences in the speech of participants as the result of conversing with a more animated and natural virtual avatar. Whilst the new version of the IVA still lacks the ability to interject or ask for follow up information on specific points raised by the participants, the enhanced gestures and body movements may serve to encourage participants to divulge more information when compared to the more static and less natural original avatar.

6.3 Concluding Remarks

This thesis introduced a novel, multidisciplinary approach to assessing levels of cognitive decline through speech. We explored conversation analysis, quantitative disfluency analysis, and machine learning approaches to this problem. The proof-of-concept study presented towards the end of this thesis confirmed that the inclusion of disfluency features into automatic cognitive decline classification systems can provide accuracy improvements. This contributes to the overall body of work that is focussed on using ACDC systems to help address current issues surrounding the diagnosis of cognitive decline.

Appendix A

Table of all group-based results from
the first manual disfluency study

	HC	FMD	MCI	ND
Average Age	69.5	55.1	63.6	68.1
Average MMSE Score	28.7	27.4	26.5	23.5
Total n. Fluent Words	8792	7542	5425	6547
Average n. Fluent Words	586	538	387	545
Total Locution Time (s)	3114.5	2878.0	2905.1	2805.9
Average Locution Time (s)	207.6	205.6	207.5	233.8
Rate of UFP per 100 Fluent Words	14.1	20.7	22.6	26.5
Average Length of UFP (s)	0.54	1.12	0.87	0.92
Rate of FP per 100 Fluent Words	7.4	6.2	8.2	9.2
Average Length of FP (s)	0.43	0.59	0.55	0.53
Rate of PRO per 100 Fluent Words	3.5	5.8	8.6	11.1
Average length of PRO (s)	0.53	0.51	0.49	0.45
Rate of ADD per 100 Fluent Words	n/a	0.013	0.018	0.076
Rate of DEL per 100 Fluent Words	n/a	n/a	0.055	0.061
Rate of SUB per 100 Fluent Words	n/a	n/a	0.092	0.030
Rate of MAL per 100 Fluent Words	0.011	0.039	n/a	0.091
PWREP rate per 100 Fluent Words	0.32	0.29	0.58	1.05
WREP rate per 100 Fluent Words	1.08	0.22	1.60	1.68
PHREP rate per 100 Fluent Words	0.29	0.13	0.31	0.48
REPA rate per 100 Fluent Words	0.95	0.74	1.27	1.86

Appendix B

Individual counts and normalised rates per 100 fluent words for part word, whole word, and phrase repetitions in the interview task.

Participant Number	PWREP		WREP		PHREP	
	Raw Count	Rate	Raw Count	Rate	Raw Count	Rate
0249	8	0.90	18	2.02	3	0.33
0250	0	0	1	0.27	0	0
0251	0	0	5	0.88	1	0.17
0252	3	0.21	46	3.24	17	1.19
0253	1	0.38	1	0.38	1	0.38
0265	1	0.10	2	0.20	2	0.20
0266	0	0	3	0.99	0	0
HC 0267	2	0.58	0	0	0	0
0270	5	0.62	6	0.75	0	0
0271	0	0	1	0.38	1	0.38
0272	0	0	1	0.25	0	0
0273	1	0.22	0	0	0	0
0274	2	0.84	0	0	0	0
0277	2	0.17	8	0.69	1	0.08
0278	4	0.84	3	0.78	0	0

	Participant Number	PWREP		WREP		PHREP	
		Raw Count	Rate	Raw Count	Rate	Raw Count	Rate
FMD	0201	3	1.05	0	0	1	0.35
	0203	1	0.07	3	0.21	4	0.29
	0206	6	1.01	5	0.84	0	0
	0207	2	1.14	0	0	0	0
	0209	1	0.71	0	0	0	0
	0210	4	0.76	1	0.19	1	0.19
	0217	0	0	0	0	0	0
	0239	1	0.18	2	0.37	0	0
	0243	0	0	1	0.54	0	0
	0259	0	0	0	0	0	0
	0268	0	0	1	0.54	0	0
	0279	2	0.09	4	0.19	2	0.09
	0280	1	0.55	0	0	0	0
	2114	1	0.18	0	0	1	0.18
MCI	0208	8	2.96	3	1.11	1	0.37
	0212	1	0.17	4	0.70	0	0
	0214	0	0	2	0.95	0	0
	0215	3	1.29	1	0.43	0	0
	0218	1	0.54	2	1.08	0	0
	0219	0	0	1	0.55	1	0.55
	0229	1	0.22	4	0.88	0	0
	0236	2	3.17	0	0	0	0
	0242	10	0.63	63	3.79	13	0.78
	0258	4	0.92	3	0.69	0	0
	0261	1	0.58	1	0.58	0	0
	0262	0	0	0	0	0	0
	2105	0	0	0	0	0	0
	2112	1	0.18	3	0.54	2	0.36

Participant Number	PWREP		WREP		PHREP	
	Raw Count	Rate	Raw Count	Rate	Raw Count	Rate
0213	1	0.49	2	0.99	1	0.49
0221	15	0.74	29	1.44	7	0.34
0222	3	0.57	3	0.57	1	0.19
0223	0	0	12	4.08	0	0
0238	1	0.29	1	0.29	0	0
0275	0	0	0	0	0	0
0281	0	0	2	0.60	0	0
2100	2	0.60	3	0.90	0	0
2104	2	2.24	0	0	0	0
2106	1	0.29	5	1.48	0	0
2110	39	5.95	34	5.19	5	0.76
2111	5	0.37	19	1.42	18	1.35

ND

Appendix C

Variations in the Intelligent Virtual Agent questions

Set One:

1. Why have you come in today, and what are your expectations?
2. Tell me, what problems have you had with your memory?
3. Who is most worried about your memory, you, or somebody else? And what did you do over last weekend? Please give as much detail as you can.
4. What has been in the news recently? Please give as much detail as you can.
5. Tell me about the school you went to, and how old you were when you left.
6. Tell me about what you did when you left school. What jobs did you do?
7. Tell me about your last job. Give as much detail as you can.
8. Who manages your finances, you, or somebody else? And has this changed recently?

Set Two:

1. Why've you come today and what are your expectations?

2. Tell me what problems you've noticed with your memory recently?
3. Who is most worried about your memory, you, or somebody else?
4. What did you do over last weekend? Giving as much detail as you can.
5. What has been in the news recently?
6. Tell me about the school you went to, and how old you were when you left.
7. Tell me what you did when you left school. What jobs did you do?
8. Tell me about your last job, give as much detail as you can.
9. Who manages your finances, you, or somebody else? Has this changed recently?

Set Three:

1. Where have you come in from today, and what are you hoping to find out?
2. Tell me what problems you've noticed with your memory recently?
3. Who is most worried about your memory, you, or somebody else?
4. What did you do over last weekend? Giving as much detail as you can.
5. What has been in the news recently?
6. Tell me about the school you went to, and how old you were when you left?
7. Tell me what you did when you left school, what jobs did you do?
8. Tell me about your last job, give as much detail as you can.
9. Who manages your finances, you, or somebody else? Has this changed recently?

Appendix D

Table of all group-based results from
the second manual disfluency
study

	HC	MCI	ND
Average Age	69.5	62.4	69.4
Average MMSE Score	28.7	26.7	23.1
Total n. Fluent Words	2152	1665	1624
Average n. Fluent Words	165	97	90
Total Locution Time (s)	880.1	838.4	984.9
Average Locution Time (s)	67.7	49.3	54.7
Rate of UFP per 100 Fluent Words	13.2	17	20
Average Length of UFP (s)	0.57	0.95	1.33
Rate of FP per 100 Fluent Words	5.4	7.2	4.9
Average Length of FP (s)	0.48	0.49	0.61
Rate of PRO per 100 Fluent Words	0.05	0.06	0.08
Average length of PRO (s)	0.57	0.44	0.33
Rate of ADD per 100 Fluent Words	n/a	n/a	n/a
Rate of DEL per 100 Fluent Words	n/a	0.06	0.06
Rate of SUB per 100 Fluent Words	0.04	n/a	n/a
Rate of CIR per 100 Fluent Words	n/a	n/a	0.06
Rate of LEX per 100 Fluent Words	0.04	n/a	0.18
PWREP rate per 100 Fluent Words	0.05	0.18	0.55
WREP rate per 100 Fluent Words	0.56	0.84	0.55
PHREP rate per 100 Fluent Words	0.37	0.30	0.43
REPA rate per 100 Fluent Words	0.88	1.74	1.73

Appendix E

Transcription conventions for conversation analysis

VA / Dr / Pt / AP	Speaker labels (e.g., VA = virtual agent, Dr = doctor, Pt = patient, AP = accompanying person)
[]	Encloses talk produced in overlap i.e., when more than one speaker is speaking
=	Links talk produced in close temporal proximity (latched talk)
> <	Talk between symbols is rushed or compressed
◦ ◦	Encloses talk which is produced quietly
<u>Underline</u>	Underlining is used to mark words or syllables which are given special emphasis of some kind
CAPS	Words or parts of words spoken loudly
s:::::::::	Sustained or stretched sound, each colon represents 10ms
.hhh	Inbreath, each “h” represents 10ms
hhh.	Outbreath, each “h” represents 10ms
(word)	Parentheses represents transcriber doubt
(this/that)	Alternative hearings
((description))	Description of what can be heard rather than transcription e.g., ((shuffling papers))
cu-	Cut-off word or sound
(0.6)	Silence in seconds
(.)	A silence of less than 20ms
↑	Indicates a marked pitch rise
↓	Indicates a marked fall in pitch
(hhenhh)	Indicates laughter while speaking (aspiration)

References

- Abbas, I., Ahmed, K., and Habib, M. (2022). Conversation analysis: A methodology for diagnosing autism. *Global Language Review*, pages 1–12. [149](#)
- Abrams, L. and Farrell, M. T. (2011). Language processing in normal aging. In *The handbook of psycholinguistic and cognitive processes*, pages 49–73. Psychology Press. [11](#), [12](#)
- Adda-Decker, M., Habert, B., Barras, C., Adda, G., Mareuil, P. B. d., and Paroubek, P. (2003). A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. In *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*. [42](#)
- Aggarwal, R. K. and Dave, M. (2011). Acoustic modeling problem for automatic speech recognition system: conventional methods (part i). *International Journal of Speech Technology*, 14:297–308. [36](#)
- Ahmed, S., de Jager, C. A., Haigh, A.-M., and Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer’s disease. *Neuropsychology*, 27(1):79. [24](#)
- Alexander, M., Blackburn, D., and Reuber, M. (2019). Patients’ accounts of memory lapses in interactions between neurologists and patients with functional memory disorders. *Sociology of Health & Illness*, 41(2):249–265. [21](#)
- Alpaydin, E. (2014). *Introduction to Machine Learning*. MIT Press, 4th edition. [43](#)

- Alsubaie, S., Grant, D., and Donyai, P. (2022). The utility of conversation analysis versus roter’s interaction analysis system for studying communication in pharmacy settings: a scoping review. *International Journal of Pharmacy Practice*, 30(1):17–27. 54
- Altinok, D. (2021). *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd. 39
- Alzheimer’s Research U.K (2023). Dementia statistics hub. 13, 14, 15
- Alzheimer’s Society (2021). How much does dementia care cost? 15
- Amatriain, X. et al. (2005). *An object-oriented metamodel for digital signal processing with a focus on audio and music*. PhD thesis, Universitat Pompeu Fabra. 37
- Ammar, R. B. and 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)*, pages 1–8. IEEE. 34
- Antaki, C. and Wilkinson, R. (2013). Conversation analysis and the study of atypical populations. *The handbook of conversation analysis*, pages 533–550. 148
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*. 131
- Arend, B., Sunnen, P., and Caire, P. (2017). Investigating breakdowns in human robot interaction: a conversation analysis guided single case study of a human-nao communication in a museum environment. *International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering*, 11(5). 55
- Aronoff, J. M., Gonnerman, L. M., Almor, A., Arunachalam, S., Kempler, D., and Andersen, E. S. (2006). Information content versus relational knowledge: Semantic deficits in patients with alzheimer’s disease. *Neuropsychologia*, 44(1):21–35. 23
- Arslan, B. and Göksun, T. (2022). Aging, gesture production, and disfluency in speech: A comparison of younger and older adults. *Cognitive Science*, 46(2):e13098. 30, 61

- ARUK (2023). What is vascular dementia? [14](#)
- Assad-Uz-Zaman, M., Rasedul Islam, M., Miah, S., and Rahman, M. H. (2019). Nao robot for cooperative rehabilitation training. *Journal of rehabilitation and assistive technologies engineering*, 6:2055668319862151. [55](#)
- Atkinson, J. M. (1985). Preference organisation. In Atkinson, J. M., editor, *Structures of Social Action*, pages 53–56. Cambridge University Press. [52](#)
- Atkinson, J. M. and Drew, P. (1979). *Order in court: the organisation of verbal interaction in juridical settings*. Macmillan Press. [53](#)
- Balagopalan, A., Eyre, B., Rudzicz, F., and Novikova, J. (2020). To bert or not to bert: comparing speech and language-based approaches for alzheimer’s disease detection. *arXiv preprint arXiv:2008.01551*. [45](#)
- Ball, H. A., McWhirter, L., Ballard, C., Bhome, R., Blackburn, D. J., Edwards, M. J., Fleming, S. M., Fox, N. C., Howard, R., Huntley, J., et al. (2020). Functional cognitive disorder: dementia’s blind spot. *Brain*, 143(10):2895–2903. [21](#)
- Bang, J., Spina, S., and Miller, B. L. (2015). Frontotemporal dementia. *The Lancet*, 386(10004):1672–1682. [14](#)
- Barczewska, K. and Igras, M. (2013). Detection of disfluencies in speech signal. *Challenges of modern technology*, 4(2):3–10. [110](#)
- Baron, N. S. (2015). Shall we talk? conversing with humans and robots. *The Information Society*, 31(3):257–264. [129](#)
- Beavis, L., O’Malley, R., Mirheidari, B., Christensen, H., and Blackburn, D. (2021). How can automated linguistic analysis help to discern functional cognitive disorder from healthy controls and mild cognitive impairment? *BJPsych Open*, 7(S1):S7–S7. [78](#), [97](#), [98](#)

- Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and 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. [133](#)
- Beckman, H. B. and Frankel, R. M. (1984). The effect of physician behavior on the collection of data. *Annals of Internal medicine*, 101(5):692–696. [144](#)
- Beier, E. J., Chantavarin, S., and Ferreira, F. (2023). Do disfluencies increase with age? evidence from a sequential corpus study of disfluencies. *Psychology and Aging*, 38(3):203. [30](#)
- Bhatt, S., Jain, A., and Dev, A. (2020). Acoustic modeling in speech recognition: a systematic review. *International Journal of Advanced Computer Science and Applications*, 11(4). [36](#)
- Bickel, C., Pantel, J., Eysenbach, K., and Schröder, J. (2000). Syntactic comprehension deficits in alzheimer's disease. *Brain and Language*, 71(3):432–448. [22](#)
- Boersma, P. and Van Heuven, V. (2001). Speak and unspeak with praat. *Glott International*, 5(9/10):341–347. [82](#)
- Bolden, G. B. and Hepburn, A. (2018). Transcription for conversation analysis. In *Oxford Research Encyclopedia of Communication*. Oxford University Press. [152](#)
- Borson, S. and Raskind, M. A. (1997). Clinical features and pharmacologic treatment of behavioral symptoms of alzheimer's disease. *Neurology*, 48(5_suppl_6):17S–24S. [16](#)
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and speech*, 44(2):123–147. [30](#), [90](#), [98](#), [188](#), [189](#), [190](#)
- Braun, A., Elsässer, N., and Willems, L. (2023). Disfluencies revisited—are they speaker-specific? *Languages*, 8(3):155. [62](#)
- Brennan, S. E. and Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of memory and language*, 44(2):274–296. [28](#)

- Bucks, R. S., Singh, S., Cuerden, J. M., and 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. 38
- Burke, D. M. and MacKay, D. G. (1997). Memory, language, and ageing. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1363):1845–1856. 11
- Burke, D. M., MacKay, D. G., and James, L. E. (2000). Theoretical approaches to language and aging. In Perfect, T. J. and Maylor, E. A., editors, *Models of Cognitive Aging*, pages 207–237. Oxford University Press. 11
- Burns, A. (2000). The burden of alzheimer’s disease. *International journal of Neuropsychopharmacology*, 3(Supplement_2):S31–S38. 15
- Burns, A., Lawlor, B., and Craig, S. (2002). Rating scales in old age psychiatry. *The British Journal of Psychiatry*, 180(2):161–167. 24, 25
- Butterworth, B. (1980). Evidence from pauses in speech. *Language production*, 1:155–176. 69
- Byrne, P. S. and Long, B. E. L. (1984). *Doctors talking to patients : a study of the verbal behaviour of general practitioners consulting in their surgeries*. Royal College of General Practitioners, Exeter. 137, 138
- Calzà, L., Gagliardi, G., Favretti, R. R., and Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, 65:101113. 40
- Campione, E. and Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In *Speech prosody 2002, international conference*. 69
- Can, E. and Kuruoglu, G. (2018). Assessment of syntactic complexity in alzheimer’s disease. In *3rd Eurasian Conference on Language and Social Sciences*, volume 517. 23

- Castro, D. M., Dillon, C., Machnicki, G., and Allegri, R. F. (2010). The economic cost of alzheimer's disease: Family or public-health burden? *Dementia & Neuropsychologia*, 4:262–267. 15
- Centre for Better Ageing (2020). Our ageing population - the state of ageing 2023-24. <https://tinyurl.com/5b7ch7j6>. Accessed: 05/01/2024. 4
- Cera, M. L., Ortiz, K. Z., Bertolucci, P. H. F., and Minett, T. (2018). Phonetic and phonological aspects of speech in alzheimer's disease. *Aphasiology*, 32(1):88–102. 66, 76
- Cera, M. L., Ortiz, K. Z., Bertolucci, P. H. F., Tsujimoto, T., and Minett, T. (2023). Speech and phonological impairment across alzheimer's disease severity. *Journal of Communication Disorders*, 105:106364. 98
- Chakraborty, R., Pandharipande, M., Bhat, C., and Kopparapu, S. K. (2020). Identification of dementia using audio biomarkers. *arXiv preprint arXiv:2002.12788*. 17
- Chertkow, H. (2002). Mild cognitive impairment. *Current opinion in neurology*, 15(4):401–407. 19
- Chertkow, H., Verret, L., Bergman, H., Wolfson, C., and McKelvey, R. (2001). Predicting progression to dementia in elderly subjects with mild cognitive impairment: A multidisciplinary approach. In *Neurology*, volume 56, pages A216–A217. LIPPINCOTT WILLIAMS & WILKINS 530 WALNUT ST, PHILADELPHIA, PA 19106-3621 USA. 18
- Chlasta, K. and Wolk, K. (2021). Towards computer-based automated screening of dementia through spontaneous speech. *Frontiers in Psychology*, 11:623237. 32
- Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., and Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology*, 9(2):14–14. 44
- Clare, L. (2004). Awareness in early-stage alzheimer's disease: a review of methods and evidence. *British Journal of Clinical Psychology*, 43(2):177–196. 15

- Clayman, S. (2013). Turn-constructive units and transition relevance place. In Sidnell, J. and Stivers, S., editors, *The handbook of conversation analysis*, pages 150–166. Blackwell Publishing. 50
- Cook, L., Souris, H., and Isaacs, J. (2020). The 2019 national memory service audit. *NHS London Clinical Networks*. 106
- Cooper, P. V. (1990). Discourse production and normal aging: Performance on oral picture description tasks. *Journal of gerontology*, 45(5):P210–P214. 26
- Cossavella, F. and Cevasco, J. (2021). The importance of studying the role of filled pauses in the construction of a coherent representation of spontaneous spoken discourse. *Journal of Cognitive Psychology*, 33(2):172–186. 97
- Croot, K., Hodges, J. R., Xuereb, J., and Patterson, K. (2000). Phonological and articulatory impairment in alzheimer’s disease: a case series. *Brain and language*, 75(2):277–309. 22
- Cummings, J. L., Benson, D. F., Hill, M. A., and Read, S. (1985). Aphasia in dementia of the alzheimer type. *Neurology*, 35(3):394–394. 23
- Cummings, L. (2019). Describing the cookie theft picture: Sources of breakdown in alzheimer’s dementia. *Pragmatics and Society*, 10(2):153–176. 104
- Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49. 80
- De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455. 44
- Deary, I. J., Corley, J., Gow, A. J., Harris, S. E., Houlihan, L. M., Marioni, R. E., Penke, L., Rafnsson, S. B., and Starr, J. M. (2009). Age-associated cognitive decline. *British medical bulletin*, 92(1):135–152. 11

- Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, 93(3):283. 65, 68, 76
- Dell, G. S. and Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of verbal learning and verbal behavior*, 20(6):611–629. 65
- Derczynski, L. (2016). Complementarity, f-score, and nlp evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266. 47
- Doherty, C. M. and Forbes, R. B. (2014). Diagnostic lumbar puncture. *The Ulster medical journal*, 83(2):93. 17
- Drew, P. (2013). Turn design. In Sidnell, J. and Stivers, S., editors, *The handbook of conversation analysis*, pages 131–149. Blackwell Publishing. 50
- Duchin, S. W. and Mysak, E. D. (1987). Disfluency and rate characteristics of young adult, middle-aged, and older males. *Journal of communication disorders*, 20(3):245–257. 107, 191
- Ekberg, K. and Reuber, M. (2015). Ekberg, k. & reuber, m.(2015). can conversation analytic findings help with differential diagnosis in routine seizure clinic interactions? *communication & medicine*, 12 (1), 13-24. *Communication & medicine*, 12(1):13–24. 148, 149
- Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis, Linköping University Electronic Press. 28, 61, 62
- Elsley, C., Drew, P., Jones, D., Blackburn, D., Wakefield, S., Harkness, K., Venneri, A., and Reuber, M. (2015). Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077. 21, 143, 149, 150, 156, 163, 195
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic

- acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202. [37](#), [38](#)
- Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. [37](#)
- Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. [36](#), [132](#)
- Faure, M. (1980). Results of a contrastive study of hesitation phenomena in french and german. *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler*, pages 287–290. [83](#)
- Feldman, R. C., Aldana, E., and Stein, K. (2019). Artificial intelligence in the health care space: how we can trust what we cannot know. *Stan. L. & Pol’y Rev.*, 30:399. [129](#)
- Fischer, K. (2006). What computer talk is and isn’t: Human-computer conversation as intercultural communication. vol. 17. *Linguistics-Computational Linguistics. AQ-Verlag*. [148](#)
- Fischer, K. (2011a). How people talk with robots: Designing dialog to reduce user uncertainty. *Ai Magazine*, 32(4):31–38. [148](#)
- Fischer, K. (2011b). Interpersonal variation in understanding robots as social actors. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 53–60. [148](#)
- Fischer, P., Jungwirth, S., Zehetmayer, S., Weissgram, S., Hoenigschnabl, S., Gelpi, E., Krampla, W., and Tragl, K. (2007). Conversion from subtypes of mild cognitive impairment to alzheimer dementia. *Neurology*, 68(4):288–291. [20](#)
- Fleming, V. B. and Harris, J. L. (2008). Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology*, 22(7-8):729–740. [24](#)

- Folstein, M. F., Folstein, S. E., and 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. 25
- Forbes, K. E., Venneri, A., and Shanks, M. F. (2002). Distinct patterns of spontaneous speech deterioration: an early predictor of alzheimer’s disease. *Brain and Cognition*, 48(2-3):356–361. 23
- Forbes-McKay, K. E. and Venneri, A. (2005). Detecting subtle spontaneous language decline in early alzheimer’s disease with a picture description task. *Neurological sciences*, 26:243–254. 22
- Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738. 27, 83
- Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer’s disease in narrative speech. *Journal of Alzheimer’s Disease*, 49(2):407–422. 132
- Fraundorf, S. H. and Watson, D. G. (2011). The disfluent discourse: Effects of filled pauses on recall. *Journal of memory and language*, 65(2):161–175. 128, 190
- Freud, S. (1989). *Psychopathology of everyday life*. WW Norton & Company. 62
- Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, pages 27–52. 65
- Gafaranga, J. and Britten, N. (2003). “fire away”: the opening sequence in general practice consultations. *Family practice*, 20(3):242–247. 140
- Gandhi, R. (2018). Support vector machine - introduction to machine learning algorithms. <https://tinyurl.com/3jc52k48>. Accessed: 03/01/2024. 45
- Gardner, R. (2004). Conversation analysis. In Davies, A. and Elder, C., editors, *The handbook of applied linguistics*, pages 262–281. Wiley Online Library. 50, 51

- Garrett, M. F. (1975). The analysis of sentence production. In *Psychology of learning and motivation*, volume 9, pages 133–177. Elsevier. 65
- Gasteiger, N., Loveys, K., Law, M., and Broadbent, E. (2021). Friends from the future: a scoping review of research into robots and computer agents to combat loneliness in older people. *Clinical interventions in aging*, pages 941–971. 146
- Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., et al. (2006). Mild cognitive impairment. *The lancet*, 367(9518):1262–1270. 19
- Geldmacher, D. S. and Whitehouse, P. J. (1996). Evaluation of dementia. *New England Journal of Medicine*, 335(5):330–336. 13
- Georgevici, A. I. and Terblanche, M. (2019). Neural networks and deep learning: a brief introduction. *Intensive Care Medicine*, 45(5):712–714. 44
- Ghai, W. and Singh, N. (2012). Literature review on automatic speech recognition. *International Journal of Computer Applications*, 41(8). 35
- Ghosh, A. (2010). Endocrine, metabolic, nutritional, and toxic disorders leading to dementia. *Annals of Indian Academy of Neurology*, 13(Suppl2):S63. 13
- Gil, D. and Johnsson, M. (2011). Support vector machines in medical classification tasks. In Boyle, B. H., editor, *Support vector machines: data analysis, machine learning and applications*, pages 81–102. Nova Science Publishers, Inc. 45
- Gill, V. (2005). Patient” demand” for medical interventions: exerting pressure for an offer in a primary care clinic visit. *Research on Language and Social Interaction*, 38(4):451–479. 53
- Gómez-Vilda, P., Rodellar-Biarge, V., Nieto-Lluis, V., de Ipiña, K. L., Álvarez-Marquina, A., Martínez-Olalla, R., Ecay-Torres, M., and Martínez-Lage, P. (2015). Phonation biomechanical analysis of alzheimer’s disease cases. *Neurocomputing*, 167:83–93. 61

- Goodglass, H. and Kaplan, E. (1983). The assessment of aphasia and related disorders, 2nd edn lea & febiger: Philadelphia. *Dictionary of Biological Psychology*, 230. 103
- Goodwin, C. and Heritage, J. (1990). Conversation analysis. *Annual review of anthropology*, 19(1):283–307. 49
- Gurbuz, N. (2017). Understanding fluency and disfluency in non-native speakers' conversational english. *EDUCATIONAL SCIENCES-THEORY & PRACTICE*, 17(6):1853–1874. 80
- Guzman, A. L. (2018). What is human-machine communication, anyway. *Human-machine communication: Rethinking communication, technology, and ourselves*, pages 1–28. 129
- Halkowski, T. (2006). Realizing the illness: patients' narratives of symptom discovery. *Studies in Interactional Sociolinguistics*, 20:86. 138, 143
- Hallowell, N., Badger, S., Sauerbrei, A., Nellåker, C., and Kerasidou, A. (2022). “i don't think people are ready to trust these algorithms at face value”: trust and the use of machine learning algorithms in the diagnosis of rare disease. *BMC Medical Ethics*, 23(1):1–14. 129
- Hamilton, H. E. (1994). *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press. 145
- Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., and Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *American Journal of Roentgenology*, 212(1):38–43. 47
- Heath, C. (1981). The opening sequence in doctor-patient interaction. *Medical work: Realities and routines*, 71:90. 140
- Heinrichs, B. and Eickhoff, S. B. (2020). Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human brain mapping*, 41(6):1435–1444. 176

- Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222. [27](#)
- Heritage, J. and Maynard, D. W. (2006). *Communication in medical care: Interaction between primary care physicians and patients*, volume 20. Cambridge University Press. [138](#)
- Heritage, J. and Robinson, J. D. (2006). The structure of patients' presenting concerns: physicians' opening questions. *Health communication*, 19(2):89–102. [138](#), [141](#), [143](#), [151](#), [154](#), [157](#)
- Hier, D. B., Hagenlocker, K., and Shindler, A. G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and language*, 25(1):117–133. [114](#)
- Hobson, J. (2015). The montreal cognitive assessment (moca). *Occupational Medicine*, 65(9):764–765. [26](#)
- Hollands, S., Blackburn, D., and Christensen, H. (2022). Evaluating the performance of state-of-the-art asr systems on non-native english using corpora with extensive language background variation. In *Interspeech 2022: Proceedings of the Annual Conference of the International Speech Communication Association*, pages 3958–3962. International Speech Communication Association (ISCA). [36](#)
- Horner, J. and Massey, E. W. (1983). Progressive dysfluency associated with right hemisphere disease. *Brain and language*, 18(1):71–85. [27](#)
- Horner, J., Norman, M., and Ripich, D. (2007). *Assessment and Management of Dementia*. Thieme Medical Publishers. [14](#), [17](#)
- Horton, W. S., Spieler, D. H., and Shriberg, E. (2010). A corpus analysis of patterns of age-related change in conversational speech. *Psychology and aging*, 25(3):708. [30](#)
- Howell, P. and Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech communication*, 10(2):163–169. [77](#)

- Iadecola, C. (2013). The pathobiology of vascular dementia. *Neuron*, 80(4):844–866. 14
- Ijäs-Kallio, T., Ruusuvuori, J. E., and Peräkylä, A. (2010). Patient involvement in problem presentation and diagnosis delivery in primary care. *Communication & Medicine*. 139
- Jaiswal, R. and Hines, A. (2018). The sound of silence: How traditional and deep learning based voice activity detection. In *Brennan, RB, Beel, J., Byrne, R., Debattista, J. and Crotti Junior, A.(eds.). Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin*. 109
- Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., and 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*, pages 27–37. 22
- Jefferson, G. (1980). On “trouble-premonitory” response to inquiry. *Sociological inquiry*, 50(3-4):153–185. 142
- Jellinger, K. A. and Attems, J. (2010). Is there pure vascular dementia in old age? *Journal of the neurological sciences*, 299(1-2):150–154. 14
- Jing, K. and Xu, J. (2019). A survey on neural network language models. *arXiv preprint arXiv:1906.03591*. 35
- Jiskoot, L. C., Poos, J. M., van Boven, K., de Boer, L., Giannini, L. A., Satoer, D. D., Visch-Brink, E. G., van Hemmen, J., Franzen, S., Pijnenburg, Y. A., et al. (2023). The screeling: Detecting semantic, phonological, and syntactic deficits in the clinical subtypes of frontotemporal and alzheimer’s dementia. *Assessment*, page 10731911231154512. 22
- Jones, D., Drew, P., Elsey, C., Blackburn, D., Wakefield, S., Harkness, K., and Reuber, M. (2016). Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders. *Aging & Mental Health*, 20(5):500–509. 21

- Juravle, G., Boudouraki, A., Terziyska, M., and Rezlescu, C. (2020). Trust in artificial intelligence for medical diagnoses. *Progress in brain research*, 253:263–282. 129
- Kaushik, M., Trinkle, M., and Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. In *Proceedings of the Australasian International Conference on Speech Science and Technology (SST)*, volume 70. 109
- Kempler, D. (2005). *Neurocognitive disorders in aging*. Sage. 22
- Kempler, D. and Goral, M. (2008). Language and dementia: Neuropsychological aspects. *Annual review of applied linguistics*, 28:73–90. 22
- Kendrick, K. H. (2015). Other-initiated repair in english. *Open Linguistics*, 1(1). 52
- Kirk, A. (2023). Cognition in normal aging—a brief review. *Canadian Journal of Neurological Sciences*, pages 1–13. 167
- Kirshner, H. S. (2014). Frontotemporal dementia and primary progressive aphasia, a review. *Neuropsychiatric disease and treatment*, pages 1045–1055. 23
- Kitzinger, C. (2013). Repair. In Sidnell, J. and Stivers, T., editors, *The handbook of conversation analysis*, pages 229–256. Blackwell Publishing. 51
- Kjellmer, G. (2003). Hesitation. in defence of er and erm. *English Studies*, 84(2):170–198. 63
- Klapi, E., Lüdeling, A., and Pompino-Marschall, B. (2011). *Disfluency Patterns: A Contrastive Analysis of L1 and L2 Speakers of German*. PhD thesis, Doctoral dissertation, Thesis. Humboldt-Universität zu Berlin. 80
- Klimova, B. and Kuca, K. (2016). Speech and language impairments in dementia. *Journal of Applied Biomedicine*, 14(2):97–103. 22
- Koller, W. C. (1983). Dysfluency (stuttering) in extrapyramidal disease. *Archives of neurology*, 40(3):175–177. 27

- Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., and Sommer, I. E. (2023). Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*, 22(2):152–160. [193](#)
- Kosmala, L. and Crible, L. (2022). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, 65(1):216–239. [97](#)
- Kothare, H., Neumann, M., Liscombe, J., Roesler, O., Burke, W., Exner, A., Snyder, S., Cornish, A., Habberstad, D., Pautler, D., et al. (2022). Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for parkinson’s disease assessment. In *INTERSPEECH*, pages 3658–3662. [193](#)
- Kramer, O. (2013). *K-nearest neighbors*, pages 13–23. Springer. [44](#)
- Laguarta, J. and Subirana, B. (2021). Longitudinal speech biomarkers for automated alzheimer’s detection. *frontiers in Computer Science*, 3:624694. [17](#)
- Larner, A. (2005). “who came with you?” a diagnostic observation in patients with memory problems? *Journal of Neurology, Neurosurgery & Psychiatry*, 76(12):1739–1739. [163](#)
- Laske, C., Sohrabi, H. R., Frost, S. M., López-de Ipiña, K., Garrard, P., Buscema, M., Dauwels, J., Soekadar, S. R., er, S., Linnemann, C., et al. (2015). Innovative diagnostic tools for early detection of alzheimer’s disease. *Alzheimer’s & Dementia*, 11(5):561–578. [22](#)
- Lebret, R. P. (2016). Word embeddings for natural language processing. Technical report, EPFL. [35](#)
- Lee, H., Gayraud, F., Hirsch, F., and Barkat-Defradas, M. (2011). Speech dysfluencies in normal and pathological aging: A comparison between alzheimer patients and healthy elderly subjects. In *ICPhS*, volume 17, pages 1174–1177. [31](#)
- Lee, S.-H. and Kim, C. W. (2015). Presentation of patients’ problems during triage in emergency medicine. *Patient Education and Counseling*, 98(5):578–587. [139](#)

- Lickley, R. (2017). Disfluency in typical and stuttered speech. *Fattori sociali e biologici nella variazione fonetica-Social and biological factors in speech variation*. 77, 107
- Lindenberger, U. and Baltes, P. B. (1994). Sensory functioning and intelligence in old age: a strong connection. *Psychology and aging*, 9(3):339. 11
- Livingston, E. H. (2004). The mean and standard deviation: what does it all mean? *Journal of Surgical Research*, 119(2):117–123. 77
- Lofgren, M. and Hinzen, W. (2022). Breaking the flow of thought: Increase of empty pauses in the connected speech of people with mild and moderate alzheimer’s disease. *Journal of Communication Disorders*, 97:106214. 96
- López, O. L. and DeKosky, S. T. (2008). Clinical symptoms in alzheimer’s disease. *Handbook of clinical neurology*, 89:207–216. 15
- Lou, P. J. and Johnson, M. (2020). End-to-end speech recognition and disfluency removal. *arXiv preprint arXiv:2009.10298*. 42
- Luz, S., de la Fuente, S., and Albert, P. (2018). A method for analysis of patient speech in dialogue for dementia detection. *arXiv preprint arXiv:1811.09919*. 24
- Luz, S., Haider, F., de la Fuente, S., Fromm, D., and MacWhinney, B. (2020). Alzheimer’s dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*. 133
- Mahl, G. F. (1959). Measuring the patient’s anxiety during interviews from ”expressive” aspects of his speech. *Transactions of the New York academy of sciences*, 21:249–257. 83
- Margallo-Lana, M., Swann, A., O’Brien, J., Fairbairn, A., Reichelt, K., Potkins, D., Mynt, P., and Ballard, C. (2001). Prevalence and pharmacological management of behavioural and psychological symptoms amongst dementia sufferers living in care environments. *International journal of geriatric psychiatry*, 16(1):39–44. 16

- Martin, R. and Kolossa, D. (2012). Voice activity detection, noise estimation, and adaptive filters for acoustic signal enhancement. *Techniques for Noise Robustness in Automatic Speech Recognition*, pages 51–85. [109](#)
- Martínez-Nicolás, I., Llorente, T. E., Ivanova, O., Martínez-Sánchez, F., and Meilán, J. J. (2022). Many changes in speech through aging are actually a consequence of cognitive changes. *International Journal of Environmental Research and Public Health*, 19(4):2137. [31](#)
- Marvel, M. K., Epstein, R. M., Flowers, K., and Beckman, H. B. (1999). Soliciting the patient’s agenda: have we improved? *Jama*, 281(3):283–287. [144](#)
- Matej, R., Tesar, A., and Rusina, R. (2019). Alzheimer’s disease and other neurodegenerative dementias in comorbidity: a clinical and neuropathological overview. *Clinical biochemistry*, 73:26–31. [13](#), [16](#)
- Matías-Guiu, J. A., Pytel, V., Cortés-Martínez, A., Valles-Salgado, M., Rognoni, T., Moreno-Ramos, T., and Matías-Guiu, J. (2018). Conversion between addenbrooke’s cognitive examination iii and mini-mental state examination. *International psychogeriatrics*, 30(8):1227–1233. [84](#)
- McDougall, K. and Duckworth, M. (2017). Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication*, 95:16–27. [63](#), [65](#), [69](#)
- McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack Jr, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., et al. (2011). The diagnosis of dementia due to alzheimer’s disease: Recommendations from the national institute on aging-alzheimer’s association workgroups on diagnostic guidelines for alzheimer’s disease. *Alzheimer’s & dementia*, 7(3):263–269. [20](#)
- Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in alzheimer’s disease: can temporal and acoustic parameters discriminate dementia? *Dementia and geriatric cognitive disorders*, 37(5-6):327–334. [22](#), [24](#)

- Mirheidari, B. (2018). *Detecting early signs of dementia in conversation*. PhD thesis, University of Sheffield. [47](#), [48](#), [49](#), [75](#), [79](#), [108](#)
- Mirheidari, B., Blackburn, D., Harkness, K., Walker, T., Venneri, A., Reuber, M., and Christensen, H. (2017). Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 58(2):373–387. [78](#), [150](#)
- Mirheidari, B., Blackburn, D., Walker, T., Reuber, M., and Christensen, H. (2019). Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79. [78](#)
- Molchan, S., Little, J., Cantillon, M., and Sunderland, T. (1995). Psychosis. In Lawlor, B. A., editor, *Behavioral complications in Alzheimer's disease*. American Psychiatric Pub. [16](#)
- Möller, H.-J. and Graeber, M. B. (1998). The case described by alois alzheimer in 1911: historical and conceptual perspectives based on the clinical record and neurohistological sections. *European archives of psychiatry and clinical neuroscience*, 248:111–122. [22](#)
- Molnar, C. (2022). Interpretable machine learning. [129](#)
- Morris, J. C., Storandt, M., Miller, J. P., McKeel, D. W., Price, J. L., Rubin, E. H., and Berg, L. (2001). Mild cognitive impairment represents early-stage alzheimer disease. *Archives of neurology*, 58(3):397–405. [18](#)
- Morris, R. G. and Mograbi, D. C. (2013). Anosognosia, autobiographical memory and self knowledge in alzheimer's disease. *Cortex*, 49(6):1553–1565. [175](#)
- Moscovitch, M. and Winocur, G. (1992). The neuropsychology of memory and aging. *The handbook of aging and cognition*, 315:372. [12](#)
- Mueller, K. D., Hermann, B., Mecollari, J., and Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and alzheimer's disease: A review of picture description tasks. *Journal of clinical and experimental neuropsychology*, 40(9):917–939. [26](#)

- Müller, N. and Guendouzi, J. A. (2005). Order and disorder in conversation: Encounters with dementia of the alzheimer’s type. *Clinical linguistics & phonetics*, 19(5):393–404. 56, 145
- Murray, L. L. (2010). Distinguishing clinical depression from early alzheimer’s disease in elderly people: Can narrative analysis help? *Aphasiology*, 24(6-8):928–939. 24
- Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699. 25
- Nasreen, S., Rohanian, M., Hough, J., and Purver, M. (2021). Alzheimer’s dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, 3:640669. 31
- Nass, C. (2004). Etiquette equality: exhibitions and expectations of computer politeness. *Communications of the ACM*, 47(4):35–37. 148, 160
- Nass, C. I. and Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge. 147
- Ness, D. E. and Ende, J. (1994). Denial in the medical interview: recognition and management. *Jama*, 272(22):1777–1781. 175
- Niebuhr, O. and Michaud, A. (2015). Speech data acquisition: the underestimated challenge. *KALIPHO-Kieler Arbeiten zur Linguistik und Phonetik*, 3:1–42. 80
- O’Brien, J. and Thomas, A. (2015). Vascular dementia. *The Lancet*, 386(10004):1698–1706. 14
- Office for National Statistics (2022). Saving for retirement in great britain: April 2018 to march 2020. 15
- O’Malley, R. P. D., Mirheidari, B., Harkness, K., Reuber, M., Venneri, A., Walker, T., Christensen, H., and Blackburn, D. (2021). Fully automated cognitive screening tool

- based on assessment of speech and language. *Journal of Neurology, Neurosurgery & Psychiatry*, 92(1):12–15. 78
- Ong, L. M., Visser, M. R., Kruyver, I., Bensing, J. M., Van Den Brink-Muinen, A., Stouthard, J., Lammes, F. B., and De Haes, J. C. (1998). The roter interaction analysis system (rias) in oncological consultations:: psychometric properties. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, 7(5):387–401. 54
- Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for alzheimer’s disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic signal to clinical reality*, pages 78–87. 45
- O’Shaughnessy, D. and Gabrea, M. (2000). Automatic identification of filled pauses in spontaneous speech. In *2000 Canadian Conference on Electrical and Computer Engineering. Conference Proceedings. Navigating to a New Era (Cat. No. 00TH8492)*, volume 2, pages 620–624. IEEE. 109
- Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–36. 31, 184, 188, 189
- O’Malley, R., Morris, L.-A., Longden, C., Turner, A., Walker, T., Venneri, A., Mirheidari, B., Christensen, H., Reuber, M., and Blackburn, D. (2020). 26 can an automated assessment of language help distinguish between functional cognitive disorder and early neurodegeneration? 78, 97
- Pakhomov, S. V., Kaiser, E. A., Boley, D. L., Marino, S. E., Knopman, D. S., and Birnbaum, A. K. (2011). Effects of age and dementia on temporal cycles in spontaneous speech fluency. *Journal of neurolinguistics*, 24(6):619–635. 31, 96
- Pakhomov, S. V., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., and Knopman, D. S. (2010). Computerized analysis of speech and language to

- identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165. 97
- Panesar, K. and de Alba, M. B. P. C. (2023). Natural language processing-driven framework for the early detection of language and cognitive decline. *Language and Health*. 67, 113
- Pastoriza-Dominguez, P., Torre, I. G., Dieguez-Vide, F., Gómez-Ruiz, I., Geladó, S., Bello-López, J., Ávila-Rivera, A., Matias-Guiu, J. A., Pytel, V., and Hernández-Fernández, A. (2022). Speech pause distribution as an early marker for alzheimer's disease. *Speech Communication*, 136:107–117. 96
- Pelikan, H. R. and Broth, M. (2016). Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 4921–4932. 147
- Pennington, C., Hayre, A., Newson, M., and Coulthard, E. (2015). Functional cognitive disorder: a common cause of subjective cognitive symptoms. *Journal of Alzheimer's Disease*, 48(s1):S19–S24. 21
- Perkins, L., Whitworth, A., and Lesser, R. (1998). Conversing in dementia: A conversation analytic approach. *Journal of Neurolinguistics*, 11(1-2):33–53. 56
- Petersen, R. C. (2003). *Mild cognitive impairment: aging to Alzheimer's disease*. Oxford University Press. 20
- Petersen, R. C. (2016). Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, 22(2 Dementia):404. 18
- Petersen, R. C., Doody, R., Kurz, A., Mohs, R. C., Morris, J. C., Rabins, P. V., Ritchie, K., Rossor, M., Thal, L., and Winblad, B. (2001a). Current concepts in mild cognitive impairment. *Archives of neurology*, 58(12):1985–1992. 20
- Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of neurology*, 56(3):303–308. 20

- Petersen, R. C., Stevens, J. C., Ganguli, M., Tangalos, E. G., Cummings, J. L., and DeKosky, S. T. (2001b). Practice parameter: Early detection of dementia: Mild cognitive impairment (an evidence-based review)[retired]: Report of the quality standards subcommittee of the american academy of neurology. *Neurology*, 56(9):1133–1142. 20
- Petti, U., Baker, S., and 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. 43
- Pinto, J. C. B. R., Schiefer, A. M., and de Ávila, C. R. B. (2013). Disfluencies and speech rate in spontaneous production and in oral reading in people who stutter and who do not stutter. *Audiology-Communication Research*, 18(2):63–70. 77
- Plug, L., Sharrack, B., and Reuber, M. (2009). Conversation analysis can help to distinguish between epilepsy and non-epileptic seizure disorders: a case comparison. *Seizure*, 18(1):43–50. 148
- Pomerantz, A. and Heritage, J. (2013). Preference. In Sidnell, J. and Stivers, T., editors, *The handbook of conversation analysis*, pages 210–228. Blackwell Publishing. 52
- Pope, B., Blass, T., Siegman, A. W., and Rahe, J. (1970). Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128. 80
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society. 34
- Proença, J., Veiga, A., Candeias, S., Lemos, J., Januário, C., and Perdigão, F. (2014). Characterizing parkinson’s disease speech by acoustic and phonetic features. In *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014. Proceedings 11*, pages 24–35. Springer. 80

- Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., and Smékal, Z. (2020). Alzheimer's disease and automatic speech analysis: a review. *Expert systems with applications*, 150:113213. 36
- Rao, S., Lane, I., and Schultz, T. (2007). Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. In *Proceedings of Machine Translation Summit XI: Papers*. 42
- Reisberg, B., Ferris, S. H., de Leon, M. J., and Crook, T. (1982). The global deterioration scale for assessment of primary degenerative dementia. *The American journal of psychiatry*. 18, 19, 20
- Rieger, C. L. (2003). Disfluencies and hesitation strategies in oral l2 tests. In *ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*. 80
- Ripich, D. N., Carpenter, B. D., and Ziol, E. W. (2000). Conversational cohesion patterns in men and women with alzheimer's disease: a longitudinal study. *International journal of language & communication disorders*, 35(1):49–64. 104
- Ripich, D. N. and Terrell, B. Y. (1988). Patterns of discourse cohesion and coherence in alzheimer's disease. *Journal of Speech and Hearing Disorders*, 53(1):8–15. 124
- Roalf, D. R., Moberg, P. J., Xie, S. X., Wolk, D. A., Moelter, S. T., and Arnold, S. E. (2013). Comparative accuracies of two common screening instruments for classification of alzheimer's disease, mild cognitive impairment, and healthy aging. *Alzheimer's & Dementia*, 9(5):529–537. 26
- Roberts, P. M., Meltzer, A., and Wilding, J. (2009). Disfluencies in non-stuttering adults across sample lengths and topics. *Journal of communication disorders*, 42(6):414–427. 29
- Robin, J., Harrison, J. E., Kaufman, L. D., Rudzicz, F., Simpson, W., and Yancheva, M. (2020). Evaluation of speech-based digital biomarkers: review and recommendations. *Digital Biomarkers*, 4(3):99–108. 17

- Robinson, J. D. (2006). Soliciting patients' presenting concerns. *Studies in Interactional Sociolinguistics*, 20:22. 140, 141
- Roddy, M. and Harte, N. (2020). Neural generation of dialogue response timings. *arXiv preprint arXiv:2005.09128*. 55
- Rohanian, M., Hough, J., and Purver, M. (2021). Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv preprint arXiv:2106.15684*. 61, 96
- Roter, D. and Larson, S. (2002). The roter interaction analysis system (rias): utility and flexibility for analysis of medical interactions. *Patient education and counseling*, 46(4):243–251. 53
- Rousseaux, M., Sève, A., Vallet, M., Pasquier, F., and Mackowiak-Cordoliani, M. A. (2010). An analysis of communication in conversation in patients with dementia. *Neuropsychologia*, 48(13):3884–3890. 145
- Royal College of Psychiatrists (2022). National audit of dementia: Memory assessment services spotlight audit 2021. <https://tinyurl.com/3ajxtysc>. Accessed: 12/04/2023. 16
- Sachs-Ericsson, N. and Blazer, D. G. (2015). The new dsm-5 diagnosis of mild neurocognitive disorder and its relation to research in mild cognitive impairment. *Aging & mental health*, 19(1):2–12. 18
- Sacks, H. (1973). The preference for agreement in natural conversation. *Linguistic Institute, Ann Arbor, Michigan*. 52
- Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organisation of turn-taking for conversation. *Language*, 50(1):696–735. 49, 50
- Salimi, S., Irish, M., Foxe, D., Hodges, J. R., Piguet, O., and Burrell, J. R. (2018). Can visuospatial measures improve the diagnosis of alzheimer's disease? *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:66–74. 16

- Salthouse, T. A. (1996). General and specific speed mediation of adult age differences in memory. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 51(1):P30–P42. 11
- Salthouse, T. A. (2019). Trajectories of normal cognitive aging. *Psychology and aging*, 34(1):17. 11
- Sánchez-Marño, N., Alonso-Betanzos, A., and Tombilla-Sanromán, M. (2007). Filter methods for feature selection—a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 178–187. Springer. 42
- Sandvik, M., Eide, H., Lind, M., Graugaard, P. K., Torper, J., and Finset, A. (2002). Analyzing medical dialogues: strength and weakness of roter’s interaction analysis system (rias). *Patient education and counseling*, 46(4):235–241. 55
- Santos, V. D., Thomann, P. A., Wüstenberg, T., Seidl, U., Essig, M., and Schröder, J. (2011). Morphological cerebral correlates of cerad test performance in mild cognitive impairment and alzheimer’s disease. *Journal of Alzheimer’s Disease*, 23(3):411–420. 22
- Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382. 51, 56
- Scheltens, P., Blennow, K., Breteler, M. M., De Strooper, B., Frisoni, G. B., Salloway, S., and Van der Flier, W. M. (2016). Alzheimer’s disease. *The Lancet*, 388(10043):505–517. 17
- Schleef, E. and Meyerhoff, M. (2010). Sociolinguistic methods for data collection and interpretation. In Schleef, E. and Meyerhoff, M., editors, *The Routledge sociolinguistics reader*, pages 1–26. Routledge. 193
- Schmidtke, K., Pohlmann, S., and Metternich, B. (2008). The syndrome of functional memory disorder: definition, etiology, and natural course. *The American Journal of Geriatric Psychiatry*, 16(12):981–988. 21

- Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*. 37
- Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F. A., Mohamed, S., Piah, H. A., and Zahari, N. I. (2012). Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*, pages 188–193. IEEE. 55
- Shao, Z., Janse, E., Visser, K., and Meyer, A. S. (2014). What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in psychology*, 5:89695. 27
- Shardlow, M. (2016). An analysis of feature selection techniques. *The University of Manchester*, 1(2016):1–7. 42
- Shattuck-Hufnagel, S. and Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, 18(1):41–55. 65
- Shriberg, E., Bates, R., and Stolcke, A. (1997). A prosody only decision-tree model for disfluency detection. In *Fifth European Conference on Speech Communication and Technology*. 110
- Singh, S., Bucks, R. S., and Cuerden, J. M. (2001). Evaluation of an objective technique for analysing temporal variables in dat spontaneous speech. *Aphasiology*, 15(6):571–583. 86, 96
- Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178. 146
- Sluis, R. A., Angus, D., Wiles, J., Back, A., Gibson, T., Liddle, J., Worthy, P., Copland, D., and Angwin, A. J. (2020). An automated approach to examining pausing in the

- speech of people with dementia. *American Journal of Alzheimer's Disease & Other Dementias*(®), 35:1533317520939773. 95
- Small, J. A., Kemper, S., and Lyons, K. (1997). Sentence comprehension in alzheimer's disease: effects of grammatical complexity, speech rate, and repetition. *Psychology and Aging*, 12(1):3. 22
- Smith, T., Gildeh, N., and Holmes, C. (2007). The montreal cognitive assessment: validity and utility in a memory clinic setting. *The Canadian Journal of Psychiatry*, 52(5):329–332. 26
- Smith, V. L. and Clark, H. H. (1993). On the course of answering questions. *Journal of memory and language*, 32(1):25–38. 129
- Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., and Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and alzheimer's disease in late life: Findings from the nun study. *Jama*, 275(7):528–532. 22
- Stanyon, M. R., Griffiths, A., Thomas, S. A., and Gordon, A. L. (2016). The facilitators of communication with people with dementia in a care setting: an interview study with healthcare workers. *Age and ageing*, 45(1):164–170. 22
- Stasak, B., Epps, J., Schatten, H. T., Miller, I. W., Provost, E. M., and Armev, M. F. (2021). Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Communication*, 132:10–20. 65, 66, 67, 68, 70
- Stemberger, J. P. (1982). *The lexicon in a model of language production*. University of California, San Diego. 65
- Stephan, B. C. M., Minett, T., Pagett, E., Siervo, M., Brayne, C., and McKeith, I. G. (2013). Diagnosing mild cognitive impairment (mci) in clinical trials: a systematic review. *BMJ open*, 3(2):e001909. 20
- Stivers, T. (2002). Presenting the problem in pediatric encounters:” symptoms only” versus” candidate diagnosis” presentations. *Health Communication*, 14(3):299–338. 139

- Stivers, T. (2013). Sequence organisation. In Sidnell, J. and Stivers, T., editors, *The handbook of conversation analysis*, pages 191–209. Blackwell Publishing. 51
- Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. (2009). Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592. 55
- Taler, V. and 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. 22
- Tang, L., Zhang, Z., Feng, F., Yang, L.-Z., and Li, H. (2023). Explainable alzheimer’s disease detection using linguistic features from automatic speech recognition. *Dementia and Geriatric Cognitive Disorders*, 52(4):240–248. 108, 109
- Tang-Wai, D. F. and Graham, N. L. (2008). Assessment of language function in dementia. *Geriatrics*, 11(2):103–110. 22
- Tauroza, S. and Allison, D. (1990). Speech rates in british english. *Applied linguistics*, 11(1):90–105. 112
- Teas-Gill, V. and Roberts, F. (2012). Conversation analysis in medicine. In *The handbook of conversation analysis*, pages 575–592. Wiley Online Library. 138
- Temple, L. (2000). Second language learner speech production. *Studia linguistica*, 54(2):288–297. 80
- Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *Plos one*, 15(7):e0236009. 40
- Therapy Box Ltd (2024). Therapy box - digital health solutions. <https://www.therapybox.co.uk/>. Accessed: 2024-10-03. 147

- Thomas, C., Keselj, V., Cercone, N., Rockwood, K., and Asp, E. (2005). Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *IEEE International conference mechatronics and automation, 2005*, volume 3, pages 1569–1574. IEEE. 40
- Thomas, M., Hollands, S., Blackburn, D., and Christensen, H. (2023). Towards disfluency features for speech technology based automatic dementia classification. In *Proceedings of the 20th International Congress of Phonetic Sciences*, volume 20, pages 3902–3906. 36, 102, 129
- Thymia Ltd (2024). Thymia - making mental health visible. <https://thymia.ai/>. Accessed: 2024-08-03. 147
- Tombaugh, T. N. and McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *Journal of the American Geriatrics Society*, 40(9):922–935. 25
- Varela-Suárez, A. (2018). The question-answer adjacency pair in dementia discourse. *International Journal of Applied Linguistics*, 28(1):86–101. 51
- Vidović, M., Sinanović, O., Šabaškić, L., Hatičić, A., and Brkić, E. (2011). Incidence and types of speech disorders in stroke patients. *Acta Clinica Croatica*, 50(4):491–493. 80
- Vincze, V., Gosztolya, G., Tóth, L., Hoffmann, I., Szatlóczki, G., Bánréti, Z., Pákáski, M., and Kálmán, J. (2016). Detecting mild cognitive impairment by exploiting linguistic information from transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–187. 41
- Vincze, V., Szatlóczki, G., Tóth, L., Gosztolya, G., Pákáski, M., Hoffmann, I., and Kálmán, J. (2021). Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild alzheimer’s disease. *Clinical Linguistics & Phonetics*, 35(8):727–742. 22, 189
- Vujović, Ž. et al. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606. 46, 47

- Walker, G., Morris, L.-A., Christensen, H., Mirheidari, B., Reuber, M., and Blackburn, D. J. (2021). Characterising spoken responses to an intelligent virtual agent by persons with mild cognitive impairment. *Clinical Linguistics & Phonetics*, 35(3):237–252. 78
- Walker, G., Walker, T., O'Malley, R., Mirheidari, B., Christensen, H., Reuber, M., and Blackburn, D. (2023). Features of answers to questions about recent events by people with mild cognitive impairment and alzheimer's disease, and healthy controls. *Journal of Interactional Research in Communication Disorders*. 123
- Walker, Z., Possin, K. L., Boeve, B. F., and Aarsland, D. (2015). Lewy body dementias. *The Lancet*, 386(10004):1683–1697. 14
- Wang, D., Wang, X., and Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018. 34
- Warnita, T., Inoue, N., and Shinoda, K. (2018). Detecting alzheimer's disease using gated convolutional neural network from audio data. *arXiv preprint arXiv:1803.11344*. 41
- Watanabe, H., Ikeda, M., and Mori, E. (2020). Primary progressive aphasia as a prodromal state of dementia with lewy bodies: A case report. *Frontiers in Neurology*, 11:510531. 23
- Weiner, J., Engelbart, M., and Schultz, T. (2017). Manual and automatic transcriptions in dementia detection from speech. In *Interspeech*, pages 3117–3121. 32
- Welsh-Bohmer, K. A. and Warren, L. H. (2006). Neurodegenerative dementias. *Geriatric Neuropsychology: Assessment and Intervention*, pages 56–88. 14
- Weninger, F., Eyben, F., Mortillaro, M., and Scherer, K. R. (2013). On the acoustics of emotion in audio: what speech, music, and sound have in common. *Frontiers in psychology*, 4:51547. 37
- Whitworth, A. (2003). The application of conversation analysis (ca) to the management of aphasia. *Travaux neuchâtelois de linguistique*, 7(38-39):63–76. 51
- WHO (2023). Dementia. 13

- Wilson, R. S., Beckett, L. A., Barnes, L. L., Schneider, J. A., Bach, J., Evans, D. A., and Bennett, D. A. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychology and aging*, 17(2):179. 11
- Wilson, R. S., Sytsma, J., Barnes, L. L., and Boyle, P. A. (2016). Anosognosia in dementia. *Current neurology and neuroscience reports*, 16:1–6. 175
- Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., Nordberg, A., Bäckman, L., Albert, M., Almkvist, O., et al. (2004). Mild cognitive impairment—beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *Journal of internal medicine*, 256(3):240–246. 18
- Wutzler, A., Becker, R., Lämmle, G., Haverkamp, W., and Steinhagen-Thiessen, E. (2013). The anticipatory proportion as an indicator of language impairment in early-stage cognitive disorder in the elderly. *Dementia and geriatric cognitive disorders*, 36(5-6):300–309. 22
- Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing. 42
- Yuan, J., Cai, X., Bian, Y., Ye, Z., and Church, K. (2021). Pauses for detection of alzheimer’s disease. *Frontiers in Computer Science*, 2:624488. 95, 189
- Yunusova, Y., Graham, N. L., Shellikeri, S., Phuong, K., Kulkarni, M., Rochon, E., Tang-Wai, D. F., Chow, T. W., Black, S. E., Zinman, L. H., et al. (2016). Profiling speech and pausing in amyotrophic lateral sclerosis (als) and frontotemporal dementia (ftd). *PloS one*, 11(1):e0147573. 96
- Zhu, Y., Tran, B., Liang, X., Batsis, J. A., and Roth, R. M. (2022). Towards interpretability of speech pause in dementia detection using adversarial learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6462–6466. IEEE. 96

- Zolnoori, M., Zolnour, A., and Topaz, M. (2023). Adscreen: A speech processing-based screening system for automatic identification of patients with alzheimer’s disease and related dementia. *Artificial Intelligence in Medicine*, 143:102624. [45](#)
- Zwarts, S. and Johnson, M. (2011). The impact of language models and loss functions on repair disfluency detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 703–711. [41](#)