Automatic Detection of Dementia and related Affective Disorders through Processing of Speech and Language

Dissertation zur Erlangung des Grades

des Doktors der Ingenieurwissenschaften (Dr.-Ing.)

der Fakultät für Mathematik und Informatik

der Universität des Saarlandes

vorgelegt von

Nicklas Maximilian Linz, M.Sc.

Saarbrücken, 2022

UNIVERSITÄT
DES
SAARLANDES

Date of the colloquium:	24.03.2023
Dean:	Professor Dr. Jürgen Steimle
Chairman of the examination board:	Professor Dr. Sven Apel
Reporter:	Professor Dr. Antonio Krüger
	Professor Dr. Josef van Genabith
Scientific Assistant:	Dr. Andrey Girenko

Abstract

In 2019, dementia is has become a *trillion dollar disorder*. Alzheimer's disease (AD) is a type of dementia in which the main observable symptom is a decline in cognitive functions, notably memory, as well as language and problem-solving. Experts agree that early detection is crucial to effectively develop and apply interventions and treatments, underlining the need for effective and pervasive assessment and screening tools. The goal of this thesis is to explores how computational techniques can be used to process speech and language samples produced by patients suffering from dementia or related affective disorders, to the end of automatically detecting them in large populations using machine learning models. A strong focus is laid on the detection of early stage dementia (MCI), as most clinical trials today focus on intervention at this level. To this end, novel automatic and semi-automatic analysis schemes for a speech-based cognitive task, i.e., verbal fluency, are explored and evaluated to be an appropriate screening task. Due to a lack of available patient data in most languages, world-first multilingual approaches to detecting dementia are introduced in this thesis. Results are encouraging and clear benefits on a small French dataset become visible. Lastly, the task of detecting these people with dementia who also suffer from an affective disorder called apathy is explored. Since they are more likely to convert into later stage of dementia faster, it is crucial to identify them. These are the fist experiments that consider this task using solely speech and language as inputs. Results are again encouraging, both using only speech or language data elicited using emotional questions. Overall, strong results encourage further research in establishing speech-based biomarkers for early detection and monitoring of these disorders to better patients' lives.

Zusammenfassung

Im Jahr 2019 ist Demenz zu einer Billionen-Dollar-Krankheit geworden. Die Alzheimer-Krankheit (AD) ist eine Form der Demenz, bei der das Hauptsymptom eine Abnahme der kognitiven Funktionen ist, insbesondere des Gedächtnisses sowie der Sprache und des Problemlösungsvermögens. Experten sind sich einig, dass eine frühzeitige Erkennung entscheidend für die effektive Entwicklung und Anwendung von Interventionen und Behandlungen ist, was den Bedarf an effektiven und durchgängigen Bewertungs- und Screening-Tools unterstreicht. Das Ziel dieser Arbeit ist es zu erforschen, wie computergestützte Techniken eingesetzt werden können, um Sprach- und Sprechproben von Patienten, die an Demenz oder verwandten affektiven Störungen leiden, zu verarbeiten, mit dem Ziel, diese in großen Populationen mit Hilfe von maschinellen Lernmodellen automatisch zu erkennen. Ein starker Fokus liegt auf der Erkennung von Demenz im Frühstadium (MCI), da sich die meisten klinischen Studien heute auf eine Intervention auf dieser Ebene konzentrieren. Zu diesem Zweck werden neuartige automatische und halbautomatische Analyseschemata für eine sprachbasierte kognitive Aufgabe, d.h. die verbale Geläufigkeit, erforscht und als geeignete Screening-Aufgabe bewertet. Aufgrund des Mangels an verfügbaren Patientendaten in den meisten Sprachen werden in dieser Arbeit weltweit erstmalig mehrsprachige Ansätze zur Erkennung von Demenz vorgestellt. Die Ergebnisse sind ermutigend und es werden deutliche Vorteile an einem kleinen französischen Datensatz sichtbar. Schließlich wird die Aufgabe untersucht, jene Menschen mit Demenz zu erkennen, die auch an einer affektiven Störung namens Apathie leiden. Da sie mit größerer Wahrscheinlichkeit schneller in ein späteres Stadium der Demenz übergehen, ist es entscheidend, sie zu identifizieren. Dies sind die ersten Experimente, die diese Aufgabe unter ausschließlicher Verwendung von Sprache und Sprache als Input betrachten. Die Ergebnisse sind wieder ermutigend, sowohl bei der Verwendung von reiner Sprache als auch bei der Verwendung von Sprachdaten, die durch emotionale Fragen ausgelöst werden. Insgesamt sind die Ergebnisse sehr ermutigend und ermutigen zu weiterer Forschung, um sprachbasierte Biomarker für die Früherkennung und Überwachung dieser Erkrankungen zu etablieren und so das Leben der Patienten zu verbessern.

Acknowledgments

First of all, I would like thank my supervisors, Prof. Dr. Antonio Krüger and Prof. Dr. Joseph van Genabith for the great support and advise they have provided me in compiling this thesis over the years.

I would like to thank my former colleagues:

- Dr. Jan Alexandersson, who has been advising me on most of the research presented in this thesis. Although we might not have always seen eye-to-eye on all issues, you have always believed in me and thereby afforded me many opportunities to outgrow myself.
- Dr. Andrey Girenko, who was always there to give good advise when needed. Over the years, I learned a lot of important lessons from you–not only academically.
- Dr. Christer Samuelsson, who was my first advisor at DFKI and tragically left us before his time. I most likely would not have become a researcher with out him!
- Dr. Johannes Tröger, who always helped bring structure intro my chaos. My life is better when you are there!
- Dr. Alexandra König, who was the clinical partner for much of this work. Thank you for always reminding me about the patient perspective and showing me that research colleagues can grow into friends.
- Prof. Dr. Philippe Robert, who kindly helped me along the journey of becoming a researcher with an always positive attitude. Just keep swimming, Philippe!

... and all of my other previous colleagues from DFKI and ki elements who allowed me to write this thesis by having my back!

Finally, I want to thank my family and friends, who dealt with me during periods of extreme work, always lending a hand when necessary, and my husband Felix and our dog Finn, who were always there for me!

Contents

Part I Introduction and Background **1**

1 Introduction **3**

1.1 Aim and Research Questions . . . 7

1.2 Results and Contributions . . . 8

1.3 Thesis Structure . . . 10

2 Foundations and Related Work **13**

2.1 Dementia . . . 13

2.1.1 Etiology . . . 13

2.1.2 Assessment methods . . . 16

2.2 Apathy, an affective syndrome . . . 18

2.3 Cognitive Testing . . . 21

2.3.1 Verbal Fluency . . . 24

2.3.2 Cookie Theft Picture Description . . . 29

2.3.3 Mini-Mental-State Examination . . . 30

2.3.4 Clinical Dementia Rating Scale . . . 31

2.3.5 MMSE and CDR as Assessment Tools . . . 32

2.3.6 Computerized cognitive testing . . . 33

2.4 Machine Learning Prerequisites . . . 35

2.4.1 Classification Performance Metrics . . . 35

2.4.2 Evaluation Methods . . . 39

2.4.3 Classification Models . . . 40

2.5 Related Work . . . 42

2.5.1 Computational approaches: New features . . . 44

2.5.2 Automated analysis pipelines . . . 46

2.5.3 Computational Analysis of SVF . . . 48

2.5.4 Language features in CTP 50
2.6 Language in Dementia 53
2.6.1 Traditional linguistic exams: AD 55
2.6.2 AD: Classic performance measures and test paradigms 55
2.6.3 The task itself matters: Reconciling conflicting results 57

Part II Automatic Detection of MCI from Speech and Language 59

3 Detection of Dementia from manual Verbal Fluency Transcripts 61
3.1 Neural Word Embeddings in analysing Semantic Verbal Fluency 61
3.1.1 Background 62
3.1.2 Methodology and Results 64
3.1.3 Discussion 66
3.2 Language Modelling in Semantic Verbal Fluency 67
3.2.1 Background 68
3.2.2 Language Modelling 70
3.2.3 Methods 73
3.2.4 Results 75
3.2.5 Discussion 76
3.3 Temporal Analysis of Semantic Verbal Fluency 77
3.3.1 Background 78
3.3.2 Methods 79
3.3.3 Results 82
3.3.4 Discussion 90
3.4 Predicting Dementia Screening and Staging Scores From Semantic Verbal Fluency 91
3.4.1 Background 93
3.4.2 Methods 95
3.4.3 Experiments 99
3.4.4 Discussion 104
3.5 Summary 105

4 Using Automated Speech Recognition to detect Dementia 107

4.1 Fully Automatic Speech-Based Assessment of Semantic Verbal Fluency 109
4.1.1 Methods 109
4.1.2 Results 113
4.1.3 Discussion 115
4.2 Telephone-based Dementia Screening through Automatic Analysis of Semantic Verbal Fluency 118
4.2.1 Background 118
4.2.2 Methods 122
4.2.3 Results 126
4.2.4 Discussion 127
4.3 Summary 128

Part III Multi- and Crosslingual Methods for Dementia Detection 131

5 Multi- and Crosslingual Methods for the Detection of Dementia from Picture Description Tasks 133
5.1 Cross-lingual Detection of Early Dementia from Speech 135
5.1.1 Background 135
5.1.2 Methodology 138
5.1.3 Results 143
5.1.4 Discussion 146
5.2 Class-based Language Modelling for cross-linguistic detection of Alzheimer's Disease 148
5.2.1 Background 148
5.2.2 Methodology 148
5.2.3 Results 152
5.2.4 Discussion 158
5.3 Summary 160

Part IV Automatic Detection of Clinical Apathy in Dementia Patients from Speech and Language 161

6 Detection of clinical Apathy in Dementia Patients 163

6.1 Speech Features for the Detection and Characterisation of Apathy . . . 163

6.1.1 Methodology . . . 164

6.1.2 Results . . . 169

6.1.3 Discussion . . . 172

6.2 Detecting late-life Apathy in Dementia using Sentiment and Psycholinguistic Analysis of Emotional Language . . . 179

6.2.1 Background . . . 179

6.2.2 Apathy Detection from Manual Transcripts . . . 182

6.2.3 Effects of Automatic Speech Recognition . . . 190

6.2.4 Discussion . . . 194

6.3 Summary . . . 197

7 Conclusions and Future Work 199

7.1 Thesis summary . . . 199

7.2 Contributions . . . 200

7.3 Future Work . . . 202

7.3.1 Standardised data collection . . . 202

7.3.2 Longitudinal data . . . 202

7.3.3 New protocols . . . 202

7.3.4 Differentiating different dementias . . . 203

7.3.5 Detection of Depression . . . 203

7.4 Closing Remarks . . . 203

A Abbreviations and Definitions 205

B Cognitive Domains in DSM-5 209

C Literature Review 213

List of Figures

1.1 Prevalence of different dementia causes. 4

1.2 Progression of cognitive decline in Alzheimer's disease. 5

1.3 Technical pipeline of machine learning experiments 7

2.1 Process of cognitive decline. 14

2.2 Prevalence of apathy in dementia and relation to depression in dementia. 19

2.3 Model of neurocognitive domains according to DSM-5. Each box represents a super-domain of cognitive functions. Subdomains are listed as items inside. 23

2.4 Example performance in a SVF task of category animals 24

2.5 Cookie Theft Picture 29

2.6 Schematic of a confusion matrix 36

2.7 Schematic plot of multiple ROC curves 38

2.8 Schematic representation of *k*-fold cross validation 39

2.9 Plot of a sigmoid function 41

2.10 Example of an optimal hyperplane in a two-dimensional feature space . 42

3.1 Boxplots of perplexity in relation to diagnostic criteria 72

3.2 Schematic of time interval analysis 80

3.3 Word Count, Word Frequency and Transition length by time interval . . 83

3.4 Correlation between 10 second WC intervals and neuropsychological test scores 88

3.5 AUC of classification models plotted against number of features 89

3.6 Histograms of MMSE and CDR-SOB scores. 95

3.7 Visualisation of feature distribution in relation to MMSE (a) and CDR (b). 97

3.8 Confusion matrix heat-map for MMSE and CDR-SOB predictions. . . . 100

3.9 Confusion Matrix heat-map for MMSE categories and CDR-SOB stages. 103

4.1 Confusion matrix for diagnosis based on normative data, automatic WC and manual WC. 115
4.2 ROC for features based on manual transcripts and on automatic transcripts. 126

5.1 Results of uni-, multi- and cross-lingual classification experiments. 153
5.2 AUC as a function of the amount of English data used in the training set 154
5.3 Visualisation of feature weights for uni- and multilingual experiments . 156

6.1 Correlation coefficient between features extracted from vocal tasks and AI subdomains. 171
6.2 ROC of classifiers trained to detect apathy from speech. 175
6.3 Boxplot group comparisons between control and apathy population. . . 188
6.4 Chord diagram showing the shift in sentiment for substitution errors from the manual to automatic transcripts. 193

List of Tables

2.1 Subtests of the MMSE 30

3.1 Hyper parameters, classification results and Pearson correlation coefficient between clustering and chaining-based features. 66

3.2 Demographic data and clinical scores by diagnostic group. 73

3.3 Classification results for different scenarios and models as F_1 scores. . . 75

3.4 Demographic information; the MMSE (Mini Mental State Exam) is a general screening test of cognitive status and has a maximum score of 30. 82

3.5 LMR model examining the effects of time interval, diagnosis, age and education on one of three variables 87

3.6 Demographic data and clinical scores by diagnostic group. 94

3.7 Pearson correlation of MMSE, CDR-SOB and computed features. . . . 98

3.8 Mean Absolute Error (MAE) and 95% confidence intervals for different regression models. 101

3.9 Mean Absolute Error (MAE) of MMSE and CDR-SOB prediction for a SVR model by diagnosis group. 101

3.10 κ for CDR-SOB staging and different cut-off strategies. 103

3.11 κ for MMSE staging and rounding to nearest integer. 104

4.1 Demographic data and clinical scores by diagnostic group. 111

4.2 Spearman correlation of automatically and manually computed features. 114

4.3 Classification results between different diagnostic groups. 114

4.4 Demographic data and clinical scores by diagnostic group. 122

5.1 Demographic data for each of the sub-corpora. 138

5.2 Speech and language features extracted from the Cookie Theft narratives. 142

5.3 AUC score for different classification scenarios and languages 144

5.4 Highly-ranked features in the unilingual and cross-lingual classification experiments 145
5.5 Demographics of participants in multilingual experiment 149
5.6 *LM* and *info* features extracted from the CTP task 150

6.1 Feature definition of acoustic markers. 168
6.2 Demographic data for population by gender and apathy. 170
6.3 Statistical group comparisons between non-apathetic and apathetic group. 174
6.4 Feature weights from L_1 regularised Logistic Regression models 176
6.5 Demographic data for population by gender and apathy. 184
6.6 Classification results for models trained on word count and sentiment features. 187
6.7 Classification results for models trained on LIWC features. 189
6.8 Count of the errors from the WER for intrusions, deletions and substitutions separated by sentiment valence. 192
6.9 Classification results for models trained on word count and sentiment features extracted from ASR transcripts. 194

B.1 Description of cognitive domains and sub-domains 212

List of Equations

3.1 Semantic cluster . . . 65
3.2 Semantic chain . . . 65
3.3 Mean semantic cluster distance . . . 65
3.4 Cluster distance cut-off . . . 65
3.5 Maximum likelihood formulation . . . 70
3.6 Conditional probability based on Markov assumption . . . 70
3.7 n-gram probability . . . 71
3.8 Perplexity of a language model . . . 71
3.9 Mean absolute error . . . 99
3.10 Cohen's κ . . . 102
4.1 Word Error Rate . . . 113

Part I

Introduction and Background

Chapter 1

Introduction

Dementia has a large economic impact on our society. In Europe, dementia results in more costs than cancer and cardiovascular disease combined. Today, 55 million people live with dementia worldwide [132]. Since most dementia causes are not deadly themselves, large proportions of these costs can be attributed to care. This is in parallel to the immeasurable amount of emotional suffering and grief caused to caregivers and families of patients. Following the World Alzheimer Report 2021, dementia became a *trillion dollar disorder* in 2019 [132], which underlines the need for effective and pervasive assessment and screening tools.

Dementia is not a disease but rather an umbrella term employed to describe a group of symptoms originated by a number of diseases and causes. It is generally understood to refer to cognitive impairments that cause an affected person to be unable to live autonomously [255]. Alzheimer's disease (AD) is a type of dementia in which the main observable symptom is a decline in cognitive functions, notably memory, as well as language and problem-solving. AD is caused by plaques and tangles in the brain that stop blood flow to cells and eventually lead to their death [30]. Because of this death of neurons, it is considered a *neurodegenerative disorder*. While AD is the most common organic cause of dementia, there are many other causes, such as vascular disorders, e.g., strokes, brain tumors, traumatic brain injuries, or frontotemporal lobe degeneration (FTLD). Additionally, there are other, mostly reversible, causes for dementia-like symptoms, such as depression and substance abuse (see also Figure 1.1).

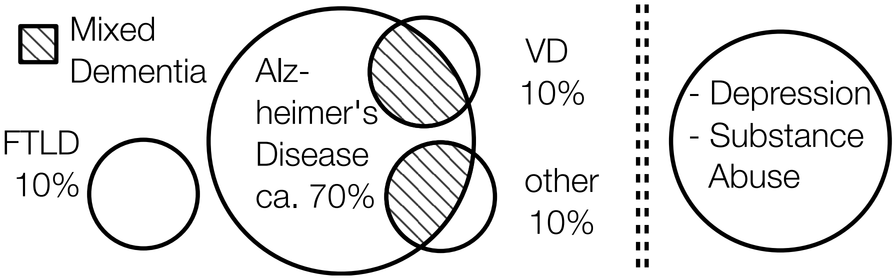


Figure 1.1: The Left panel shows the types of dementia, according to their cause, including Fronto-Temporal Lobar Degeneration, and Vascular Dementia (VD); the dotted areas indicate those cases where more than one cause underlies the disorder. The right panel shows other, mostly reversible, causes for dementia-like symptoms.

Due to the degenerative nature of most dementia causes, experts agree that early diagnosis is vital. Figure 1.2 depicts the progress of Alzheimer's Disease over time. A decline in cognitive ability, and therefore autonomy that is beyond normal ageing is characteristic for AD. In the debate of staging dementia severity, the concept of Mild Cognitive Impairment (MCI) was introduced [100]. It generally describes the early stages of, e.g., AD, where only a minimal cognitive impairment is visible, and most people are still capable of living on their own. As of now, no effective medical treatment to stop the progression of AD has been discovered. There are however measures that allow slowing the progression if applied early [40, 332, 136]. There is a broad agreement, that detection at an MCI or even preclinical stage is key to effective treatment as well as the development of drugs [212].

AD and its predecessor MCI are currently diagnosed and detected through a combination of imaging, biomarker, and functional assessments. Imaging techniques, such as magnet resonance tomography (MRI), are capable of producing visual scans of the brain. Using these scans, different dementia causes are easily identifiable and separable. A clear statement about the affected brain areas can be made from them. Albeit they are highly predictive, they also require expensive equipment (an MRI scanner can easily cost a million euros). This limits their suitability for screening applications. Additionally, the injection of a radioactive tracer would make the repeated screening with these techniques dangerous in itself.

The functional assessment is usually administered by a psychologist using a series of

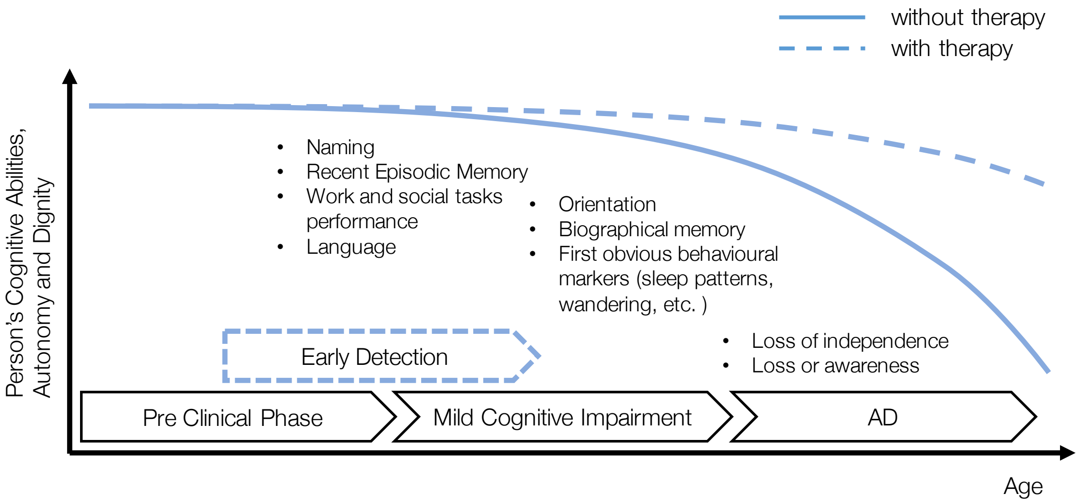

without therapy
with therapy
Person's Cognitive Abilities, Autonomy and Dignity
Naming
Recent Episodic Memory
Work and social tasks performance
Language
Orientation
Biographical memory
First obvious behavioural markers (sleep patterns, wandering, etc.)
Early Detection
Loss of independence
Loss or awareness
Pre Clinical Phase
Mild Cognitive Impairment
AD
Age

Figure 1.2: Progression of cognitive decline in dementia using the example of Alzheimer's disease. Autonomy and dignity are plotted against age. At each stage, common symptoms are presented.

questionnaires and so-called *cognitive tests*. These tests are, often speech-based, short cognitive exercises that allow a clinician to assess a particular part of patient's cognitive functions. According to the Diagnostic Manual for Psychologists (DSM-V) there are six known superdomains of human cognitive ability. A speech-based cognitive test is not only able to evaluate language functions but, e.g., also memory or executive function. In this context, speech can be seen as a window to human cognition [404]. The kind of stimuli speech is being provoked with, play a significant role in both possibilities of predicting a persons' cognitive state as well as in the analysis method that is used to do so. Generally, this can be seen as a trade-off between usefulness for prediction of cognitive functions and ecological validity. Some open speech tasks show high ecologic validity, as they assess everyday skills people need, but do not put enough cognitive load on the assessed to notice the more subtle changes in early-stage dementia (e.g., picture description tasks). Some speech tasks require high cognitive effort from a participant but do not directly relate to critical every-day skills (e.g., verbal fluency tasks).

While a full assessment of cognitive function requires a trained clinician, the increasing prevalence of dementia and milder forms of cognitive impairment warrant large-scale screening of the population, even in high-income countries, as many as 50% of all relevant cases remain undiagnosed [320]. To address this problem, we need new tools that are fast, do not need a laboratory, and can automatically indicate which patients might need to be referred to a specialist [378, 213, 352]. Such tools are highly scalable and can be made accessible to health care professionals with little to no specialized training in old age psychiatry. Ideally, it should be possible to administer them remotely, and they should integrate easily with existing telehealth and telecare solutions for older patients. Automated analysis of speech, in particular speech that is produced during a standard clinical assessment, is a prime candidate for such a tool [203, 340, 168].

Recent advances in the fields of artificial intelligence—particularly in natural language / speech processing, machine learning, as well as speech recognition—have rendered the idea of automatic screening systems possible. Previous authors have proven that the analysis of speech and language can be used to detect dementia, with a particular focus on AD, e.g., from transcribed speech samples produced in picture description tasks [128]. Some authors even report fully automatic systems for dementia detection [374]. However, only a few researchers have explored the question of detecting early stages of AD, such as MCI, from speech samples [125]. Additionally, due to the cost of collecting clinical data, only a few resources of speech recordings from these patient groups exist. This lack of available data is especially challenging for the training of well-performing machine learning models, that could be used to screen for MCI. Multilingual analysis methods that would allow the combination of these scarce resources from different languages are a potential solution but have not been explored so far. Finally, when constructing a real-world screening system based on speech and language, one has to acknowledge the complexity of MCI. Consequently, it has to be considered that about 65% of cases do not only suffer from dementia, but also from an affective disorder, such as depression, caused by dementia [1]. Seeing as these disorders are not only responsible for large proportions of patient's suffering [412, 256], but are also good predictors of conversion from MCI to AD stages [360, 359], their detection is equal in importance. In contrast to speech-based depression detection, which has received a substantial amount of attention in the literature [87], the speech-based detection of clinical apathy in MCI patients has not received any attention.

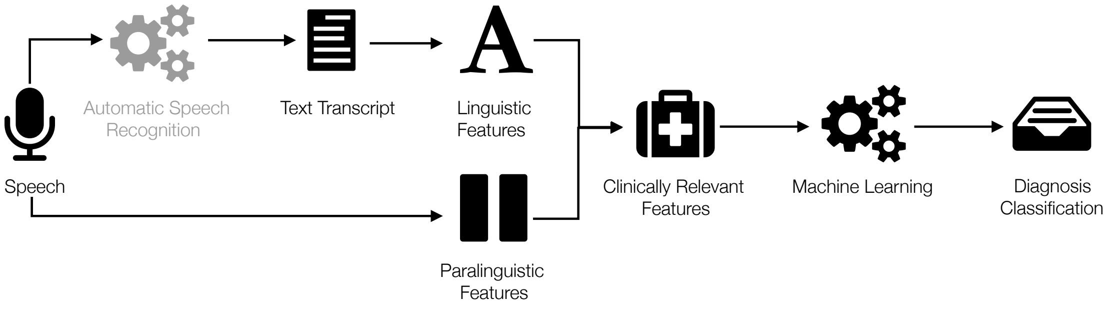


Figure 1.3: Technical pipeline of machine learning experiments. Patient generated speech and language data is fed into linguistic and para-linguistic feature extractions, with the potential of using automatic speech recognition. Clinically relevant features are selected and used to train machine learning models, with the goal of classifying patient populations.

1.1 Aim and Research Questions

The goal of this thesis is to explore how computational linguistic techniques can be used to process speech and language samples produced by patients suffering from dementia or related affective disorders, to the end of automatically detecting them in large populations using machine learning models. Figure 1.3 describes the general architecture of these experiments. Multiple research question are going to be worked on to the end of advancing this field with concrete contributions. The main questions are

1. **Can Mild Cognitive Impairment be automatically detected from concise speech recordings?**
   The construction of automatic diagnostic models for MCI from speech samples will be investigated. The presented approach will focus on recordings of verbal fluency tasks, in which patients are asked to name as many words according to a given rule as possible in a given time frame (e.g., as many animals as possible in 60 seconds). Clinical performance in these test is usually assessed as the number of correct words and has been shown to be highly predictive for MCI. We will introduce and validate novel and extended automatic analysis methods and show that they improve the diagnostic ability of these tasks. Both the functionality to detect and stage dementia will be explored. In addition to validation experiments on manual transcripts, fully automatic experiments using automatic

speech recognition will be carried out. These will be used to validate the utility of such analysis and assessment methods for real-world broad population screening applications—i.e., over the telephone.

2. **How can data resources from different languages be leveraged in multi- and cross-lingual dementia detection?**
   Multilingual analysis methods that allow increasing the productivity of models in under-resourced languages will be explored. The most widely available speech data in most language are picture descriptions of the Boston Cookie Theft Picture. The use of English data to improve productivity in other languages will be explored. In addition to general domain adaptation methods, novel multi-lingual analysis methods are used.

3. **Can affective disorders in dementia be automatically detected based on speech recordings?**
   Methods for the detection of clinical apathy in dementia patients from speech and language will be explored. To this end, a subpopulation of cognitively matched patients telling positive and negative stories will be analysed. Speech and signal processing, as well as sentiment and psycholinguistic language analysis, will be considered as a diagnostic marker and validated using machine learning.

1.2 Results and Contributions

This thesis explores how computational linguistic techniques can be used to process speech and language samples produced by patients suffering from dementia or related affective disorders, to the end of automatically detecting them in large populations using machine learning models. A strong focus is laid on the detection of early stage dementia i.e., MCI, as most clinical trials today focus on intervention at this level. To this end, novel automatic and semi-automatic analysis schemes for Verbal Fluency tasks are explored and evaluated to be an appropriate screening task. Furthermore, world-first multilingual approaches to detecting dementia are introduced in this thesis. This is mainly motivated due to the sparsity of data in most languages and a wish to combine this resources. Results are encouraging and clear benefits on a small French dataset become visible. Lastly, the task of detecting these people with dementia who also suffer from an affective disorder called apathy is explored. Since they are more likely to convert into later stage of dementia faster, it is crucial to identify them. These are the fist experiments that consider this task using solely speech and language as inputs. Results are again encouraging, both using only speech or language data elicited using emotional questions.

The following publications resulted directly from work at this dissertation.

[223] N. Linz, J. Tröger, J. Alexandersson, and A. König. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, 2017 Chapter 3

[222] N. Linz and J. Tröger. Language modelling for the clinical semantic verbal fluency task. In D. Kokkinakis, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA) Chapter 3

[220] N. Linz, K. L. Fors, H. Lindsay, M. Eckerström, J. Alexandersson, and D. Kokkinakis. Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: Reconciling Outcomes. Computational Linguistics and Clinical Psychology Workshop (CLPsych-2019), 6th, located at 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), June 6, Minneapolis,, MN, USA*. o.A., 2019 Chapter 3

[224] N. Linz, J. Tröger, J. Alexandersson, M. Wolters, A. König, and P. Robert. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728. IEEE, 2017 Chapter 3

[378] J. Tröger, N. Linz, A. König, P. Robert, J. Alexandersson, J. Peter, and J. Kray. Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131:53 – 61, 2019 Chapter 3

[201] A. König, N. Linz, J. Tröger, M. Wolters, J. Alexandersson, and P. Robert. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders*, 45(3-4):198–209, 2018 — Chapter 4

[377] J. Tröger, N. Linz, A. König, P. Robert, and J. Alexandersson. Telephone-based dementia screening I: Automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. International ICST Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health-2018), May 21-24, New York, USA*. ACM, 2018 — Chapter 4

[123] K. C. Fraser, N. Linz, B. Li, K. L. Fors, F. Rudzicz, A. König, J. Alexandersson, P. Robert, and D. Kokkinakis. Multilingual prediction of alzheimer's disease through domain adaptation and concept-based language modelling. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2019), June 2-7, Minneapolis,, Minnesota, USA*. o.A., 2019 — Chapter 5

[202] A. König, N. Linz, R. Zeghari, X. Klinge, J. Tröger, J. Alexandersson, and P. Robert. Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *Journal of Alzheimer's Disease*, 69(4), 2019 — Chapter 6

[221] N. Linz, X. Klinge, J. Tröger, J. Alexandersson, R. Zeghari, P. Robert, and A. König. Automatic detection of apathy using acoustic markers extracted from free emotional speech. In *Proceedings of the 2nd Workshop on AI for Ageing, Rehabilitation and Independent Assisted Living (ARIAL)*, pages 17–21, 2018 — Chapter 6

1.3 Thesis Structure

Below, the contents and goals of each chapter are outlined:

Chapter 2: Foundations

In this chapter, a general introduction to the functional assessment of dementia and its causes is given. A special focus is laid on the introduction of cognitive testing. Furthermore, machine learning techniques are introduced. The chapter also contains a list of abbreviations and definitions used throughout the thesis.

Chapter 3: Detection of Dementia from manual Verbal Fluency Transcripts

This chapter starts with an introduction to the clinical application of verbal fluency tasks, with a particular focus on their neuropsychological properties and associated brain regions. A review of clinical literature is presented to examine the state-of-the-art in analysis methods. Afterward, novel analysis methods for both the semantic and phonemic version of verbal fluency are introduced, and their predictability for cognitive impairment is examined in classification experiments. First, language modelling and neural word embedding techniques are both explored as analysis methods on manually created transcripts. Novel features extracted from manual verbal fluency task transcripts are used to predict clinically relevant dementia screening and staging scores.

Chapter 4: Detection of Dementia from automated Verbal Fluency Transcripts

In this chapter, analysis methods introduced and evaluated in Chapter 3 are transferred to two real-time applications using automated speech recognition (ASR). The effects of errors introduced by ASR as a statistical system are examined in a clinical use case, where automated analysis can be used either as part of a regular assessment to save time and generate additional features. To enable potential remote screening applications, the performance of novel markers when extracted from automatically created transcripts from telephone-quality audio is considered.

Chapter 5: Multi- and Cross-lingual Methods for the Detection of Dementia from Picture Description tasks

This chapter considers multi- (or even cross-)lingual solutions to the problem of data-sparsity observed in dementia detection. Multiple experiments are carried out with a number of datasets, analysis methods, and languages. First, features mentioned in previous work are extracted from transcripts of picture description tasks, from two small

French, Swedish and English corpora. Classifiers are trained multilingually using domain adaptation to separate early-stage dementia from controls, with the goal of improving performance on each single data set. Second, we introduce novel features for the multilingual analysis of picture description tasks, based on concept-language models. Their performance is evaluated by augmenting a small French with a larger English Alzheimer dataset and training classifiers multi- and cross-lingually. Lastly, multi- and cross-lingual experiments are performed on Swedish, French and German transcripts of verbal fluency performances from early-stage dementia patients.

Chapter 6: Detection of clinical Apathy in Dementia Patients

This chapter starts with an introduction to clinical apathy observed in dementia patients, its consequences and relation to depression. Recordings of answers to positive and negative questions are used to provoke speech from dementia patients. The chapter first investigates the detection of apathy from para-linguistic analysis directly on the speech-signal of these recordings and later examines the feasibility of detecting clinical apathy in dementia patients by psycholinguistic and sentiment analysis of manual as well as automatic transcripts.

Chapter 7: Conclusions and Future Work

This chapter sums up the main contribution of all previous chapters. It discusses and concludes their implications and potential applications in different clinical contexts. Future directions of clinical speech and language analysis are discussed.

Chapter 2

Foundations and Related Work

This chapter introduces the reader to concepts from both clinical psychology and computer science. Accordingly, not all sections might be relevant to a single reader. The chapter closes with an alphabetical listing of clinical and technological abbreviations used throughout the thesis and their definitions.

2.1 Dementia

This section serves as an introduction to the dementia syndrome, the underlying pathologies that cause it, current assessment methods and how cognition is modelled in neuropsychology. Figure 2.1 show the process of cognitive decline over time and the clinical stages.

2.1.1 Etiology

Cognitive disorders negatively affect the mental capability of a person. Causes vary widely between chronic neurodegenrative disease–e.g. Alzheimer's and Parkinson's disease– to mostly reversible causes–such as drug or alcohol abuse–to trauma-induced brain damage, that is usually the result of a single event, e.g. a stroke, and therefore non-progressive and partly reversible [317]. The symptoms experienced by people suffering from cognitive disorders are referred to as ”dementia”.

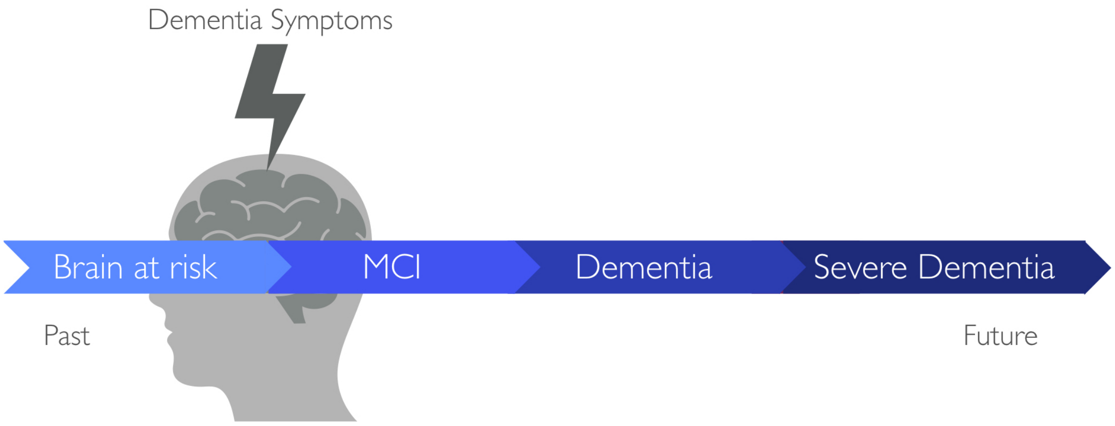


Figure 2.1: Process of cognitive decline.

2.1.1.1 Dementia syndrome

Dementia is a syndrome–not a disease–comprised of several symptoms caused by an underlying cognitive disorder. Not all symptoms are present in every case of dementia, but impairments in the following cognitive functions are common:

Memory and Learning Patients experience a gradual loss of old information and an inability to store new memories. The loss affects memories (i.e., people do no longer recognise friends and family) as well as factual and procedural knowledge (e.g., how to prepare coffee). The trouble with storing new memories classically manifests itself in the repetition of tasks of conversation topics.

Language and Communication Demented persons often have trouble with producing meaningful utterances with a certain fluency. Problems are either due to an inability to control the motor aspects of speech production (i.e., Dysarthria) or in forming semantically complex sentences (i.e., Aphasia). Language in dementia is often perceived as being ”empty”.

Focus and attention Patients show trouble with complex tasks or activities that require selective or prolonged attention. Driving a car is a good example of a task that requires all these skills.

Executive function Demented persons often have problems with planning tasks, calculations and decision making. Inhibition reflexes are also impaired [20].

Perceptual-motor functions Patients show problems coordinating visual inputs and motor functions. Copying simple figures with a pen can be problematic for people with dementia.

Social Cognition Demented persons often have trouble recognising or understanding emotions in others. They also often show delusional or even aggressive behaviour [20].

Affected persons' life quality and the effort required to care for these people is dependant on the severity of symptoms. The progressing losses in cognition, quality of life and dignity, lead to an eventual dependence on other people. Progressive types of dementia are lethal, often due to complications from impaired primary motor functions [57, 320].

2.1.1.2 Alzheimer's Disease

The most common organic cause of dementia is Alzheimer's disease, responsible for about 60% to 80% pf cases [19]. Prevalence increases with age, as 95% of AD patients develop the disease aged 65 or older [18]. With an increasingly ageing population, the amount of affected people in developed countries continues to rise. Worldwide, 46 million people suffered from AD in 2015. This number is expected to triple by 2050 [320].

The specific cause for this type of dementia is known by the name of psychiatrist and neuropathologist Alois Alzheimer (1864 – 1915), who first documented a case of this disease. While the specific destructive mechanisms of AD are still under investigation, its biochemical aftereffects are well documented. AD causes fragmentation, missfolding and clumping of tube structure permanently connected to neurons, referred to as *amyloid precursor proteins*. This transition of an essential protein leads to the death of the neural cell. The spread of this decay is heterogeneous, but often starts in regions like the temporal lobe. The dementia symptoms observed in AD are directly linked to the atrophy in these areas [307].

The loss of short-term memory is the most-common and therefore most well-known early symptom. In addition, difficulties in language production are also observed across early to late stages of the disease. Beginning with less fluent speech patterns and a reduced vocabulary, the symptoms deteriorate into a stage in which a person is incapable to verbally express themselves.

2.1.1.3 Mild Cognitive Impairment

Mild Cognitive Impairment describes a stage of cognitive decline in which patients experience first obvious symptoms of cognitive impairment, but are not yet hindered in their everyday life [310]. Impairment is usually specific to one cognitive domain. If memory is affected, this is referred to as amnestic Mild Cognitive Impairment (aMCI). MCI is often seen as a predecessor to Alzheimer's disease. Although there are cases where this form early impairment can not be linked to neurocognitive decline, patients with MCI convert to AD with a rate of 10% to 15% in comparison to the normal population with 1% to 2% per year [310].

2.1.2 Assessment methods

There are several methods to diagnose cognitive disorders, that vary in their scaleability, intrusiveness and effectiveness. The rising number of dementia cases require us to find a diagnostic method that is accessible and applicable to large populations, has low costs, is simple to administer and is effective in identifying early stages of disorders like AD.

2.1.2.1 Biological and physiological markers

One, very accurate, was of differential diagnosis is through brain imaging. In these methods, images and models of a persons brain allow detailed accounting of causes for observed dementia symptoms. This is especially useful for determining different stages of the disease.

There are multiple techniques of brain imaging, such as magnetic resonance imaging, positron emission tomography (PET) or computer tomography (CT) [282]. In theses procedures, a patients head is placed in a machine and scanned externally. The resulting images of the brain are mostly accurate and unambiguous in identifying organic causes of dementia, such as AD [269]. There main disadvantage lies in the practicality for continuous application. While the scanners do not come in direct physical contact wit the patient, a contrast agent has to be injected to trace blood flow in the brain. In case of PET and CT scans, these contrast agents are radioactive and therefore a health hazard in themselves. A single PET scan can lead to 2-5 times higher radiation exposure than the average normal exposure per annum [157]. Non-radioactive agents that are used in MRI can cause extreme allergic reactions and agents containing gadolinium are thought to cause complications in patients with low kidney functions. Additionally, severely claustrophobic patients or those with magnetisable materials inside their bodies are excluded from this procedure [155]. Scanners are also extremely expensive and require specialised personnel to operate [275]. These health risks and limiting factors make brain imaging infeasible for use in broad population screening.

Brain imaging methods are used to observe the physical damages caused by neurodegenerative disease. Those damages however are also indicated by the symptoms that they cause (e.g. damage in the temporal lobe results in memory loss). By testing for typical symptoms, a conclusion can sometimes be derived. Methods to quantify cognitive abilities will be discussed in the next section.

2.1.2.2 Behavioural markers

The medical practice to derive a diagnosis from the symptoms which a patient reports is obviously nothing new. Fever usually indicates an infection the body fights against, acute localized pain most likely results from overtaxing of the respective body parts. Such simple diagnoses usually do not require a specific blood test or leg-CT, nor would most patients put up with a time consuming procedure if an improvement of the symptoms is likely. When talking about the diagnosis of diseases like Alzheimer's, there are several issues with asking a patient for their symptoms. Firstly, neurodegerative disease do not cause any direct pain like headaches, therefore it is not helpful to ask a patient "where it hurts". The symptoms might also be not be significant enough at first, and often get attributed to just getting old. When eventually visiting a doctor, the damage could already be severe and - as discussed - permanent. Given the nature of the symptoms like memory loss that Alzheimer's causes, the patients might not even notice loss of memory at first, because they forgot it [19]. People at an age where they are prone to suffer from the disease would need regular testing for symptoms, and those tests would need to be efficient enough for this purpose.

So-called *Cognitive Assessments* are the method by which the mental capabilities of a patient are tested with the purpose of detecting cognitive disorders. They are short, game-like tasks of multiple categories a patient has to perform. Those tasks challenge a patients performance in six major cognitive domains defined by the fifth edition of the DSM-5 [26]. Those are Language, Executive Function, Learning and Memory, Complex Attention, Perceptual-Motor Function and Social Cognition [335].

To test for example a persons memory and consequentially function of the temporal lobes, a patient would have to remember a sequence of numbers or words, and recount them after a short waiting period. There are dozens of clinically approved tests over all categories. Depending on the type of assessment, a predetermined sequence (battery) of differing tests and varying length is administered [298]. The Mini-Mental State Examination takes only about 5-10 minutes, and is a commonly used tool for early assessment first published in 1975 [299, 116]. These cognitive assessments can help show symptoms of dementia in its early stages due to their accessibility. They provide good ground work for future diagnostic methods, are clinically approved and commonly used by doctors and neuropsychologists. Generally, doctors are aware of the issue that Alzheimer's represents, and will administer cognitive tests if an unexpected decline of cognitive functions is noticeable [291]

The disadvantages of these tests is that they can show at most symptoms of a potential disease, not detect the disease itself. Since dementia can be caused by different disorders with different treatment methods but overlapping symptoms, a final diagnosis may still be determined via brain imaging. Even though those cognitive tests are relatively easy to administer, they still require guidelines to follow. The tests and their rating scales are standardized, and any distraction could skew the results. The tests can be administered by any neuropsychologist who sits down with the patient. If those tests are administered for the first time, demographic data of the patient has to be raised if not available already. The patients age, years of education, gender or preexisting conditions are relevant in the evaluation of the patients performance.

2.2 Apathy, an affective syndrome

Apathy can be described generally as a syndrome comprising a reduction in goal-directed behaviours, reduction of interests and emotional blunting [276]. Study findings suggest that a disruption of mechanisms underlying the way in which reward is processed to motivate behaviour could be the potential cause [43, 32]. Consequently, it can be seen primarily as a motivational disorder present in several psychiatric and neurological conditions such as stroke [62], traumatic brain injury [407], major depres-

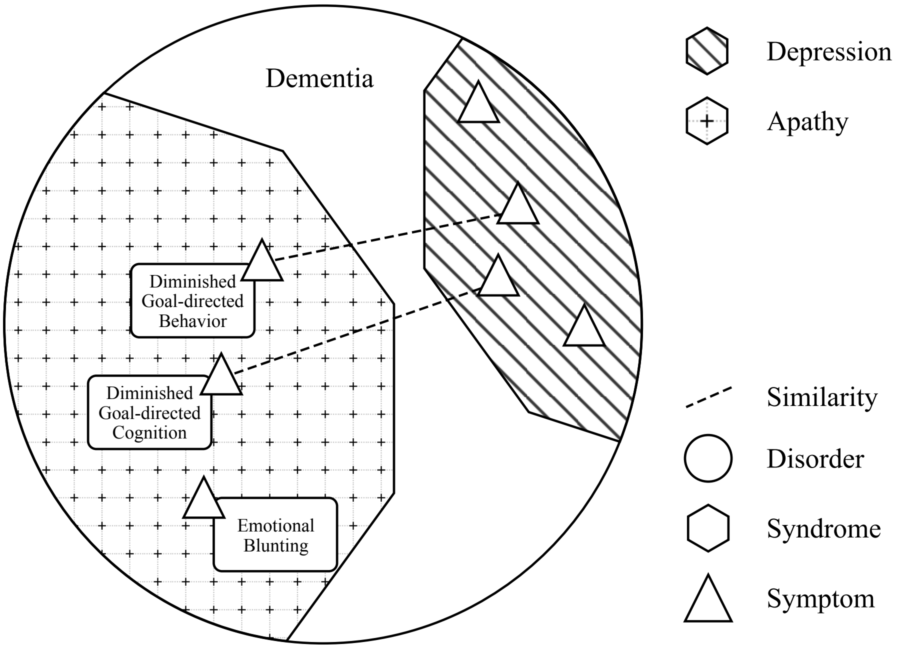


Figure 2.2: Prevalence of apathy in dementia and relation between symptoms of apathy and depression in dementia. Both affective syndromes can occur in dementia (disorder). Symptoms of diminished goal-directed behaviour and cognition can overlap between apathy and depression.

sion [415] or schizophrenia [411], as well as in neurodegenerative diseases including Alzheimer's disease [1] or Parkinson's disease [292]. Although there seems to be a lack of consensus in the definition across different pathologies, with different terms employed interchangeably according to patient groups, Cathomas *et al* [66] proposes that for research purposes it may be helpful to regard it as one concept to a large extent, applicable across traditional nosological categories, to be considered a ”trans-diagnostic clinical phenotype”.

Figure 2.2 depicts the relation between dementia, apathy and depression.

The presence of apathy visibly and significantly affects the patient's and caregiver's quality of life [412, 256]. In neurodegenerative disorders, apathy is associated with faster cognitive and functional decline [359] representing a risk factor for the conversion from early stages to AD [300]. Thus, identifying apathy timely in disease progression is considered a clinical and research priority.

Current assessment methods for apathy rely mostly on scales or interview-based self-reports such as the Apathy Inventory [328] or the Neuropsychiatric Inventory [86], which might not always reflect the actual activities and motivational states during the time a patient is being evaluated [148]. Furthermore, their application for early detection is rather limited because of their dependency on human observers as well as frequent impaired capacity for self-observation [199, 72]. Recently, a task force redefined the apathy diagnostic criteria for better operationalization in clinical and research practice, stipulating the presence of quantitative reduction of goal-directed activity either in the behavioural, cognitive, emotional or social dimension in comparison to the patient's previous level of functioning. These changes may be reported by the patient him/herself or by the observation of others [327]. Furthermore, information and communications technologies (ICT) may supplement these classical tools with additional objective measures and potentially provide more continuous endpoints in clinical trials.

König *et al* [199] performed a review of ICT for the assessment of apathy and concluded that no one had previously used ICT specifically in this context, but that techniques seemed promising. Since apathy seems to affect emotion-based decision making, attempts to measure it through games were made such as the Philadelphia Apathy Computerized Task (PACT) [115], detecting impairments in goal-directed behavior including initiation, planning, and motivation. Reward and effort mechanisms have been explored along with physical effort discounting through paradigms such as the one developed by Pessiglione *et al* [308]. Studies in Schizophrenia have shown that actigraphy and the measurement of motor activity provides a promising readout for quantifying apathy [195]. Actigraphy has been used as well to measure physical changes in dementia patients with apathy [91]. Apathy has also been explored using eye-tracking in AD patients [69] with the result that apathetic patients tend to less fixate social images than the non apathetic patients [211]. Despite these efforts to find alternative objective measurements of apathy, a cheap and fast method which could help with early, remote and non-intrusive screening is still urgently needed.

Recent advances in computer linguistics and language processing have led to the use of automatic speech analysis in the assessment of various clinical manifestations [110]. Semantic and acoustic features automatically extracted from speech tasks seem highly sensitive to cognitive decline and potential conversion risk to dementia [204]. Significant associations were found between negative symptoms in schizophrenic patients and variability in pitch and speech proportion, even in different languages [42]. Strong correlations were obtained between negative symptom severity and phonetically measured reductions in tongue movements during speech [79]. Since it is well recognized that aprosody – flattened speech intonation – represents a negative symptom of schizophrenia and that there is a clear resemblance to apathy, we can hypothesise that this technology could represent a promising candidate for measuring and tracking its severity even in different types of population [75].

In depression, it is notable by ear, that patients show a reduced prosody spectrum and sound rather monotonous which could serve as an indicator, if objective measurements can quantify these observations. Until now several groups investigated the use of automatic analysis of speech as an additional assessment tool with an extensive review published by Cummins *et al* [87] outlining the interest of using speech as a key objective marker for disease progression. Prosodic, articulatory and acoustic features of speech seem affected by depression severity and thus can be easily identified and used for continuously monitoring patients. With a considerable overlap of symptoms between depression and apathy, namely the lack of interest and goal-oriented behaviour, we anticipate similar results when applying speech technology methods to apathy with a slightly different pattern in regards to emotionally triggered speech.

2.3 Cognitive Testing

Cognitive tests are concise exercises that are designed to measure specific cognitive abilities of an assessee. An example is a wordlist learning task, in which a patient is read a list of word repeatedly and has to remember them after a certain amount of time. They belong to the category of functional assessments, as they offer an external perspective on the cognitive capabilities of a person. Exercises are often speech or pencil based and administrable in a few minutes. They are used extensively in the diagnosis of different dementia types. Tests usually result in one or more outcome variables that are used as a measure of performance of the assessee (in the wordlist example, the number of correctly recalled words). The outcome variables (*scores*) alone are often not directly interpretable as a point of reference is needed to interpret what constitutes a normal or an impaired performance. This is why tests are usually distributed with so called *norms*. These norms are used to normalized the obtained *scores* against a control population of healthy individuals. To this end, large groups of people who are cognitively healthy perform these tests to establish what a ’normal’ performance looks like. Since cognitive functioning generally declines with age and is increased by a higher education level, these demographic variables are used to stratify norm groups. Formally, norms are encoded in tables and can be created and applied in multiple variants:

- ***z*-score**
  A statistical representation of a outcome score that expresses it in terms of mean and standard deviation of the reference population. Let x be the score obtained in the test, μ the mean of the applicable age and education level stratified norm group and σ their standard deviation. Then the z-score can be expressed as

  $$z = \frac{x - \mu}{\sigma}$$

  It is a way to express the outcome score on a scale where 0 represents the population mean and in terms of standard deviations. Consequently, a z-score of e.g., -1 would be interpreted, as a performance that is one standard deviation below the reference populations mean.

- **Percentiles**
  Percentiles represent a value in relation to how many percent of observations in a population are below it. So performance in the 20th percentile is equal to 20% of the population showing a worse performance. Where the z-score represents a value using the population mean and standard deviation, percentiles relate values to the population median. This is preferable, especially if populations are are not normally distributed and therefore not well represented by mean and standard deviation.

- **Cut-Offs**
  For easy decision support, some norms are transformed into cut-off values that clearly state at which point a performance can be counted as impaired. Although practical, these representations often lack in describing the differences in performance experienced by people in a single cut-off category.

Norms are an intuitive way to compare performance in a cognitive test against a group of healthy individuals to identify abnormal behaviour. Problems arise mainly when a persons demographic information is not well represented in a norm population. Either

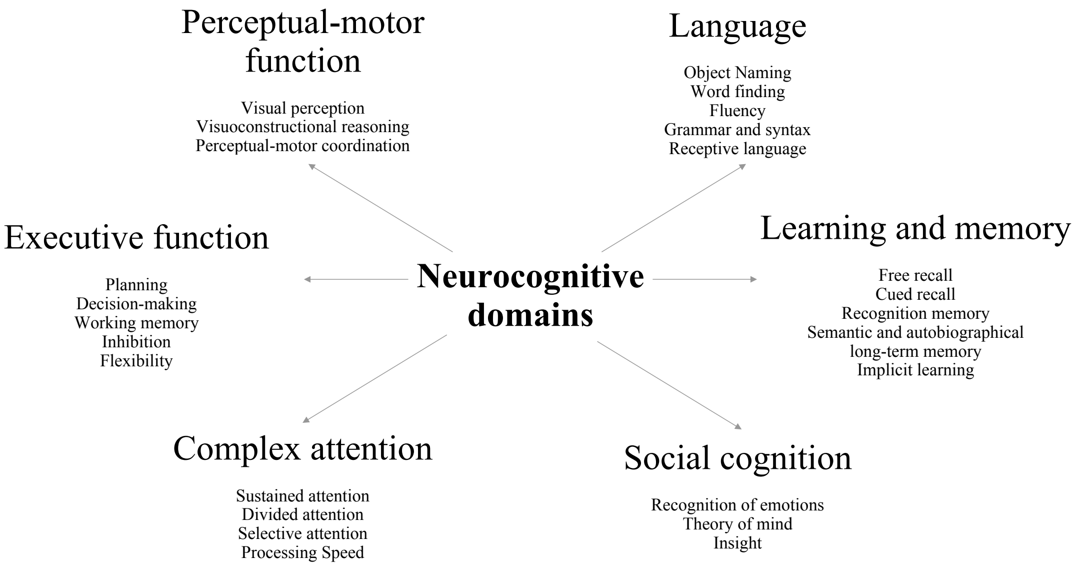


Figure 2.3: Model of neurocognitive domains according to DSM-5. Each box represents a super-domain of cognitive functions. Subdomains are listed as items inside.

because of a unusual demographic structure (very high/low age/education level) or a lack of data variability in the population the norm was created on. In this case, single time measurements of cognitive functioning are often note sufficient, since people with a very high education level might show above average performance, while still experiencing cognitive decline. Comparisons are clearer when assessing the same person over a period of time. The focus then shifts away from performance at individual time points to longitudinal performance developments.

As previously mentioned, cognitive tests are often defined in a way to measure a specific cognitive ability. This can be especially important in differential diagnosis. Multiple types of dementia show different progression patterns in the brain, which translate to different declining cognitive deficits. Furthermore, a certain cognitive impairment might be linked to a physical injury of a closely associated brain area (e.g., after an accident). Human cognition can be modelled in several dimensions and there are multiple approaches to do so. In this work, we utilise the model introduced in the DSM–5. This diagnostic manual is the gold-standard for psychological assessment methods and diagnostic criteria around the world. It models a total of six super domains of cognitive function, which each contain multiple subdomains. See Figure 2.3 for an overview and

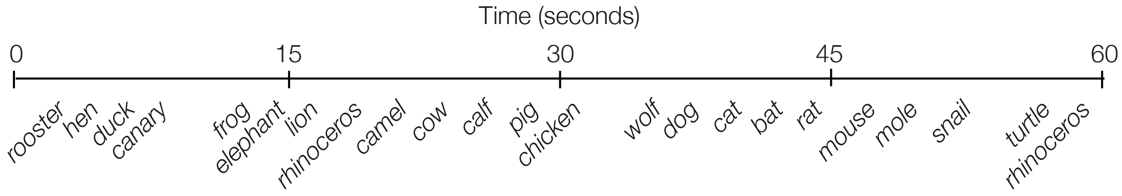


Figure 2.4: Example performance in a Semantic Verbal Fluency task of category animals. The position of each word on the timeline indicates when the assessee named the animal in the 60 seconds.

Appendix B for detailed descriptions.

Cognitive tests are usually designed to measure a certain cognitive subdomain. Many of the tests are speech-based exercises, although that does not necessarily mean they measure language function. Take a word list learning task for example. A list of words is read out loud to an assessee and he is asked to verbally recall them after some time. Although the complete test is speech based, it clearly does not measure language function, but learning and memory ability. This distinction is important, as consequentially patient generated speech and language can be used as a means to measure more than just language functions. Other tests are not performed using speech, but rather paper and pencil. Although these tests also offer wide room for improvements through digitalisation and advanced analysis of the so captured data, they are not the topic of this thesis.

If multiple tests are combined into a package, which often comes with a specific scoring scheme, this is referred to as a *test battery*. Through the combination of a variety of tests, a broad part of cognition can be assessed. These batteries can either be used as screening instruments or to stage the progression of cognitive decline. In the following, some tests and test batteries which are analysed or referred to throughout the thesis are introduced in more detail.

2.3.1 Verbal Fluency

Verbal fluency (VF) is amongst the most widely adapted neuropsychological standard tests and is routinely applied in the assessment of neurocognitive disorders. Its subform, category fluency or *semantic verbal fluency*, demands the assessed person to produce as many different items from a given category as possible within a given time interval, e.g., “as many animals as possible in 60 seconds”. A substantial number of clinical studies confirm the discriminative power of SVF for brain pathologies including such as Alzheimer’s disease [293, 323, 28, 139, 160], Parkinson’s disease [159], psychiatric disorders such as schizophrenia [329], Primary Progressive Aphasia (PPA) and its subforms [49, 244], as well as focal lesions [379]. Generating words according to a given semantic category involves multiple cognitive processes including (1) lexical retrieval, (2) systematic lexical search, (3) holding active generated words, and (4) inhibition of automatic erroneous responses [81].

Especially the latter ones impose high executive-control demand on the subject due to the efficient organisation of the actual retrieval, as well as self-monitoring aspects. However, semantic verbal fluency depends primarily upon the integrity of semantic associations within the lexicon; deficits in this test above all indicate semantic retrieval problems. Considering the involvement of two distinct cognitive processes, Troyer et al. [379] first introduced a systematic framework to calculate measures for both processes from the response behaviour of a subject. In general, production of words is organised in *spurts*—temporal clusters—followed by pauses, implying the lexical search (2) for semantic fields or subcategories between clusters, and retrieval/production of words (1) within the cluster [151, 379]. This means, that between temporal clusters, executive search processes—switching (2)—and within temporal clusters, semantic memory retrieval processes—clustering (1)—are engaged. The underlying notion is, that temporal clusters, correspond to semantic clusters; in other words, ”words that comprise these temporal clusters tend to be semantically related” [379, p. 139]. From the transcript of the produced succession of words (e.g. animals), qualitative measures can be calculated, following the suggested approach by [379]. Hereby, multiple subjectively defined taxonomic subcategories, based on main categories, are used to determine whether successively generated words belong to the same subcategory—i.e. form a cluster.

Temporal analysis of verbal fluency tasks shows, that subjects’ performance follows a logarithmic curve asymptotically flattening out in the second half of the, typically, 60 seconds of the fluency task [112, 405]. Factor analysis based on 10-second-intervals fluency performance reveals two main components which have been interpreted by [112], as *semi-automatic rapid retrieval*—during the first three intervals—and *relatively slow effortful search/retrieval*—during the second 30 seconds; only the first half scores have been shown to effectively discriminate between dementia, MCI and healthy controls

[113]. Beyond this classic fixed interval temporal analysis, temporal clustering as indicator for cognitive processes is typically not investigated, although the spurt character of the performance shown in such a task has been well documented. In fact, the association of retrieval time and semantic memory organisation goes back to Collins et al. [74] and has been investigated early on by Grünewald et al. [151], temporally fitting a two-stage model (fast semantical retrieval and slow executive search) revealing a strong correlation between the temporal clusters and the semantic clusters. However, current practice in neuropsychology does not require to record the SVF task performance, which results in a loss of the temporal dimension of the data.

Additionally, there is strong evidence for the multifactorial nature of the SVF from neuroimaging studies. Investigating the correlation of underlying neural structures and behavioural patterns in VF tasks, researchers have been putting forward the hypothesis of a distributed neural system for semantic retrieval. In neuroimaging studies, pooling functional magnetic resonance imaging (fMRI) and behavioural data, overall SVF performance, i.e., word count, has been found to be related to the activation of distributed functional systems, comprising specific brain regions, such as inferior frontal gyrus (associated with executive retrieval processes, correlated with working memory), medial and lateral fusiform gyrus (semantic storage), as well as occipital regions, e.g., lingual gyrus (visual processing) [388, 245].

Beyond these investigations on the macro level of SVF, i.e., the overall word count in the task, some brain lesion studies investigate the underlying neural structures of qualitative measures of the SVF task on a micro level, i.e., switching and clustering behaviour. Most prominently, [380] compared semantic clustering and switching variables between frontal-lobe lesion and temporal-lobe lesion patients. However, results are not conclusive, as the authors report (1) significant main effects in switch-counts for frontal-lobe lesion patients compared to healthy controls and no cluster size effects, but (2) found not significant main effects for clustering among temporal lobe patients and the to the contrary ”on semantic fluency switching there were overall group differences” [380, p. 502]. Additionally, frontal-lobe patients’ SVF scores have been found to improve significantly, as soon as they are provided with external subcategory cues, e.g., by the interviewer. Given the frontal brain regions’ importance for executive processes [383], this serves as evidence for the distinct cognitive functions involved in the semantic fluency task.

To summarise, the overall involvement of different distinct cognitive processes (i.e. executive control processes and semantic memory retrieval processes) in the performance of the SVF task has been well documented in the literature. This is mainly due to evidence from lesion and neuroimaging studies. The evidence from behavioural studies is less conclusive, especially when it comes to the dissociation of selective functional impairments and their predictive power for neurocognitive impairments like found in multiple forms of dementia.

2.3.1.1 SVF in Alzheimer's Dementia

On the behavioural macro level, evidence for the involvement of a distributed neural network during semantic fluency task and the corresponding behavioural impairments which are obvious in most dementias' progressions is well documented. However beyond this general/macro-level SVF impairment, both clustering and switching processes could be impaired and correlated with the symptom. Accordingly, evidence from literature on the corresponding semantic fluency measures with Alzheimer's patients—and dementia in general—shows inconsistent results: Longitudinal studies report a significant decline in switching index, driving the semantic fluency performance's decline, [323], whereas other longitudinal studies identify mean cluster size as propelling measure [273]. Other cross-sectional studies report an impairment of both measures discriminating between Alzheimer's patients and healthy age-matched controls [279, 139, 380] or neither one of these measures [293]. However, across multiple studies semantic fluency scores highly correlate with both the number of semantic clusters, as well as switches [329, 139]. This sometimes leads to the conclusion, that those measures do not provide additional incremental validity, arguing that ”examination of qualitative aspects of fluency is time consuming and does not appear to improve diagnostic discriminability beyond that achievable using quantitative aspects of fluency” [139, p. 774]. This is in line with the notion, that the hybrid character of the SVF renders it difficult to interpret in clinical applications [346].

Nevertheless, SVF performance can be modelled as a combination of two cognitive sub-functions, i.e., executive search processes and semantic memory retrieval processes, manifesting in two respective measures, i.e., switching index and cluster size. Assuming that AD is marked by a vast corruption of the respective neural distributed network, both of the derived qualitative fluency measures should be equally affected, resulting in basically no more than simply an impaired semantic fluency count. The correlation/connection between both number of switches (NOS) and mean cluster size (MCS) and the overall semantic fluency count (SVF-count) necessarily follows from their formal relation:

$$\text{SVF-count} = (\text{MCS} * (\text{NOS} + 1)) - \text{repetitions} - \text{intrusions}$$

The fact that some studies report different cluster sizes, can be explained through the subjective clustering criterium [379], which leaves room for interpretation regarding the clustering, thereby directly affecting both measures, NOS and MCS.

Following the traditional approach, words, e.g., animals, can belong to one or more predefined subcategories. There are about 25 subcategories based on three main categories: ”living environment”, ”zoology”, and ”human use”. A cluster is then defined as successively generated words belonging to the same subcategory. If a word can be assigned to two consecutive clusters, it is counted as belonging to both. A cluster contained by another cluster is not counted. To better understand the ambiguity incorporated in this rating scheme, one should consider the following example from the SVF with animals:

frog - dolphin - donkey - monkey - gorilla - tiger - panther - aardvark - ant - crane

There are multiple ways how these utterances could be clustered following the traditional subcategory-based approach:

- (*frog - dolphin - donkey - monkey - gorilla - tiger - panther - aardvark - ant - crane*) [all African animals]
- (*frog - dolphin*) [water animals], (*donkey*) [animals of burden], (*monkey - gorilla - tiger*) [jungle animals], (*tiger - panther*) [felines], (*aardvark*) [insectivores], (*ant*) [insects], (*crane*) [birds]
- ... (*monkey - gorilla - tiger - panther*) [jungle animals], ...

Additionally, there are multiple non category-based cluster possibilities which are not catered for: (1) phonemically similar words (e.g. *donkey* & *monkey*), or (2) concepts that occur together in popular culture (e.g. *panther*, *crane* & *aardvark*, as in the cartoon series *The Pink Panther*). One could also argue, that having only one subcategory for *water animals*—having *frog* and *dolphin* appear in the same semantic cluster—may not capture the variance accurately.

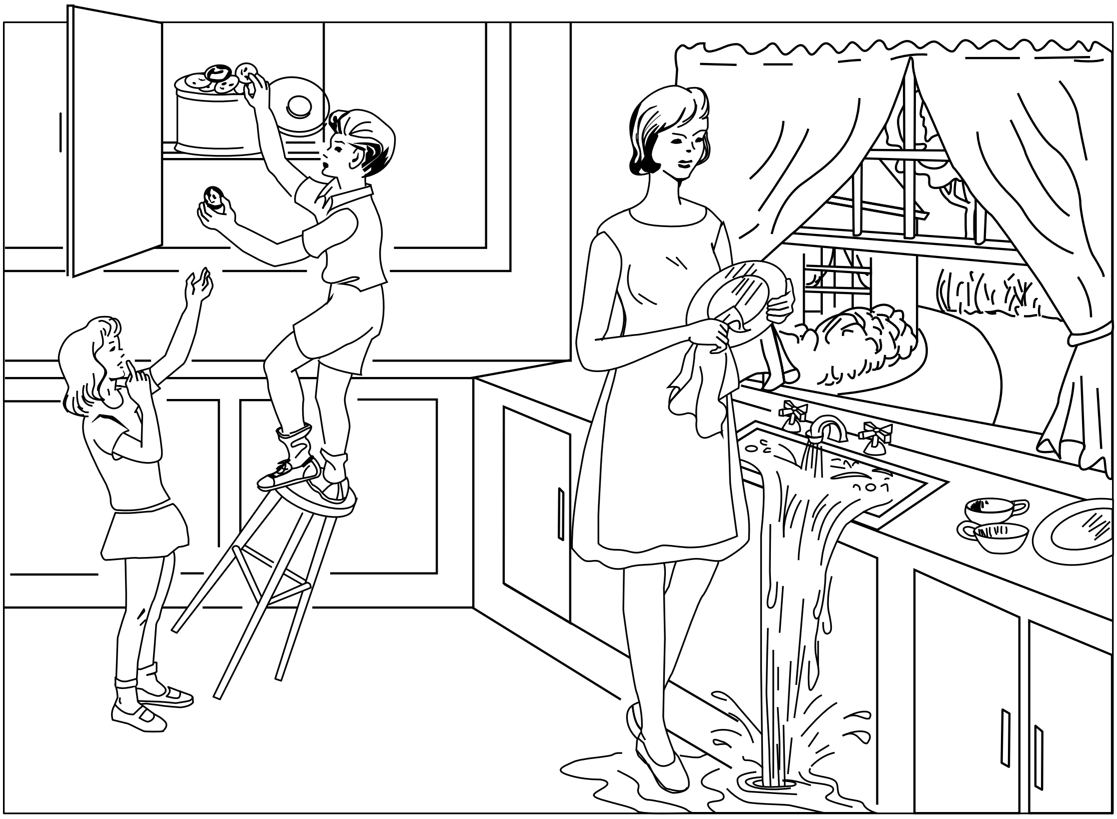


Figure 2.5: The Cookie Theft Picture used to elicit speech in description tasks.

However, beyond the comprehensive qualitative analysis of the SVF category *animals*, the literature does not consider other parallel versions of the SVF featuring different semantic categories than animals, at all. This might be due to the fact that the traditional qualitative scoring scheme is based on manual categorisation and hand-crafted taxonomies. As a matter the practical application of qualitative SVF analysis is highly restricted.

2.3.2 Cookie Theft Picture Description

A widely-used language assessment is the *Cookie Theft* picture (CTP) description from the Boston Diagnostic Aphasia Examination [143]. In this task, participants are shown a line drawing of a kitchen scene, where a boy can be seen standing on a stool, trying to reach a cookie jar, while a woman is preoccupied washing dishes. They are asked to describe everything they see going on in the picture, in as much detail as possible.

Cognitive domain	**Task**	**Points**
Concentration	Serial subtraction or spell 'WORLD' backwards	5
Language	Repetition (single complex sentence)	1
	Comprehend instructions	2
	Follow written instructions	1
	Write a short sentence	1
	Recognise and name two common objects	3
Memory	Word list learning	3
Orientation	Questions about the current time, date and location	10
Visuospatial	Copying of intersecting pentagons	1
Working memory	Repeat three words	3

Table 2.1: Subtasks of the Mini Mental State Examination, what cognitive function they measure and point distribution among them.

The line drawing is presented in Figure 2.5. The task has been used as a way to elicit free speech from people suffering from Alzheimer's disease [128], although its original application is for the detection of aphasia.

2.3.3 Mini-Mental-State Examination

The Mini Mental State Examination (MMSE) is a common test battery used as a screening tool for dementia. It takes around ten minutes and requires a minimally trained assessor, consists of a series of tasks that cover different forms of cognitive functions, such as memory and attention [347, 116], and is designed to be used as a global screening tool. Table 2.1 gives an overview of its tasks and scoring scheme. Patients score between 0 and 30 points and generally, the score can be interpreted as

30-25	degree of impairment is questionably significant
25-20	Mild cognitive impairment
20-10	Moderate cognitive impairment
10-0	Severe cognitive impairment

Its validity has been proven and it is widely translated and used [372]. The MMSE is unfortunately sometimes misunderstood as a diagnostic tool, when it is actually a screening test with relatively modest sensitivity in detecting a mild degree of cognitive impairment. It has floor and ceiling effects and limited sensitivity to change which is becoming a particularly important issue with the recent increased focus of researchers on the milder stages of AD [381]. Despite all these drawbacks, the MMSE is still an incredibly widely used screening instrument around the world. All of the few available speech and language corpora of MCI populations contain the MMSE as a general score for cognitive health. Due to the difference in diagnostic criteria used between different studies, it can be used to make patient populations comparable.

2.3.4 Clinical Dementia Rating Scale

The Clinical Dementia Rating scale (CDR) [285] represents internationally the most widely applied staging tool for assessing dementia's global severity. It is a mix of questionnaire and cognitive test battery that encompasses six domains of cognitive and functional performance: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care [174]. The assessment is conducted in the form of semi-structured interviews with the affected person and an affiliated person/co-interviewee, e.g., a family member. The CDR is relatively time-consuming - interviews can take up to 90 minutes - depending on the availability of a co-interviewee and requires significant training of the raters in order to achieve good reliability [270].

The Inter-rater reliability for CDR is excellent (correlation coefficient 0.89) [271] and content validity can be assumed, as the six cognitive domains rated by the CDR are linked to validated clinical diagnostic criteria for AD [114]. Each domain is rated on a 5-point scale of functioning as follows: 0, no impairment; 0.5, questionable impairment; 1, mild impairment; 2, moderate impairment; and 3, severe impairment (personal care is scored on a 4-point scale without a 0.5 rating available). The global CDR score is computed via an algorithm that weighs memory more heavily than the other categories. The CDR Sum of Boxes (CDR-SOB) score is obtained by summing each of the domain box scores, with scores ranging from 0 to 18 [285]. In general, the higher the score, the greater the severity of dementia. CDR-SOB scores are used to rate dementia severity as follows :

0	Indicates normal cognitive functioning
0.5–4.0	Indicates questionable cognitive impairment
3.0–4.0	Indicates very mild dementia
4.5–9.0	Indicates mild dementia
9.5–15.5	Indicates moderate dementia
16.0–18.0	Indicates severe dementia

The CDR-SOB score has been considered a more detailed quantitative general index than the global score and provides more subtle information than the global CDR score in patients with mild dementia, and a suitable tool for measuring the response to treatment in clinical trials of AD [236]. The advantages of the SOB method include that the CDR-SOB scores can be treated as interval data in statistical analyses, whereas global CDR scores are ordinal by the nature of the algorithm approach to condensing the data. Finally, the most significant advantage of using CDR-SOB scores for staging dementia severity is the increased precision, allowing for tracking changes over time [285].

2.3.5 MMSE and CDR as Assessment Tools

Both, MMSE and CDR-SOB, global assessment measures are the most widely used in clinical and research settings for dementia screening and staging its severity. Staging dementia is crucial for clinical trials and the development of effective pharmacological interventions. They are administered and interpreted by specially trained healthcare clinicians in order to provide appropriate patient care and to identify the effectiveness of prescribed treatment interventions in patients with dementia. Inter-rater reliability for CDR is excellent (correlation coefficient 0.89) [271] and content validity can be assumed, as the six cognitive domains rated by the CDR are linked to validated clinical diagnostic criteria for AD [114].

Each domain is rated on a 5-point scale of functioning as follows: 0, no impairment; 0.5, questionable impairment; 1, mild impairment; 2, moderate impairment; and 3, severe impairment (personal care is scored on a 4-point scale without a 0.5 rating available). The global CDR score is computed via an algorithm that weighs memory more heavily than the other categories[1]. The CDR-SOB score is obtained by summing each of the domain box scores, with scores ranging from 0 to 18 [285]. In general, the higher the

[1] http://www.biostat.wustl.edu/~adrc/cdrpgm/index.html

score, the greater the severity of dementia.

The CDR-SOB score has been considered a more detailed quantitative general index than the global score and provides more subtle information than the global CDR score in patients with mild dementia, and a suitable tool for measuring the response to treatment in clinical trials of AD [236]. The advantages of the SOB method include that the CDR-SOB scores can be treated as interval data in statistical analyses, whereas global CDR scores are ordinal by the nature of the algorithm approach to condensing the data. Finally, the most significant advantage of using CDR-SOB scores for staging dementia severity is the increased precision, allowing for tracking changes over time [285].

The MMSE encompasses a variety of questions, requires minimal training and takes around 10 min. The questions are typically grouped into seven categories, representing different cognitive functions: orientation to time (5 points), orientation to place (5 points), registration of three words (3 points), attention and calculation (5 points), recall of three words (3 points), language (8 points) and visual construction (1 point) [347, 116]. Patients score between 0 and 30 points, and cutoffs of 23/24 have typically been used to show significant cognitive impairment.

Its validity has been proven and it is widely translated and used [372]. The MMSE is unfortunately sometimes misunderstood as a diagnostic tool, when it is actually a screening test with relativelxy modest sensitivity in detecting a mild degree of cognitive impairment. It has floor and ceiling effects and limited sensitivity to change which is becoming a particularly important issue with the recent increased focus of researchers on the milder stages of AD [381].

2.3.6 Computerized cognitive testing

Digital tests that seek to assess cognitive functions, briefly and globally, are being developed with the aim to be administered remotely [54]. The exhibited advantages of these tests are standardization of administration and stimulus presentation as well as the measures (e.g. reaction times and latencies) are more accurate: performances can be compared to established norms [402] allowing the clinician to concentrate on a personalized analysis of the patients' needs.

For instance, the CogState Brief Battery (CogState) is a brief computerized test which assesses reaction and processing speed, episodic memory, attention, working memory, learning, and decision-making. [93] examined the specificity and sensitivity of the CogState test for the diagnosis of mild cognitive deterioration, comparing it with classical pen and paper tests with the result that it reaches similar discrimination level as traditional tests.

CANTAB, one of most known cognitive screening tools, offers specialized AD test battery versions for assessing prodromal states, or mild dementia. The batteries measure motor skills, executive function, episodic memory, visual memory information processing and sustained attention. CANTAB has been shown to be highly sensitive to cognitive dysfunction and ties in closely with current neurobiological models for MCI [121, 103].

The TDAS (Touch Panel-type Dementia Assessment Scale) [175] based originally on the pen and paper ADAS-cog test [331], measures word recognition, instruction compliance, temporal orientation, visuospatial skills, recognition of object use, naming, planning of the writing process, money computation, and recognition of the time indicated by an analogue clock. This digital test can be administered in 30 minutes, just two-thirds of the time that ADAS-cog requires.

The CNSVS (CNS Vital Signs) [152] is a digital screening test, assessing working memory, mental flexibility, psychomotor speed, verbal and visual memory, set shifting and inhibition and vigilance and sustained attention. The authors studied test-retest reliability as well as concurrent and discriminant validity concluding that it can be used as a reliable screening tool in medical contexts.

Phone-based screening has been investigated by Castanho et al. (2016) comparing the delayed recall task and a classical neuropsychological assessment with the Telephone Interview of Cognitive Status (TICS) in a population of older adults. The TICS consists of 13 items evaluating spatial, temporal and personal orientation, working memory, attention, and verbal and semantic memory. TICS showed high correlation levels with global scores of classical tests as well as a satisfactory internal consistency. This method could allow faster access to assessment for people living in rural areas producing similar results as the usual pencil and paper screening tests.

2.4 Machine Learning Prerequisites

One main focus of this work is to use markers extracted from speech and spoken language to construct computational models that are able to automatically label new samples with a diagnostic group. The area of computer science concerned with building such predictive models is referred to as *Machine Learning* (ML).

Given a group of *samples* (i.e. recordings of patients), a set of specific properties for each sample, referred to as *features* (i.e. variables calculated from each recording) and an assignment of a *label* to each sample (i.e. a diagnosis), a machine learning model tries to use inductive inference to predict a samples label based on the given feature representation. If the predicted set of labels is discrete (as is the case for diagnosis), these models are referred to as *classifiers*. The features extracted for each sample are encoded in a numerical *feature vector* x, which contains a single feature per row. A set of samples can be encoded by appending these features vectors horizontally, which leads to a feature matrix referred to as X. The label of single sample is denoted by y, labels of a set of samples are gathered in a vector Y.

Generally, the construction and evaluation of a classifier can be split in a *training* and *testing* phase. In the *training* phase, the classifier is presented with features from labeled samples. Its task is to find structures in the given features, that allow it to predict the label of a given sample well. These structures should *generalise* well to unseen samples, since in practical application, the task of a classifier is to label unseen samples. The performance of classifiers is evaluated in the *testing* phase. To this end, the classifier is first trained on a separate *training set* and its performance is evaluated by letting it predict the labels of samples from a separate *test set*. Given the actual labels of samples from the test set and the predicted ones, a performance score can be calculated.

2.4.1 Classification Performance Metrics

Assume a binary classification problem, where the possible labels $y \in \{0, 1\}$. Let Y denote the true labels of a set of samples and Y' the labels predicted by a classifier. Then the results of classification can be visualised in a so called confusion matrix. This matrix visualises the agreement of true and predicted labels and the type of error made in prediction (i.e., confused 0 with 1; confused 1 with 0). Figure 2.6 gives a schematic example of a confusion matrix. Samples are counted to one of four categories in accor-

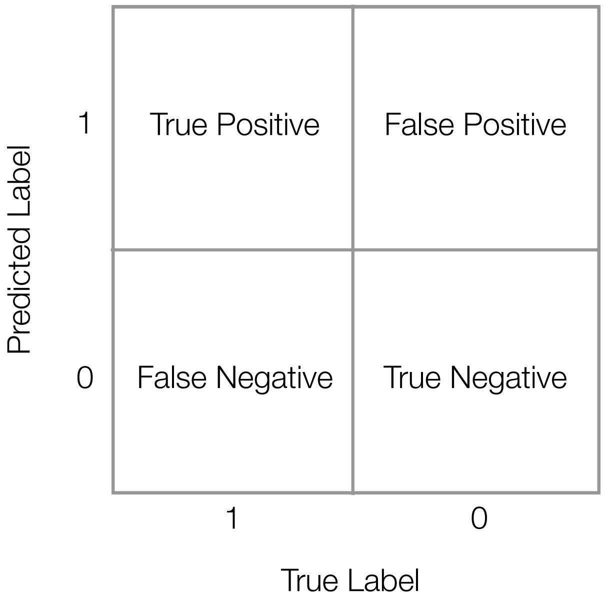


Figure 2.6: Schematic of a binary confusion matrix. Samples are counted into one of four categories depending on their true and predicted labels. The column is determined by the true label and the row by the predicted one.

dance with their true and predicted label. These categories are:

- *True Positive* (TP), for samples with a true and predicted positive label
- *True Negative* (TN), for samples with a true and predicted negative label
- *False Positive* (FP), for samples with a true negative and predicted positive label
- *False Negative* (FN), for samples with a true positive and predicted negative label

Multiple performance metrics can be directly calculated form the counts of these categories.

Accuracy

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy is a general performance metric that counts the instance of correctly classified samples in relation to the number of overall samples. It does not distinguish between the types of errors made (i.e., False Positives vs. False Negatives) which is problematic in many application scenarios where one type of error has much more severe consequences than another (i.e., failing to treat a condition). Furthermore, accuracy is easily skewed by class imbalances (i.e., when one class only occurs in 1% of samples, a classifier that always predicts the other class has 0.99% accuracy).

Sensitivity

$$SENS = \frac{TP}{TP + FN}$$

Sensitivity (or how it is more often referred to in the ML community, Recall) measures the ratio of actual positives that are correctly classified as such (e.g., ratio of people with a disease, who are correctly classified as having it). Especially in the medical field, this metric is often considered to evaluate the goodness of a diagnostic method. For example in screening scenarios, it might be considered most important that people suffering from a disease are correctly identified for further treatment.

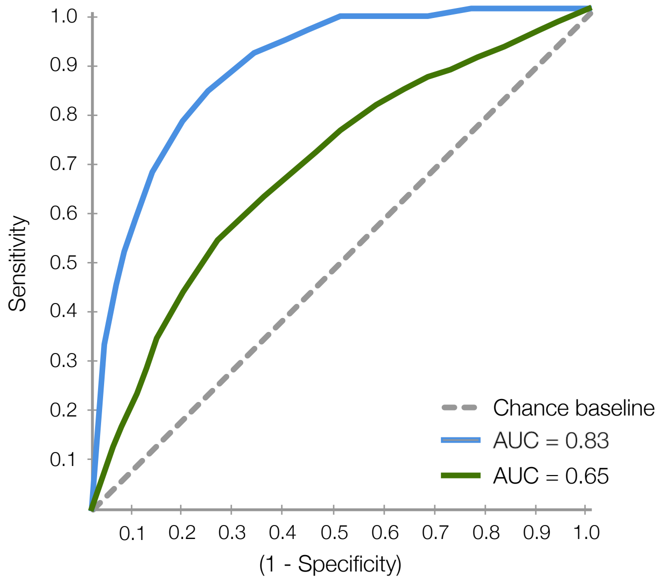


Figure 2.7: Schematic plot of multiple ROC curves. The chance baseline is indicated by the dotted diagonal. Two ROC curves with different AUC values are displayed by a green and blue curve.

Specificity

$$SPEC = \frac{TN}{TN + FP}$$

Specificity is the inverse of Sensitivity. It measures the ratio of actual negatives that are correctly classified as such (e.g., ratio of healthy controls, who are correctly classified as such). In the medical domain, it relates a diagnostic instruments ability to correctly reject healthy controls without a certain disease. Especially for treatments that are dangerous or harmful themselves, high sensitivity in choosing the subjects to which to administer these procedures is desirable.

Area Under the Receiver Operating Characteristic Curve

The Receiver Operating Characteristic Curve (ROC curve) is a plot of Sensitivity as a function of (1 - Specificity) for different decision threshold settings of a classifier. It

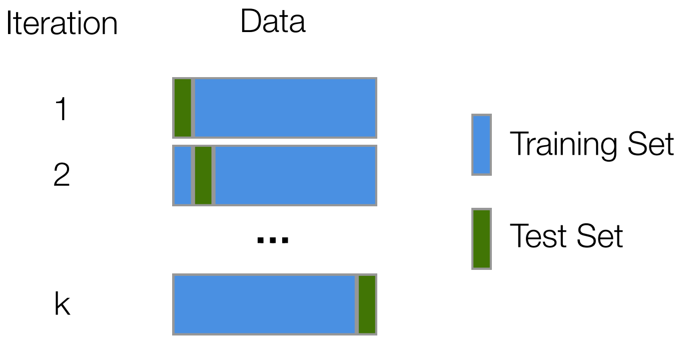


Figure 2.8: Schematic representation of k-fold cross validation. The training set (blue) and test set (green) in each fold is displayed.

allows for a visual representation of different trade-offs of sensitivity and specificity to which a classifier can be configured. The name originates from its invention by electrical engineers in World War II, where it was used to detect enemy objects on the battle field.

Figure 2.7 shows a schematic plot of multiple ROC curves. The stepped appearance indicates the possible threshold configurations and their respective sensitivity and specificity levels. Generally, a curve that leans toward the upper left of the plot is indicative of better performance, since the tradeoff between Sensitivity and (1 - Specificity) is less severe. This fact can be specifically encoded into a variable by measuring the area under this curve (AUC). AUC can be interpreted as the probability of the classifier ranking a randomly chosen positive sample higher than a randomly chosen negative one. In Figure 2.7 the ROC curves for three different classifiers are given. The random chance baseline is indicated by the diagonal curve. Since AUC is rather robust against imbalances in the label distribution, the AUC baseline is always at 0.5. The classifier with better performance is represented with a blue line (more to the upper left, higher AUC) and the other one by the green line (closer to the chance baseline, smaller AUC).

2.4.2 Evaluation Methods

There are multiple ways to estimate the performance of a classifier. The easiest way is to keep a hold-out set of test samples which a trained classifier can predict to evaluate model performance. This set is usually split off the whole available data (often about 15% of it). Especially in domains where sample sizes are small already (as is the case for us), separate test sets are problematic. They further limit the amount of data models can be trained on and often do not accurately represent the statistical distribution of the data. In these cases, the preferred alternative is to use so called *cross validation* (CV) approaches.

In CV, a random portion of the data is dynamically split off to serve as a test set. This procedure is repeated multiple times and performance is estimated through averaging results. The most famous form of CV is *k-fold cross validation*, where $\frac{1}{k}$th of the data is split off in each iteration (where k is classically 10). The general process is visualised in Figure 2.8. The most extreme form of k-fold CV is when k is chosen as the number of samples, meaning that only a single sample is held out in each iteration, which is referred to as *Leave-One-Out cross validation* (LOOCV). AUC can not be computed without bias on a single data point in each iteration [9], which is why the concept of *Leave-Pair-Out cross validation* (LPOCV) was introduced. Since the hold-out set now consists of a pair of samples, this procedure is often applied to all possible pairs of data points, where the two labels of each point not to be different.

Although more suitable to smaller datasets, CV is computationally more expensive. In k-fold CV, k–instead of only one–models have to be trained and evaluated. For LPOCV the number is exponentially higher which makes it infeasible for larger datasets or learning methods that require longer training or prediction times.

2.4.3 Classification Models

A small description of ML algorithms used in classification is given. In recent years, *deep learning* has been very popular and applied successfully to a number of problems. These models usually require large amounts of data for training, but often outperform classical models on a number of tasks. A drawback of deep learning is, that models usually lack explainability, meaning that a trained classifiers is not able to reason about which markers or which combination of markers was helpful to classify a given sample. Since the data sets used in our experiments are very small and explainability is often a requirement in medical applications, we only use simple classification models.

2.4.3.1 Logistic Regression

Logistic Regression (LR) is a discriminative classifier which tries to explicitly model the boundaries between two classes. As the name suggests, it is closely related to classical

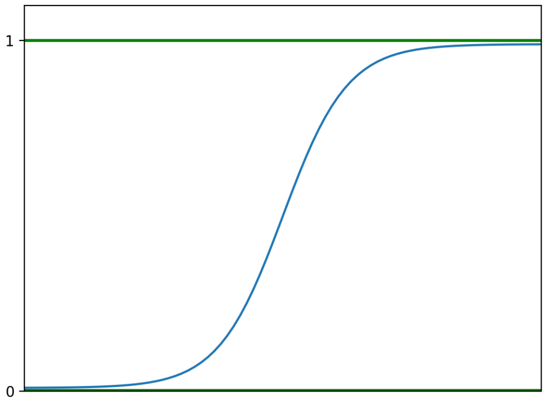


Figure 2.9: Example of a sigmoid function plot used in Logistic Regression to convert the output into a probability value

linear regression models, with the difference that the logistic *sigmoid* function is used to make a discriminatory distinction between two classes, as shown in Figure 2.9. Mathematically, LR models use a so called *regularisation function*, a mathematical addition to the goal function of these algorithms, that force the to prefer feature structure with high generalisability. LR models usually use either the so called L_2 regulariser, which uses the L_2 norm or the L_1 regulariser, which is best on the L_1 norm (in this setting, the LR model is often referred to as *lasso*). Especially the L_1 regularised model has a nice property, in that it performs implicit selection of features by setting the importances of all other features to 0. The results can be interpreted to understand what features are important to a trained classifier and their discriminative between the groups.

2.4.3.2 Support Vector Machines

Support Vector Machines (SVM) are also discriminative classifiers, but are in contrast to LR models, what is referred to as maximum margin classifiers. This means that they are constructed to find a decision boundary in an N-dimensional space that maximises the margin between the classes. In two-dimensional space, the hyperplane is a line, in three dimensional space, it is a plane. In Figure 2.10 the optimal hyperplane for the illustrated two-class problem in a two-dimensional space is highlighted in blue. Specifically, SVMs try to maximise the distance between the nearest data points of both

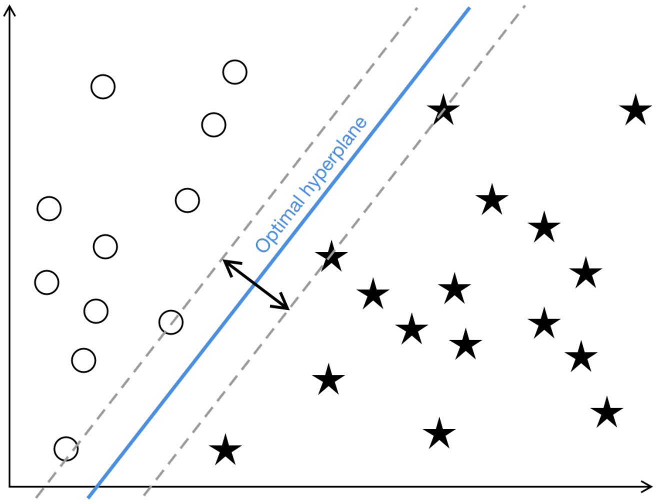


Figure 2.10: The optimal hyperplane for the given two-class problem in a two-dimensional feature space is highlighted in blue. The arrow points out the maximum margin.

classes and the decision boundary. In cases where the points are not linearly separable, the condition is relaxed to maximise the margin and, at the same time, minimise the misclassification rate. These data points closest to the decision boundary are referred to as *support vectors*, hence the name of the method. In non-linearly separable cases, SVMs can use non-linear so called *kernel* functions to implicitly map their inputs into a high-dimensional feature space. When using a linear kernel, an interpretation of the feature weights is also possible.

2.5 Related Work

Recently, progress has been made in approaching the detection of neurodegenerative diseases using computer-supported techniques. The degree of manual intervention varies. Some researchers even report completely automatic analysis pipelines: [374] use automatic speech recognition to extract features and employ machine learning methods to separate subjects with MCI from HC yielding significant discrimination results. [128] obtain state-of-the-art classification accuracies in distinguishing subjects with AD from HC based on short speech samples elicited with a picture description task. Again, when looking at these recent studies, it is important to carefully analyse which target groups they address, which corpora and tasks are being used and which features are being extracted. When discriminating between cognitive impairment and HC and using speech data from paragraph-recall and semantic verbal fluency tasks, [413] develop support vector classification models and, with the best model, achieve 90% accuracy. Note that their cognitive impairment class consists of participants diagnosed with either some form of amnestic MCI or with dementia. Other authors report classifiers discriminating between AD and HC [60, 369, 254], between MCI and HC [374, 146, 205] and between MCI and AD [205]. Some related work has dealt with the automatic analysis of language in other clinical pathologies and dementia types [325, 130, 127].

While many of these studies conduct experiments with small sample sizes, corpora such as the *DementiaBank* [39] are attractive in that they comprise speech samples, partially also transcripts, for larger populations of specific target groups. Speech samples originate from different formats and may include conversations as well as tasks that are drawn from classic, validated test batteries such as the *Boston Diagnostic Aphasia Examination*, which also includes the CTP. Apart from target groups, corpora and tasks, the features that are extracted can be used to structure related computational approaches. Generally, these features draw on the rationale that informs traditional linguistic exams and the observations being made in other scientific disciplines. As such, they apply and expand those features introduced in the previous sections to quantify semantic, morpho-syntactic, pragmatic and/or acoustic language impairment. Interestingly, especially temporal and acoustic features have been proven to be powerful in discriminating different target groups – specifically being sensitive to MCI, early-AD and HC [205, 194, 254].

This subchapter complements the previous one by first introducing additional features that have been used with respect to picture description tasks. To provide a sound overview, it will then broaden the scope to include not only CTP experiments but experiments in general that are based on natural language data, computer-supported feature extraction and machine-learning classifiers. The majority of these experiments treat the detection of signs of neurodegenerative diseases as an n-ary, typically binary, classification problem while a relatively small subset of studies predicts clinical scores from speech samples.

2.5.1 Computational approaches: New features

2.5.1.1 Content: Number of information units

Given that manually annotating speech samples for information units is time-consuming, there have been several approaches to scoring information content computationally. Prud'homme- aux and Roark (2015) generate a source description and use a graph-based alignment method to determine the degree to which any other description in their data recalls the source elements. When doing so, the description with the highest average pairwise BLEU score is selected from the HC data to act as source narrative. This candidate then is manually tuned to only contain information portrayed in the picture. Using only the content words in the source narrative, the authors achieve a best accuracy of 82% on a subset of the DementiaBank corpus when distinguishing AD and HC. Other automated estimates of informativeness search for lists of keywords or n-grams: [295] manually predefine lists of unigrams, bigrams, trigrams and 4-grams that represent correct information units and account for their morhological variants. Similarly, [154] rely on unigram recall to score picture descriptions from elderly HC. Drawing on the last two papers, [128] extract binary information units features from CTP descriptions taken from the DementiaBank. Using WordNet, they semi-automatically generate sets of possible synonyms for the list of information units introduced by [83]. Additionally, they compute integer-valued frequency features for the relevant information units counting each unigram occurrence as contributing to a key word separately. Although this approach allows for a higher degree of underspecification than the original list proposed by [83], reference resolution is a potential problem: Reasonable but unpredictable word choices could be omitted and discounted, whereas false positive counts wrongly sanction matching words even when applied to refer to a different entity or when used off-topic.

2.5.1.2 Content: Temporal perspective

Low density in conveyed information over time manifests itself in perseverative behaviour [283, 38]. For example, [373] report that AD patients are more likely to repeat ideas in a picture description task than HC. While these authors manually annotate ideational repetitions, [128] automatically compute semantic similarity between each pair of utterances in a transcript using a bag-of-words model. After stopword removal, they calculate five features using this information, namely average cosine distance, minimum cosine distance, zero cosine distance and the proportion of utterances falling below some threshold. Finally, [33] identify repeated speech segments in recorded data using motif discovery techniques. As a proof of concept, they carry out a test on five HC reading scripts containing short, embedded, repeated questions and statements and achieve a maximum accuracy of 71%.

2.5.1.3 Morpho-syntactic form

Previous computational studies of AD have used *moving-average type-token ratio*, *Brunet's index* and *Honoroé's statistic* as alternative measures to assess lexical diversity [60, 153, 369, 128]. Bucks et al. [60] borrow the latter two measures from stylometric studies of text to overcome the shortcomings of standard type-token ratio, namely that it positively correlates with text length. *Brunet's index* is calculated as $B = N^{V^{(-0.165)}}$, where N is the number of tokens and V the number of types or the vocabulary. The lower this value is, the richer is the speech sample. *Honoroé's statistic* is based on the notion that lexical complexity is high if there is a high number of word types only used once. It is calculated as $R = \frac{(100logN)}{(1-V_1/V)}$, where V_1 is the number of types occurring only once. The higher this value is, the richer is the vocabulary. Drawing on [111], Fraser et al.,[128] additionally use *moving-average type-token ratio* (MATTR) as an unbiased metric of lexical complexity. A window of fixed size is moved through the sample and the TTR is computed for every position of the window. The MATTR is the mean of all TTRs in a sample. Unlike taking the average of a fixed number of segments, the MATTR thus offers a higher resolution view into the sample. The same authors also dive deeper into the complexity structure on a single-word basis by rating each content word on existing psycholinguistic norms for familiarity, imageability and age-of-acquisition. The authors hypothesize that poor lexical complexity manifests itself in an increased reliance on words that are highly familiar, have a strong association to their mental image and are acquired early in life. With respect to syntactic complexity, computational approaches have added features that exploit parse tree information such as the height of the tree or measures of its embeddedness or that built on the constituents of the parse tree by quantifying them in sets of applied context-free grammar rules [67, 128].

2.5.1.4 Acoustic features

Several computational studies have augmented purely textual features extracted from the transcribed speech sample with acoustic features that can be calculated directly from the audio file: [295] use a semi-automated system to evaluate speech characteristics in patients with FTLD. Among other variables under investigation, they look at the total duration of speech in the sample, normalized pauses of different lengths and pause-to-word ratio, where a pause is defined conservatively as a silent segment longer than 150ms. Especially the latter measure turns out to be able to discriminate between some of the FTLD variants. [326] additionally use measures of phonation to quantify the amount of time in the sample that contains speech. Totally relying on acoustic and temporal speech features, [254] show that measures such as variations in the percentage of voice breaks, number of periods of voice, number of voice breaks, shimmer and noise-to-harmonics ratio characterise people with AD with an accuracy of 84.8%. Similarly, [205] distinguish between HC, MCI and AD solely relying on acoustic and temporal features.

2.5.2 Automated analysis pipelines

2.5.2.1 Detecting clinical populations as a classification problem

Studies that use computational approaches are novel in the sense that they apply computational techniques to quantify signs of neurodegenerative impairments, i.e they transfer technical methods to research questions and data that are fundamentally clinical. Mostly, the detection of impairments is treated as a classification problem, where the number of output variables is typically small and labels correspond to a highly restricted number of potential pathologies or stages of a disease. [128] use linguistic and para-linguistic features as introduced above to discriminate patients with AD from HC. With logistic regression, they achieve an accuracy of 81% with the 35 top-ranked features, which remains relatively constant until a feature set of size ¿50 is chosen. This stresses the need to do feature selection. Unlike previous studies, the authors use a much larger sample size drawn from *DementiaBank*: The corpus has 240 samples from AD patients and 233 data points for HC. [153] use conversations between 31 AD patients and 57 HC. Comparing different learning algorithms, both the decision tree and support vector machine algorithms achieve high accuracies, namely 79.5% and 75% respectively. The authors find fluency measures to be statistically more significant than measures of morpho-syntactic complexity. [176] use acoustic features, POS-based features and psycholinguistic features as given by the software *Linguistic Inquiry Word Count* which, with more than 80 output variables, provides quantifications for psychological and human-made constructs on top of standard linguistic dimensions [305]. When distinguishing between semi-structured interviews between 9 AD patients and 9 HC, they obtain a maximum accuracy of 88%.

The approach of [392] is solely based on n-gram models with subsequent evaluation of perplexity. Applied to picture description tasks from *DementiaBank*, they achieve an accuracy of 77.1% separating AD patients and HC. With the same target groups and also using picture descriptions form *DementiaBank* but extracting para-linguistic features, [11] compare the performance of four different machine leaning algorithms under different running configurations. The authors achieve their highest classification accuracy of 94.71% when using a Bayes Net classifier using pre-processing and the top 22 features. The study by [205] mentioned above trains three different classifiers with varying accuracy on data from MCI and AD patients as well as HC performing different vocal tasks (counting down, picture descriptions, sentence repetition, semantic verbal fluency). The results point to the assessment utility of automatic audio analyses: Between HCs and those with MCI, 79% ± 5%; between HCs and those with AD, 87% ± 3%; and between those with MCI and those with AD, 80% ± 5%. With temporal and acoustic features extracted from a sentence-reading task, [254] distinguish between 30 AD patients and 36 age-matched HC. The spectographic measures the authors use are subjected to linear discriminant analysis with diagnosis as the dependent variable obtaining an accuracy of 84.8%.

2.5.2.2 Predicting clinical scores as a regression problem

There has been little work on the prediction of scores from screening or diagnostic tests instead of predicting adherence to a disease category. Most of the existing work stems from the image processing community, where authors have successfully predicted clinical scores from brain imaging features [370, 418, 173]. Using a set of 477 lexicosyntactic, acoustic and semantic features extracted from 393 speech samples in *DementiaBank*, [408] predict clinical MMSE scores. They obtain a mean absolute error (MAE) of 3.83 in predicting MMSE which they improve to 2.91 for patients where longitudinal data is available. This highlights the importance of longitudinal data collection. [224] investigate the relation between the performance on a semantic verbal fluency task and scores from the standard dementia screening tool MMSE as well as the standard dementia staging tool CDR. Over a set of 179 patients with different degree of dementia, they are able to train a regression model on linguistic and vocal features with a MAE of 2.2 for the MMSE and 1.7 for the CDR.

2.5.3 Computational Analysis of SVF

Recently, computational approaches to analyse SVF have been proposed. There goal is twofold: (1) to avoid the above-mentioned shortcomings, statistical methods have been applied in order to obtain semantic clusters and (2) automated methods are reliable and fast, rendering the extraction of more detailed measure feasible in a clinical scenario.

To verify their adaptation of Troyer's method, Ledoux et al. [214] use Latent Semantic Analysis (LSA), based on the LSA website[2] to compute similarity within clusters and between clusters. They recorded VF production of 153 healthy adults between 18 and 63, transcribed the responses and time aligned the transcript to the audio signal. Afterwards, clusters were determined manually by the Troyer method and a newly developed scoring system. The quality of clustering approaches was validated by examining the LSA distance between and within clusters, as well as the time between words in and between a cluster. Generally, the newly developed rule-based method showed higher separations in all metrics between words inside and between clusters. Ledoux et al. mainly used computational semantic as well as temporal analysis to verify a novel manual scoring system. Although the new system seems more effective in determining clusters, it still suffers from the problems of inter-rater reliability, inter-study reliability and automatability that are intrinsic in manual rule-based systems.

Woods et al. [405] use Explicit Semantic Analysis (ESA) [129]—a vector embedding trained on co-occurrence of words in Wikipedia articles—to identify chaining behaviour for different demographics based on pairwise cosine similarity. They studied a population of 180 control subjects between the ages of 18 to 82 years performing semantic verbal fluency for the category *animals*. In addition to the traditional performance metric of SVF—the word count—they compute semantic clusters and switches based on

[2] http://lsa.colorado.edu/

the pairwise cosine distance in an ESA model, based on a fixed percentage of the average ESA distance. Moreover, novel metrics, such as the mean word frequency, mean word length, as well as temporal decline to the end of the task were examined. Strong correlations between ESA clusters and ones determined using the Troyer method were found.

Clark et al. [71] proposed novel semantic measure based on graph theory; most prominently they put forward graph-based coherence measures which compare the patient's created sequence/path of words with the ”shortest” possible path through the fully connected weighted graph of all patient's words. Although the [71] approach is quite similar to the idea of LSA/ESA or word-embeddings in general—comparing a patient's actually produced sequence with an independent/averaged global world/representation—they normatively provide the graphs' weights (orthographic, phonological, and semantic similarity) and thereby influence the global representation, whereas distributional semantic and word embeddings directly learn spacial representations of words from large corpora; the latter methods allow for the same sort of coherence measures without the need of normatively constructing the global space representation.

Pakhomov et al. [293] examined the decline of several indices of semantic verbal fluency performance in longitudinal study of ageing. To this end, they included 53 patients diagnosed with AD, 71 with MCI and 46 age and education matched controls. All performed SVF tests of the *animals* category. The authors then go on to use part of the healthy population to built a model of semantic relatedness of words by looking at co-occurrence of word pairs. Through this, clusters are determined and the mean cluster size, the cumulative relatedness, repetition density and mean word frequency are recorded. All metrics but the mean cluster size, show significant differences in between the three groups.

Para-linguistic features also have been shown to be of value in the analysis of SVF. [413] used the pseudo-syllable rate and average pause lengths for the analysis of SVF. [404] analysed pauses, speech rate and disfluencies in SVF. In order to differentiate between multiple pathologies, the above mentioned qualitative measures have been established which serve as additional markers next to the raw fluency word count [151, 379]. There is a broad agreement that these measures serve as indicators for underlying cognitive processes. Pauses can occur both within clusters, as participants search their mental lexicon for more examples of a specific group, and between clusters, at the time of a switch, when a participant is searching for the next potentially productive subcategory. In the first 10-15 seconds of the task, pauses tend to be rare, and they typically become more frequent, and longer towards the end of the test.

In summary, using statistical analysis of the SVF task, one could obtain powerful qualitative semantic measures which are relatively independent from the actual macro-level performance (SVF-count) within this task. This is enabled through decoupling the SVF task performance measures from the semantic/lexical measures and interpreting them within an external framework of large text-corpora based semantic representations.

2.5.4 Language features in CTP

2.5.4.1 Content: Quantity

Turning to the *Cookie-Theft-Picture*, the hallmark stimulus, features that capture content are the most prominent and most frequently cited indicator of clinical language alterations. Notions of content are quite diverse with respect to the linguistic surface phenomena they address. In the most basic interpretation, content has been conceptualized as the total amount of words, i.e. mere quantity. [83, 135, 225] report that AD descriptions are always shorter than control texts. The latter two studies as well as [58] do not find a difference in quantity when comparing other groups - [135] compared three subgroups of AD severeness, [225] used two severness subgroups as did [58] who additionally included an MCI group. Note that all these studies make use of different languages with respect to the speech data (English, Portuguese, German) and, more importantly, rely on different diagnostic criteria to assign patients to the respective severity groups. It is thus hard to interpret the results across studies although they all point to a lack of discriminant validity when it comes to staging.

2.5.4.2 Content: Number of information units

The same studies do however identify the number of information-carrying units to be a salient variable which differentiates between the different groups. Again, all of these studies label different entities as information units. [135], for example, use the term information unit to refer to the smallest, non-redundant piece of information – in their understanding synonyms are discounted while grammatical morphemes such as tense -ed or plural -s count as one. [58] have sets of words and phrases that refer to persons, objects, localizations or actions in the *Cookie-Theft-Picture*. Their measures discriminate between different AD stages but they do not distinguish between MCI and HC. Finally, [225] use a set of words that are relevant to the picture. This approach is common in studies that have examined the *Cookie-Theft-Picture*. Focusing only on those studies that include MCI and/or AD pathologies, [164, 389] assess content using a list of eight relevant observations and find the failure to make relevant observations to be more pronounced in the AD group – when contrasting it with HC but also when dividing the AD group in two stages. [120] come to the same conclusion when dividing their AD population into three subgroups. Starting with [83], numerous studies have used an extended set of information units that thus is more liberal with respect to the patient's performance [63, 185, 225].

2.5.4.3 Content: Temporal perspective

It is worth noting that applying a similar construct to a different picture stimulus, [351] do not find a difference in the number of pictorial themes but in the conciseness of the description, i.e. the sample duration and the amount of syllables needed to convey these pictorial themes. Several other studies explored the relation between content words and the length of the sample as indexed by the total number of words in the sample and/or its duration [164, 251, 338]. The general assumption is that AD speech is less specific and more empty. Low density in ideas is thought to result from word-finding difficulties that increase the proportion of low-content phenomena such as repetitions, circumlocations, indefinite umbrella terms and deictic expressions [283]. Most of the studies mentioned above find AD speech to be corrupted. Those studies focusing on early AD stages point to the fact that dysfluent speech tends to be more pronounced later in the disease continuum. Another area of research has investigated the compensatory meta-strategies patients apply when producing errors or have difficulties in finding target words. [118, 120] find significant differences between HC and minimal AD. In assessing patient performance, they use several scales with which raters score each individual sample manually – a single step which costs at least 30 minutes per sample according to the authors.

2.5.4.4 Morpho-syntactic form: Lexical complexity

Apart from content features, measures of lexical variation and lexical diversity have widely been applied. The type-token ratio, which is obtained by dividing the total number of different words by the total number of words occurring in a sample, is the measure of choice in studies that do not have a strictly computational approach [225, 280, 338]. [8, 6, 5] additionally calculated proportional frequencies of open-class (nouns, verbs, descriptive terms) and closed-class (grammatical function) words. They find a significant change in lexical complexity that is largely driven by the proportion of pronouns that is different between MCI and moderate AD groups. The assumption that informs measures such as the *animia index* is that pathological speech is less specific. The anomia index, which is the number of nouns divided by the number of nouns plus pronouns, captures the specificity of referents.

2.5.4.5 Morpho-syntactic form: Syntactic complexity

Several studies have found evidence for a decrease in syntactic complexity. Measures of syntactic complexity are: Words per clause [185], phrase length [338], degree of subordination [83], counts for distinct morphological forms or syntactic constructions such as tense, voice, coordination [120, 210, 164, 185] and, prominently, mean length of utterance, which is the number of morphemes divided by the number of utterances [6, 8, 164, 280]. Some studies also identified morpho-syntactic errors such as agreement errors [185, 8, 338]. Although, in sum, these studies point towards significant changes in syntactic complexity with disease progression, the effect of both MCI and AD on syntax is controversial and feeds the debate on whether observed changes are due to working memory limitations or reflect difficulties due to increasingly comprised semantic processing.

2.5.4.6 Pragmatics and prosody

The measures mentioned above when discussing semantic content can also be seen as pragmatic compensatory strategies as they mark the patients effort when confronted with word-finding difficulties, communication problems and uncertainty. Few studies have analysed other aspects of pragmatics: [68] find important differences in coherence between HC, AD and a third group of old-elderly using a manual annotation methodology based on frame analysis. [238] evaluated the discourse of neurologically normal adults doing a picture description task by manually rating the presence, accuracy and completeness of seven concepts they consider to be central to the *Cookie-Theft-Picture*. In an equally labor-intense manual annotation process, [119] score prosodic features according to subjective notions of the appropriateness of the melodic line. Their finding is of interest here as the authors compare a simple and a complex picture. The *Coolie-Theft-Picture* as a simple stimulus comes with less pictorial themes than the complex stimulus they use. Indeed, there is a complexity effect: Groups only show differences in melodic line in the complex condition.

2.5.4.7 Objectivity and reliability

Especially the hand-crafted measures that assess pragmatics and prosody introduced above risk to be subjective and can hardly be reproduced in subsequent studies. This lack in objectivity also holds for the measures of semantic content that consist in manually created referent lists. These run the risk of being biased towards the assumptions of the list author. Another limitation of most of these studies is that guidelines for transcription and validation are not made explicit. Few are those studies that report scores of inter-rater reliability for both transcription and scoring. Seen from this perspective, studies that take different versions of the same assessment task into account, such as by using a complexity condition within a picture description scenario, benefit from higher transparency as they provide explicit measures of parallel forms reliability. In an earlier study, [120] use two picture stimuli per condition and find high correlations on both measures for both responses to the simple stimuli and responses to the complex stimuli. Using one stimulus per complexity condition, [104] find their AD group producing more content units in the simple condition than in the complex condition relative to HC, i.e. there is a significant group and complexity interaction.

2.6 Language in Dementia

Linguistic ability is impaired in persons with MCI, and the impairment mirrors the decline of language in AD, albeit to a lesser degree. The semantic level of language is typically most affected, resulting in problems with word-finding and naming, whereas there is no clear evidence of syntactic impairment [365]. Computational modelling of linguistic impairment for different pathologies is motivated by the empirical knowledge about distinctive language deficits. Language is considered a sensitive and highly informative marker for cognitive assessments–not only a symptom–in a number of neurodegenerative disease [109, 356, 14]. Studies have been conducted on several levels of linguistics, i.e., phonetics, phonology, semantics, morpho-syntax and pragmatics. There is, however, a large variance in study design, regarding patient populations, used elicitation tasks, as well as the choice and operationalisation of linguistic variables. A generalisation of findings is therefore difficult.

There are a number of ways to elicit narrative or connected speech in research studies and clinical practice, including semi-structured interviews, story-telling tasks, or asking the speaker to describe a picture or series of pictures. The exact nature of the task has been shown to affect various properties of the speech that is produced [338]. One widely-used task is the *Cookie Theft* picture description (see Section 2.3.2). Due to the widespread use of this task in multiple languages, CTP narratives will be the basis of multilingual experiments in Chapter 5.

Distinct linguistic profiles have been described for a number of psychiatric and neurodegenerative dissorders, including schizophrenia [208], multiple sclerosis [134], amyotrophic lateral sclerosis (ALS) [31, 76, 217, 382], Huntington's disease (HD) [178, 281, 314], Parkinson's disease (PD) [138, 150, 306, 385] and primary progressive aphasia [353].

PPA is a selective, progressive language disorder, therefore language impairment can be assessed without any other major cognitive deficits or behavioural disturbances [144, 145]. Three different sub-types of PPA, with distinctive language impairments, can be separated: the *agrammatic* variant, is characterised by non-fluent speech and morpho-syntactic problems (i.e., missing function words). In the *semantic* variant, patient examine troubles in single word comprehension and often produce speech that is perceived as ”empty”. The last variant, the *logopenic* one, is characterised by mispronunciations and word finding difficulties.

Since language impairment is not the major symptom in the other listed pathologies, they have received little attention in research. As PD mainly affects the motor system, speech impairments are often visible at an acoustic and phonetic level [333]. A reduction of verbal fluency and use of action verbs has also been observed [163]. Speech and language patterns in ALS and HD have rarely been studied and only along few linguistic domains [52].

2.6.1 Traditional linguistic exams: AD

Unlike PPA, AD cannot be characterized by a selective language disorder. Its pathology is more global and affects a wide range of cognitive processes [192]. A vast amount of literature exists on the observed decline in memory, the hallmark domain of AD [59]. Symptoms typically involve difficulties in remembering recent events and encoding new information. Episodic memory impairment is the cornerstone of clinically probable AD. Diagnosing dementia and identifying the underlying cause typically involves screening for behavioral and biological markers. At an early stage, executive functioning is frequently affected translating into difficulties with regards to planning or decision-making. As the disease progresses, changes become more substantial and symptoms significantly affect instrumental activities of daily living [25]. Although the most prominent early symptom is episodic memory impairment, language impairment is another characteristic symptom of AD. Within the language domain, particularly manifestations along the semantic dimension have gained scientific attention and given rise to many linguistic tasks by which the deficient performance of AD patients should be demonstrated. Distinct tools and test batteries such as the MMSE have been developed in order to make the screening for AD more light-weight. Although the MMSE does not exclusively assess language function and also includes non-linguistic tests (orientation, attention, memory, drawing), this expresses a general conceptual shift in research priorities: It stresses the identification of AD criteria that precede obvious mid-stage symptoms. The aim to detect potentially incipient dementia in prodromal and preclinical stages, i.e. before the impairment is severe enough to be classified as dementia, also explains the increased attention paid to MCI [131].

2.6.2 AD: Classic performance measures and test paradigms

Word-finding difficulties have long been recognized [193, 36, 192]. Mostly, word-finding ability is evaluated by tests that belong either to the the paradigm of *Verbal Fluency Tasks* or to the paradigm of *Confrontational Naming*. Evidence suggests that category naming fluency and letter naming fluency impose different demands on semantic memory and executive function and that category naming fluency tasks tend to be more sensitive as they rely more heavily on the integrity of semantic memory [339, 268, 158, 4]. Picture-naming tasks such as the *Boston Naming Task* [142] require the patient to name drawings on visual confrontation. Apart from the conceptual integrity of semantic memory and retrieval, such tasks thus involve processing of the visual stimulus. Studies reporting difficulties in the AD population mostly use nouns as target concepts [37, 167]. [330] used frequency-matched pairs of nouns and verbs that were homophonic and homographic. They found verb naming to be significantly more difficult within the AD population relative to HC. As a result of the decline in lexico-semantic abilities and additional anomias and conceptual substitutions, i.e. specific words are replaced by semantically empty words of the same word category, the language of AD patients maintains fluency but lacks information content [283, 187]. *Perseverations*, the persistent repetition or continuation of a response, are other phenomena that are common in AD and likely to be observed using one of the paradigms above. For category naming fluency, [302] distinguish three different types of perseverations two of which, recurrent (e.g. ”cat, dog, mouse, cat”) and continuous perseveration (e.g. ”cat, cat, cat”), correlate with disease severity.

Higher-order language functions such as thematic coherence have also been reported to be impaired in AD. However, the notion of coherence that informs studies that investigate discourse tend to be pretty vague and thus to rely strongly on the impression of the rater that manually scores transcripts of speech output. Using an interview task, [137] find a reduction in global coherence in their AD population. Global coherence is understood as the extent to which a response provides substantive information directly related to the designated topic, where the topic is the question asked by the interviewer. [46] hand-coded the speech acts they identify in open-ended interviews between five AD patients and their spouses. According to them, the AD group produced significantly more turns and topic shifts than did their healthy spouses. Also focusing on macro-linguistic language functions, in their review article, [324] stress that studies predominantly report a deficit in the comprehension of nonliteral, figurative language such as metaphors, proverbs, idioms, irony or sarcasm in AD populations and come to the conclusion that nonliteral language is a worthwile test tool in everyday clinical routine that requires further exploration.

The effect of AD on phonetics and syntax is controversial. At early stages, phonological and syntactic processing appear to be relatively spared [185, 120]. Contrary to this claim, [84] find phonological and articulatory deficits for a variety of tasks. Using picture description tasks, [83] and [104] provide evidence for a decrease in syntactic complexity as operationalized by the number of subordinate clauses and functors for the former study and utterance length for the latter, respectively. By contrast, [188] and [137] show a similar level of complexity and range and frequency of syntactic constructions for both AD and HC using an interview task. [338] directly compare these two most commonly used tasks to elicit connected speech, namely structured interview and picture description. They present evidence for a phonological and syntactic decline in patients at the mild stages of AD and, more importantly, show that the tasks are not fully interchangeable: While the structured interview task is more sensitive towards morpho-syntax, the picture description task is more sensitive towards semantic performance measures.

2.6.3 The task itself matters: Reconciling conflicting results

The discrepancies inherent to AD literature can partially be disentangled by the differences in study design and the statistical evaluation of the results obtained. Small sample sizes are usually employed in the studies mentioned above. Taken together, they critically point to potential effects of the elicitation task itself. Generalizations from one paradigm to another neglect such an effect [52]. To give an example: While pauses in semantic category fluency are well researched and generally understood as being indicative of a semantic memory problem, pauses in tasks that go beyond the single-word level require a multifactorial explanation as connected speech involves interactions between diverse cognitive processes and representational levels. Besides, populations that are labelled the same in different studies may show different impairments due to disease progression. Language decline in AD has been shown to be heterogeneous [101]. [166] assess episodic and semantic memory in three AD groups (minimal, mild, moderate) that were formed according to disease severity as assigned by their performance on the MMSE. While all patients show a significant deficit in episodic memory, the minimal and mild group show a considerably heterogeneous performance.

The general consensus is that language impairment becomes more pervasive as the disease progresses. In the moderate to severe stages, deficits in both production and comprehension become more severe ultimately leading to communication breakdowns and a decline in the quality of everyday interactions and relationships [186, 44, 343]. Patients with MCI show language impairments that are similar to those of early AD. Again, there is considerable variability. Conflicting results in literature are further complicated by varying, co-existing diagnostic criteria. Language tasks that selectively assess specific language functions, prominently semantic category fluency and confrontation naming, provide evidence that deficits appear at pre-AD stages [4, 181].

The take-away of this sub-chapter is that no language test by itself will be able to identify all cases of MCI and/or AD while being at the same time specific. It is in the combination of a variety of methods and the careful collation of their results that language performance can be assessed. Low scores on standard language tests do by no means reflect everyday language performance and competence [334, 274, 60]. Picture description elicitation tasks as introduced more closely in the next sub-chapter mediate between the richness of free speech and the idiosyncrasy that is inherent to unregulated, everyday communication.

Part II

Automatic Detection of MCI from Speech and Language

Chapter 3

Detection of Dementia from manual Verbal Fluency Transcripts

This Chapter presents analysis methods and experiments detecting MCI using transcripts of verbal fluency tests. In these common neuropsychological tests, people are asked to name as many words as they can in a short time frame. In clinical scenarios, a handwritten transcript is often the only form of recording preserved in such a tests. The tests are highly predictive for core cognitive functions, such as lexical retrieval and executive control. Clinical performance in these test is usually assessed as the number of correct words and has been shown to be highly predictive for MCI. We will introduce and validate novel and extended automatic analysis methods and show that they improve the diagnostic ability of these tasks. Both the functionality to detect and stage dementia will be explored. Results of this Chapter will be the basis and justification for fully automated approaches presented in Chapter 4.

3.1 Neural Word Embeddings in analysing Semantic Verbal Fluency

Multiple studies investigating the same subject group report a great variance of cluster sizes and switch counts in VF. This can be explained through the subjective clustering criterion [379] which leaves some room for interpretation regarding the clustering and thereby directly affecting both measures, switches and cluster size. Statistical semantic analysis automatically and reliably providing clusters is a powerful solution to this problem.

This section explores the possibility of using distributional semantics in the analysis of SVF tasks with a focus on clustering and switching patterns. This is in contrast to *taxonomic models* which are based on predefined subcategories and might not be able to capture the full complexity of semantic connections made by humans. We investigate the application and performance of word2vec [257] by which words are embedded into a vector space and where the cosine distance in this space is used as a metric for semantic similarity. This allows for an automatic identification of semantic clusters as well as the computation of switches and cluster size. To indicate the feasibility of this approach within the particular scenario of automated SVF analysis for clinical MCI detection, we compare a set of statistical classification experiments building upon multiple variations of word2vec models to an implementation of the taxonomic approach provided by [379].

3.1.1 Background

The concepts of clustering and switching have been previously discussed (see Section 2.3.1) and related computational work has been explored (see Section 2.5.3). Here, the general differences between *subcategory-based* and *statistical* clustering methods are explored in further detail.

3.1.1.1 Subcategory-based clustering

[379] described a taxonomy-based semantic clustering approach, which despite obvious shortcomings is still extremely popular within clinical research [380, 139, 49]. In this approach words, i.e., animals, can belong to one or more predefined subcategories. There are about 25 subcategories based on three categories “living environment”, “zoological categories”, and “human use”. A cluster is then defined as successively generated words belonging to the same subcategory. If a word can be assigned to two consecutive clusters, it is counted as belonging to both. A cluster contained by another cluster is not counted. Several adaptations have been suggested, e.g., extending the inclusion rules [214], the minimal cluster size [329], and the handling of repetitions and intrusions [273]. However, the fundamental mechanisms remain the same and

some prominent limitations are: (1) recognising non-category based associations is not catered for: phonemically similar words (e.g. *donkey* & *monkey*) or animals that occur together in popular culture (e.g. *panther*, *crane* & *aardvark*, as in the cartoon series *The Pink Panther*); (2) human-made taxonomies are error prone and likely to be incomplete. In the [379] system, there is only one category for *water animals* and therefore, *frog* and *dolphin* appear in the same semantic cluster which may not capture the differences between both animals well; (3) there is a high effort to build a model for a new category which leads to usage within a single category. However, availability of different semantic categories (e.g., tools & supermarket) is of high clinical value for re-testing patients as it prevents confounding training effects, see also [405].

3.1.1.2 Statistical clustering and chaining

To avoid the above-mentioned shortcomings, statistical methods have been applied in order to obtain semantic clusters. However, careful revision of these approaches reveals that many do not actually implement semantic clustering, but rather what we would call *semantic chaining*. In semantic chains, the semantic chain adherence decision is solely based on the previous word.

chain: (*cat* - *dog* - *wolf*) - (*cow*) vs. cluster: (*cat* - *dog*) - (*dog* - *wolf*) - (*cow*)

To our knowledge, [165] are the only authors who explicitly differentiates between a *static* and *fluid* switch model—a clustering and a chaining model. In this study, the model of [379] is used to evaluate clustering and chaining models. A chaining model is built on the basis of the BEAGLE [180] model, a holographic word embedding trained on Wikipedia. To the best of our knowledge, there has been no research into building a clustering model instead of a chaining one based on distributional semantics.

In summary, though very powerful for automation of SVF tasks, statistical approaches are only as good as the linguistic material they are trained on. Most approaches discussed above were trained on Wikipedia articles. However, this might not be the most suitable training material for a model that should capture semantic associations made by humans. Therefore, we compare the discriminative performance of qualitative SVF parameters derived from statistical models based on word2vec to the approach by [379] as prominent baseline and subcategory-based approach. Additionally, we investigate the performance of two different text corpora as basis—the common Wikipedia-based approach vs. a less organised and less academic corpus. We also explore the performance of semantic clustering and semantic chaining implementations.

3.1.2 Methodology and Results

Also we are left with a lack of hard metrics to reliably compare the performance of semantic similarity models. [257] propose a benchmark task for evaluating word2vec models, but it is not suitable to judge the applicability to our task. To get around this conundrum, we adhere to the following line of reasoning: Whatever approach performs best at our task at hand, that is discriminating between MCI and healthy subjects, is the approach we should use in analysis. This method is obviously limited by the amount of data that is available for evaluation and results have to be interpreted with this in mind. Below, we compare two different distributional semantic models with different hyper parameters.

3.1.2.1 Data

The corpus used consists of 100 samples from older persons[1]: 53 patients diagnosed with MCI (M_{Age}=76.8 ±7.2; 28F/ 25M; $M_{FluencyCount}$=14.63 ± 5.76) and 47 healthy control subjects (HC) with a subjective memory complaint (M_{Age}=72.4 ±7.9; 40F/ 7M; $M_{FluencyCount}$=18.86 ± 5.57). Patients are given 60s to name as many animals as they can. All performances have been recorded and transcribed.

3.1.2.2 Models

We compare a set of models, all of them learned using word2vec [257]. word2vec is a word-embedding based on a shallow, two-layer neural network trained to embed words in a vector space, where the cosine distance is a measure for semantic similarity. We compare models trained on two different linguistic corpora: (1) models based on the FraWac corpus [34], a large corpus collected by a web crawler and (2) models based on a dump of the French Wikipedia. Pre-trained models are taken from here[2]. All varying word2vec hyper parameters are reported in Table 3.1. For all models, the context window was set to 5 tokens and negative sampling was used.

[1] Data collected in the context of the Dem@Care project [183]

[2] http://fauconnier.github.io/

3.1.2.3 Clustering and Chaining

On the basis of these models and the cosine distance in the resulting vector space we compute semantic clusters/chains in the following way:

Let a_1, a_2, …, a_n be the sequence of animals produced by patient p. Let $\vec{a_1}$, $\vec{a_2}$, …, $\vec{a_n}$ be their representations in the vector space and let a_1, …, a_{n-1} form a semantic cluster/chain. a_n is part of this cluster/chain if

Cluster

$$|\frac{\langle \vec{\mu}, \vec{a_n} \rangle}{\|\vec{\mu}\| \cdot \|\vec{a_n}\|}| > \delta_p \tag{3.1}$$

Chain

$$|\frac{\langle \vec{a_{n-1}}, \vec{a_n} \rangle}{\|\vec{a_{n-1}}\| \cdot \|\vec{a_n}\|}| > \delta_p \tag{3.2}$$

with

$$\vec{\mu} = \frac{1}{n-1} \cdot \sum_{\vec{x} \in \{\vec{a_1}, \ldots, \vec{a_{n-1}}\}} \vec{x} \tag{3.3}$$

$$\delta_p = \frac{n!}{(n-2)!} \cdot \sum_{\vec{x}, \vec{y} \in \{\vec{a_1}, \ldots, \vec{a_n}\}} |\frac{\langle \vec{x}, \vec{y} \rangle}{\|\vec{x}\| \cdot \|\vec{y}\|}| \tag{3.4}$$

One of the main problems of using distributional semantic models to determine clusters/chains is finding a sensible cut-off value δ. We decided to use the mean distance between any animal produced by a subject. An ad-hoc global cut-off value would be hard to determine, since similarity scores tend to vary a lot.

3.1.2.4 Prediction

We train different classifiers, one for each model using SVMs with a radial basis kernel: this is mainly because we only have two features [171]. Moreover, since our data set is small, we perform a stratified 10-fold cross validation. As features we use the mean size of clusters identified and the number of switches between clusters. For results, see Table 3.1.

		Hyper parameters			Chain			Cluster			Correlation	
Model	Size	Algorithm	Cutoff	Dimensionality	Pre	Rec	F1	Pre	Rec	F1	r_{Switch}	r_{Size}
FraWac	1.6 B	CBoW	100	200	0.75	0.79	**0.77**	**0.73**	0.80	0.76	0.90	0.87
		Skip	100	200	0.66	0.75	0.70	0.70	0.83	0.76	0.90	0.85
		Skip	100	500	0.72	0.72	0.72	0.68	0.68	0.68	0.91	0.88
		Skip	200	500	0.71	0.72	0.69	0.71	**0.84**	**0.77**	0.90	0.75
Wiki	600 M	CBoW	100	1000	0.67	0.71	0.69	0.67	0.75	0.71	0.99	0.95
		CBoW	200	1000	**0.77**	0.74	0.75	0.71	0.69	0.70	0.96	0.87
		Skip	100	1000	0.68	**0.80**	0.74	0.71	0.72	0.72	0.91	0.84
		Skip	200	1000	0.70	0.74	0.72	0.67	0.76	0.71	0.84	0.77
Troyer	-	-	-	-	-	-	-	0.71	0.74	0.72	-	-

Table 3.1: Hyper parameters of trained word2vec models (CBoW=Continous Bag of Words; Skip=Skip-Gram Model), classification results for chaining and clustering implementations (Pre=Precision; Rec=Recall; F1=F_1 Score; highest values are marked in bold) and Pearson correlation coefficient between clustering and chaining-based features (switch counts=r_{Switch}; mean cluster size=r_{Size}).

3.1.3 Discussion

This section set out to compare the discriminative performance of qualitative SVF parameters derived from statistical models based on neural word embeddings with the traditional subcategory-based approach by [379]. We thus implemented Troyer's approach as a baseline deriving the semantic clustering criterion from predefined subcategories. We compared this to a group of statistical approaches based on a patient-dependent clustering criterion derived from word2vec models. We automatically calculated mean cluster size and number of switches based on transcripts of two groups' SVF recordings: MCI and healthy controls. In order to examine both approaches' feasibility within the given scenario, we trained classifiers, showing results clearly in favour of the statistically derived feature set. This is in line with reported feasibility benefits of this approach in [405, 165]. However, to the best of our knowledge, no study so far compared both approaches based on the discriminative performance they achieve given a clinical classification scenario; so far, either one of both approaches has been used to validate the features derived by the other approach and vice versa. Nonetheless, perhaps the most straight forward way of comparing both approaches is by applying them to a relevant clinical scenario—which SVF has actually been designed and used for—and deciding based on their performance in the classification task at hand.

Additionally, we investigate the performance of two different text corpora as basis for the word2vec models. Our results show that the classifiers using features based on the FraWac corpus models [34] achieve higher F1 scores than the ones based on the Wikipedia models. Although it is difficult to derive a conclusion from this rather exploratory result, possible explanations might be that the FraWac corpus is simply larger, or that it represents a less (artificially) academic and therefore more natural linguistic resource.

Finally, considering different effects through semantic chaining vs. semantic clustering, we yield no interpretable results favouring either one of the implementations. Our experiments yield throughout high correlation indices between both implementations across both SVF dependent variables/features: switch count & mean cluster size. This is in line with [165], who also did not succeed in finding clear patterns.

To conclude, this section presents a clinical application of neural word embeddings rendering a statistical approach to the traditionally manual analysis of semantic verbal fluency tasks. Our results demonstrate the feasibility and therefore economic validity of such an approach, having especially relevant implications for remote automatic screening applications as in [376]. The strong dependency between both qualitative SVF measures, switch count & mean cluster size, and simple word count performance, still remains a challenge for understanding their respective diagnostic values.

The next section explores the utility of natural language modelling as a parallel approach to modelling production strategy in SVF tasks.

3.2 Language Modelling in Semantic Verbal Fluency

In this Section we use statistical language models (LMs) as a tool for modelling production of SVF responses of healthy patients, those with a diagnosis of MCI and Alzheimer's disease or related dementia (ADRD). LMs intuitively model production of words in SVF, as production of the next word depends on the previously produced words. Given a corpus of SVF performances, we use LMs to learn these probabilities from data, and then test the model, by estimating the likelihood of a patient's SVF performance. We use the LM's perplexity of a given SVF performance — a score for how well the model is able to predict a given sequence — as a feature for classification of a person's cognitive health.

This section is structured as follows: Section 3.2.1 discusses prior work on clinical applications of language models and perplexity scores. Section 3.2.2 introduces language models. Section 3.2.3 describes the data for further experiments, how the language models were trained and evaluated in a classification experiment. Section 3.2.4 presents results of the conducted experiments. Lastly, Section 3.2.5 discusses implications and concludes this section.

3.2.1 Background

There is a growing body of research using language modelling and perplexity scores for classification of neurocognitive disorders including Alzheimer's disease, varying types of dementia, and FTLD. In previous work, perplexity scores have been used to automatically classify between AD patients' and healthy controls' speech [392]. Language models were built on transcripts from spontaneous speech of subjects describing the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination battery. The resulting language models based on AD speech and control subjects' speech were then used to compute different perplexity scores per patient including perplexity of an AD language model given an AD speech sample and perplexity of an AD language model given a control speech sample. The authors conclude that perplexity in such a free speech task is higher for AD samples than healthy controls, which could be interpreted as evidence for the deterioration of expressive language capabilities over the course of AD. Using free speech from autobiographic interviews — a more liberal scenario for natural language — Weiner et al. used perplexity scores to automatically discriminate between general dementia patients and healthy controls [396]. Multiple-hour interviews (98 subjects, 230 hours) were cleaned of experimenter speech interventions and transcribed both manually and by an automatic speech recognition system. Based on the raw audio signal and transcripts, the authors compared classification results using both automatically and manually generated feature sets divided into acoustic features, linguistic features and ASR features. Perplexity scores were reported as ASR features, differentiated into within and between subject perplexity. The authors concluded that automatic classification is feasible and report within/between speaker perplexity as two of their best performing features. Similarly to [392], other researchers used manual transcripts from speech of the Cookie Theft Picture description task and language models built on healthy controls' speech to differentiate between different forms of FTLD [296]. Results show that perplexity scores discriminate well between different subforms of FTD: behavioural variant of the FTLD and semantic dementia. This is in line with the notion that the behavioural FTD variant manifests not primarily in corrupted language but semantic dementia does. The authors also correlated perplexity scores with results from common neuropsychological tests, such as SVF: the free speech task perplexity scores negatively correlate with the SVF task. This is perfectly in line with the semantic retrieval problems in semantic dementia, manifesting in a very low SVF word count (i.e., high perplexity due to corrupted free speech and low SVF score).

The underlying latent objective of free speech tasks is, by nature, to produce syntactically correct speech. Using a language model trained on healthy controls, perplexity measures how people are not able to produce such an output following the given objective. In the semantic verbal fluency task however, the inherent objective is to produce as many items as possible which necessarily requires to exploit deeper semantic stock. As the objective is also to not produce repetitions, to be successful one has to produce sequences of increasingly rare items to maintain a high production rate towards the end of the task; this follows as the common easy-to-access semantic items are typically produced at the beginning of the timed task. There is broad evidence, proving that demented persons have significant difficulties in the SVF task which manifests not only in a lower SVF raw count, but also in inefficient semantic stock exploitation strategies. In other words, demented patients are, especially towards the second half of such a task, not able to produce rare/repetition-free sequences of correct item responses. This lack of strategic semantic memory exploitation can be observed through multiple computational approaches [405], allowing to automatically compute semantic exploitation measures which compare the patient's sequence of words to a global semantic representation inferred from large text corpora leveraging either graph theory [71] or neural word embeddings (see Section 3.1).

Work presented in the previous section, on the qualitative computational analysis of the SVF in demented patients shows that features based on neural word embeddings discriminate well between healthy controls and dementia types. Especially semantic density—the lexical coverage of a patients semantic exploitation—and word frequency—the lexical rareness of a patients produced items—have been shown to be very predictive and highly significant features in this task (see Section 3.4). In general, demented persons are less successful in the SVF task as they are less able to systematically exploit a large distributed semantic stock and produce sequences of relatively rare items.

Therefore the aim of this section was to explore the possibility of a SVF language model to detect inefficient SVF production strategies, thus dementia. This represents a novel approach, as to the authors' knowledge, perplexity has so far only been used to detect language corruption.

3.2.2 Language Modelling

Statistical Language Models are a common tool for representing the probability distribution of language data, in either written or spoken form. After computing these models, they can be used to determine the probability of a given sequence of words.

To train a model, a corpus is split into a list of n-grams, a sequence of words of length n, $N = (w_1 \ldots w_n)$. The probability of the ngram, N, is determined using maximum likelihood estimation (MLE):

$$P(N) = P(w_n | w_1 ... w_{n-1}) = \frac{P(w_1 ... w_n)}{P(w_1 ... w_{n-1})} \tag{3.5}$$

The model stores the counts of all the n-grams in the corpus, thus 'training' it. To evaluate the probability of getting a certain sequence of words of length m, $S = (w_1 ... w_m)$, from our model, based on the Markov assumption, we can multiply the probability of each ngram in the sequence.

$$P(S) = \prod_{i=1}^{m} P(w_i | w_1 ... w_{i-1}) \tag{3.6}$$

Unigram models are simple models where the probability of every type, or unique word, is equivalent to the relative frequency of the word in the training set. Because unigrams assume that every word does not depend on any of the previous words, they does not capture the relationships between words. This is why we continue with the bigram and trigram models, where conditional probabilities are used in training.

One challenge of language modelling is data sparsity as we will never encounter every possible combination of n-gram that can be generated during training. Data sparsity makes it likely that our model will encounter unseen n-grams during testing and assign them a probability of zero, causing $P(S) = 0$. To counter this, language models employ a technique known as smoothing, in which some of the probability mass of seen n-grams is shifted to unseen n-grams. Lidstone smoothing [219] is an additive smoothing technique in which an 'unknown' token is added, as a placeholder, to our training set. Then, a predetermined α is added to every n-gram count. Any n-grams that appear in testing, and that were not seen in training, will be accounted for by the 'unknown' token. The counts of the n-grams are then normalized by adding the count of the n-gram's history, $C(w_1...w_{n-1})$, to the size of the vocabulary of the n-gram's history, V, multiplied by α. After smoothing, the probability of an n-gram is represented by:

$$P(w_n|w_1...w_{n-1}) = \frac{C(w_1...w_n) + \alpha}{C(w_1...w_{n-1}) + V\alpha} \tag{3.7}$$

After calculating the smoothed probability distribution of a training set, language models can be evaluated on a test sample using a measure called perplexity. Perplexity is a score that shows how well a trained model predicts a test sample by taking the probability of the test sample and normalizing it by the number of words in the test sample. Perplexity is computed by the following equation:

$$PPL(S) = \frac{1}{\sqrt[m]{\prod_{n=1}^{m} P(w_n|w_1...w_{n-1})}} \tag{3.8}$$

Perplexity and probability are inversely related, so when perplexity is minimized, probability is maximized. This means a low perplexity indicates that the model fits the test sample well.

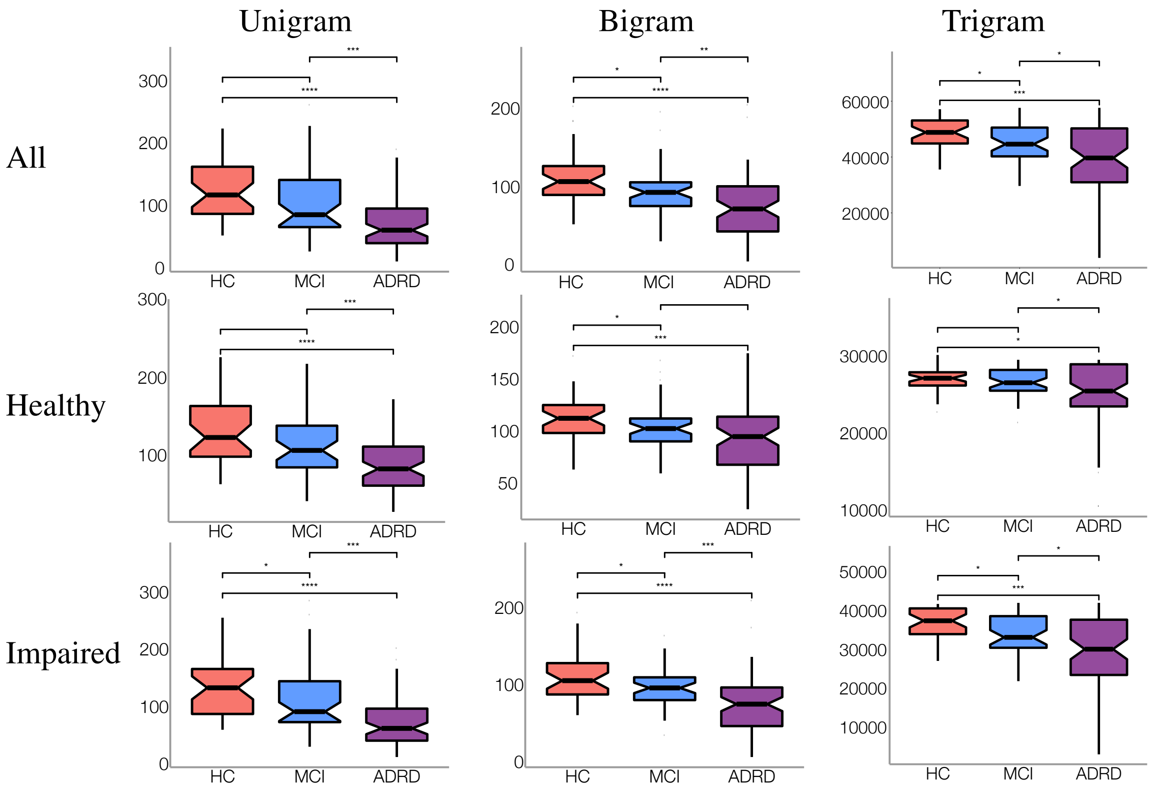


Figure 3.1: Boxplots of perplexity in relation to diagnostic criteria for all three sets of language models. The HC group is depicted in red, the MCI group in green and the ADRD group in blue. Horizontal brackets indicate group comparisons by a Wilcoxon-Mann-Whitney test ($* : p \leq 0.05$, $** : p \leq 0.01$, $*** : p \leq 0.001$, $**** : p \leq 0.0001$).

	HC	MCI	ADRD
N	40	47	79
Age	72.65 (8.3)	76.59* (7.6)	79.0* (6.1)
Sex	8M/32F	23M/24F	39M/40F
Education	11.35 (3.7)	10.81 (3.6)	9.47* (4.5)
MMSE	28.27 (1.6)	26.02* (2.5)	18.81* (4.8)
CDR-SOB	0.47 (0.7)	1.68* (1.11)	7.5* (3.7)

Table 3.2: Demographic data and clinical scores by diagnostic group; mean (standard deviation); Significant difference ($p < 0.05$) from the control population in a Wilcoxon-Mann-Whitney test are marked with *; HC='Healthy control', MCI='Mild cognitive impairment', ADRD= 'Alzheimer's disease and related disorders'; MMSE='Mini-Mental-State-Examination'; CDR-SOB='Clinical Dementia Rating Scale - Sum of boxes'.

3.2.3 Methods

3.2.3.1 Data

The data used for the following experiments was collected during the *Dem@Care* [183] and ELEMENT [376] projects. All participants were aged 65 or older and were recruited through the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Speech recordings of elderly people were collected using an automated recording app on a tablet computer and were subsequently transcribed following the CHAT protocol [239]. Participants completed a battery of cognitive tests, including a 60 second animal SVF test. Furthermore, all participants completed the MMSE [116] and CDR [270]. Following the clinical assessment, participants were categorised into three groups: Control participants (HC) diagnosed healthy after assessment, patients with MCI and patients that were diagnosed as having Alzheimer's Disease or related disorders. AD diagnosis was determined using the NINCDS-ADRDA criteria [250]. Mixed/Vascular dementia was diagnosed according to ICD 10 [406] criteria. For the MCI group, diagnosis was conducted according to Petersen criteria [310]. Participants were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, or psychotic or aberrant motor behaviour. Demographic data and clinical test results by diagnostic groups are reported in Table 3.2.

3.2.3.2 Language Modelling

Based on our three patient populations (HC, MCI, ADRD), we construct three LMs: (1) trained only on the healthy population, (2) trained only on the impaired population (MCI + ADRD) and (3) trained on all patient data, regardless of diagnosis. For each training set we build unigram, bigram and trigram models. We stop at trigrams, since given our vocabulary (n=238) the possible number of trigrams is 13,481,272 and our corpus only contains 2,203 trigram tokens, leading to extreme sparsity. We apply Lidstone smoothing to the model with $\alpha = 1$. Due to the nature of our training samples, lists of animals, and leave one out method of cross validation, we have a small vocabulary and do not expect a high amount of unseen tokens in the testing sequence, compared to natural language, making this a justifiable method of smoothing on this data set.

Perplexity is calculated as described in Equation 3.8. For models (1) and (2) we discriminate between the training population and the rest. Let $A_t = a_1, ..., a_m$ be the training population and $A_r = a_{m+1}, ..., a_n$ the rest of the samples. Then we perform leave-one-out cross validation on A_t, generating one perplexity value for the held-out sample a_i and each sample in A_r, per iteration. In the end, every sample in A_t has one perplexity value and every sample in A_r has m perplexity values. Averaging the m values per sample, leaves us with one perplexity value per sample. For (3) we perform a simple leave-one-out cross validation on the complete set $a_1, ..., a_n$, yielding one perplexity value per patient.

3.2.3.3 Prediction

To confirm the diagnostic power of perplexity, we perform a simple classification experiment. Each person in the database was assigned a label relating to their diagnosis (HC, MCI and ADRD). Perplexity values from different models were used as input to classification models. All features were normalised using z-standardisation.

In all scenarios we use SVMs 2.4.3.2 implemented in the scikit-learn framework [301]. We use a radial bases kernel, since there is only one feature [171] and 10-fold cross validation was used for testing. To find a well-performing set of hyperparameters, parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration was performed. Performing cross validation on small data sets only once leads to performance fluctuations between different iterations. To work

Scenario	Model	F_1
HC vs. MCI	U_{all}	0.62
	B_{all}	**0.71**
	T_{all}	0.67
HC vs. ADRD	U_{all}	**0.83**
	B_{all}	0.81
	T_{all}	0.72
MCI vs. ADRD	U_{all}	0.75
	B_{all}	**0.76**
	T_{all}	0.69

Table 3.3: Classification results for different scenarios and models as F_1 scores. U_{all} = Unigram model trained on all samples; B_{all} = Bigram model trained on all samples; T_{all} = Trigram model trained on all samples.

around this problem, cross validation was performed multiple times and then the mean of all performance metrics was calculated.

3.2.4 Results

Figure 3.1 displays boxplots of perplexity values by diagnostic groups. Each column corresponds to either uni-, bi- or trigram models. Rows indicate the training scenario. In general the perplexity decreases with disease progression - from HC, to MCI, to ADRD.

People with ADRD have significantly smaller perplexity values compared to the HC population, regardless of the context history length considered and training material. The same is true for people with ADRD in comparison to the MCI population. A significant difference between the HC and the MCI population for unigrams is only visible in the 'Impaired' model, (3). Bigram models all show significant differences between both populations. Trigrams only show this effect for models trained on the whole population or the impaired part. Overall, trigrams show less differences between populations and high perplexity values, which can be attributed to the extreme sparseness of these models given our small data set.

Table 3.3 shows classification results for different models and scenarios. Following inspection of Figure 3.1, only models trained on all samples in the population were used in classification experiments, as the inter-group effects seem consistent between different training material. Between the HC and the ADRD group, as well as the MCI and ADRD populations, the unigram and bigram model show comparable performance. For classification of the HC and the MCI population the bigram model clearly shows the best performance.

3.2.5 Discussion

A general result of this study is that people with MCI or dementia show significantly lower perplexity values in SVF compared to a healthy population, meaning the n-gram LMs, regardless of training corpus, are more suited to model a demented person's speech versus that of a healthy person. Thus people with dementia are more predictable in their production of words in the SVF task. This differs from findings about perplexity of demented patients in free speech tasks, where perplexity values of demented speech have been shown to be higher than that of healthy controls [392]. This can be explained by the different scenarios where language modelling is applied: on natural language, a LM and its resulting perplexity can be interpreted as a measure for syntactic normality/correctness. When training on and predicting SVF performances, in which production of word sequences is motivated semantically, the perplexity can be viewed as a measure for effective semantic retrieval strategy.

Furthermore, we found word production in SVF differed in advancing stages of dementia syndromes. Unigram perplexity approximated on the SVF task, can be seen as a measure of predictability of word choice. Perplexity values of unigram models were found to be good indicators to separate the ADRD group from the HC group, but not the MCI population from the HC. Thus, word choice in SVF is more predictable in late stage dementia and not in early stage. Perplexity of bigram models trained on SVF productions—and for that matter any ngram where $n \geq 2$—can be seen as a measure for predictability of production strategy in the task. Both ADRD and MCI groups show significant differences in perplexity of bigram models to the HC group. Consequently, both populations show more predictable production strategies. When modelling with trigrams, we would expect to see effects of context length—such as people with dementia using less contextual information. Unfortunately, this study is limited in the conclusions that can be drawn about the trigram models as it lack sufficient amounts of

SVF data and therefore those models are severely undertrained.

In future experiments, we would like to gather more data to generate well-trained trigram models and possibly draw a more definitive conclusion on the effects of context length in SVF. We would also like to try different smoothing techniques, possibly interpolated methods such as Witten-Bell, that are not as coarse as the Lidstone technique. Based on the trends shown in the unigram and bigram models, demented patients show significantly lower perplexity values, regardless of training data, and are therefore more predictable. Furthermore, persons in advanced stages of dementia differ in predictability of word choice — as shown by the unigram models — and production strategy — as shown by the bigram models — where as people with mild cognitive impairment only show significant predictability in their production strategy. Perplexities from both the unigram and bigram models also function as adequate diagnostic features in classification tasks where the unigram model differentiates the best between HC and ADRD and the bigram model differentiates best between the more fine-grained distinctions of MCI versus the healthy controls or more severely demented patients.

Previously presented approaches to extract qualitative features from SVF were largely language dependant. The next Section will introduce a temporal analysis method that analysis the production of words over the task to reason about cognitive processes. This analysis method is largely language independent.

3.3 Temporal Analysis of Semantic Verbal Fluency

In this section, we examine SVF results of three groups of Swedish participants; those with Subjective Cognitive Impairment (SCI), with MCI and healthy controls (HC). By analysing the data temporally, we are able to reveal differences that are not evident when looking at the SVF as a whole. This section is structured in the following way: An overview of related work is given, with a focus on performance on the SVF by persons with MCI and SCI. Then the dataset and methodology are described as well as the features that were extracted. Finally, the results of our analyses and machine learning experiments are presented and discussed in tandem with other relevant neuropsychological metrics.

3.3.1 Background

Performance of SVF tasks in healthy older adults tends to decline with age, and is partially attributed to a decrease in processing speed, rather than a diminished verbal knowledge [105]. In line with this reasoning, [366] found that the performance of Swedish speakers on SVF is negatively correlated with age and positively correlated with years of education. Healthy participants in the age range 65-89 with $\leq$12 years of education produced a mean of 14.9±6.4 animals, whereas those in the same age range but with an education of $>$12 years produced 19.4±5.6 animals in the same task. The deterioration of cognition in MCI, with impairment both in processing speed and switching attention [24], results in persons with amnestic MCI (aMCI) producing smaller clusters and fewer switches than healthy controls [309]. This reduction across strategy generalises to persons with aMCI producing significantly less categorical words [319, 273]. [284] found categorical differences between naming animals and vegetables when comparing participants with SCI and HC on the SVF test. While the animal category revealed no differences, persons with SCI generated significantly fewer vegetables, specifically in the later 30 seconds. Participants with SCI produced smaller clusters and made more switches in the animal category. The groups did not differ significantly on any demographic variables (age, education, gender) or on the MMSE.

Throughout the SVF, word production rate decreases regardless of the presence of cognitive impairment. To further explore the performance of persons with MCI and healthy controls, [96] divided and analyzed the task into three 20-second sections with two substantial findings; both groups declined over time and generated more words in the first time span. However, persons with MCI performing within normal limits produced fewer words in the first time interval. Slow initiation of lexical search process suggests that MCI inhibits early semi-automatic word retrieval processes. This is in line with previous research showing that the last 30 seconds of the verbal fluency task does not differ between participants, whereas the first 30 seconds contain discriminating information [113]. When performing an even finer-grained temporal analysis based on ten second intervals, [113] found that intervals 1 and 2 were useful in distinguishing persons with AD and MCI, and interval 3 made it possible to differentiate between persons with MCI and SCI, and MCI and AD respectively.

3.3.2 Methods

3.3.2.1 Recruitment and Data Acquisition

All the participants in the current study on ”Linguistic and extra-linguistic parameters for early detection of cognitive impairment” were recruited from the Gothenburg MCI study [390]. All participants were speakers of Swedish, selected according to detailed inclusion and exclusion criteria [198]. Data collection took place in a quiet lab environment where participants were fitted with a lapel microphone (AudioTechnica ATR3350) and digitally recorded with a Zoom H4n Pro recorder (44.1 kHz sampling rate; 16bit resolution). The following instruction was given in Swedish: ”Your task is to think of words. I want you to tell me all the different *animals* you can think of. You have 60 seconds. Do you have any questions? Are you ready? Go ahead and start.” If the participant seemed unsure, they were told ”any animals are okay: big ones, little ones, etc.”. At the end of the 60 seconds, a timer would go off and the test leader would let the participant know that 60 seconds had passed. The resulting audio files were manually transcribed and manually time aligned in Praat [94]. All animals named were transcribed on a separate tier.

A future follow-up visit at the memory clinic in 2019, after a second round of language tests, will include a renewed GDS (Global Deterioration Scale) classification and neuropsychological tests. The study was approved by local ethical committee (ref. number: 206–16, 2016 and T021-18, 2018).

3.3.2.2 Clinical Assessments

Participants in the Gothenburg MCI study were classified as having SCI, MCI, or dementia, and the controls were recruited separately and evaluated to ascertain that they were cognitively healthy. The classification is based on the GDS, where level 1 codes for cognitively healthy, level 2 SCI, level 3 MCI and level 4 and above dementia [27, 390]. Participants were further evaluated with neuropsychological tests, MRI, blood samples, and spinal fluid samples [390].

Compared to the other study participants, the persons with SCI were relatively young, had higher levels of education, higher prevalence of stress conditions and depressive symptoms as well as a family history of dementia [102].

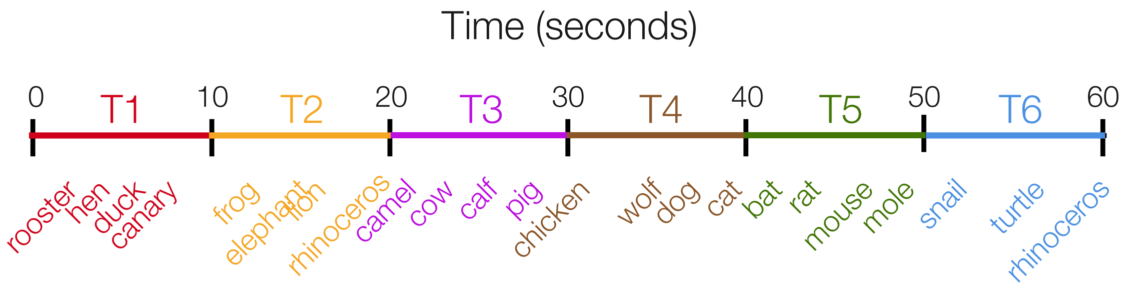


Figure 3.2: The schematic of a time interval analysis for a verbal fluency sample. Produced words are presented on a timeline. Color of a word indicates the affiliation to a time interval.

3.3.2.3 Traditional features

From the manual transcripts, traditional SVF performance metrics were automatically extracted. The word count was determined as the number of unique, correctly named animals. Clusters and switches were determined based on a temporal metric proposed by [378]. In this approach, the cluster structure is solely determined by the temporal position of words in the recording. Consecutive words are clustered if the transition time between them is shorter than then average transition time over the sample. This threshold is furthermore scaled over the process of the task to account for the decline in production speed. The mean number of clusters and the number of switches between them is extracted.

3.3.2.4 Temporally resolved features

To explore different cognitive processes engaged over the course of the one minute task, SVF performance is examined in 10 second steps. Words in the transcript were assigned to a temporal interval based on their onset, as depicted in Figure 3.2. Word count is determined for each interval, disregarding repetitions from earlier intervals. Lexical frequency of words were determined using the KORP collection of Swedish corpora [51]. Transition times between consecutive words were defined as the difference between the end of the current word and the onset of the next. Word frequency and transition times are reported as the average over each interval.

3.3.2.5 Statistical analysis

Statistical analysis was performed using R (software version 3.4.0). For group comparisons of traditional measures, linear models with the measure as a function of diagnostic group were examined. Temporally resolved measures were examined with separate linear mixed effects analysis, one for each response variable –word count, lexical frequency and transition time– using the *lme4* [35] package. Each time interval is modelled as a single data point and with age and education level, as well as the interaction between the time interval (T) and diagnosis, as fixed effects. The participant identifier was modelled as a random intercept. Spearman correlations between the interval word count and neuropsychological scores were examined. Age and education were chosen as demographic variables. As neuropsychological correlates, the following scores were used: the Trail Making Test Part A (TMT-A), as an indicator for processing speed; the Boston Naming Test (BNT; [182]), which assess language ability with a spectrum of high to low frequency words as a proxy of vocabulary size; and the Wechsler Adult Intelligence Scale Similarities (WAIS-Similarities), which measures abstract thinking, concept formation and verbal reasoning [394].

3.3.2.6 Prediction

The predictive power of the proposed temporal and semantic features were validated with machine learning experiments for the HC and MCI populations. For each transcribed speech sample, the features described in Section 3.3.2.3 and 3.3.2.4 were extracted and label in accordance to their diagnostic category. LR 2.4.3.1 and SVM 2.4.3.2 models, as implemented by the scikit-learn [301] framework, were trained as binary classifiers to separate the groups. First, models were trained with only word count, to establish a baseline, and then, on the complete feature set, utilizing univariate feature selection. AUC is reported as the evaluation parameter. Due to the small size of the dataset, we used leave-pair-out cross validation, which has been shown to produce an unbiased estimate for AUC on small datasets [9]. We also computed the standard deviation in AUC as described by [326].

Feature scaling and hyper-parameter optimisation were done on the training set in each fold. Features were scaled using min-max scaling between 0 and 1. For both SVMs and LR, C was optimised between $C \in [10^{-4}, ..., 10^{4}]$ using a grid search. LR models were trained with both L1 and L2 loss; for SVM a linear and an rbf kernel were used. For the

	HC	SCI	MCI
N	32	19	24
Sex (M/F)	12/20	8/11	11/13
Age (years)	68.1 (7.2)	66.0 (6.7)	70.8 (5.6)
Education (years)	13.2 (3.5)	16.0 (2.3)	13.8 (3.5)
MMSE (max 30)	29.7 (0.5)	29.6 (0.8)	28.5 (1.4)

Table 3.4: Demographic information; the MMSE (Mini Mental State Exam) is a general screening test of cognitive status and has a maximum score of 30.

extended feature set, feature selection based on χ^2-tests was applied to the training set in each fold. The number of selected features was scaled between 1 and the maximum of 30.

3.3.3 Results

3.3.3.1 Demographic information

Demographic information by diagnostic group is reported in Table 3.4. The SCI group is slightly younger and has a higher education level than the other two groups. The MMSE, a general index of cognitive status with a maximum score of 30, is lower in the MCI group. With an average MMSE of 28.5, this MCI population is still quite functional in comparison to other MCI populations (mean MMSE score can vary between 23 and 29 in the MCI group) [226]. Note that cut-off points for MMSE may vary slightly: for Swedish, a cut-off value between 25 and 27 indicates possible cognitive impairment which should be further evaluated [297] while other studies consider an ”abnormal” MMSE score to be lower or equal to 25 [416].

3.3.3.2 Traditional measures

A linear model of word count as a function of diagnosis revealed a significant main effect ($F(2,72) = 8.57, p < 0.01$). Compared to the control group ($WC = 24.06 \pm 6.37$), the SCI group ($WC = 27.84 \pm 5.6$) had a significantly increased word count ($3.78 \pm 1.8, p < 0.5$); the MCI group ($WC = 20.12 \pm 6.08$) a significantly lowered one ($-3.94 \pm 1.6, p < 0.5$). No significant effects for the size of temporal clusters

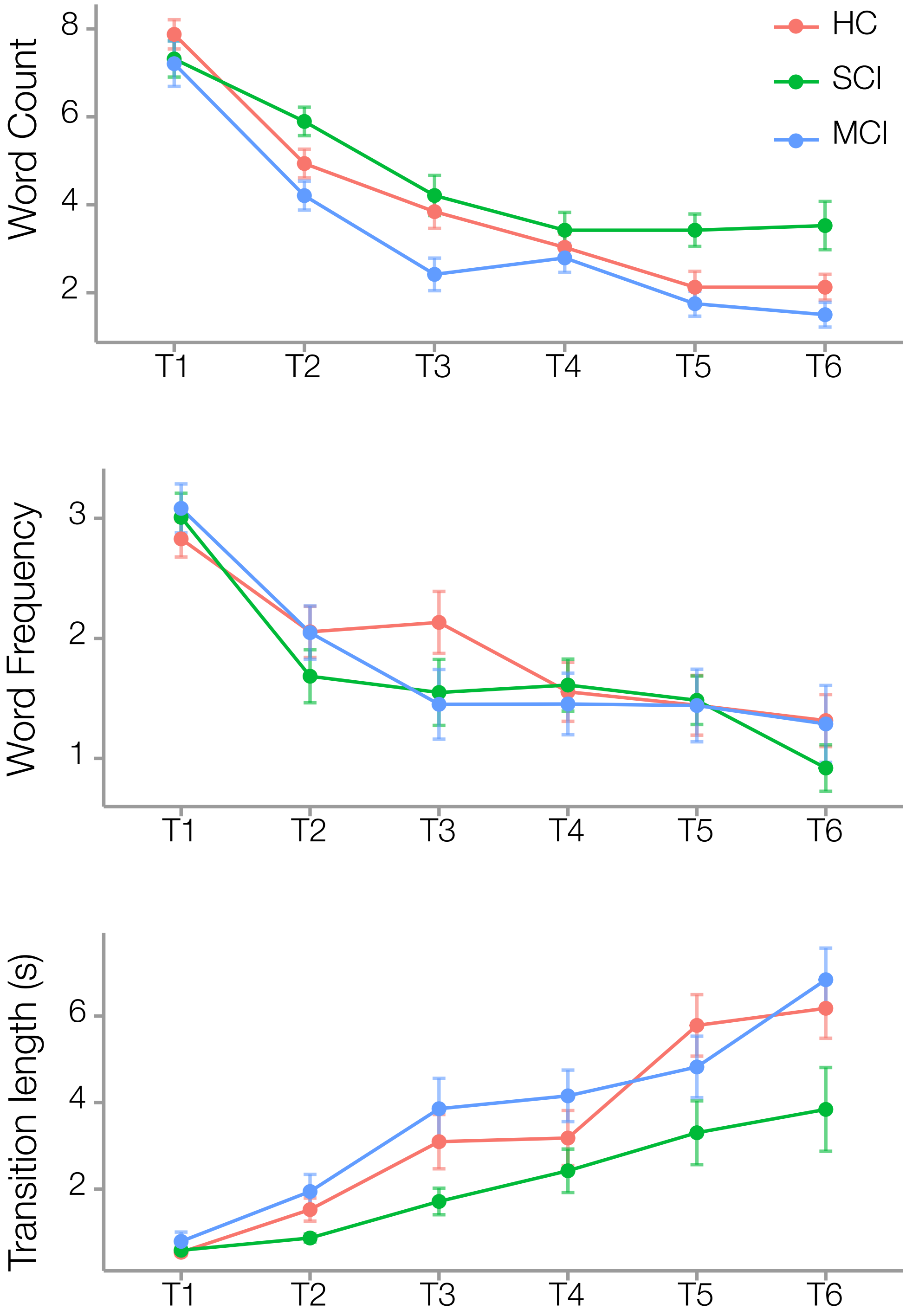


Figure 3.3: Word Count, Word Frequency and Transition length by time interval and for each group separately. Error bars display standard error.

($F(2,72) = 2.59, p = 0.08$) or the number of temporal switches ($F(2,72) = 1.64, p = 0.2$) as a function of diagnosis are found.

3.3.3.3 Temporally resolved measures

Word count, lexical word frequency and transition times by 10 second intervals is visualized in Figure 3.3 and the results of linear mixed random effects models are presented in Table 3.5.

A general decline in the word count for each time interval is visible and reflected in the model, regardless of diagnostic group. A significant effect for age is present, implicating that higher age leads to a reduced word count. For the SCI group, there is a significant interaction between the diagnostic group and the decline in WC_{T2},WC_{T5} and WC_{T6}. In these intervals, the decline of the SCI group is less severe. The MCI diagnostic group shows a significant interaction with the decline in WC_{T3}, with a stronger decline in word count than the other groups.

For lexical word frequency, again, a significant decline over time is visible, regardless of diagnostic group, which means that participants produce more common words at the start of the task, and less common words towards the end. Older participants produce words that are significantly more frequent. The MCI group has a significant interaction with WF_{T_3}, indicating this group uses lower frequency words in this time interval.

Starting from the third interval, a significant increase in word transition times is visible. A significant interaction between the SCI group and the fifth and sixth interval, indicates the SCI group shows significantly lower transition times in these intervals.

3.3.3.4 Correlation analysis

Spearman correlations between the word count by time interval, neuropsychological scores and demographic information is displayed in Figure 3.4. Only significant correlations are displayed.

Significant positive correlations between the BNT score and the word count in the last three time intervals are observed. The WAIS Similarity score shows positive correlations with the word count of the last two intervals. Negative correlations are observed between TMT A and the second and third interval, as well as between age and these two

Variable	Estimate	t	95% CI	p-Value
$\mathbf{WC_{T_1-T_2}}$	-0.456	-6.196	[-0.529, -0.382]	$< .01$
$\mathbf{WC_{T_1-T_3}}$	-0.698	-7.898	[-0.787, -0.61]	$< .01$
$\mathbf{WC_{T_1-T_4}}$	-0.937	-8.681	[-1.046, -0.83]	$< .01$
$\mathbf{WC_{T_1-T_5}}$	-1.301	-8.675	[-1.452, -1.152]	$< .01$
$\mathbf{WC_{T_1-T_6}}$	-1.290	-8.690	[-1.439, -1.142]	$< .01$
Age	-0.011	-3.294	[-0.014, -0.008]	$< .01$
Education	-0.003	-0.411	[-0.010, 0.004]	.68
SCI	-0.086	-1.128	[-0.164, -0.010]	.26
SCI x T				
SCI x $\mathbf{WC_{T_1-T_2}}$	0.247	2.161	[0.133, 0.361]	$< .03$
SCI x $WC_{T_1-T_3}$	0.155	1.102	[0.014, 0.296]	.27
SCI x $WC_{T_1-T_4}$	0.180	1.068	[0.012, 0.349]	.29
SCI x $\mathbf{WC_{T_1-T_5}}$	0.543	2.738	[0.345, 0.742]	$< .01$
SCI x $\mathbf{WC_{T_1-T_6}}$	0.575	2.959	[0.381, 0.770]	$< .01$
MCI	-0.041	-0.602	[-0.111, 0.028]	.55
MCI x T				
MCI x $WC_{T_1-T_2}$	-0.088	-0.724	[-0.210, 0.034]	.47
MCI x $\mathbf{WC_{T_1-T_3}}$	-0.383	-2.176	[-0.559, -0.207]	$< .05$
MCI x $WC_{T_1-T_4}$	-0.015	-0.089	[-0.189, 0.158]	.93
MCI x $WC_{T_1-T_5}$	-0.101	-0.396	[-0.354, 0.153]	.69
MCI x $WC_{T_1-T_6}$	-0.299	-1.046	[-0.585, -0.013]	.30

(a) Word Count

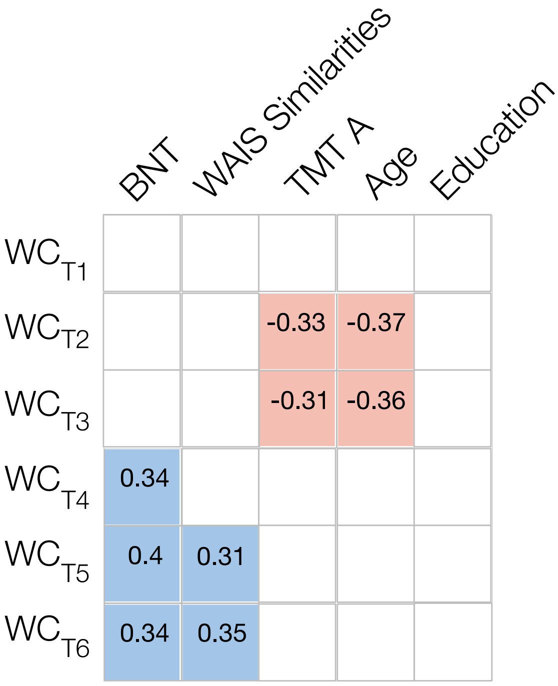
Variable	Estimate	t	95% CI	p-Value
$\mathbf{WF_{T_1-T_2}}$	-0.774	-2.558	[-1.077, -0.472]	$<.05$
$\mathbf{WF_{T_1-T_3}}$	-0.696	-2.298	[-0.999, -0.393]	$<.05$
$\mathbf{WF_{T_1-T_4}}$	-1.274	-4.208	[-1.577, -0.971]	$<.01$
$\mathbf{WF_{T_1-T_5}}$	-1.386	-4.578	[-1.689, -1.083]	$<.01$
$\mathbf{WF_{T_1-T_6}}$	-1.514	-5.000	[-1.816, -1.211]	$<.01$
Age	0.023	2.600	[0.014, 0.032]	$<.05$
Education	0.000	0.003	[-0.018, 0.018]	0.99
SCI	0.228	0.642	[-0.127, 0.582]	.52
SCI x T				
SCI x $WF_{T_1-T_2}$	-0.549	-1.108	[-1.045, -0.053]	.27
SCI x $WF_{T_1-T_3}$	-0.763	-1.539	[-1.259, -0.267]	.12
SCI x $WF_{T_1-T_4}$	-0.123	-0.248	[-0.619, 0.373]	.80
SCI x $WF_{T_1-T_5}$	-0.138	-0.279	[-0.634, 0.358]	.78
SCI x $WF_{T_1-T_6}$	-0.575	-1.159	[-1.071, -0.079]	.25
MCI	0.193	0.588	[-0.135, 0.521]	.56
MCI x T				
MCI x $WF_{T_1-T_2}$	-0.261	-0.564	[-0.723, 0.202]	.57
MCI x $\mathbf{WF_{T_1-T_3}}$	-0.936	-2.025	[-1.399, -0.474]	$<.05$
MCI x $WF_{T_1-T_4}$	-0.356	-0.769	[-0.818, 0.107]	.44
MCI x $WF_{T_1-T_5}$	-0.256	-0.554	[-0.719, 0.206]	.58
MCI x $WF_{T_1-T_6}$	-0.282	-0.610	[-0.745, 0.180]	.54

(b) Word frequency

Variable	Estimate	t	95% CI	p-Value
$L_{T_1-T_2}$	0.986	1.460	[0.311, 1.662]	.15
$\mathbf{L_{T_1-T_3}}$	2.557	3.786	[1.882, 3.233]	$<$ .01
$\mathbf{L_{T_1-T_4}}$	2.641	3.911	[1.966, 3.317]	$<$ .01
$\mathbf{L_{T_1-T_5}}$	5.245	7.766	[4.570, 5.921]	$<$ .01
$\mathbf{L_{T_1-T_6}}$	5.641	8.352	[4.965, 6.316]	$<$ .01
Age	0.028	1.029	[0.001, 0.055]	.31
Education	-0.074	-1.355	[-0.129, -0.019]	.18
SCI	0.311	0.365	[-0.541, 1.163]	.72
SCI x T				
SCI x $L_{T_1-T_2}$	-0.703	-0.635	[-1.81, 0.404]	.53
SCI x $L_{T_1-T_3}$	-1.429	-1.291	[-2.536, -0.322]	.20
SCI x $L_{T_1-T_4}$	-0.803	-0.726	[-1.910, 0.303]	.47
SCI x $\mathbf{L_{T_1-T_5}}$	-2.528	-2.284	[-3.634, -1.421]	$<$ .05
SCI x $\mathbf{L_{T_1-T_6}}$	-2.384	-2.154	[-3.490, -1.277]	$<$ .05
MCI	0.22	0.281	[-0.564, 1.004]	.78
MCI x T				
MCI x $L_{T_1-T_2}$	0.167	0.162	[-0.865, 1.198]	.87
MCI x $L_{T_1-T_3}$	0.510	0.494	[-0.522, 1.542]	.62
MCI x $L_{T_1-T_4}$	0.724	0.702	[-0.308, 1.756]	.48
MCI x $L_{T_1-T_5}$	-1.212	-1.175	[-2.244, -0.18]	.24
MCI x $L_{T_1-T_6}$	0.41	0.397	[-0.622, 1.441]	.69

(c) Transition Length

Table 3.5: Linear Mixed Random Effects model examining the effects of time interval, diagnosis, age and education on one of three variables, while controlling random effects per subject. Significant values ($p < .05$) are indicated in bold.



	BNT	WAIS Similarities	TMT A	Age	Education
WC_{T1}					
WC_{T2}			-0.33	-0.37	
WC_{T3}			-0.31	-0.36	
WC_{T4}	0.34				
WC_{T5}	0.4	0.31			
WC_{T6}	0.34	0.35			

Figure 3.4: Spearman correlation between 10 second word count (WC) intervals and neuropsychological test scores. Only significant correlations are shown. Positive correlations in blue, negative ones in red.

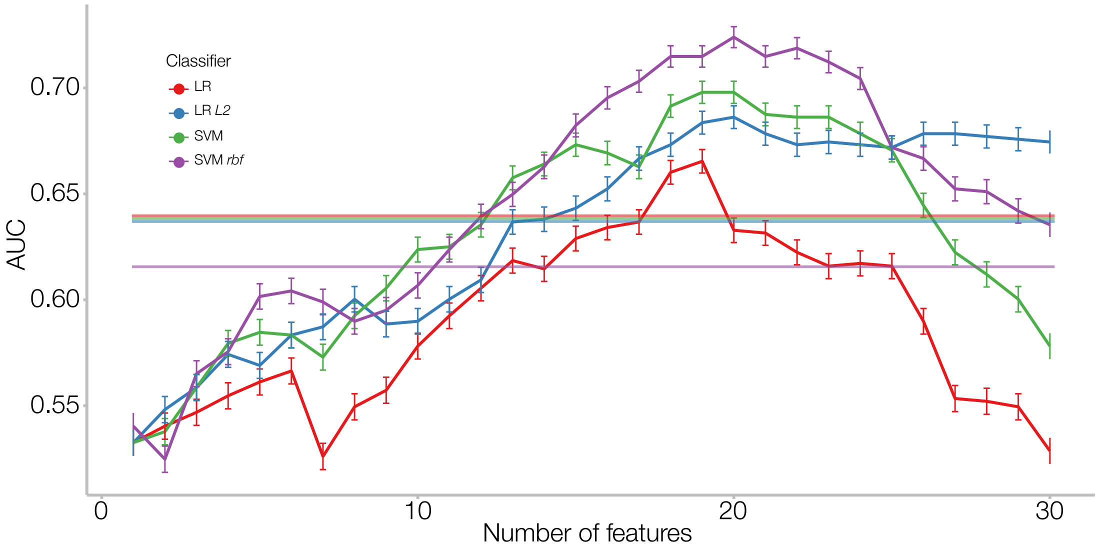


Figure 3.5: AUC of different classification models separating HC and MCI, plotted against number of features selected through univariate feature selection. Horizontal lines show the performance of models solely trained on the word count. Error bars indicate standard deviation of performance.

intervals (for the TMT A a lower score indicates a better performance).

3.3.3.5 Prediction

Figure 3.5 displays the results of the machine learning experiments. AUC is plotted, while varying the number of features chosen in feature selection, using different classifiers.

The baseline performances of models using just the word count is $AUC = 0.64$ for LR, both with $L1$ and $L2$ loss, and the linear SVM. The SVM with an rbf kernel only achieves $AUC = 0.62$ with the word count feature. Generally, the models using all features outperform the baseline. The best performance of $AUC = 0.72$ is observed for a linear SVM with 20 features. Generally, the linear and rbf SVM and the LR with $L1$ loss show similar performance patterns, across all number of features. The LR with $L2$ shows steadily increasing performance. The SVM with rbf kernel outperforms the other models with a lower number of features.

3.3.4 Discussion

Reviewing the overall performance on the SVF, a significant difference in word count was found between the groups, but no differences in cluster size or number of temporal clusters. The temporally resolved measures showed that the MCI, SCI and HC group follow similar trends with regard to word count, word frequency and transition length: word count and word frequency generally decrease over time, while average transition times increase. Significant differences between the MCI group and the other two groups were found mainly for the third interval, where the participants in the MCI group produce fewer and less frequent words. For the word count, this is in line with previous findings from [113], and the lower word frequency in the third interval indicates that persons with MCI have to resort to low frequency words earlier in the task, switching from semi-automatic retrieval of more common words to effortful retrieval at an earlier point than the other groups.

The persons with SCI showed an increased word count in the second, fifth and sixth interval, and reduced transition times in the fifth and the sixth interval. This suggests that they were able to sustain a continuous production for longer. The words they produced in the last intervals did not differ in frequency from the other groups, but the persons with SCI seemed to have access a larger store of words. Participants in the SCI group had a longer education than the general population, and one possibility is that the participants with SCI in the Gothenburg MCI study perform better because of higher premorbid functioning [102].

Correlation analysis with additional psychometric data lends a deeper understanding of the results, and significant correlations showed that higher BNT and WAIS similarities scores were associated with a higher word count in the latter part of the SVF. This suggests that having a broader vocabulary, as measured by the BNT, predicts a higher word count in the second half of the SVF. When reviewing the word count graph in Figure 3.3 and comparing the groups, it is evident that the ability of participants with SCI to sustain performance in the later time intervals can be explained by the access to a larger vocabulary as measured by the BNT. Age and TMT-A both show significant negative correlation with the second and third time intervals of the SVF. TMT-A is a measure of processing speed, and it decreases with increasing age. A decrease in processing speed seems to specifically inhibit production in the second and third interval. [96] suggested a semi-automatic retrieval phase at the beginning and a more effortful retrieval at the end of the task. Our findings support the notion of these phases occurring over the course of task, where the first phase is more influenced by processing speed and the later benefits more strongly from a larger vocabulary.

The benefits of temporal analysis were apparent in the increase of the ability to correctly classify participants as HC or MCI, compared to a classification based solely on word count. In the best case, the performance of the SVM with rbf kernel improved from $AUC = 0.62$ to $AUC = 0.72$ with temporal analysis.

This section introduced a novel, interval-based temporal analysis method for SVF tasks. The resulting outcome revealed distinct patterns that differentiated the groups: persons with SCI had a higher word count and sustained lexical frequency level during the last intervals, while persons with MCI had a steeper decline in both word count and lexical frequencies during the third interval. Correlations with neuropsychological scores suggested that the superior performance of the SCI group could be attributed to vocabulary size. Classification results improved when adding the novel features ($AUC = 0.72$), supporting their diagnostic value. This increase over the baseline performance underlines the value of using novel methods in addition to clinical standards. The results of group comparisons and correlations are in line with previous findings about phases of production in SVF. The special role of the third time interval in discriminating MCI patients is also supported by previous research.

The previous sections have looked at diagnosis as a classification problem. In the next section, SVF data will be used to to predict diagnostic scores of healthy, MCI and AD patients.

3.4 Predicting Dementia Screening and Staging Scores From Semantic Verbal Fluency

In order to quantify dementia's severity and prepare for its potential impact on a patient's environment, staging and screening tools have been developed. The Clinical Dementia Rating scale represents internationally the most widely applied staging tool for assessing the disease's global severity. It encompasses six domains of cognitive and functional performance: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care [174]. The assessment is conducted in the form of semi-structured interviews with the affected person and an affiliated person/co-interviewee, e.g., a family member.

The CDR is relatively time-consuming - interviews can take up to 90 minutes - depending on the availability of a co-interviewee and requires significant training of the raters in order to achieve good reliability [270]. The CDR is often used in combination with the Mini Mental State Examination, a common screening tool for dementia. It takes around ten minutes and requires a trained assessor, consists of a series of tasks that cover different forms of cognitive functions, such as memory and attention, and is designed to be used as a global screening tool. However, in some applications the MMSE lacks sensitivity; especially for early stages, its items are considered to be relatively easy and are highly likely to result in ceiling effects [372]. Moreover, it has been shown, that the standard MMSE might lack sufficient intra- and interrater reliability [267].

While there are many screening tools, a reliable diagnosis of probable dementia can only be made through in-depth assessments, and a comprehensive combination of behavioural (e.g., psychometric tests) and in vivo organic assessment (e.g., functional brain imaging). Behavioural assessments typically consist of structured interviews and can also include a number of well-defined tasks to assess particular aspects of cognition, such as memory and executive function. We argue that qualitative analysis of SVF allows for the deduction of corresponding dementia staging and screening scores which would allow to objectify and underpin CDR and MMSE scores, as well as to mitigate some of their afore-mentioned methodological caveats. In this Section, we present an analysis method that uses SVF data to predict two test scores, MMSE and CDR - Sum of Boxes, which has been used as a quantitative approximation of the CDR scale itself [285].

After introducing background information in Section 3.4.1, we outline our approach in Section 3.4.2. In Section 3.4.3, we compare regression models for prediction of the MMSE and CDR-SOB. We interpret predicted scores according to common clinical thresholds and report Cohen’s κ as a reliability measure. In Section 3.4.4, we discuss how our algorithm can be leveraged for medical human-computer interaction applications for dementia screening, and conclude by outlining further work.

3.4.1 Background

Background on the MMSE is provided in Section 2.3.3, on the CDR in Section 2.3.4. Both measures are compared in Section 2.3.5.

3.4.1.1 Diagnosis as a Classification Problem

The common approach for detecting signs of neurocognitive diseases from speech is to treat it as a classification problem, which is either binary or n-ary (with small n for a highly restricted number of potential diseases). The degree of manual intervention varies, from approaches that rely on manual transcriptions to a completely automated speech-based screening pipeline yielding significant discrimination results [374]. Work in this direction usually differs in means of the analysed corpora (free speech vs. cognitive tests vs. conversation), classification scenario (healthy vs. impaired or healthy vs. mildly impaired vs. severely impaired) and extracted features (linguistic vs. para-linguistic).

[128] worked on recordings of picture descriptions of the *Cookie Theft Picture Description Task*, extracted from the *DementiaBank* corpus [240]. They discriminate individuals with AD from healthy, age-matched, controls (HC) with an accuracy of 81% using linguistic and para-linguistic features. [392] uses language modelling techniques to calculate the perplexity of picture description tasks from *DementiaBank* to separate AD and HC individuals with an accuracy of 77.1%. [11] extracts para-linguistic features (e.g., pauses, pitch & jitter) of picture descriptions from *DementiaBank* to discriminate between AD and HC with an accuracy of 94.7%. [203] use para-linguistic markers from recordings of people performing different spoken cognitive tests (countdown, picture descriptions, sentence repetition and SVF) to classify individuals into three groups: early AD, MCI and HC. They train three binary classifiers with varying accuracies (HC vs. MCI: 20% ± 5; AD vs. MCI: 19% ± 5; HC vs. AD: 13% ± 3). [374] analyses spontaneous speech collected in a clinical setting through extracting temporal and para-linguistic features to separate HC from MCI patients. The resulting classifier yields an F_1 score of 86.2% and an accuracy of 82.4%.[254] extracted vocal features from a sentence reading task to discriminated between age-matched AD and HC patients with an accuracy of 84.8%. [413] uses phonetic features collected from a SVF and the East Boston memory test (EB) to discriminate between HC and MCI groups with an accuracy of 86.5%. Our own group previously extracted vocal features from cognitive tests

	HC	MCI	AD	MD	VD	Other	Total
N	42	47	33	37	10	10	179
Age	72.5 ± 8.3	76.6 ± 7.7	79.2 ± 5.0	78.8 ± 7.5	78.6 ± 4.6	78.1 ± 7.2	76.8 ± 7.5
Sex	8M/34F	23M/24F	12M/21F	19M/18F	8M/2F	10M/10F	80M/109F
MMSE	28.3 ± 1.6	26.0 ± 2.5	18.9 ± 5.0	18.5 ± 4.7	20.2 ± 4.1	23.7 ± 4.8	23.2 ± 5.5
CDR-SOB	0.48 ± 0.68	1.68 ± 1.11	7.52 ± 3.95	8.05 ± 3.31	5.50 ± 4.16	3.03 ± 3.63	4.02 ± 4.16

Table 3.6: Demographic data and clinical scores by diagnostic group (mean ± standard deviation). HC='Human control', MCI='Mild cognitive impairment', AD= 'Alzheimer's disease', MD='Mixed dementia', VD='Vascular dementia'.

(counting down numbers and Cookie Theft picture description) to identify patients with AD from HC with an accuracy of 89% ±3 [376].

3.4.1.2 Diagnosis as a Regression Problem

Neurocognitive diseases are complex and vary in their exact symptoms from person to person and from stage to stage. Therefore, it might be more useful to predict scores on screening or diagnostic tests than predicting a raw diagnosis. This makes it easier for clinical practitioners to integrate findings from an automatic analysis tool with the overall clinical picture, in particular when it comes to distinguishing between different potential causes for the same symptoms.

To our knowledge, there has been very little work on prediction of clinical scores from audio samples. [408] used semantic, acoustic and lexiosemantic features extracted from *DementiaBank* to predict MMSE scores. Using a bivariate dynamic Bayes net they achieved a mean absolute error of 3.83, which they improved to 2.91 for patients where longitudinal data is available. The topic has received more attention in the image processing community and multiple authors have predicted clinical scores from brain imaging features, e.g., average regional grey matter density and tissue volume of MRI [370, 418]. As an example, [173] uses a Random Forest Regressor to predict clinical scores, including the MMSE and CDR-SOB, based on imaging data. This leads to a best Mean Absolute Error of 1.68 for the MMSE and 0.69 for the CDR-SOB.

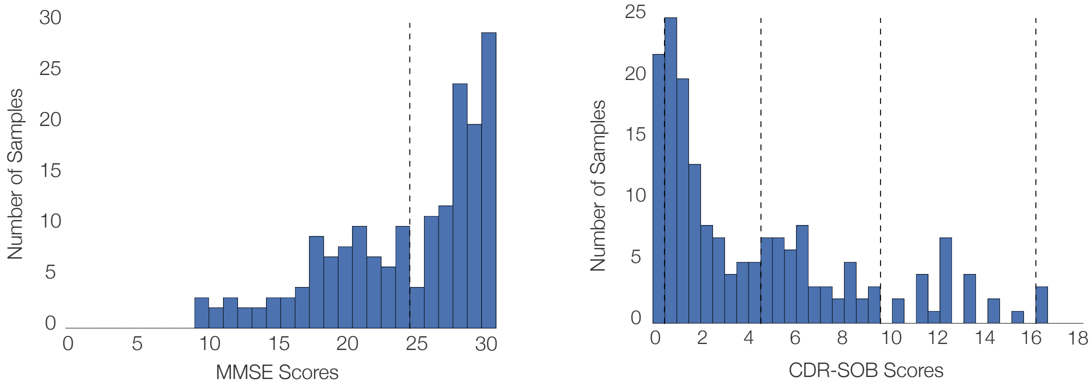


Figure 3.6: Histograms of MMSE and CDR-SOB scores with cut-off values for staging are indicated by dotted lines.

3.4.2 Methods

3.4.2.1 Data

The data used for the following experiments was collected during the *Dem@Care* [183] and ELEMENT [376] projects. All participants were aged 65 or older and were recruited through the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Speech recordings of elderly people were collected using an automated recording app on a tablet computer and were subsequently transcribed following the CHAT protocol [239]. Participants were asked to perform a battery of cognitive tests, including a 60 second animal SVF test. Furthermore all participants completed the MMSE and CDR. Following the clinical assessment, participants were categorised into three groups: Control participants that complained about having subjective cognitive impairment (SCI) but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI and patients that were diagnosed as suffering from Alzheimer's Disease and related disorders (ADRD). AD diagnosis was determined using the NINCDS-ADRDA criteria[250]. Mixed/Vascular dementia were diagnosed according to ICD 10 [406] criterea. For the MCI group, diagnosis was conducted according to Petersen criteria [310]. Participants were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behaviour. Demographic data and clinical test results by diagnostic groups are reported in Table 3.6.

The distribution of clinical scores in the data is shown in Figure 3.6. The left figure shows MMSE scores, which range from 0 (worst) to 30 (best). The most commonly used cut-off in the literature for possible dementia is 24. Other cut-offs include 17, 18, 19, 23, 25, and 26 [82]. Fewer than 10 of our participants fall below the lowest cut-off, while roughly half of them are below the traditional cut-off. The right figure shows CDR-SOB scores, which range from 0 (normal) to 18 (worst). Again, most subjects are staged as normal or having possible impairment, and only few have moderate or severe dementia.

3.4.2.2 Features

In the following we describe which features have been computed for each sample. We compute features from three different categories: Statistical Clustering and Switching, Word Frequency Features, and Vocal Features.

Let a_1, a_2, $\ldots$, a_n be the sequence of animals produced by patient p, with $a_i \in \mathbb{A}$ and $\mathbb{A}$ being the set of all animals.

Word Count

$$WC = n$$

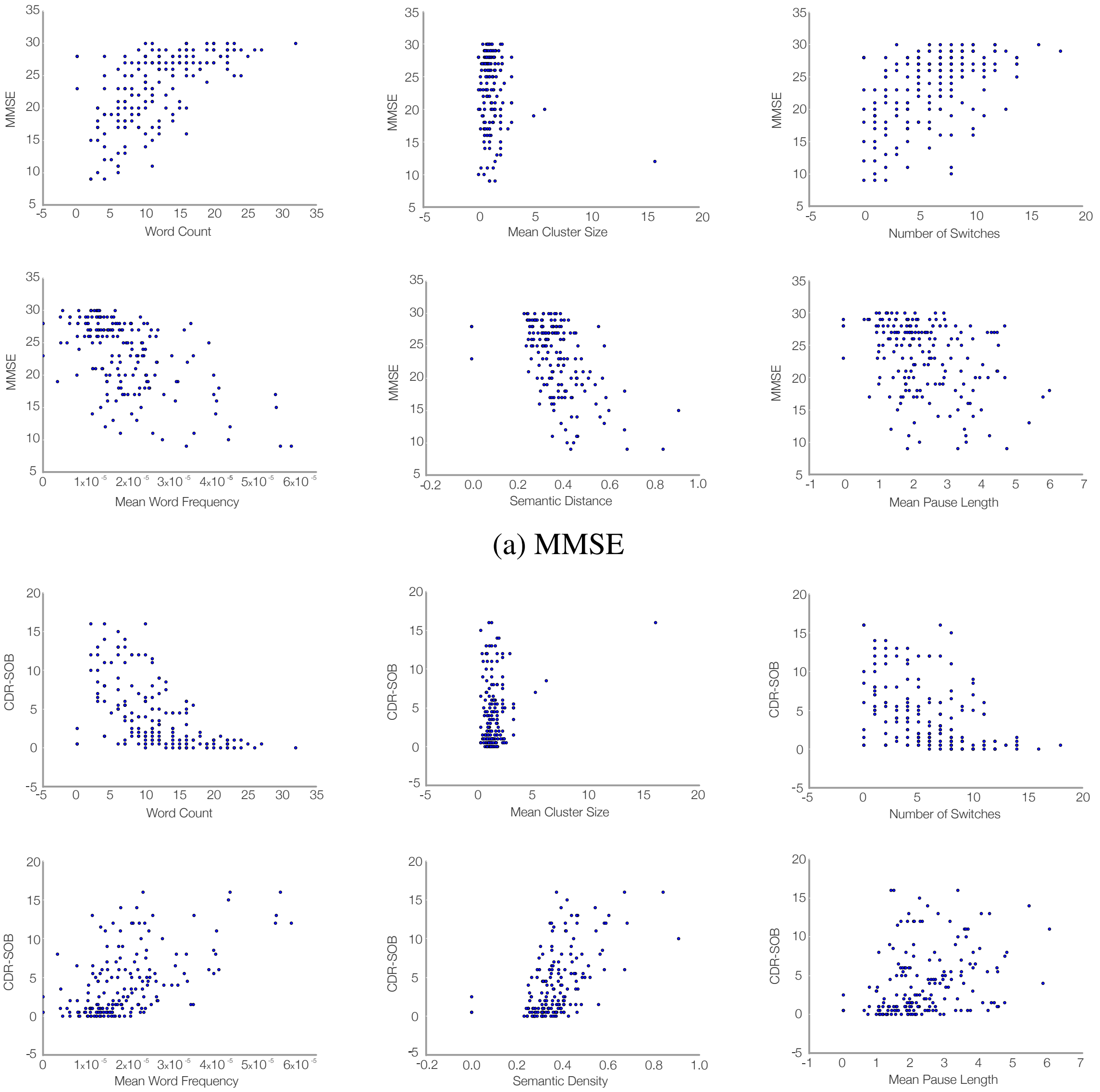
Statistical Clustering and Switching

We compute features based on semantic clusters, which are determined using the Equations 3.1, 3.3 and 3.4 from Section 3.1.2.3.

Let c_1,c_2, $\ldots$, c_m be the sequence of clusters, determined as described above and let $|c_i|$ be their size. We compute the following metrics:

Semantic Density

$$SD = \delta_p$$



(a) MMSE

(b) CDR-SOB

Figure 3.7: Visualisation of feature distribution in relation to MMSE (a) and CDR (b).

	MMSE	CDR-SOB	WC	MCS	NOS	MWF	SD	MPL
MMSE	1.000	-0.834***	0.602**	-0.176	0.486*	-0.560**	-0.552**	-0.352*
CDR-SOB		1.000	-0.569**	0.226	-0.464*	0.550**	0.553**	0.306*
WC			1.000	-0.123	0.838***	-0.514**	-0.538**	-0.398*
MCS				1.000	-0.335*	0.191	0.339*	0.006
NOS					1.000	-0.370*	-0.511**	-0.376*
MWF						1.000	0.642**	0.311*
SD							1.000	0.399*
MPL								1.000

Table 3.7: Pearson correlation of MMSE, CDR-SOB and computed features.

* $|\sigma| > 0.3$ ** $|\sigma| > 0.5$ *** $|\sigma| > 0.7$

Mean Cluster Size

$$MCS = \frac{1}{m}\sum_{i=1}^{m} |c_i|$$

Number of Switches

$$NOS = m - 1$$

Word Frequency

We approximate word frequency of animals using the Python *wordfreq* package [355], which combines resources such as Wikipedia, news and book corpora and Twitter. Let $f : \mathbb{A} \rightarrow \mathbb{R}$ be the function mapping a word to its frequency.

Mean Word Frequency

$$MWF = \frac{1}{n}\sum_{i=1}^{n} f(a_i)$$

Vocal Features

Let $p_1, p_2, \ldots, p_s$ be the pauses in the audio sample, determined using the Praat software [94] as intervals of absence of sound longer than 250 ms. Let $|p_i|$ be the length of a pause.

Mean Pause Length

$$MPL = \frac{1}{s}\sum_{i=1}^{s} |p_i|$$

3.4.2.3 Evaluation Criterion

For evaluation of the quality of prediction of regression models there are many different metrics. Popular for its mathematical sophistication and severe punishment for large errors is the *Root Mean Squared Error* (RMSE).

In our case the use of the *Mean Absolute Error* seems more appropriate. It delivers interpretable results on the real error made by the predictive model, scaled in the same way the clinical scores are. Let y_i be the actual value of sample i, let $\hat{y}_i$ be the regression models prediction and N the number of samples. The MAE is defined as

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{3.9}$$

In the following we will describe the results of regression models and discuss the implications of their predictions.

3.4.3 Experiments

In order to determine the importance of and relationship between MMSE, CDR-SOB and the computed features we examine correlations, reported in Table 3.7, and look at scatter plots of features and MMSE/CDR-SOB in Figure 3.7. Correlations smaller than 0.3 are considered as weak, greater than 0.5 as moderate and greater than 0.7 as strong. Both MCS and MCL have weak correlations to MMSE and CDR-SOB. Looking at their respective scatter plot, MCS does not seem to have any predictive power for either score, whereas the MPL seems to have at least some. Therefore, we exclude MCS from our

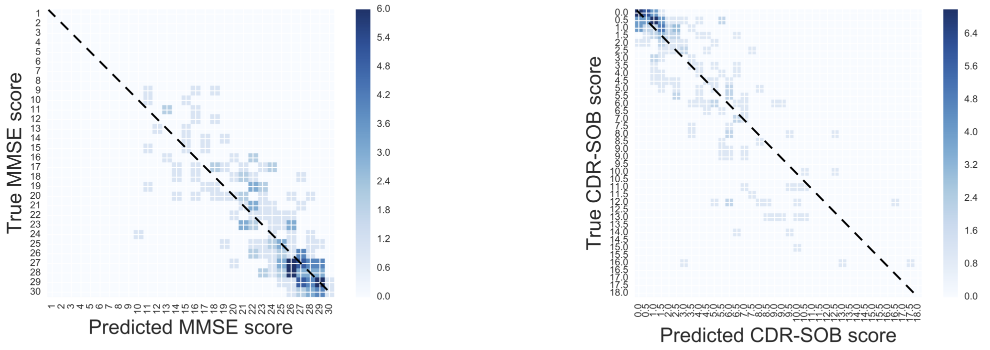


Figure 3.8: Confusion matrix for MMSE and CDR-SOB predictions, as heat-map, obtained using a SVR model and rounding predictions to the nearest scale values.

feature set for all further analysis. WC, MWF and SD have correlations greater than 0.5 with both MMSE and CDR. Inspection of their respective scatter plots shows a near linear relationship. To predict the CDR-SOB and MMSE, we train different regression models and evaluate their performance using MAE.

3.4.3.1 Prediction

Regression models are trained including Support Vector Regression (SVR), Lasso (Linear Regression with L_1 regularisation), Ridge Regression (Linear Regression with L_2 regularization), Elastic Net (EN) and a Random Forest Regressor (RFR). Their implementations are provided by the *scikit-learn* python framework [301] and all are trained with the features described in Section 3.4.2.2 excluding MCS. Features are normalised by subtraction of their mean and division through their standard deviation. Because of the small data set size (n=179) we can not use a separate validation/test set. Instead we rely on averaging multiple shuffled k-Fold cross validations, with k set to 5. Hyper parameters are determined using a cross validation based grid search on the training folds in each iteration of the outer cross validation loop.

Results of the regression are reported in Table 3.8. The RFR shows the worst performance of all tested regression models. For prediction of the CDR-SOB all other models (SVR, LR-L_1, LR-L_2, EN) show similar performance with overlapping 95% confidence intervals. For the MMSE the RFR also has the worst performance and the other regressors' performance is comparable again. Especially because of the small data set we are

	MAE_{MMSE}	$MAE_{\text{CDR-SOB}}$
SVR	**2.205** [1.920, 2.490]	**1.670** [1.433, 1.907]
LR - L_1	2.274 [1.988, 2.560]	1.683 [1.454, 1.912]
LR - L_2	2.289 [1.997, 2.581]	1.715 [1.485, 1.945]
EN	2.286 [1.993, 2.579]	1.688 [1.456, 1.920]
RFR	2.363 [2.073, 2.654]	1.728 [1.469, 1.986]

Table 3.8: Mean Absolute Error (MAE) and 95% confidence intervals for different regression models. Best performance indicated in bold.

	μ_{MMSE}	MAE_{MMSE}	$\mu_{\text{CDR-SOB}}$	$MAE_{\text{CDR-SOB}}$
SCI	28.244 ± 1.523	1.205 [0.905, 1.505]	0.489 ± 0.687	0.808 [0.589, 1.027]
MCI	25.679 ± 2.759	2.175 [1.678, 2.672]	1.708 ± 1.121	1.328 [1.030, 1.626]
DCI	18.914 ± 4.882	2.781 [2.311, 3.251]	7.556 ± 3.843	2.372 [1.955, 2.789]

Table 3.9: Mean Absolute Error (MAE) of MMSE and CDR-SOB prediction for a SVR model by diagnosis group.

not able to identify any clear best performing model.

In contrast to normal regression, our predicted value is bound to a discrete scale, we are able to draw a confusion matrix for each score by rounding predictions to the nearest value on the respective scale (1 steps for MMSe and 0.5 steps for CDR-SOB). Figure 3.8 shows the confusion matrices for MMSE and SDR-SOB using predictions from the SVR model. For predictions of the MMSE score, there seems to be an underestimation for patients with an MMSE > 24 and an overestimation for patients with MMSE ≤ 24. Predictions of the CDR-SOB are overestimating for a CDR-SOB ≤ 3 and underestimating for CDR-SOB > 3.

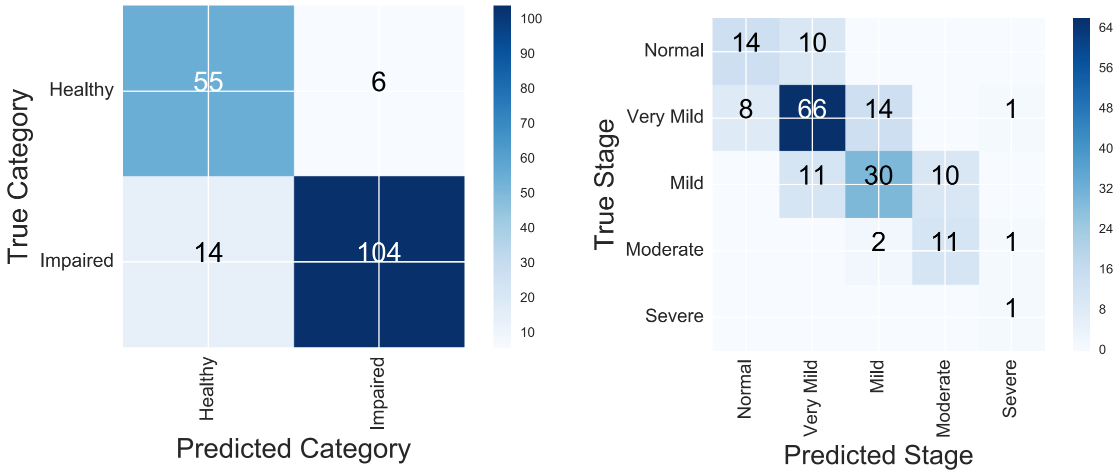
To better understand the results we examine the MAE by diagnosis group. We define three different groups: SCI, MCI and dementia (DCI). SCI and MCI are diagnosis groups appearing in our dataset. Anyone with a confirmed diagnosis of Alzheimer's disease, Vascular Dementia or Mixed Dementia is put into the DCI group. The results are listed in Table 3.9. At first glance it seems like the prediction error is growing with impairment of patients. But looking at the mean of each diagnosis group, one can observe that the standard deviation grows as well - meaning the values are spread further apart. This increases the complexity of the regression problem and accounts for the increased error.

3.4.3.2 Clinical Interpretation of Predictions

In practice, clinicians will interpret predicted scores relative to the interpretation framework they use for the actual tests. Therefore, we translated the continuous predicted test scores into categorical judgements and compared these judgments to those made on the original values using Cohen's unweighted κ [73] to measure agreement. For each case, we used the predicted value where the case was part of the test cross validation fold, not the training folds. A total of 179 cases with predicted CDR-SOB and MMSE values were available. κ was computed using the R package `psych`, Version 1.7.5.

$$\kappa = \frac{agreement_{observed} - agreement_{expected}}{1 - agreement_{expected}} \tag{3.10}$$

Since the predicted scores are continuous, we devised two strategies for mapping them



(a) (b)

Figure 3.9: Confusion Matrix for MMSE (a) categories and CDR-SOB (b) stages, as heat-map.

Cut-Off	κ
Strict	0.47 [0.36, 0.57]
Rounded	**0.52** [0.41, 0.62]

Table 3.10: κ for CDR-SOB staging, different cut-off strategies. Estimated value with 95% Confidence Interval. Best performance indicated in bold.

onto the discrete scores required for decision making. For the MMSE, we used a strict cut-off, where all values smaller than the boundary value indicate possible dementia, and a cut-off that rounds the predicted value to the nearest integer. For CDR-SOB, we used a strict cut-off that mapped values in between two category boundaries onto the category indicating less impairment, and a cut-off where values are rounded to the nearest 0.5.

Reliability for CDR-SOB is not very high—the best agreement is 0.52, and there is a lot of overlap in the 95% confidence intervals (Table 3.10). As the confusion matrix shows, this is due to a tendency to slip into the next higher or next lower category. While this does not seem critical at first, in clinical practice, misdiagnosis in either direction can be highly problematic [403].

Threshold	κ
17	0.59 [0.40, 0.78]
23	**0.76** [0.66, 0.86]
24	0.74 [0.64, 0.84]
26	0.64 [0.53, 0.75]

Table 3.11: κ for MMSE staging, rounding to nearest integer. Estimated value with 95% Confidence Interval. Best performance indicated in bold.

For the MMSE, however, agreement is much better. Depending on the threshold and the cut-off mechanism used, κ varies between 0.59 (95% CI [0.4, 0.77]) for a threshold of 17 and 0.76 (95% CI [0.66, 0.86]) for a threshold of 23. Table 3.11 shows agreement values for four thresholds, 17 (lowest), 23 (best), 24 (traditional), and 26 (highest), using the rounding strategy to match thresholds. As we can see from the confusion matrix, decisions based on the MMSE scores estimated by our approach would lead to slightly more people being screened.

3.4.4 Discussion

In principle, it is desirable to detect dementia at an early stage, so that the person with the disease and their family can take steps to maximise their quality of life. However, coming to terms with a diagnosis of dementia can be very difficult [64, 21]. Even if a person is referred for additional screening on the basis of a test such as the MMSE, and is found to be healthy, there can be negative consequences, such as people taking screening results less seriously, or becoming more anxious to bother their doctor for nothing [249, 403]. Therefore, once we have established that a machine learning approach has promise, we need to consider how it is best integrated into practice to avoid unnecessary harm.

While SVF clearly contains some information that can be useful when establishing the stage of a person's dementia, the most promising results are those for predicting MMSE scores. This makes sense clinically, as SVF does not reflect all of the dimensions on which people with dementia can be impaired, and the trajectory of decline can be very different depending on the person and the subtype of dementia they have. At the moment, for the MMSE, we achieve good agreements with traditional judgements using manual features. Problems might arise when automating the scenario. [294] saw the performance of their classifiers deteriorate when using ASR but this is likely to improve as ASR modules are specially developed for clinical data. Since administering SVF requires minimal training, this makes the test ideal for deployment in a telehealth scenario. Recordings of patients can be obtained by carers, case workers, social workers, and nurses, and they can take place in a quiet room in the patient's home or a convenient clinic room. After automatic analysis, the results can be sent automatically to the patient's General Practitioner and their specialist geriatrician or old age psychiatrist. It is even possible to fully automate the SVF test as part of an in-home kiosk or tablet app. However, for this use case, algorithms would need to be calibrated with additional training data, as people with moderate to severe dementia may find it difficult to follow the instructions of an automated app.

In this Section, we explored the possibility to predict MMSE and CDR-SOB scores based on linguistic and vocal features extracted from a SVF task. We were able to train a regression model with a MAE of 2.2 for the MMSE and 1.7 for the CDR-SOB. We discussed how these predictions could be used in clinical practice and that the agreement of MMSE predictions and real scores were high enough for a potential use as a screening tool. For predictions of the CDR-SOB the SVF task does not seem to capture all dimensions of impairment found in dementia. These promising results are first steps in the direction of formulating diagnosis and cognitive assessment as a regression problem. To additionally reliably predict severity of dementia progression, in-depth analysis of more than one cognitive test might be needed.

3.5 Summary

This chapter presented different novel approaches to the analysis of verbal fluency tasks on the basis of text transcripts to the end of automatically classifying dementia. Applications of these novel scores for prediction of dementia screening and staging scores were also explored.

Section 3.1 presented a novel approach to calculate clusters in semantic verbal fluency utilising neural word embeddings. The resulting clusters had high correlations, with previous taxonomic approaches and showed promising performance as input features in a classification task. Advantages, such as the generalisability and validity of this approach to their semantic categories than *animals*, were discussed.

Section 3.2 used language modelling techniques to model production strategies in SVF. LMs were trained in a Leave-One-Out fashion, on all three groups (HC, MCI, ADRD). Demented patients show significantly lower perplexity, thus are more predictable. Persons in advanced stages of dementia differ in predictability of word choice and production strategy - people in early stages only in predictability of production strategy. Consequently, unpredictability is an important factor for good performance in SVF.

Section 3.3 described experiments on Swedish verbal fluency data from healthy controls, MCI and subjective cognitive complaints, using a novel temporal analysis method. A general decline in word count and lexical frequency over the course of the task is revealed, as well as an increase in word transition times. Persons with subjective cognitive impairment had a higher word count during the last intervals, but produced words of the same lexical frequencies. Persons with MCI had a steeper decline in both word count and lexical frequencies during the third interval. Additional correlations with neuropsychological scores suggest these findings are linked to a person's overall vocabulary size and processing speed, respectively. Classification results improved when adding the novel features ($AUC = 0.72$), supporting their diagnostic value.

Section 3.4 applied the previously validated metrics to predict the standard dementia screening tool MMSE and the standard dementia staging tool CDR. A mean absolute error of of 2.2 for MMSE (range 0–30) and 1.7 for CDR-SOB (range 0–18) were achieved. True and predicted scores agreed with a Cohen's κ of 0.76 for MMSE and 0.52 for CDR-SOB.

Novel analysis methods introduced in this chapter, and their application in classification experiments, validated verbal fluency as a speech task that can be used to predict early stage dementia or MCI automatically. All experiments performed in these chapter used manually created transcripts of SVF. In real world applications, it is not always feasible to first have to transcribe a patients answers to this task. This is why Chapter 4 will explore the possibility to integrate these metrics into a fully automated pipeline using automatic speech recognition.

Chapter 4

Using Automated Speech Recognition to detect Dementia

The urgent need to identify a treatment that can delay or prevent AD has increased the number of preventional trials targeting disease modifying risk factors for which early screening of subjects at risk to develop cognitive impairment is highly relevant [10]. Recent research has shown that prevention at prodromal stages targeting disease mechanisms show promising results and are more likely to be effective [349]. Many challenges remain detecting these 'silent' stages, where clinical signs are not yet very obvious since our understanding of the pathological mechanism is still quite limited [106] and current tools may lack sufficient sensitivity to detect subtle but meaningful changes.

This approach has led to the current discussion on creating and approving more clinically relevant measures for early population based screening with low-cost tests of high sensitivity and lower specificity [100]. For instance, currently, just 50% of cases are diagnosed in Europe and the US [320]. This can be attributed to effective screenings for early signs of dementia (mild neurocognitive disorder) having not reached sufficient coverage. Especially in areas with low population density, clinical facilities and experts are too distributed to screen populations effectively, as this is still done in a face-to face manner today. Many clinical trials suffer from high drop out rates partly due to visit frequency and study length [149]. This translates into a medical supply resource problem and highlights the opportunities for telemedicine applications.

It has been put forward that new tools may address this need fast, require neither laboratory setup nor external material, and automatically evaluate and indicate potential clinically relevant persons. Therefore, research should focus on innovative computerized tools that reveal robust psychometric properties for early detection of neurocognitive disorder significantly decreasing the workload of expert clinicians, which represent a very rare resource in most cases. Thus, automatic, inexpensive and remote solutions allowing a broad frontline screening of cognitive abilities in the general population should be developed.

There is growing evidence for the feasibility of automatic speech analysis in addressing exactly this need [216, 168, 374]. Speech-based solutions can be remotely administered via telephone and therefore have minimal technical user interface requirements. This makes them a very attractive solution in the mentioned frontline screening context.

Neuropsychological studies comparing a video and telephone based psychometric dementia screening with a face-to-face assessment, reported good ecological validity for the telemedicine application [278]. However, such studies do not fully exploit the combined opportunities of telemedicine neuropsychological screening empowered by automatic speech analysis and machine learning classification.

This chapter focuses of the automatic analysis of Verbal Fluency tasks, on the basis of transcripts produced by automatic speech recognition. In Chapter 3, different analysis methods for VF tasks were explored. Good results were achieved, both in the classification of individuals with dementia from controls and in the separation of different dementia stages, both through classification and regression of diagnostic staging scores. The goal of the chapter is twofold: (1) in a clinical scenario, to provide further analysis of these tests, and (2) as a screening tool, to automatically determine people in need of medical attention.

ASR can be seen as a mature technology, that is offered as a service for integration into many different consumer applications. However, ASR as a statistical system still produces errors. Since acoustic models in ASRs are hardly ever trained on older individuals, performance is generally worse for them [176]. Furthermore, VF tasks are very unusual settings that might be confusing to language models, seeing as they are trained to give higher probabilities to word sequences that occur in natural speech. All this posses the question in how far fully automatic analysis of VF is actually feasible, since even small errors in ASR might obstruct important information about production strategy.

4.1 Fully Automatic Speech-Based Assessment of Semantic Verbal Fluency

Any analysis of SVF data that goes beyond word counts is too time consuming for daily clinical practice, especially for general practitioners and family physicians, who are typically the first point of contact for people who suspect that they or one of their loved ones has a cognitive impairment. In addition, any analysis strategy that is based on fixed, pre-defined categories is open to subjective judgement. This might explain some of the variation in cluster sizes and switch counts reported in the literature [380, 279, 139].

While automatic analysis introduces its own systematic biases, it is objective, replicable and yields almost immediate results for clinicians to act on. In this Section, we describe an automated analysis method for the fine-grained analysis of SVF data in terms of clusters and switches and validate it for the category of animals. Clusters and switches, determined by the tool correlate well with clusters and switches that were determined manually using a strict annotation procedure. Both manually and automatically derived statistics were successful in distinguishing between healthy controls, people with mild cognitive impairment and people with ADRD.

4.1.1 Methods

4.1.1.1 Recruitment

Within the framework of a clinical study carried out for the European research project *Dem@care*, and the EIT-Digital project *ELEMENT*, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University hospital in Nice, France. The Nice Ethics Committee approved the studies (ID RCB Dem@care ID RCB 2012-A00175-38, ELEMENT ID RCB 2017-A01896-45). Each participant gave informed consent before the assessment. Speech recordings of participants were collected using an automated recording app which was installed on an iPad. The app was provided by researchers from the University of Toronto, Canada and the company Winterlight Labs.

4.1.1.2 Clinical Assessments

Each participant underwent an assessment including: MMSE, phonemic verbal fluency (letter 'f'), semantic verbal fluency (animals), and the CDR. Following the clinical assessment, participants were categorised into three groups: Control participants (HC) that complained about having subjective cognitive impairment but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI, and patients that were diagnosed as suffering from Alzheimer's Disease and related disorders. For the AD group, the diagnosis was determined using the NINCDS-ADRDA criteria [250]. Related mixed / vascular dementia was diagnosed according to the ICD 10 [406]. For the MCI group, diagnosis was conducted according to Petersen criteria [310]. Participants were excluded if they had any major hearing or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behaviour.

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinicians who operated the mobile application. For the Dem@care data, the vocal tasks were recorded with an external microphone attached to the patients shirt and for the ELEMENT data, with the internal microphone. Instructions for the vocal tasks were pre-recorded by one of the psychologist of the center ensuring standardised instruction over both experiments. Administration and recording were controlled by the application and facilitated the assessment procedure.

Relevant demographic characteristics of the HC group (n = 40, age 72.65 years, MMSE 28.27, CDR-SOB 0.47), the MCI group (n = 47, age 76.59 years, MMSE 26.02, CDR-SOB 1.68), and the ADRD group (n = 79, age 79 years, MMSE 18.81, CDR-SOB 7.5) are presented in Table 4.1. The total number of participants was 166. Healthy controls were younger than the MCI and ADRD population. 75% of the HC population was female, compared to 50% of the MCI and the ADRD sample. The AD sample also had fewer years of education.

4.1.1.3 Speech data processing and transcription

Recordings of patients were analysed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances

	HC	MCI	ADRD
N	24	47	24
Age	76.12 (4.41)	76.59 (7.6)	77.7 (3.99)
Sex	5M/19F	23M/24F	8M/16F
Education	10.50 (4.05)	10.81 (3.6)	9.75 (4.69)
MMSE	28.21 (1.82)	26.02* (2.5)	18.83* (4.99)
CDR-SOB	0.46 (0.67)	1.68* (1.11)	7.5* (3.7)

Table 4.1: Demographic data and clinical scores by diagnostic group; mean (standard deviation); HC='Healthy control', MCI='Mild cognitive impairment', ADRD= 'Alzheimer's disease and related disorders'. Significant difference ($p < 0.05$) from the control population in a Wilcoxon-Mann-Whitney test are marked with *.

following the CHAT protocol [239] and aligned the transcriptions with the speech signal using PRAAT [48]. For the automatic transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google's ASR service, which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

4.1.1.4 Features

Word count was defined as the number of animal names produced minus the number of repetitions.

Clusters were determined based on statistical word embeddings, a commonly used technique in computational linguistics, which is discussed in detail in [223]. Mean cluster size was computed as the average number of words per cluster, and the number of switches was the number of clusters - 1.

4.1.1.5 Prediction

In order to evaluate the feasibility of the automatic approach, we performed two analyses that aimed to replicate existing results in the literature on differences in semantic verbal fluency performance between people with no impairment, mild neurocognitive impairment/MCI, and major neurocognitive impairment/AD [279, 357]. The first used a staging approach using validated normative data provided by [357], and the second used machine learning classifiers.

Automatic norm-based neurocognitive evaluation For simulation of a real world clinical application scenario, word counts from manual and automatic transcripts were compared using normative data for SVF. First, normative equations [357] were used to determine a z-value, based on manual word counts, age and education level, and people were staged in accordance with diagnostic categories of DSM-5 ($z > -1$ = no impairment, $z > -2$ = minor impairment, $z <= -2$ = major impairment). In a second step, people were staged using the normative equations, based on automatic word count, age and education level. The first staging was considered the ground truth and the second was compared to the first using classification metrics.

Machine learning based classification To give an idea of how the collected features could be combined to make a diagnostic decision, a ML classifier was trained. Each person in the database was assigned a label relating to their diagnosis (HC, MCI and ADRD). The features described in Section 4.1.1.4 were used, either calculated from automatic or manual transcripts, depending on the scenario. All features were normalised using z-standardisation.

In all scenarios we use SVMs implemented in the scikit-learn framework [301]. 10-fold cross validation was used for testing. In this procedure the data is split into 10 equally sized subsets (”folds”). For each of the folds, the classifier is trained on the 9 remaining folds and evaluated on the held out fold. To find a well-performing set of hyperparameters, parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration was performed.

Performing cross validation on small data sets only once, leads to performance fluctuations between different iterations. To work around this problem, cross validation was performed multiple times consecutively and then the mean of all performance metrics was calculated.

4.1.1.6 Performance Measures

The performance of ASR systems is usually determined using Word Error Rate (WER) as a metric. WER is a combination of the mistakes made by ASR systems in the process of recognition. Mistakes are categorised into substitutions, deletions and intrusions. Let S, D and I be the count of these errors respectively, and N be the number of tokens in the ground truth. Then

$$WER = \frac{(S + D + I)}{N} \tag{4.1}$$

We only calculated WER for words describing animals, not for off-task speech, which also occurs in our data. We refer to this metric as VFER (Verbal Fluency Error Rate).

As performance measures for prediction of each class in the ML classification experiment, sensitivity, specificity, accuracy, F_1 Score and AUC are reported.

4.1.2 Results

4.1.2.1 Automated Speech Recognition

Evaluation of all samples in the corpus yielded a VFER of 19.4 %. Since not all types of errors might have the same impact on analysis (e.g. word count is not influenced by substitutions), the proportion of types of error made are considered. 75.9 % of all errors were deletions, 13.3 % were substitutions and 10.8 % were insertions.

4.1.2.2 Correlation

The relationship between features extracted from automated transcripts and manual ones was examined. Since one is the prediction of the other, their relationship is linear. Consequently, Pearson's correlation coefficient was computed. All relationships are reported in Table 4.2. The correlation between manual and automatic SVF analysis was strong across all three relevant features with a correlation of r = 0.90 for the main clinical feature in this task, the word count.

		Manual		
		WC	*MCS*	*NOS*
Automatic	*WC*	0.921^{***}	0.436^{*}	0.547^{**}
	MCS	0.404^{*}	0.862^{***}	-0.161
	NOS	0.612^{**}	-0.142	0.797^{***}

* $|\sigma| > 0.3$ ** $|\sigma| > 0.5$ *** $|\sigma| > 0.7$

Table 4.2: Spearman correlation of automatically and manually computed features. WC='Word Count', MCS='Mean Cluster size', NOS='Number of switches'

		ACC	*SENS*	*SPEC*	F_1	*AUC*
HC vs. ADRD	Manual	0.882 ± 0.042	0.898 ± 0.065	0.728 ± 0.137	0.882 ± 0.042	0.822 ± 0.069
	Auto	0.877 ± 0.043	0.914 ± 0.058	0.664 ± 0.134	0.873 ± 0.042	0.784 ± 0.072
HC vs. MCI	Manual	0.756 ± 0.079	0.816 ± 0.108	0.612 ± 0.139	0.758 ± 0.078	0.708 ± 0.093
	Auto	0.778 ± 0.075	0.846 ± 0.102	0.624 ± 0.139	0.779 ± 0.073	0.736 ± 0.079
MCI vs. ADRD	Manual	0.770 ± 0.046	0.867 ± 0.078	0.362 ± 0.136	0.772 ± 0.046	0.643 ± 0.077
	Auto	0.793 ± 0.048	0.861 ± 0.073	0.485 ± 0.137	0.795 ± 0.048	0.696 ± 0.081

Table 4.3: Classification results. (± 95% confidence interval); HC='Healthy control', MCI='Mild cognitive impairment', ADRD= 'Alzheimer's disease and related disorders'; ACC='Accuracy', SENS='Sensitivity', SPEC='Specificity', AUC='Area under the curve'

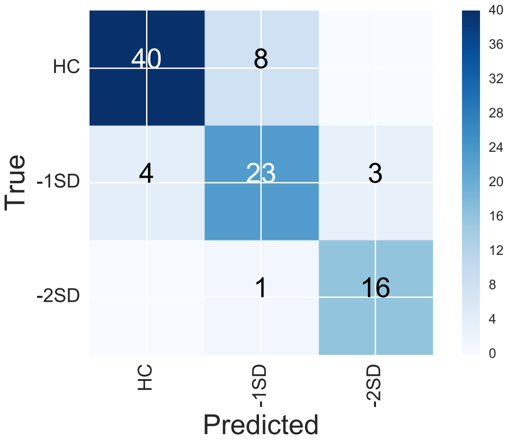


Figure 4.1: Confusion matrix for diagnosis based on normative data, automatic WC and manual WC; (no='$z > -1$ no impairment', minor='$z > -2$ minor impairment', major='$z <= -2$ major impairment')

4.1.2.3 Automatic norm-based neurocognitive evaluation

Neurocognitive disorder evaluations (no impairment, minor and major impairment) determined with the automatic word count, agree with labels based on the manual WC with an accuracy of 0.783, weighted precision of 0.81, weighted recall of 0.78 and F_1 of 0.78. When looking at sensitivity and specificity in a one versus all scenario, using HC as the negative class, the model achieves a sensitivity of 0.988 and a specificity of 0.736. A detailed confusion matrix is depicted in Figure 4.1.

4.1.2.4 ML automatic diagnosis classification

Classification measures for all scenarios are reported in Table 4.3. Classifiers trained on automatic measures perform as well as ones trained on manual features.

4.1.3 Discussion

In this research, we set out to investigate whether fully automatic analysis of the SVF task can be (1) considered as reliable as the manual one, (2) can be used for automatic qualitative assessment of neurocognitive impairment within this task and the corresponding domain and (3) in the end could be used as a valid fast and scalable screening tool, based on ML classification.

4.1.3.1 Automated Speech Recognition

Considering the reliability of the fully automated pipeline, ASR is often considered to be the main limiting factor [374]. Our results show an overall low error rate of 19.4 % for the automated system, compared to the manual transcripts. This in itself represents an improvement over results of other authors using ASR systems for evaluating the SVF tasks [294, 216]. In line with previous research, more word errors are produced by the ASR for ADRD patients, compared to healthy subjects, which can be explained by age-related speech erosion. Closely looking at the types of errors, insertions and deletions are both problematic for further analysis. Both skew the raw word count, which still is the single most predictive performance indicator in SVF for dementia detection. Substitutions only affect qualitative measures such as the mean size of clusters and the number of switches between clusters, but do not effect the word count.

4.1.3.2 Automatic norm-based neurocognitive evaluation

Even though the ASR produced word errors, mainly deletions, which negatively affect the overall word count and thereby the main clinical measure of SVF, the correlation between the automated and manual systems is very strong, i.e. 0.90. This shows that although the ASR system introduces some errors, it does not greatly affect the overall clinical measure, since the errors are not correlated to cognitive status. In the first experiment, we benchmarked the automatic pipeline for a norm-based neurocognitive evaluation. The performed neurocognitive evaluation based on automatic word count agreed strongly with labels based on the manual word count. The confusion matrix (see Figure 4.1) shows that the automatic approach tends to systematically underestimate the performance of a person in the SVF task. This can be attributed to the deletions of the ASR. Thus, the automatic pipeline can be considered conservative, showing high sensitivity, which is of great importance to its use as a screening tool.

4.1.3.3 Automated ML diagnosis classification

No significant difference can be seen between the models trained on manual and automatic transcripts. In each scenario, the 95% confidence intervals overlapped. The difference of the previous experiment can be explained by the flexibility of ML models to learn decision boundaries, in contrast to pre-determined diagnostic norms. ML models are also able to accommodate the previously mentioned systematic errors of ASR.

A similar approach has been suggested by [71], studying the utility of an automatic SVF score for the prediction of conversion with the result that higher prediction accuracy was obtained with the classifiers trained on all scores, rather than on manual scores. Overall, it can be stated that using automatic analysis of the SVF task allows immediate access to reliable and clinically relevant measures such as the word count, switches and clusters. This is potentially useful for differentiating between deficits in either executive or semantic processing. The automation of recording, transcription and analysis streamlines test administration and ultimately leads to more robust, reproducible data.

In addition to the assessment of cognitive decline, these qualitative measures extracted from the SVF performances may be of great interest as well for other neurocognitive disorders affecting verbal ability and executive control such as frontotemporal dementia or primary progressive aphasia [41].

Costa et al. [78] states that we are far from having available reliable tools for the assessment of dementias, since one of the main problems is the heterogeneity of the tools used across different countries. Therefore, a working group of experts recently published recommendations for the harmonisation of neuropsychological assessment of neurodegenerative dementias in Europe with the aim to achieve more reliable data on the cognitive-behavioural examination. Automated speech analysis of the SVF could be one potential tool to assist in harmonising test procedures and outcomes. It also provides additional quantitative measurements extracted from speech signals for cognitive screening without increasing time, costs or even workload for the clinician. Such a tool could be used as an endpoint measurement in clinical trials to assess intervention outcome and monitor disease progress, even remotely over the phone.

4.1.3.4 Limitations

A few limitations of this study should be considered. Significant differences in education level and age were found between our study populations. The age differences are sufficiently small not to have influenced SVF performances particularly under neutral instruction condition [345]. The data set for this study is only in French, thus, limiting transferability of its results to other languages. A major goal for future work is the collection of SVF recordings in multiple languages and within the framework of longitudinal studies.

4.1.3.5 Conclusion

To conclude, the study demonstrates the feasibility of automatic analysis of SVF performances in elderly people to assess and monitor cognitive impairment. Furthermore, new measures beyond simple word counts such as word frequencies could be investigated in the future, possibly giving way to a deeper understanding of underlying cognitive functions and changes due to neurodegenerative disease.

4.2 Telephone-based Dementia Screening through Automatic Analysis of Semantic Verbal Fluency

The goal of this section is to validate technology with which raw speech data can be processed via the telephone—facilitated by computational linguistic techniques and machine learning—in order to give a simple risk assessment for dementia. Instead of using free, unconstrained speech, we hope to achieve better performance and shorter assessment times, through analysing performances of cognitive tests.

Therefore we benchmark a solution processing raw telephone quality SVF data suitable for inclusion in a fully automated dementia frontline screening for global risk assessment.

4.2.1 Background

The following section gives an overview of efforts aiming at the automated detection of dementia based on multiple different sensor solutions. For this paper, we would like to differentiate between solutions based on classic *pervasive sensing* such as home monitoring systems and speech analysis as a special subcategory of pervasive sensing.

4.2.1.1 Automated Screening Based on Pervasive Sensing

Manifold research has been done into the feasibility of home monitoring systems for modelling domestic circadian activities (activity patterns following a biological 24h rhythm). As such rhythms are typically disturbed by dementia—especially nocturnal activity patterns—these techniques provide a useful basis for automatic dementia detection/screening. Using infrared sensors to monitor nocturnal activities, studies have found significant differences between dementia patients and healthy controls (e.g. [361]). Similarly, the same technical setup has been shown to effectively model daily routines [122]. Following the same rationale and technique [200] leveraged automatic detection of instrumental activities of daily living (IADL) in patients with MCI and healthy participants. Besides promising results, such studies are often carried out with very small sample sizes ($N < 50$) and focus mainly on the automatic classification of activities rather than the actual neurocognitive disorder. Moreover, the installation of home-monitoring systems require significant resources and a person's consent to be monitored in their private life; two issues that render such a solution unrealistic in broad population frontline screening.

Also focusing on circadian rhythm monitoring but using less complex wrist-worn technology, [290] found significant correlations between sleep patterns and common dementia staging scales. However, similar to the above-mentioned studies, sample size is relatively small and the main automatic analysis effort was spent on activity monitoring rather than prognostic classification problems.

Beyond such passive sensing approaches, there is also research on the diagnostic use of pro-active sensing situations: situations that are framed by some task/instruction producing more diagnosis related variance. Leveraging virtual reality technology, [367] used a realistic virtual reality (VR) fire evacuation task to predict amnestic Mild Cognitive Impairment; often considered as the precursor of dementia, Alzheimer's disease (AD) and controls from task performance reaching AUC values of more than 80%. Though very sensitive, the classification setup requires a lot intervention from technicians to analyse the VR task performance. Moreover, the VR screening setup has similar limitations as the classic neurological assessment: it requires the expensive VR laboratory and test persons have to leave their home.

Other studies combine gait and balance analysis through a hip-/foot-worn accelerometer and specific walking tasks [172, 70]. Such approaches take advantage of classic geriatric assessments showing age-/dementia-related gait irregularities when confronted with a simple straight-line walking task or dual task paradigms (e.g. walking and mental arithmetic task).

These pervasive sensing approaches reveal several *shortcomings* for our use case. They are either very technology-heavy, which implies significant investments, and rely heavily on activity recognition which represents an ongoing classification research challenge in itself. Alternatively, they have to be done in laboratories far away from peoples' homes. Conversely, automatic speech analysis recently has reached a technical readiness level that renders it very attractive for speech based pervasive solutions. Moreover, the only technical requirement is a working telephone which can be considered as ubiquitous in most countries even for an aged population such as the dementia screening target group.

4.2.1.2 Automated Screening Based on Speech

Authors have reported studies on automated dementia screening with possible applications in phone-based telemedicine scenarios. [376] extracted paralinguistic features from speech based cognitive tests and trained classifiers to discriminate between healthy controls and patients with AD. Furthermore, [216] used ASR to extract features from a story retelling task and was able to discriminate between MCI and healthy controls with an AUC score of 80.9%. [340] used four spoken cognitive tests (Countdown, Picture description, Repetition and SVF), extracted paralinguistic features to discriminate individuals with MCI, early AD and healthy controls (HC). Trained models achieve an accuracy of 87% for early AD vs. HC and 81% for MCI vs. HC. Not focusing on dementia detection but on Parkinson's Disease, [196] report an application which is phone-based and acts as a passive listener to monitor speech over time. However, as soon as an anomaly is detected the app also uses classic cognitive speech tasks to elicit richer and more controlled variance (i.e. a psychomotor task: continuously repeating *pa-ta-ka* during a given period of time)

Multiple studies report approaches that are less feasible in phone-based screening scenarios but provide strong evidence for the effectiveness of speech-based screening for dementia patients, including early stages. Overall, reported work either uses speech from conversations, spontaneous speech tasks, reading or repetition tasks, and fluency tasks.

The most liberal setting consists of conversations with clinicians. Audio files of spontaneous speech from conversations [99, 191], or classical autobiographic patient interviews [168] have been used in small setups, yielding significant effects. For such data, considerable effort has to be spent on preprocessing the data (e.g. annotating turns or trimming the audio file) in order to prepare it for further computational learning.

Tasks, eliciting spontaneous speech, are slightly more restricted and therefore easier to process; descriptions of the Cookie Theft Picture or comparable visual material, allows for extracting a wide variety of features and yields very good results [128, 11, 203, 288]. Similarly, some researchers report positive results from speech samples based on an animated film free recall task [146].

Reading or repetition tasks are the most handy to deal with, in the sense of automated processing, as they need little transcription and provide an inherent ground truth. Simple sentence reading has been shown to provide enough variance to effectively discriminate between AD and HC with an accuracy of 84% [254].

Verbal fluency tasks, such as the semantic animal fluency task, have produced rich variance to discriminate between AD patients and HC [203, 413, 223]. The benefits of semantic vs. phonemic fluency tasks have been discussed in multiple publications and there is a large body of neuropsychological evidence reporting dementia patients' difficulties in semantic fluency tasks, concluding that dementia patients and MCI patients have significant more difficulties in semantic, e.g., animal, fluency tasks compared to other psychometric standard tests.

In summary, speech analysis provides a powerful opportunity to broad dementia screening as it has minimal technical requirements and leverages a mature technology—ASR—and can be done remotely in almost all geographic areas. Sensitivity can even be increased through the use of specific psychometric speech tasks, such as the semantic verbal fluency task. Therefore, our aim is to benchmark an entirely automatic pipeline for dementia screening using telephone-quality audio recordings of a classic dementia screening speech task, ASR and machine learning classifiers on top.

	SMC	MCI	D
N	40	47	79
Age	72.65 (8.3)	76.59 (7.6)	79.0 (6.1)
Sex	8M/32F	23M/24F	39M/40F
Education in years	11.35 (3.7)	10.81 (3.6)	9.47 (4.5)
MMSE	28.27 (1.6)	26.02 (2.5)	18.81 (4.8)
CDR-SOB	0.47 (0.7)	1.68 (1.11)	7.5 (3.7)

Table 4.4: Demographic data and clinical scores by diagnostic group; mean (standard deviation); SMC='Subjective Memory Complaints', MCI='Mild Cognitive Impairment', D= 'Dementia', MMSE='Mini Mental State Examination', CDR-SOB='Clinical Dementia Scale - Sum of Boxes'.

4.2.2 Methods

In order to address the above-mentioned challenges, this section will elaborate on the technical pipeline of the proposed system and provide evidence for its feasibility. In the following, the telephone-based speech data processing and the machine learning experiment will be described.

4.2.2.1 Participants

Within the framework of a clinical study carried out for the European research project *Dem@care*, and the EIT Digital project *ELEMENT*, speech recordings were conducted at the Memory Clinic located at the Institut Claude Pompidou and the University hospital in Nice, France. The Nice Ethics Committee approved the study. Each participant gave informed consent before the assessment. Speech recordings of elderly people were collected using an automated recording app which was installed on a tablet computer. Participants underwent a clinical assessment including a battery of recorded speech-based tasks.

Each participant went through an assessment including: MMSE, the phonemic and semantic verbal fluency [371], and the CDR. Following the clinical assessment, participants were categorised into three groups: control participants that complained about having subjective cognitive impairment (SMC) but were diagnosed as cognitively healthy after the clinical consultation, patients with MCI and patients that were diagnosed with dementia (D), including AD. For the AD group, the diagnosis was determined using the NINCDS-ADRDA criteria [250]. Related mixed/vascular dementia was diagnosed according to the ICD 10 [406]. For the MCI group, diagnosis was conducted according to Petersen criteria [310]. Participants were excluded if they had any major audition or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behaviour.

Each participant performed the SVF task during a regular consultation with one of the Memory Center's clinician who operated the mobile application which was installed on an iPad tablet. Instructions for the vocal tasks were pre-recorded by one of the psychologist of the center ensuring a standardised instruction over the experiment. Administration and recording were controlled by the application and facilitated the assessment procedure.

4.2.2.2 Speech Data Processing

Speech was recorded through a mobile tablet device using the built-in microphone. The recordings were digitised at 22050 Hz sampling rate and at 16 bits per sample. To simulate telephone conditions, the recordings were downsampled to a 8000 Hz sampling rate, using the Audacity[1] software. Since the tablet device's microphone is used in mobile phones, no further transformations were applied.

Recordings of patients were analysed manually and automatically. For manual analysis, a group of trained speech pathology students transcribed the SVF performances following the CHAT protocol [239] and aligned the transcriptions with the speech signal using PRAAT [94]. For the automatic transcription, the speech signal was separated into sound and silent parts using a PRAAT script based on signal intensity. The sound segments were then analysed using Google's Automatic Speech Recognition (ASR) service[2], which returns several possible transcriptions for each segment together with a confidence score. The list of possible transcriptions was searched for the one with the maximum number of words that were in a predefined list of animals in French. In case of a tie, the transcription with the higher confidence score was chosen.

[1] http://www.audacityteam.org/

[2] https://cloud.google.com/speech/

4.2.2.3 Features

We extracted a variety of features from the generated transcripts. All hereunder reported features are either clinically accepted (i.e. word count), have been proven to have diagnostic power based on previous medical research (i.e. clusters and switches) or proved to have diagnostic power based on research in the field of computational linguistics (i.e. semantic metrics). Moreover, all features are firmly based on clinical research and therefore explicable and understandable by medical experts.

4.2.2.4 Word Count

The count of distinct correct responses (animals), excluding repetitions, is the standard clinical measure for evaluation of SVF. Its diagnostic power for even early stages of cognitive impairment has been shown in countless studies.

4.2.2.5 Clusters and Switches

Many previous researchers [379, 151, 323, 223] have shown that production in SVF is guided by so called clusters—clusters of words that are produced in rapid succession and often shown to be semantically connected. We determine clusters in multiple ways—taxonomy-based [379] and statistical [223] semantic, as well as temporal analysis [113]—and compute mean cluster size and number of switches between clusters as features.

4.2.2.6 Semantic Metrics

Many purely semantic metrics have been suggested for the analysis of SVF, that look at the type of words produced. We include frequency norms (see Section 3.4.2) estimated from large text corpora and computed as the mean frequency of any produced word and semantic distance (see Section 3.4.2) approximated using neural word embeddings trained on external text resources. We include the mean semantic distance between any produced word, the overall mean of means of semantic distances inside a temporal cluster and the the mean semantic distance between any temporal cluster.

4.2.2.7 Prediction

In order to evaluate the feasibility of using SVF in a telephone screening scenario, we performed a machine learning experiment. We built classifiers that discriminate the healthy population from the impaired samples. People were counted into the impaired population, when they belonged to either the MCI or dementia groups. First we established a performance baseline, training models based on features extracted from manual transcripts. After that we used the transcripts from ASR to extract features and constructed models.

In all scenarios we used SVM 2.4.3.2 implemented in the scikit-learn framework [301]. Due to our limited amount of data—166 samples—we could not keep a separate hold-out set for testing and instead used leave-one-out cross validation. For each sample, the data is split into a training-set—all samples but the one—and a test-set—the one held-out sample. The classifier is trained on the test set and evaluated on the held-out training set. To find a well-performing set of hyperparameters for the SVM (i.e., kernel, C, γ), we performed parameter selection using cross-validation on the training set of the inner loop of each cross validation iteration.

4.2.2.8 Performance Measures

The performance of ASR systems is usually determined using Word Error Rate as a metric. WER is a combination of the types of mistakes made by ASR systems in the process of recognition. Mistakes are categorized into substitutions, deletions and intrusions. Let S, D and I be the count of these errors and N the number of tokens in the ground truth. Then WER is calculated as described in Equation 4.1.

Since WER considers all utterances, including off-task speech which is not reflected in any of our features, we used a slightly adapted version. Instead of comparing the ground truth annotation of the recording and the ASR results, we transformed both into a list of animals and calculate the WER for these sequences. We refer to the result as the Verbal Fluency Error Rate (VFER) in further discussion.

As performance measures for prediction of each class in the ML classification experiment, we report the receiver operator curve, as different tradeoffs between sensitivity and specificity are visible. We also report AUC as an overall performance metric.

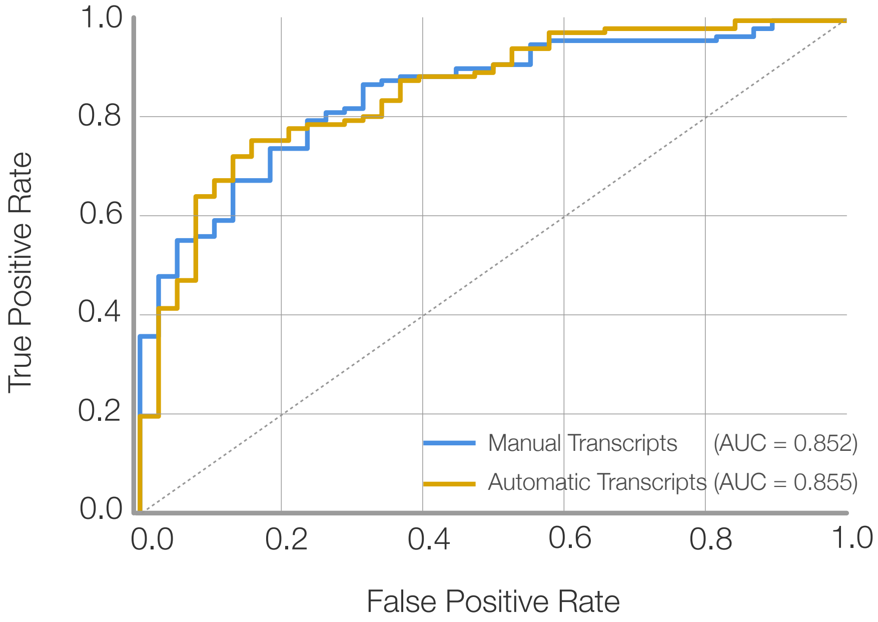


Figure 4.2: Receiver Operator Curve for features based on manual transcripts (green) and on automatic transcripts (red). AUC is reported in the legend.

4.2.3 Results

We first evaluate the VFER on the automatic transcript, which is determined to be 33.4%. Of the errors made by the ASR, 69% are deletions, 22% are substitutions and 9% are intrusions. Substitutions are the least problematic error, since they only skew the word count—the single most predictive feature—in rare cases, where a word is substituted with a previously named one.

Figure 4.2 shows the receiver operator curve—a plot of true positive rate vs. false positive rate, see Section 2.4.1—for both classification experiments. Models based on features extracted from manual transcripts have an AUC of 0.852 and models built on features extracted from automatic transcripts show an AUC of 0.855. Since a high sensitivity is key for screening applications, a sensible sensitivity-specificity trade-off for the automatic model could be at a sensitivity of around 0.85 and a specificity of 0.65.

4.2.4 Discussion

The results of our experiments show, that (1) the fully automated analysis of phone-based SVF is feasible for dementia screening, (2) the phone-based pipeline produces classification results comparable to the gold-standard manual transcription based classifiers and (3) the word error rate for the ASR approach is acceptable despite the reduced telephone bandwidth and the aged population.

In general, regarding screening scenarios, high sensitivity scores are important. Our classification experiment based on the fully automated pipeline shows a good AUC and for screening scenario a good sensitivity of 0.85 and decent specificity of 0.65. For achieving better specificity results, it may be necessary to include additional tasks, especially focusing on the differentiation of MCI and healthy controls. Nevertheless, this is not the main goal for broad screening, as false positives are less expensive for a health-care system than false negatives.

In our experiments, the automated ASR-/phone-based screening pipeline and the pipeline based on manually transcribed speech reach comparable classification results. This is very encouraging, as the transcription of speech is the number-one resource-straining factor, showing that an automatic speech-based system has become a powerful alternative to manual analysis of speech-based psychometric tests.

ASR is often considered to be the main weakness in speech based automatic screening approaches [374]. Our results show an overall error rate of 33.4 % for the automated system, compared to the manual transcripts. This result represents an improvement over results of other authors using ASR systems for evaluating the SVF tasks [294, 216]. In line with previous research, more word errors are produced by the ASR for dementia patients, compared to healthy subjects, which can be explained by age-related speech erosion. Considering the types of errors, insertions and deletions are both problematic for further analysis, as they skew the raw word count, the single most predictive performance indicator in SVF for dementia detection. Substitutions affect the word count less, only in rare cases, where a word is substituted with a previously named one, generating a false repetition.

In this Section we set out to benchmark a telephone-based analysis of SVF for inclusion into a fully automated dementia frontline screening for global risk assessment. Our results show that SVF is a prime candidate for inclusion into an automated pipeline, providing decent sensitivity and specificity scores. Additionally, we show that the phone-based classification is as effective as the gold-standard manual transcription based classifier displaying an acceptable ASR word error rate despite telephone setup and the aged sample for the experiments.

Further research will be directed into finding additional tests, that offer increased sensitivity and specificity in combination with SVF. The idea of this series is to validate and construct a system, that solely based on the telephone as a technological interface and administrable in less than 10 minutes, perfectly fits the need of broad dementia screening tools. It should also serve epidemiological research studies and inclusion for pharmaceutical trials, which aim at including representative shares of the population by cost-effective screening for persons with early onset neurocognitive impairments.

4.3 Summary

This chapter showed the potentiell applications of automated VF in clinical and automated screening scenarios.

Section 4.1 explored the applicability of automatic SVF analysis in a clinical scenario. Although a WER of 19.4% was present, manual and automatic analysis of SVF showed high correlations in the achieved word count. Comparing the two through usually applied clinical norms showed a potentiell for high sensitivity in the automatic approach, since the word count was mostly underestimated by ASR. In a classification scenario, no differences between classifiers trained on manual or automatic features were found in distinguishing between HC vs. AD and HC vs. MCI. Only for the case of MCI vs. AD, differences were visible with the manual approach outperforming the automatic one. This underlines the applicability of automatic SVF analysis for low-level screening with general practitioners.

Section 4.2 introduced experiments for the validation of an automatic SVF pipeline as a telephone-screening tool. Recorded speech was transformed to telephone-quality and transcribed using ASR. The WER increased drastically to 33.4%. Features form Chapter 3 were extracted from the newly created automatic and manual transcripts. Classifiers were trained to separate healthy controls and people suffering from different severity levels of dementia (MCI, ADRD). No significant differences between models trained on manually and automatically extracted features were found. A good AUC of 0.85 was achieved in both cases.

In summary, experiments in this chapter validated both the technical feasibility of automatic SVF analysis based on ASR technology, as well as its applicability in real world clinical assessment and non-clinical screening applications. Chapter 3 and Chapter 4 have focused on analysing a constrained neuropsychological task. The next chapter is going to focus on the multi- and cross-lingual detection of dementia based on more open tasks using transcripts and through acoustic analysis of speech.

Part III

Multi- and Crosslingual Methods for Dementia Detection

Chapter 5

Multi- and Crosslingual Methods for the Detection of Dementia from Picture Description Tasks

The fact that the elicitation task by itself has an effect on the dependent variables under investigation, i.e. that different tasks make different contributions to the assessment of different linguistic dimensions, rewards further consideration. Indeed, the task may even influence the most salient variables such as word retrieval [184] or perseverations [38] – both have widely been described as reliable indicators of cognitive impairment and a common characteristic of conditions such as AD. As tasks that test isolated language functions such as confrontation naming [37] neglect the full spectrum of everyday language performance, tasks to which the patient responds by producing connected language have been conceived as an ecological approximation of real language in use [52]. Research indeed suggests that picture description measures show weak correlations with measures obtained from the former testing paradigms, i.e. that picture description tasks may potentially help to shed light on additional cognitive processesses [272]. By design, these tasks rely on diverse cognitive processes and, at the same time, place a relatively low burden on episodic and autobiographical memory. Connected language analysis has its roots in developmental psychology and was originally concerned with language development in children [53]. Its study has received much scientific attention in AD research. Yet literature on MCI and the status of picture description tasks in MCI is comparatively small.

Comparing different elicitation methods, picture description tasks have several advantages: Unlike spontaneous speech or semi-spontaneous speech tasks such as story narration or interview [22], picture description tasks place restrictions on the context and the expected response. As a consequence, clear hypotheses can be stated with respect to the realizations to be observed along various linguistic dimensions. At the same time, these restrictions make responses comparable in between- and within-group settings. These empirical claims do not necessarily hold for (semi-)spontaneous speech in which subtle changes may not be apparent and less standardization is induced by the weak manipulation as given by the elicitation task [384]. From the patient's point of view, a cognitive support is provided by the picture stimulus itself and the task moderates well between heavier testing scenarios and completely free production prompts [135]. However, picture description tasks have one big disadvantage: From a clinical point of view, they require a significant amount of time as they rely on transcriptions and a multidimensional analysis. This becomes evident when turning to the measures that have been brought forward with respect to the *Cookie-Theft-Picture*.

Machine learning experiments using speech and language for the detection of dementia or related disorders have been conducted in many languages, including English [326, 259, 128, 23], French [376, 204], German [397], Hungarian [362, 387], Spanish [254], Greek [342], Swedish [233, 125], Japanese [348], Portuguese [17], and Mandarin Chinese [210]. Most of these studies acknowledge the small size of the data sets as a limitation of analysis, and describe the difficulties in gathering more data; these include the challenges in patient recruitment, the expense of running clinically-based studies, and the manual effort required for transcription and annotation. Here, we consider whether it could be possible to increase the amount of available data by augmenting a data set in one language by data from other languages and thus increasing prediction performance.

The previous Chapters 3 and 4 have focused on the automatic detection of early dementia/MCI through processing of speech and language generated during a constrained cognitive task, namely verbal fluency. This chapter uses a different, more open task, in which patients are shown an image and asked to describe what they see. The image used here is the Cookie Theft Picture from the Boston Diagnostic Aphasia Examination [143] (see Figure 2.5). This is mainly due to the fact, that for most languages it is the only available form of patient generated speech. First, background on the CTP task and related work is presented.

5.1 Cross-lingual Detection of Early Dementia from Speech

We examine English data from the DementiaBank corpus which is part of the TalkBank project [240], French data from the EIT-ELEMENT project [376] and Swedish data from the ALZ-RJ project [197].

Speech and language features are extracted from the speech samples in each data set, and through z-scaling features by language, the data sets are made comparable cross-linguistically. For each language, we first build language-specific classifiers, and then consider the effect of augmenting the training set with data from other languages, using domain adaptation. For each feature, predictability of dementia per language is considered and effects are compared across languages. Per language classification experiments are carried out, where data from other languages is used to augment the training set of each language. This work represents a first step towards the goal of multilingual dementia detection.

The Section is structured in the following way: Section 5.1.1 presents background information about how multilingual approaches have been realised in other NLP subdomains. Section 5.1.2 describes the data, extracted features and set-ups used in the classification experiments. Section 5.1.3 presents the results of the classification experiments and explores the relationships of extracted features in the examined languages. Section 5.1.4 closes the paper by discussing the results and their implications, as well as the limitations of the study and possible future directions.

5.1.1 Background

As mentioned above, several recent studies have used natural language processing and machine learning to analyse speech samples from people with dementia and other cognitive disorders. Most relevant, here, are those which focus on picture description tasks in English, French, or Swedish.

One of the most commonly-used corpora for this task is DementiaBank, which contains primarily English data. This large database of semi-openly available patient speech data is unique in the field. As a result, there have been a number of papers automatically classifying participants with and without AD on the DementiaBank Cookie Theft data [11, 409, 350]. Only a small subset (around 5%) of the DementiaBank participants are labelled as having MCI; [248] reported an F-score of 0.64 classifying MCI vs. healthy controls (HC), which was improved to 0.71 by including the AD data and using domain adaptation techniques.

Language analysis of Cookie Theft data from other sources has also been used to differentiate between different underlying pathologies in AD [325], and variants of frontal lobar temporal degeneration [295]. Other English-language work on detecting MCI/dementia using speech and language processing has focused on other tasks, such as immediate and delayed story recall [326], semi-structured interviews [176, 23], and conversation [259].

In French, picture description was one of multiple tasks used to elicit speech for the classification of participants with MCI and AD reported by [203] and [204]. In the first study, 15 controls and 23 MCI participants were distinguished with a best accuracy of 79% using purely acoustic features. In the second study, with a greater number of tasks contributing to the classification, 44 participants with MCI were distinguished from 56 participants with subjective cognitive impairment (i.e. a subjective sense of impairment in the absence of a clinically measurable cognitive deficit) with 86% accuracy. The features which distinguished the groups best on the picture description task were related to the duration of silent and voiced segments.

In other related French-language work, [376] achieved an accuracy of 89% distinguishing between AD participants and controls on the basis of speech features only, extracted from three story-telling tasks and a counting task.

In Swedish, [233] considered only syntactic features extracted from Cookie Theft narratives, and reported an F-score of 0.68 on the task of distinguishing MCI from control participants. [125] included a wider range of linguistic features, but also included cognitive test scores in their classification. A best result of $F = 0.81$ was achieved on the MCI vs. HC task using the combined feature types.

5.1.1.1 Multilingual classification

To our knowledge, there is no prior work on multilingual or cross-lingual dementia classification. However, since problems of unilingual data sparsity arise in other NLP domains, we briefly outline some previous work on multi- and cross-lingual classification more generally.

[391] used English sentiment analysis training data to augment the sparse available Chinese resources. Using a machine translation service, unlabelled Chinese data was transferred to English and labeled English data was transferred to Chinese. Finally, both views, original and translated, of a single datum were used for training a classifier using the co-training algorithm [47]. Although the single-language baseline was overcome, we do not consider translation to be an ideal approach, since subtle, yet important, information may be lost around word choice and sentence structure.

[318] also considered the task of cross-lingual sentiment classification, building on the 'structural correspondence' learning algorithm of [45]. This method involves identifying *pivot features* that generalize across languages, and then inducing correspondences between features across the languages, based on their correlations with the pivot features within languages. Using English as the source language, they were able to train sentiment classifiers in German, French, and Japanese, without any labeled data from the target languages.

Here, in contrast, we have labels for all the available data. One simple and popular method of supervised domain adaptation was proposed by [90]. The method involves augmenting the feature space with copies of each feature, such that one copy is specific to the target domain (i.e. the domain of the test set), one copy is specific to the source domain (i.e. the domain of the extra available training data), and one copy combines information from both domains. We use a similar approach here by considering each language to be a different domain, such that for each target language we have two additional source languages with which to augment the training data.

Note also that although we label each domain by its language (i.e. French, English, or Swedish) and we expect the differences between languages to be the main source of variation, there are other relevant differences between the data that must be taken into account by the adaptation procedure, such as recording conditions and participant demographics.

5.1.2 Methodology

5.1.2.1 Data collection

All data were collected as part of the projects referenced in Section 5.1, and detailed information about the protocols for each study can be found in the cited papers. The demographics for the participants in each language are shown in Table 5.1. The MMSE score is a global measure of cognitive health. In all studies, the participants were asked to perform the Cookie Theft picture description task in their respective languages. In English and Swedish, the image was shown on paper and speech was digitally recorded, while in the French study, the image was displayed on a tablet and speech was recorded via the tablet microphone.

	English		French		Swedish	
	HC	MCI	HC	MCI	HC	MCI
N	34	34	22	27	36	31
Age (years)	67.4 ± 6.9	68.4 ± 8.1	75.8 ± 7.4	75.6 ± 7.2	67.9 ± 7.2	70.1 ± 5.6
Education (years)	14.4 ± 2.7	15.7 ± 2.5	13.5 ± 2.3	13.0 ± 3.1	13.2 ± 3.4	14.1 ± 3.6
Sex (M / F)	26 / 8	26 / 8	6 / 16	13 / 14	13 / 23	15 / 16
MMSE (/30)	29.0 ± 1.3	27.9 ± 1.4	28.5 ± 1.5	25.4 ± 2.6	29.6 ± 0.6	28.2 ± 1.4

Table 5.1: Demographic data for each of the sub-corpora.

5.1.2.2 Feature extraction

The English and French audio samples were manually transcribed using the CHAT protocol [239]. The Swedish transcriptions were manually produced by a professional transcription company according to provided guidelines similar to the CHAT protocol. Thereafter, a wide variety of lexical and acoustic features were extracted using automated methods. In English and French, part-of-speech (POS) tagging and dependency parsing was done using the Stanford CoreNLP package [243]; in Swedish, the Sparv tool was used [50]. In all languages, acoustic features were extracted using a combination of Matlab [55], openSMILE [108] and the Syllable Nuclei Praat script [48, 94]. A detailed description of the extracted features and their motivation is given in Table 5.2.

Dependency distance	**Dependency distance** is a measure of the complexity of sentences, and has been shown to be significantly reduced in MCI participants on a story retell task [326]. We compute the mean dependency distance for each word in an utterance, and then calculate the mean, median, and maximum over all utterances.
Part-of-speech tags	**POS counts** are computed by mapping the POS tags to the Universal Tag Set [311] and then normalising the raw counts by the total number of words in the narrative. **POS ratios** are also computed in some cases. Based on the previous literature, we compute the ratio of nouns to verbs, pronouns to nouns, determiners to nouns, and open-class words to closed-class words [326, 5].
Lexical features	**TTR** is the type-token ratio, calculated by dividing the number of unique word types by the total number of tokens in the narrative. Vocabulary size may be diminished in early cognitive impairment [22]. **Word frequency** We compute the mean, median, and max frequency of all tokens, and of the relevant information-bearing words (below). Word frequency estimates are calculated using the `wordfreq` Python library, which is available in English, French, and Swedish [354]. People with early stage cognitive impairment have more difficulty naming low-frequency items [7, 193]

Information units	**Information unit occurrence** Using the list of information units from [83], we manually construct a set of synonyms for each information unit, for each language (e.g. in English, the boy can be referred to as *boy*, *son*, or *brother*, in French he can be referred to as *garçon*, *fils*, *fiston*, or *frère*, and in Swedish as *pojke*, *son*, or *bror*). For each of the 23 information units, we have a binary indicator for presence of each concept. Deficits in information content in the Cookie Theft Task have been observed in dementia [83, 117, 120] as well as MCI [5], and even in the asymptomatic stage [85] **Information unit ratios** are computed for each of the information units, by dividing the total number of times a given information unit is mentioned by the total number of words in the narrative. **Content density** and **information density** are computed by dividing the total number of information units mentioned (out of 23) and the total number of information unit counts by the total number of words in the narrative. **Content efficiency** and **information efficiency** are computed by dividing the total number of information units mentioned (out of 23) and the total number of information unit counts by the total time taken to produce the narrative.
Fluency features	**Narrative length** is the length of the narrative, measured in the total number of words produced (excluding filled pauses, but including unintelligible words and false starts), number of sentences, number of syllables, and total time. Cookie Theft narratives in AD tend to be shorter than control narratives [83].

	Sentence length is the length of each sentence, in words; we calculate the mean, median, and maximum. **Word length** is the character length of each word; we calculate the mean, median, and maximum. **Speech rate** is measured in words per minute (total words divided by total time) and number of syllables divided by duration. Persons with AD have been shown to have a lower speech rate [133]. **Articulation rate** is measured as number of syllables divided by phonation time. Persons with moderate to severe AD show lower articulation rates than healthy controls [168] **Average syllable duration** is measured as speaking time divided by number of syllables. **Pause features** We measure the mean duration of pauses, number of pauses, the pause rate (number of pauses divided by total time), the phonation rate (time spent in speech divided by total time), and the ratio of silent to non-silent segments. Speakers with word-finding difficulties may pause longer and more often as they think about what to say next [133, 215].
Acoustic features	**Fundamental frequency** (F_0) measures the number of periods per second in voiced segments of speech. We measure the mean, minimum, maximum and standard deviation of fundamental frequency, the smoothed fundamental frequency contour and the envelope of the smoothed fundamental contour. Voicing probability of the final fundamental frequency candidate offers a measure of confidence in the selected F_0 curve. Persons with AD tend to have smaller variations in fundamental frequency [119, 190].

	Jitter is a measure of the frequency variation between consecutive periods, and has been found to be a useful feature in distinguishing between persons with AD and healthy controls. [263, 11]
	Shimmer is a measure of the amplitude variation between consecutive periods. Persons with mild AD have been found to present with more shimmer in their voices than healthy controls [254, 263].
	Periodicity measures the regularity of the speech signal. Features measuring periodicity include the mean, maximum, and minimum cross-correlation. [203] found that periodicity measurements helped differentiate AD, MCI, and control participants in a picture description task.
	Intensity is measured directly via the PCM loudness, and indirectly via the logarithmic power of Mel frequency bands (log of the rate of energy release in the lower Mel-scaled frequency bands). Patient with dementia tend to have lower intensity and less control of airflow than healthy individuals [254]
	MFCC (Mel-Frequency Cepstral Coefficients) 0-14 are the discrete cosine transforms of logarithms of spectral power, and they are commonly used in speech recognition to separate the speaker-dependent characteristics from the linguistic information in the speech signal. MFCCs have been found to be useful in machine learning classifications of persons with AD and healthy controls [11, 15].
	Linear predictive coding (LPC) provides an estimate of the source of the speech signal. We measure statistics of the LPC coefficients as well as the line spectral pair (LSP) frequencies.

Table 5.2: Speech and language features extracted from the Cookie Theft narratives.

5.1.2.3 Cross-lingual prediction

As a baseline, classification models are built only on the target language. For each target language, classification models are also trained on combinations of data sets from different languages and evaluated on samples from the target language.

Leave-one-out cross validation is performed on the target language. Data normalisation is always applied inside each loop of cross validation, based only on the training set. To account for differences between languages, features are normalised inside a single language data set using z-scaling. This has the advantage that it scales features between languages while preserving effects between the MCI and HC groups inside a language.

For the cross-lingual case, they are then combined according to method described by [90]: the feature matrix X^t_{train} containing all training samples from the target language for this fold and X^t_{test} containing the test sample from the target language for this fold are expanded by a copy of themselves. The feature matrix X^s_{aug} containing all samples from a given source language is expanded by an equally sized matrix containing only zeros. Afterwards, both the expanded X^t_{train} and the expanded X^s_{aug} are stacked to create the training set. The first half of the resulting feature matrix can be seen as containing the set of *combined* features, the second half only the set of *target* language specific ones.

To help interpret which features transfer well between languages, logistic regression with L_1 regularisation is used as a classifier, implemented in the Python scikit-learn framework [301]. The L_1 regularisation has the effect of setting many of the feature weights to zero, therefore acting as a method of feature selection. The parameter C is chosen from the following set $\{0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$ using grid search with an inner loop of cross validation on the training set, optimising AUC. Only samples from the target language were used in test folds of this inner cross validation procedure. As an evaluation metric for final classification, AUC was recorded.

5.1.3 Results

5.1.3.1 Cross-lingual prediction

Classification results of cross-language prediction are reported in Table 5.3. The languages used in training and testing are given as rows. Each column shows different settings for augmentation of the training set. In every case, there is an augmented set-

	Baseline	+F	+S	+E	+SE	+SF	+ EF
French (F)	0.553	/	0.493	**0.580**	**0.703**	/	/
Swedish (S)	0.344	**0.477**	/	**0.524**	/	/	**0.641**
English (E)	0.640	**0.672**	0.636	/	/	**0.654**	/

Table 5.3: AUC score for different classification scenarios and languages. Rows correspond to languages tested on, columns to languages training sets are augmented with. '/' shows scenarios which are not possible. Improvements over the single-language baseline are indicated in bold.

ting that outperforms the baseline of training on the target language alone.

Swedish clearly shows the worst baseline performance – under the 0.5 AUC mark of random performance. French is just above the overall baseline and English shows decent performance. Augmenting training sets with data from other languages overall shows a positive effect. Adding either French or English data leads to clear improvements over the baseline. Adding Swedish data does not increase classifier performance. Adding data from *both* other languages leads to the best performance for French and Swedish, while in the case of English it is better to augment only with French. Adding English and French to Swedish data shows the highest improvement, nearly doubling performance.

5.1.3.2 Feature analysis

To examine which features drive the classification results in each case, we look at the weights assigned to each feature by the classifier. For each task, we first rank each feature according to the proportion of folds in which it is assigned a non-zero weight (i.e. selected), and then by the absolute value of the weight it is assigned, averaged across folds. The top 10 features in the unilingual case, and the case of training on all three languages, are given in Table 5.4.

The acoustic features generally dominate the feature rankings. The only linguistic features to be ranked in the top 10 in the unilingual cases are: the number of times the *woman* information unit is mentioned, divided by the total number of words (Swedish), the POS count for X, indicating an unknown or unintelligible word (English), and the binary indicator for the *sink* information unit (French). In the multilingual cases, the

Feature	Proportion	Weight
voicingFinalUnclipped_sma_quartile1	0.15	0.05
mfcc_sma_de[13]_skewness	0.15	0.03
mfcc_sma_de[3]_quartile2	0.15	0.03
ratio_concept_woman	0.15	0.02
F0final_sma_de_minPos	0.13	0.09
mfcc_sma[11]_stddev	0.13	0.08
lspFreq_sma_de[5]_kurtosis	0.13	0.08
logMelFreqBand_sma[3]_percentile1.0	0.13	0.07
lspFreq_sma[3]_linregc2	0.13	0.06
lspFreq_sma_de[4]_amean	0.13	0.06

(a) Swedish unilingual.

Feature	Proportion	Weight
voicingFinalUnclipped_sma_quartile1_C	1.00	0.28
median_word_length_C	1.00	0.26
mfcc_sma_de[6]_percentile1.0_C	1.00	0.21
mfcc_sma_de[4]_percentile99.0_T	1.00	0.18
mfcc_sma_de[0]_skewness_C	1.00	0.18
lspFreq_sma[5]_linregc2_C	1.00	0.17
lspFreq_sma[3]_minPos_C	1.00	0.15
mfcc_sma_de[4]_amean_C	1.00	0.12
logMelFreqBand_sma[7]_maxPos_C	1.00	0.11
voicingFinalUnclipped_sma_skewness_C	1.00	0.11

(b) Swedish augmented with English+French.

Feature	Proportion	Weight
mfcc_sma[10]_quartile3	1.00	0.24
mfcc_sma_de[7]_linregc2	1.00	0.18
mfcc_sma[10]_minPos	0.99	0.66
pcm_loudness_sma_de_maxPos	0.99	0.62
mfcc_sma_de[1]_minPos	0.99	0.60
mfcc_sma_de[6]_amean	0.99	0.59
ratio_pos_X	0.99	0.56
mfcc_sma[12]_quartile1	0.99	0.51
lspFreq_sma[6]_iqr1-2	0.99	0.50
mfcc_sma[4]_pctlrange0-1	0.99	0.50

(c) English unilingual.

Feature	Proportion	Weight
voicingFinalUnclipped_sma_quartile1_C	1.00	0.39
logMelFreqBand_sma[7]_maxPos_C	1.00	0.33
lspFreq_sma[3]_minPos_C	1.00	0.28
median_word_length_C	1.00	0.24
lspFreq_sma[5]_linregc2_C	1.00	0.23
lspFreq_sma_de[1]_quartile2_C	1.00	0.23
mfcc_sma_de[3]_quartile2_C	1.00	0.19
mfcc_sma_de[6]_percentile1.0_C	1.00	0.18
mfcc_sma_de[0]_skewness_C	1.00	0.18
voicingFinalUnclipped_sma_upleveltime75_C	1.00	0.16

(d) English augmented with Swedish+French.

Feature	Proportion	Weight
mfcc_sma_de[14]_quartile3	1.00	0.52
mfcc_sma_de[13]_iqr2-3	0.98	0.40
mfcc_sma_de[13]_quartile3	0.92	0.29
mfcc_sma_de[12]_quartile1	0.86	0.35
mfcc_sma[13]_linregc1	0.84	0.68
mfcc_sma_de[11]_skewness	0.84	0.65
lspFreq_sma[5]_maxPos	0.84	0.47
has_concept_sink	0.84	0.46
jitterDDP_sma_linregc1	0.84	0.40
logMelFreqBand_sma[4]_maxPos	0.84	0.40

(e) French unilingual.

Feature	Proportion	Weight
voicingFinalUnclipped_sma_quartile1_C	1.00	0.23
median_word_length_C	1.00	0.20
mfcc_sma_de[11]_skewness_T	1.00	0.18
mfcc_sma_de[0]_skewness_C	1.00	0.17
lspFreq_sma_de[1]_quartile2_C	1.00	0.17
mfcc_sma_de[14]_quartile3_T	1.00	0.17
voicingFinalUnclipped_sma_upleveltime75_C	1.00	0.15
lspFreq_sma[5]_linregc2_C	1.00	0.13
lspFreq_sma[3]_minPos_C	1.00	0.12
mfcc_sma_de[4]_amean_C	1.00	0.11

(f) French augmented with Swedish+English.

Table 5.4: Highly-ranked features in the unilingual and cross-lingual classification experiments. Features are ranked first by the proportion of folds in which they are assigned a non-zero weight and then, in the case of a tie, by the average absolute value of the weights, across folds. In the domain-adapted cases, the suffix _*C* indicates the feature contains data from the target and source data combined, while _*T* indicates a feature from the target dataset only.

median word length appears in all three experiments.

While the interpretation of the acoustic features is more difficult, we can observe broadly that the selected features tend to be related to the MFCC, LSP, and voicing features. To the extent that the voicing features indicate prosodic markers, these may provide a proxy for lexical emphasis or grammatical structure. By contrast, the relation of voicing (and MFCCs) to articulatory differences is not yet clear.

Looking at the proportion of folds in which the top-ranked features were selected helps illuminate the benefit of the cross-lingual training set. Particularly for Swedish (Table 5.4a and Table 5.4b), where the unilingual performance was very poor, we can see that there was very little consistency in the features selected across folds, leading to overfitting and poor generalisability. In the cross-lingual training scenario, the same features are selected in every fold. Another indicator of the success of the multilingual training set is that most of the top selected features are from the *combined* set (marked with 'C') rather than only from the *target* set ('T').

5.1.4 Discussion

Overall, the classification results are encouraging, as we are able to improve over the baseline for each language by including training data from other languages. However, even in the best-case scenario, the data sets are still small ($n \leq 183$) compared to modern machine learning problems, which may limit generalisability of results.

We were somewhat surprised to see that the most frequently-selected and highly-weighted features were almost exclusively acoustic. While other researchers have reported on the utility of such features for dementia detection [11, 15], a clinical assessment of speech for signs of cognitive impairment typically focuses on basic prosodic features (such as rate of speech) and linguistic markers. As a result, those features are better-supported by the medical and psychological literature. One benefit of the acoustic features is that they can be extracted using the same script in all languages, limiting the variability caused by using language-specific tools such as parsers and POS-taggers. However, we note that even in the unilingual case, the majority of the top-ranked features are acoustic.

These results also raise the question of interpretability: while we use a simple classifier and manual feature engineering, the resultant classifiers are not likely to be interpretable by a clinician in a downstream application. Further work is needed to better understand how these paralinguistic features relate to clinically observable symptoms. Additionally, success in achieving performance improvements through training set augmentation seem to be language—or at least data set—specific. Therefore no claims about transferability to other languages or resources can be made.

There are many sources of variability between our three subcorpora. The main difference, of course, is language. The z-scaling served to normalise feature values across languages (e.g., the ratio of determiners to nouns was much lower in Swedish than in English or French, due to morphological differences in how definiteness is expressed in those languages). Differences in feature extraction were also accounted for in the scaling (e.g., the word frequencies were, by necessity, computed from different corpora in the different languages). Potentially harder to account for are the minor differences in patient demographics and diagnostic procedures. If we consider the MMSE score as a proxy for the severity of cognitive decline, we can see that the mean MMSE in the Swedish MCI group is approximately 3 points higher than in the French MCI group. Therefore learning methods trained on one data set with lower MMSE might be sensitive to different forms of impairment compared to methods trained on data sets with higher MMSE. However, we consider that when augmenting a training set of one language with data from another, these effects of differences between corpora (i.e., MMSE, age) can be considered to have implicit effects of regularisation, perhaps actually increasing generalisability to the test data.

It is also possible that the nature of a picture description task might not be the most fitting for detection of MCI. Many studies have shown that features extracted from this task have high predictability in separating people with Alzheimer's disease from controls [128], which has lead to a comparably high number of available data and research. However, it would appear that many of the previously used features do not transfer to MCI, as—depending on the progression—key linguistic abilities are still preserved [365].

Finally, technical challenges aside, collaborations of this nature can be difficult due to the sensitive nature of the data, and the need to respect ethical guidelines and participant consent when sharing and storing data. With this in mind, we recommend to other researchers working in similar domains to consider from the outset whether their data could eventually be shared, and to make suitable provisions in their ethics protocols and participant consent forms. We look to DementiaBank as a model for this kind of data-sharing and openness, and hope that researchers can continue to find ways to share resources of this nature.

5.2 Class-based Language Modelling for cross-linguistic detection of Alzheimer's Disease

Here, we consider whether it could be possible to increase the amount of available data by augmenting a corpus in one language with data from another language, and thus improve predictive performance without the need for new data collection. Specifically, we consider augmenting a relatively small French dataset with a much larger English one. The two aims of this study are: (1) to identify a set of features that are both useful for the detection of dementia and that we expect to transfer across different languages, and (2) to improve classification results on the French dataset by augmenting the training set with English data.

5.2.1 Background

In contrast to the previous work on AD classification, we measure not only which information units are mentioned, but also the order in which they are mentioned. Our approach has some similarity to class-based language models [56], in which words are first grouped into classes (or clusters), and then the language model is trained on the classes rather than the individual words. One benefit to this approach is improved generalisability [169], and another is the ability of classes to span different languages [363]. In other applications, the biggest challenge in applying such methods has been the generation of appropriate word clusters, but here it is a natural extension to our procedure for extracting information units, which already maps words to concepts.

5.2.2 Methodology

5.2.2.1 Data

Data were taken from two corpora: a small French dataset ($n = 57$), collected at the Memory Clinic and Research Centre of the University Hospital Nice, and the Pitt sub-

	English		French	
	HC	AD	HC	AD
N	241	309	25	33
Gender	154F/87M	189F/120M	19F/6M	22F/11M
Age	64.8 (7.7)	71.4 (8.4)	75.4 (7.0)	79.2 (6.6)
Education	14.2 (2.6)	12.8 (3.0)	14.0 (2.6)	11.3 (4.0)
MMSE (/30)	29.1 (1.1)	19.8 (5.7)	28.6 (1.4)	18.9 (3.9)

Table 5.5: Demographics of participants, where AD indicates Alzheimer's disease, and HC indicates healthy control. The Mini Mental State Examination (MMSE) is global measure of cognitive status.

corpus of DementiaBank, containing 550 English samples[1]. Detailed information about the protocols for each study can be found in Tröger et al., 2017 [376] and Becker et al., 1994 [39]. In both cases, ethics approval for the data collection was obtained from the local governing bodies.

The demographics for the participants in each language are shown in Table 5.5. In both studies, the participants were asked to perform the CTP task in their respective languages. In English, the image was shown on paper and speech was digitally recorded, while in the French study, the image was displayed on a tablet and speech was recorded via the tablet microphone.

5.2.2.2 Features

The English and French audio samples were manually transcribed using the CHAT protocol [239]. A set of pre-defined information units found in the CTP was determined as an extension to [83], and is given in Table 5.6a. Mentions of information units were determined using keyword-spotting (based on manually-constructed word lists specific to each language), and used to translate the full narratives to sequences of information units. As an example, the English *A boy is standing on a stool* and French *Le garçon est sur un tabouret* would both be mapped to the sequence BOY STOOL.

[1] In this analysis, we included all participants in the Dementia subfolder, regardless of specific diagnosis, to maximize the size of the source data.

Actions STEAL, FALL, WASH, OVERFLOW, GIRL'S ACTION, WOMAN'S INDIFFERENCE
Actors BOY, GIRL, CHILD(REN), WOMAN
Places KITCHEN, EXTERIOR
Objects COOKIE, JAR, STOOL, SINK, DISHCLOTH, WATER, WINDOW, CUPBOARD, DISH, CURTAIN, COUNTER

(a) Information units.

has_*unit* Binary feature indicating presence or absence of each information unit (23 features)
ratio_*unit* For each information unit, the number of times that unit was mentioned, divided by the total number of words in the original narrative (23 features)
unique_concept_density Total number of information units which were mentioned at least once, divided by the total number of words in the original narrative (1 feature)
unique_concept_efficiency Total number of information units which were mentioned at least once, divided by the duration of the sample in seconds (1 feature)
total_concept_density Total number of words referring to information units, divided by the total number of words in the original narrative (1 feature)
total_concept_efficiency Total number of words referring to information units, divided by the duration of the sample in seconds (1 feature)

(b) *info* features

perplexity_*class*_*n*-gram The perplexity assigned to the sample by each of the eight language models, where $n = 2,3,4,5$, and the models are trained on data from either the AD or HC class. (8 features)
score_*class*_*n*-gram The log probability assigned to the sample by each of the eight language models. (8 features)
max_perplexity_*class*_*n*-gram The maximum perplexity, computed over all n-grams in a sample, for each of the eight language models. (8 features)
min_score_*class*_*n*-gram The minimum log probability, computed over all n-grams in a sample, for each of the eight language models. (8 features)

(c) *LM* features

Table 5.6: Top, the information units extracted from CTP narratives. Bottom, the *info* and *LM* features that are computed from the resulting sequence of information units.

Features relating to the occurrence of each distinct information unit comprise the *info* feature set, described in Table 5.6b. Additionally, new features are derived from language models build on the sequence of information units. To this end, concept-based language models are trained for English and French in a leave-one-out fashion, using the `kenlm` framework [156]. Models up to 5-grams were constructed. For each participant, two language models are constructed for each n: one trained on the healthy control

(HC) population and one trained on the AD population. The participant is left out of the model built on their associated diagnostic group. The trained language models are then applied to the held-out participant's sequence of information units and various language model features are extracted (Table 5.6c).

5.2.2.3 Unilingual prediction

To evaluate the performance of the three proposed feature sets (*info*, *LM*, and *info+LM*), we first train classifiers to distinguish between HC and AD participants within a given language. To examine the importance of certain features, we restrict ourselves to more explainable linear models, namely LR 2.4.3.1 and linear SVMs 2.4.3.2 [301]. In both cases, we use L_1 regularisation to promote sparsity in the feature weights.

AUC is reported as the evaluation parameter. Due to the small size of the French dataset, we use leave-pair-out cross validation, which has been shown to produce an unbiased estimate for AUC on small datasets [9], and has also been used in related work [326]. However, since LPO-CV is computationally very costly, we instead use 10-fold cross-validation (10-CV) for English, making sure that any samples for a given participant occur in either the training set or the test set, but not both. For LPO-CV we compute AUC and its standard deviation as described by [326]; for 10-CV we compute the AUC in each test fold and then report the average and standard deviation over folds.

Feature scaling and hyper-parameter optimisation is done on the training set in each fold. Features are scaled using Maximum-Absolute Scaling to preserve the binary nature of the *info* features. For both SVMs and LR, C was optimised between $C \in [10^{-4}, ..., 10^{4}]$ using a grid search.

5.2.2.4 Multilingual prediction

Our goal is to improve classification in French, by incorporating training data from English. To this end, we examine multiple ways to combine data from both English and French in the training set.

We first consider *domain adaptation*, where we treat French as the target domain and English as the source domain. We implement the AUGMENT method of [90], which involves augmenting the feature space with source-specific, target-specific, and combined versions of all the original features, allowing the classifier to assign a higher weight to the combined version when that feature transfers well across domains, while also retaining source- and target-specific information where appropriate.

We consider as well as the baseline methods described in [90]: WEIGHT, in which the samples from the source domain are assigned reduced weights in the model; PRED, in which the prediction made by the source classifier is used as an additional feature in the target model; LININT, in which the predictions from the source and target models are linearly interpolated; and ALL, in which target and source data are simply combined in a single training set. Due to the limited size of our data, we do not optimise the weighting factors in WEIGHT and LININT, but rather assume the two languages should be given equal importance, and use a weighting factor of 0.1 in WEIGHT (since the English data is 10 times the size of the French data), and 0.5 in LININT.

Another option is to combine the French and English datasets before extracting features. Specifically, we first replace the word-level transcripts with the sequence of information units, and then combine the two datasets and train the language models over the multilingual corpus, thus generating *multilingual language models*.

5.2.2.5 Cross-Lingual prediction

To understand how well a trained classification model in one language could be applied to another, we also perform cross-lingual experiments. For this, we train language and classification models in one language and test it on the other.

5.2.3 Results

The results of the classification experiments are presented in Figure 5.1.

5.2.3.1 Unilingual prediction

In French, for both LR and SVM, using *LM* features leads to higher AUC than the *info* features, and the combination of features is more effective than either feature set alone. In the English case, the *LM* and *info* features lead to equivalent performance individually, but the AUC is again marginally improved when the feature sets are combined, suggesting that they are capturing at least somewhat complementary information.

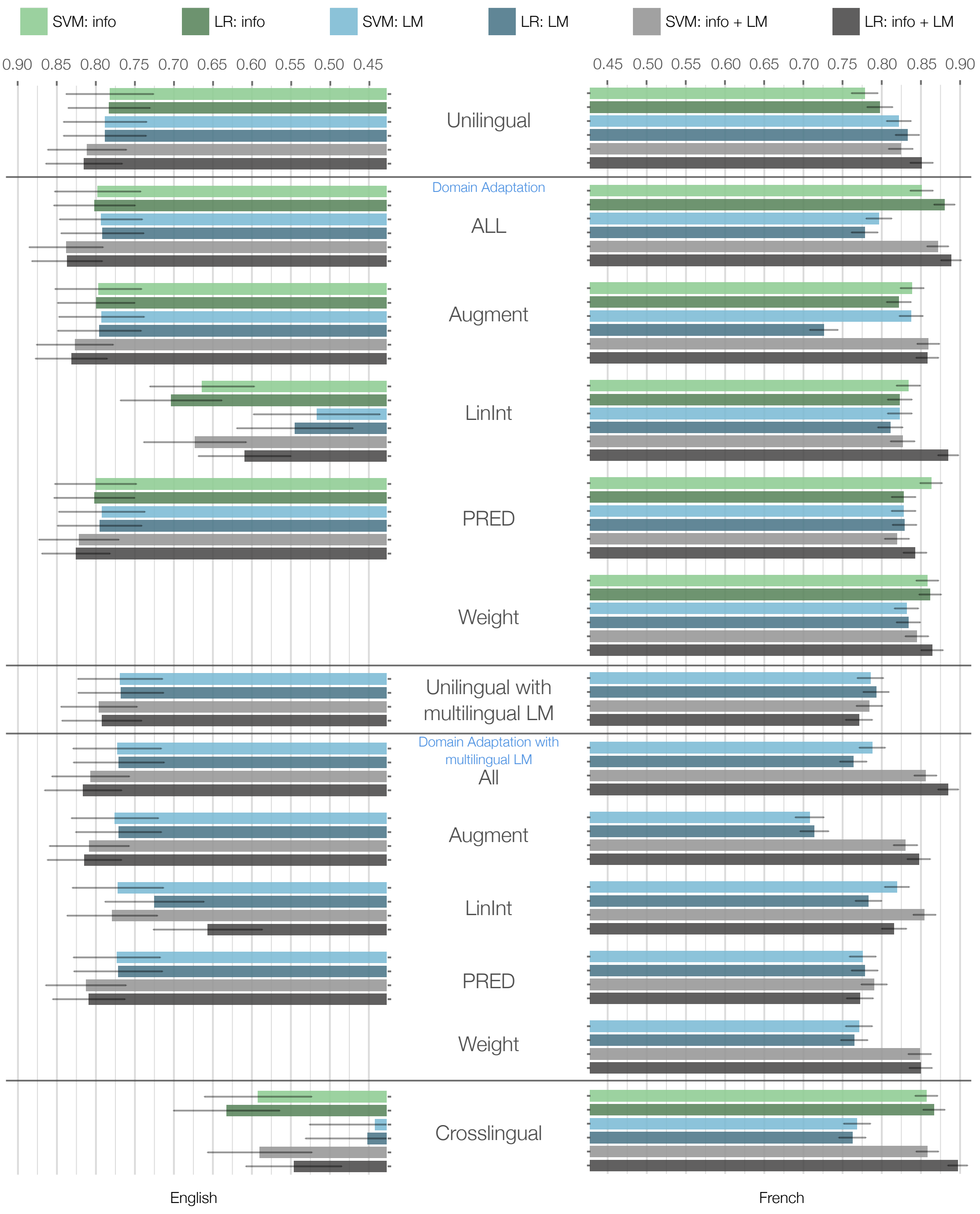


Figure 5.1: Results of uni-, multi- and cross-lingual classification experiments. Left panel displays results for English, right panel for French. Labels in the middle indicate the classification scenario and method of domain adaptation. Colours indicate the feature set and classifier. Bars indicate the AUC; error bars represent standard deviation.

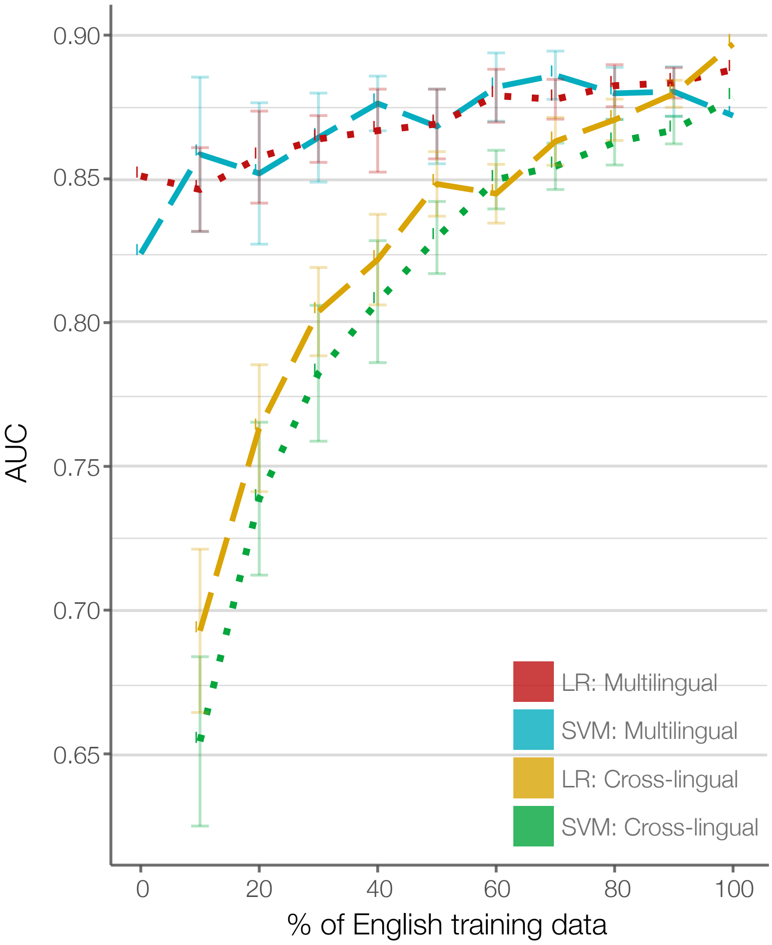


Figure 5.2: AUC as a function of the amount of English data used in the training set, for both multi- and cross-lingual cases. Error bars indicate 95% confidence intervals.

5.2.3.2 Domain adaptation results

For French, the *LM* features generally do not benefit from domain adaptation, with equivalent or poorer AUC relative to the unilingual case. The best result with the *LM* features is achieved in the AUGMENT scenario, where the classifier can select the French *LM* features only (although this result holds only for the SVM classifier). In contrast, the *info* features do benefit from the additional data available through domain adaptation, and lead to better results than the unilingual baseline. The best overall result of AUC = 0.89 is achieved by combining the feature types in the ALL configuration.

For English, we do not expect to see much benefit from including the (much smaller) French dataset. The WEIGHT adaptation technique is not feasible when the source data is smaller than the target data, and the LININT technique performs poorly, as it assigns too much importance the smaller and out-of-domain dataset. However, we do see marginal improvements using ALL and AUGMENT, reflecting the value of increasing the training set size by roughly 10%. The best result of AUC = 0.84 is achieved in the ALL condition, using the combined feature set.

5.2.3.3 Multilingual LM results

Using the multilingual LM does not affect the *info* features, and therefore Figure 5.1 shows only the *LM* and *info+LM* results. Clearly, the multilingual LM approach does not work well here. Unlike in domain adaptation, combining the datasets using this method assumes that information units will be produced in the same order in the two languages. While French and English are similar in this respect, there are many possible counter-examples, such as *cookie jar* (COOKIE JAR) versus *boîte à biscuits* (JAR COOKIE).

5.2.3.4 Cross-lingual prediction

When training entirely on English data and testing on French, the results using *info* and *info+LM* features are significantly improved over the unilingual baseline, while the *LM* results are reduced, once again indicating that the *info* features transfer better across languages. The results are very similar to those using the ALL technique for domain adaptation, suggesting that in that case, model training is dominated by the English data.

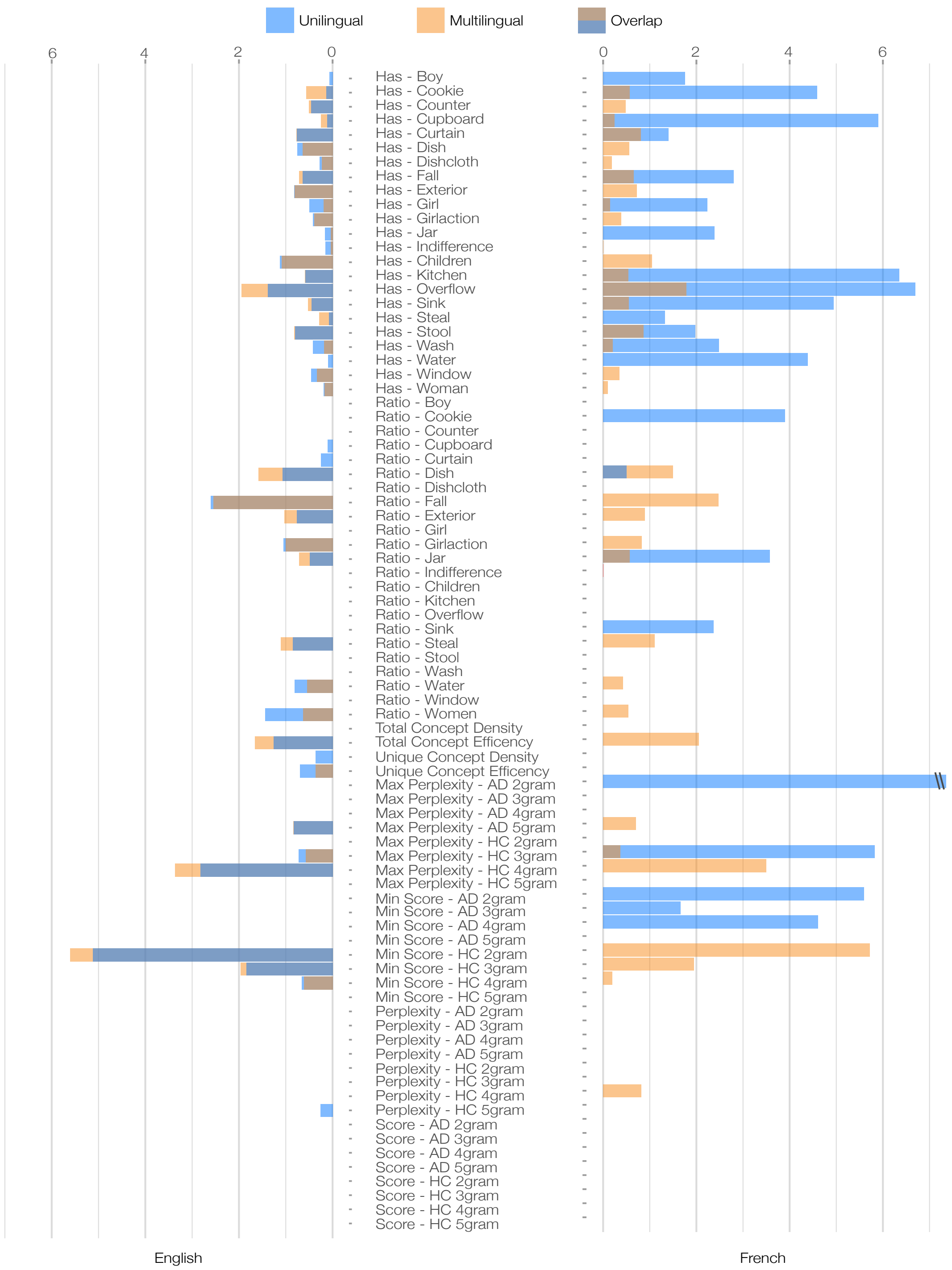


Figure 5.3: Visualisation of feature weights for uni- and multilingual experiments. Median feature importances over LPO- and 10-CV are displayed. The left panel displays the English and the right panel the French data sets. Unilingual experiments are given in blue and multilingual in yellow.

To further explore the similarity in performance in the ALL and cross-lingual cases, we examine the effect of incrementally increasing the amount of English data in the training set, when testing on French data. Figure 5.2 displays the classification performance of SVM and LR classifiers trained either using the ALL method of domain adaptation or cross-lingually with increasing amounts (10% at a time) of the English data. Considering first the ALL method (red and blue), at $x = 0$ there is no English data, and so we recover the French unilingual baseline. As we increase the amount of English data in the training set, performance slowly increases, eventually reaching the values reported in Figure 5.1. Considering next the cross-lingual case (yellow and green), we see that training on only 10% of the English data (55 samples) results in much poorer AUC values. However, each further 10% increases the classification performance. At 80% of English data (440 samples) the multi- and cross-lingual cases converge in performance. Thus, it would appear that domain adaptation is more data-efficient, as we achieve close to optimal results with a smaller proportion of English data, but that the cross-lingual approach can be equally effective, given a large enough corpus.

5.2.3.5 Feature analysis

Finally, we examine the features to determine which features are most useful to the task of dementia detection, and to compare the selected features in the unilingual and multilingual cases. Figure 5.3 shows the median absolute value of the weights assigned to each feature, for English and French, in the unilingual and multilingual ALL condition. The L_1 regularisation serves to set many feature weights to zero.

As a high-level observation, in both the uni- and multilingual cases, relatively more *info* features are selected, and relatively fewer *LM* features. Of the *LM* features that are selected, those which relate to the maximum perplexity or minimum probability appear to be more useful. These features capture locally anomalous speech patterns, relative to either the AD or control language models.

In the unilingual case, the French models show a preference for the binary "has" features (indicating whether or not an information unit has been mentioned). Only 4 of the "ratio" features and none of the density or efficiency features have a median value greater than zero. However, these features *are* relevant to the task, and potentially more generalisable (e.g., total concept efficiency differs between the French AD and HC groups with $p < 0.001$ on a t-test, and represents an aggregate score rather than depending on the presence or absence of a single information unit). Such features are selected more often in the multilingual case, and lead to improved performance. One explanation for this could be that in the small French training set, spurious correlations due to noise can overpower the real signal, and lead to less relevant features being assigned high weights, while correlated (but perhaps actually more relevant) features are suppressed. By increasing the size of the training set with English data, the signal-to-noise ratio is improved, and a better set of features is selected.

Generally, the feature values (not shown) support the intuition that controls mention more of the information units in the image (higher "has" feature values), convey information more efficiently, with fewer off-topic words (higher density and efficiency scores), and organize the narrative in a more predictable way (narratives have lower perplexity and higher probability) than the AD participants. Again, these trends are more apparent in the English data than the French data, likely due to the relatively larger number of samples.

5.2.4 Discussion

One perhaps surprising result of this study was that naively combining features in the ALL condition led to better results than the AUGMENT algorithm. However, this is in line with the original findings of [90], where he identified a set of tasks where AUGMENT performed sub-optimally: specifically, those cases where training on source-only data was better than training on target-only data. This is precisely the case we have here, as training cross-lingually (on English source data) leads to better results than training unilingually (on French target data). The explanation offered by Daumé III is, "If the domains are so similar that a large amount of source data outperforms a small amount of target data, then it is unlikely that blowing up the feature space will help." In some sense, then, these results are confirmation that we have indeed identified a set of features over which the two languages (i.e. domains) are very similar.

The fact that the ALL configuration is optimal in both French and English has an added practical benefit: since there is no distinction between source and target features, the resulting classifier is language-agnostic. This means that test data could come from either language, in a hypothesized future screening application.

There are no previously published results on this French dataset; however, we note that the classification results presented for the English data do not exceed the state-of-the-art. The DementiaBank dataset has proven difficult to benchmark, due to different (but poorly documented) releases, decisions by researchers to exclude participants with multiple samples, samples with fewer than 100 words, and so on. Our best English result is AUC=0.84, which corresponds to an accuracy of 75% and F_1 score of 0.77. This improves over the reported results of AUC=0.83 [321], F_1=0.74 [409], and F_1=0.75 [350], but does not improve on the accuracy of 82% reported by [128]. Further complicating matters, our result exceeds the AUC=0.79 reported by [162], but is worse than their accuracy of 79% and F_1 of 0.81. Regardless, the main contribution of the paper is not to push the state-of-the art on the DementiaBank dataset, but to use that resource to improve classification in a lower-resource language.

In this work, we have shown that there are features which can both distinguish AD patients from healthy controls with a high degree of accuracy, and also generalize across languages. By incorporating a large English dataset, we were able to improve the AUC on the French dataset from 0.85 to 0.89. We also developed a new set of features for this task, using concept-based language modelling, which improved AUC from 0.80 to 0.85 in the unilingual case, and 0.88 to 0.89 in the multilingual case.

Future work will involve extending the set of features involved, incorporating data from other languages, and testing whether similar techniques can be effective for detecting earlier stages of cognitive decline, such as MCI. Other work from our group has also begun to explore the use of unsupervised methods and out-of-domain data sources [218].

Technical challenges aside, collaborations of this nature can be difficult due to the sensitive nature of the data, and the need to respect ethical guidelines and participant consent when sharing and storing data. With this in mind, we recommend to other researchers working in similar domains to consider from the outset whether their data could eventually be shared, and to make suitable provisions in their ethics protocols and participant consent forms. We look to DementiaBank as a model for this kind of data-sharing and openness, and hope that researchers can continue to find ways to share resources of this nature.

5.3 Summary

This chapter examined several methods to combine data resources from different languages to improve the predictability of classifiers in under-resourced languages.

Section 5.1 examined the feasibility of combining data sets (i.e., French, Swedish and English) of MCI patients performing a picture description task. Several linguistic and acoustic features from previous work were extracted for each language separately. Classifiers were trained in a multi-lingual fashion using domain adaptation techniques. Overall, we are able to improve over the baseline for each language by including training data from other languages. The most frequently selected features were of acoustic nature, suggesting that linguistic feature did not transfer well between languages. Demographic and clinical differences in patient populations across languages were found to help generalisability of classifiers. We concluded that a picture description task, although being the most available form of speech-recorded language data in dementia and across languages, might not be the ideal task to capture cognitive impairment in early dementia stages.

Section 5.2 examined the multi- and cross-lingual detection of Alzheimer's disease from picture description transcripts using novel information content measures, specifically created to transfer across languages. Transcripts were transformed into a language independent abstract sequence of mentioned information units and concept based language models were trained on this sequence, both inside and between languages. Classifiers were trained in multi- and cross-lingual settings using domain adaptation techniques. Results show a clear benefit of adding a large amount of English data to a small French data set. By gradually adding a percentage of the English data to both multi- and cross-lingual models, we observed that although having in-domain French data to start with gave the multi-lingual classifier an advantage, a sufficient amount of English data led to similar performance cross-lingually. A regularisation effect of adding English data to the small French data set was observed, leading the classifier away from giving high weights to binary features.

The next chapter will move away focus from automatically separating different stages of dementia from healthy controls, to separating between patients inside a diagnostic group based on their affective status.

Part IV

Automatic Detection of Clinical Apathy in Dementia Patients from Speech and Language

Chapter 6

Detection of clinical Apathy in Dementia Patients

Dementia syndromes often occur with other comorbidities. The most likely co-occurring syndrome is Apathy, an affective disorder characterised by flattened affect and loss of motivation–see Section 2.2 for details. As apathy is the biggest risk factor for conversion from MCI to dementia and because it is treatable, its identification is a large priority on slowing disease progress.

This Chapter describes experiments on the automatic detection of apathy syndrome from speech recordings of dementia patients. In comparison to previous Chapters, speech considered here is not produced in standard cognitive assessments but rather during open questions. Analysis also focuses more heavily on acoustic aspects of speech production.

6.1 Speech Features for the Detection and Characterisation of Apathy

The current chapter intends to investigate the feasibility of automatic analysis utilizing paralinguistic speech features extracted during a short free speech task as a potential candidate for clinical apathy assessment (characterisation) and broad screening (detection) in elderly patients with cognitive impairment.

6.1.1 Methodology

6.1.1.1 Participants

60 patients with cognitive disorders were included in this study. Participants underwent a clinical assessment including the Mini-Mental State Examination (MMSE) [116], the Clinical Dementia Rating Scale [285] , the Apathy Inventory (AI) [328] and the Neuropsychiatric Inventory (NPI) [86]. Apathy was diagnosed based on the AI total score ($\geq$ 4). According to this assessment, participants were categorised into either non-apathy or apathy groups and matched for age and MMSE per gender group. Subjects with scores above 4 on other NPI items except from ’Apathy’ were excluded. Speech features (e.g. pitch) vary naturally between males and females and previous work found differences depending on gender in the effects of apathy [221], as well as depression [89] and the effectiveness of classifiers for its detection [232]. This is why this study considers males and females separately.

All participants were recruited through the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Participants were all native speakers of French and excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, psychotic or aberrant motor behavior, or history of drug abuse. Written informed consent was obtained from all subjects prior to the experiments. The study was approved by Nice Ethics Committee (ELEMENT ID RCB 2017-A01896-45, MoTap ID RCB 2017-A01366-47) and was conducted according to the Declaration of Helsinki.

6.1.1.2 Speech Protocol

Free and natural speech tasks require low cognitive effort and are capable of eliciting emotional reactions (or a lack thereof) [88] by asking to describe events that triggered recent affective arousal [237]. To this end, people were asked to perform two tasks: (1) talk about a positive event in their life and (2) to talk about a negative event in their life. Instructions (”*Pouvez-vous me raconter en une minute d’un évènement positif/négatif?*”) for the vocal tasks were prerecorded by one of the psychologist and played from a tablet computer ensuring standardised instruction over both experiments. The vocal tasks were recorded with the internal microphone. Administration and recording were controlled by the application and facilitated the assessment procedure. To increase comparability, all recordings were sampled at 22.050 kHz and encoded with 16 Bit in the *wav* format.

6.1.1.3 Features

Features are extracted directly and automatically from the audio signal. This increases the applicability of results in a clinical scenario, seeing as no prior processing, such as transcription of what has been said, is required. For each speech task, features are extracted separately. Overall, acoustic features were extracted from four different main areas: *prosodic*, relating to perceived stress, intonation and rhythm in speech (e.g. perceived pitch), *formant*, features carrying information about the acoustic resonance of the vocal tract and its use, *source*, relating to measures of air flow through the glottal speech production system (e.g. measures of voice quality), as well as *temporal*, describing measures of speech proportion (e.g. length of pauses). Table 6.1 gives a detailed overview, definition and explanation of all extracted acoustic features. All features from the *tempo* category as well as F_0 features were extracted using the Praat software [48]. *Jitter*, *Shimmer* were determined using the openSmile software [108]. A Matlab script was used to extract *HNR* and statistics over the first three formants.

6.1.1.4 Statistical Analysis

All statistical analysis were run using R software version 3.4.0[1]. This study computed the Wilcoxon signed-rank and ranked-sum tests for dependent and independent sample testing respectively and Spearman's ρ for correlations. For the characterisation of apathy, differences in acoustic measures are examined between the apathy and non-apathy group inside a gender. The goal being to find correlations between acoustic markers and the AI apathy sub-scales, as well as between acoustic markers, ultimately deriving properties of apathetic speech.

6.1.1.5 Prediction

Machine learning experiments are carried out to validate the diagnostic power of extracted markers. For this, classifiers are always trained within a gender, to differentiate people with and without apathy. As classifiers, simple Logistic Regression models implemented in the scikit-learn framework [301] were used. The L_1 loss was used as a

[1] https://www.r-project.org

penalty, as it is capable of performing implicit feature selection by reducing weights of unimportant features to zero. This is especially useful, since the number of used features is larger than the number of samples. Because of the small data set, they are trained and evaluated in a leave-one-out cross-validation scenario. Here, all but one sample are used in training of the classifier and its performance is evaluated on the held out sample. This is repeated for all samples and results are averaged. Features are normalised using z-standardisation based on the training set in each fold. As a performance metric we report AUC to be able to reason about possible specificity and sensitivity trade-offs.

Category	Feature	Definition	Intuition
Prosodic	F_0	Mean, Max, Min, Range, Variance and Standard deviation of F_0	Statistics over the perceived auditory pitch (speech melody)
	Periodicity	Mean, Max and Min cross-correlation of speech signal	Measure of the regularity of the speech signal
Formant	$F_1 - F_3$	Mean and Variance of the first three formant frequencies	Indicative of the class of speech sound
Source	*Jitter*	Average absolute difference between consecutive signal periods, divided by the average period length	Indicative for a lack of control for vibration of the vocal cords
	Shimmer	Average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude	Indicative for reduction of glottal resistance
	Harmonics-to-Noise-Ratio (HNR)	Ratio between periodic components and aperiodic components comprising voiced speech	Measure of voice quality
Temporal	*Sounding segments*	Mean, Max and Standard Deviation of sounding segment lengths determined based on intensity	Statistics over length of connected speech segments
	Pause segments	Mean, Max and Standard deviation of silence segment lengths determined based on intensity	Statistics over length of continuous pause segments

Duration	Total length of recording	Total length of recording
Speech duration	Total length of all sounding segments	Amount of speech
Pause duration	Total length of all silence segments	Amount of pause
Speech proportion	Ratio of *Speech duration* and *Duration*	Proportion of recording participant is talking
Speech Rate	Ratio of number of syllables, detected using [94], and *Duration*	Measure of information density
Articulation Rate	Ratio of number of syllables, detected using [94], and *Speech duration*	Measure of speech tempo

Table 6.1: Feature definition of acoustic markers. Name, definition and intuition of features sorted by category is presented.

6.1.2 Results

6.1.2.1 Demographics

Demographic data is provided in Table 6.2. After matching for MMSE and age, 24 male subjects and 36 female subjects were included in the final analysis and divided into equal groups of apathy and non-apathy subjects. No significant differences were present between the groups except for the results on the apathy scales.

6.1.2.2 Correlation

Figure 6.1 presents Spearman correlation coefficients between extracted features and the AI sub-scales (i.e. *affective*, *interest*, *initiative*), split by gender. Only significant correlations are presented.

The male population shows overall comparable correlations between the positive and the negative story. Generally, more significant correlations are observed for temporal features. In the positive story, correlations between these markers and all AI subdomains are present. Only a small negative correlation between F_0 Range and the *affective* domain is observable ($\rho = -0.47$). For the negative story, temporal features again dominate, while only showing correlations with the *interest* and *initiative* subdomains. Correlations with the *affective* domain are observed for both F_0 Max ($\rho = -0.61$) and F_0 Range ($\rho = -0.69$).

The female population shows more correlations in the positive story. Strong correlations are observed between all three subdomains and features relating to pause lengths. Features relating to sound length and speech tempo correlate significantly with the *interest* and *initiative* domain. In the negative story, nearly no correlations between temporal variables and any subdomain are present. Weak correlations are present between variables relating to mean Jitter (*affective*: $\rho = 0.28$; *interest*: $\rho = 0.29$) – which is consistent with correlations in the positive story – minimum Shimmer (*interest*: $\rho = -0.40$; *initiative*: $\rho = -0.31$) and minimum (*interest*: $\rho = 0.46$; *initiative*: $\rho = 0.50$) and maximum Periodicity (*interest*: $\rho = -0.47$; *initiative*: $\rho = -0.41$).

	Male		Female	
	N	A	N	A
N	12	12	18	18
Age	78.25 (4.33)	79.58 (5.45)	77.83 (6.12)	79.50 (5.86)
MMSE	22.66 (3.11)	19.42 (4.17)	22.33 (4.02)	19.56 (5.52)
AI total	1.7 (1.23)	6.0*** (1.60)	0.56 (0.99)	5.33*** (1.97)
AI-Intr	0.75 (0.87)	2.42*** (0.90)	0.17 (0.38)	2.33*** (0.91)
AI-Init	0.83 (0.94)	2.67*** (0.89)	0.39 (0.69)	2.33*** (1.19)
AI-Affect	0.08 (0.29)	0.92**(0.90)	0.00 (0.00)	0.67** (1.08)
NPI-Apathy	1.67 (2.01)	6.50*** (3.73)	0.44 (0.70)	5.44*** (3.01)
NPI-Depression	0.50 (0.90)	1.50 (2.68)	0.16 (0.51)	1.50 (2.91)
NPI-Anxiety	1.50 (2.06)	2.75 (3.33)	0.94 (1.16)	3.11 (3.61)

Table 6.2: Demographic data for population by gender and apathy; Mean (standard deviation); Significant difference from the control population in a Wilcoxon-Mann-Whitney test are marked with $^{*} : p < 0.05,^{**} : p < 0.01,^{***} : p < 0.001$.

Abr.: N='No Apathy', A='Apathy', MMSE='Mini Mental State Examination', AI='Apathy Inventory', AI-Intr='AI domain Interest', AI-Init='AI domain Initiative', AI-Affect='AI domain affective', NPI='Neuropsychiatric Inventory', NPI-Apathy='NPI domain apathy', NPI-Depression='NPI domain depression', NPI-Anxiety='NPI domain anxiety'

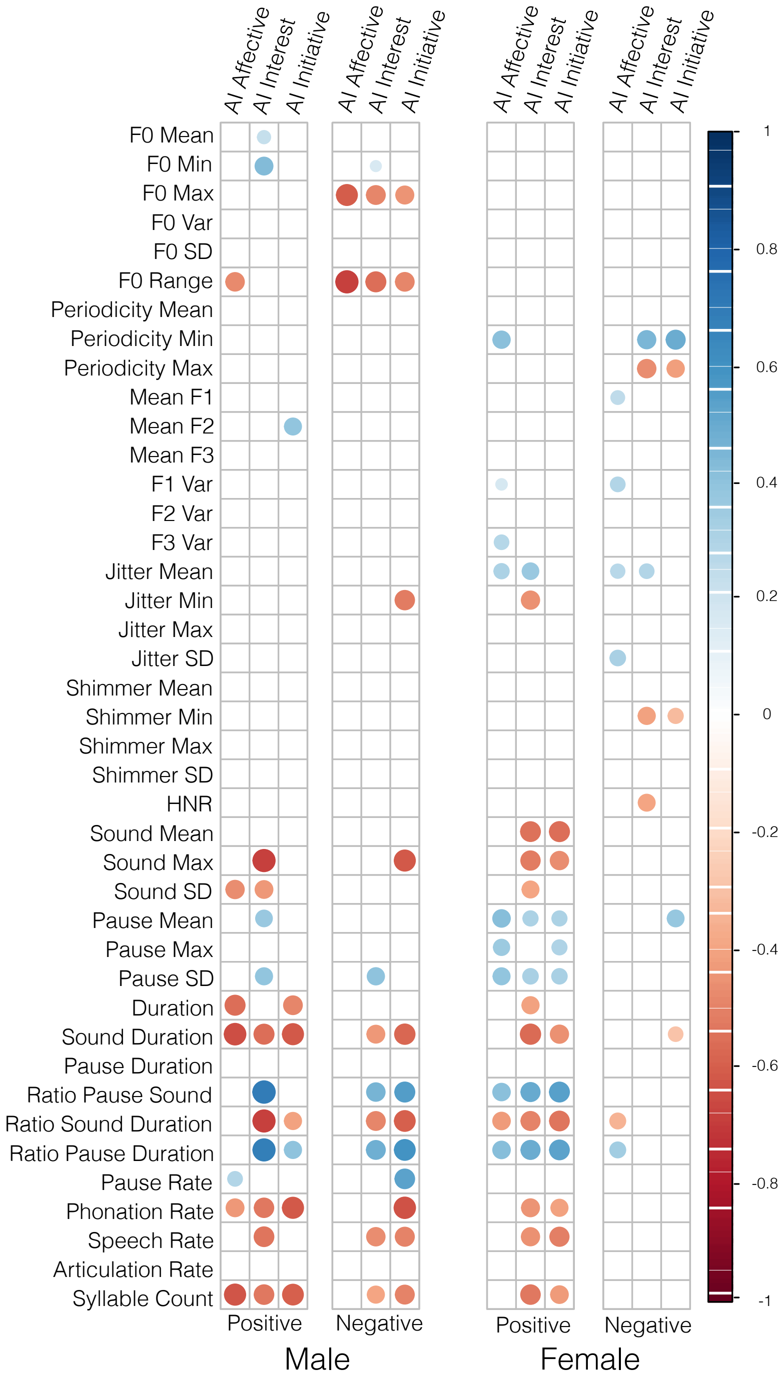


Figure 6.1: Spearman correlation coefficient between features extracted from vocal tasks and AI subdomains. One correlation matrix is presented per speech task and gender. Only significant correlations ($p < 0.01$) are displayed.

6.1.2.3 Group comparison

Statistical comparisons between the apathetic and non-apathetic groups are presented in Table 6.4a for the male population and in Table 6.4b for the female population. Only significant values are reported.

Overall, features relating to temporal aspects of speech dominate. Some features show significant differences regardless of gender (i.e. Speech Rate, Ratio Pause Duration, Ratio Sound Duration, Ratio Pause Sound, Sound Max, Sound Duration), but for the female population only in the positive story. Males show significant differences in F_0 Range and F_0 Maximum in the negative story. Females show significant differences in HNR across both tasks. Females show differences in the negative story only in voice quality markers (Periodicity, Jitter and HNR). For the male population, the largest effect in the positive story is the Sound Duration ($\rho = 0.61$) and for the F_0 Range in the negative story ($\rho = 0.52$). For the females, the largest effects are in the Ratio Sound Duration for the positive story ($\rho = 0.54$) and the HNR for the negative story ($\rho = 0.51$).

6.1.2.4 Prediction

Results of classification are reported in Figure 6.2. *AUC* is far over the chance baseline of 0.5 for both male and female populations. The classifier trained on the male population achieves an *AUC* of 0.88 and the one trained on the female population an *AUC* of 0.77. The ROC visualises a trade-off between sensitivity and (1 - specificity). For the male population, the classifier could be configured to achieve a good sensitivity of 0.91 and a reasonable specificity of 0.68. For the female population, a sensitivity of 0.85 and specificity of 0.72 can be configured. Feature weights from L1 regularised Logistic Regression models are reported in Table 6.4.

6.1.3 Discussion

Early detection of apathy in older adults has reached high clinical relevance because of an increased risk of incident of dementia and the danger to be easily overlooked by clinicians, which could lead to premature withdrawal from care [386].

The current study is the first one of its kind demonstrating clearly that certain paralin-

Origin	**Feature**	**Significance**	**Statistic χ^2**	**Effect size ρ**	**Direction**
Positive	Duration	*	5.60	0.39	↓
	Ratio Pause Duration	**	6.75	0.43	↑
	Ratio Sound Duration	**	6.75	0.43	↓
	Ratio Pause Sound	*	5.60	0.39	↑
	Sound Max	*	6.45	0.42	↓
	Sound Mean	*	5.33	0.38	↓
	Sound Duration	***	13.23	0.61	↓
	Pause Mean	*	4.56	0.36	↑
	Syllable Count	***	11.81	0.57	↓
	Speech Rate	**	9.36	0.51	↓
Negative	Ratio Pause Duration	*	6.16	0.41	↑
	Ratio Sound Duration	*	6.16	0.41	↓
	Ratio Pause Sound	*	5.60	0.39	↑
	Sound Duration	*	6.45	0.42	↓
	Sound Max	*	4.08	0.34	↓
	Pause SD	*	6.45	0.42	↑
	Pause Mean	*	4.32	0.35	↑
	Pause Max	*	4.08	0.37	↑
	Syllable Count	*	3.85	0.33	↓
	Speech Rate	*	5.88	0.40	↓
	F_0 Range	**	9.72	0.52	↓
	F_0 Max	**	9.36	0.51	↓

(a) Comparison for male population

Origin	Feature	Significance	Statistic χ^2	Effect size ρ	Direction
Positive	Ratio Pause Duration	**	10.62	0.54	↑
	Ratio Sound Duration	**	10.62	0.54	↓
	Ratio Pause Sound	**	9.61	0.52	↑
	Sound Max	**	6.73	0.43	↓
	Sound Mean	**	8.29	0.48	↓
	Sound Duration	**	8.66	0.49	↓
	Pause Mean	**	6.73	0.43	↑
	Pause Max	*	5.48	0.39	↑
	Pause SD	**	7.06	0.44	↑
	Syllable Count	**	7.23	0.45	↓
	Speech Rate	**	8.11	0.47	↓
	Jitter Mean	*	5.33	0.38	↑
	HNR	**	6.73	0.43	↓
Negative	Periodicity Min	**	8.11	0.48	↑
	Periodicity Max	**	7.93	0.47	↓
	Jitter Min	*	5.41	0.39	↓
	Jitter Mean	*	5.05	0.37	↑
	Jitter SD	*	5.33	0.38	↑
	HNR	**	9.42	0.51	↓

(b) Comparison for female population

Table 6.3: Statistical group comparisons between non-apathetic and apathetic group using Kruskal-Wallis tests. Features with $p < 0.05$ are reported. Vocal task of origin, p-value, test statistic (χ^2), effect size (ρ) and direction of effect in the apathetic group in comparison to the non-apathetic group are reported.
* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$

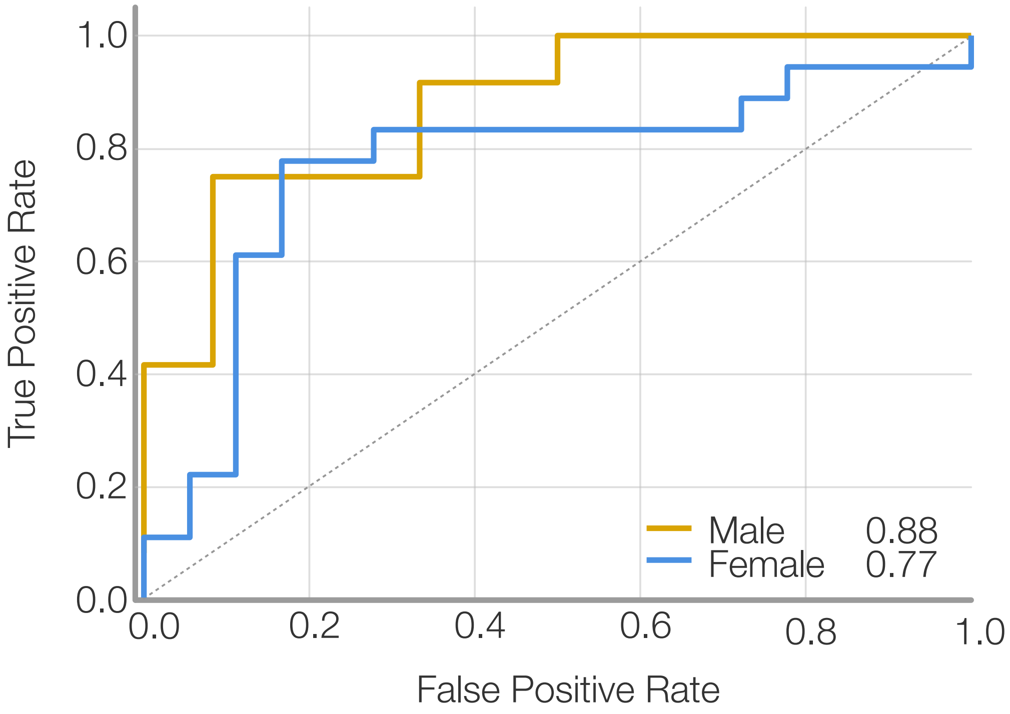


Figure 6.2: Receiver Operator Curve (ROC) of classifiers trained to detect apathy from speech. The blue and red lines represent classifiers trained and evaluated on the male and female populations respectively. *AUC* is reported in the legend.

Feature	Source	Mean	SD	Proportion
Sound Duration	Pos	.99	.24	1.00
F_0 Max	Neg	.77	.25	.96
F_3 Mean	Pos	.71	.21	.96
Ratio Pause Duration	Pos	.26	.22	.96
Pause Max	Neg	.41	.16	.92
Shimmer SD	Neg	.33	.17	.92
Ratio Sound Duration	Pos	.26	.19	.92
Articulation Rate	Neg	.25	.16	.88
Jitter SD	Pos	.08	.08	.54
Mean Pause	Neg	.04	.12	.25

(a) Male

Feature	Source	Mean	SD	Proportion
HNR	Neg	.79	.18	1.00
Jitter Min	Neg	.23	.11	.97
Periodicity Min	Neg	1.21	.31	.97
HNR	Pos	.49	.23	.97
Pause Rate	Pos	.39	.13	.97
Sound Mean	Neg	.45	.20	.94
Duration	Pos	.79	.23	.94
Shimmer Max	Pos	.35	.12	.94
Shimmer Min	Pos	.46	.14	.94
Ratio Pause Duration	Pos	.31	.18	.89

(b) Female

Table 6.4: Feature weights from L_1regularised Logistic Regression models. Mean of weights, standard deviation (SD) of weights and proportion of folds with an above zero weight are presented over all folds, for the top 10 features according to proportion.

guistic features correlate significantly with levels of apathy severity. Thus, automatic speech analysis could be a promising new tool for its assessment.

Overall, the strongest correlations were found between the subdomains *interest* and *initiative* of the AI and temporal speech features. The *affective* subdomain, which represents the emotional blunting in apathy, seems more associated with prosodic speech features which is in line with previous findings on depressed speech with mainly prosodic speech abnormalities such as reduced pitch resulting often in a dull and 'lifeless' tone [87]. Similar observations were made in patients of this study with presence of emotional blunting. Thus, it seems that through speech features, distinct profiles can be characterized confirming what previous neuroimaging analyses revealed, namely that apathy is multidimensional and different subdomains are associated with different brain regions and circuits; the affective one with the ventral prefrontal cortex; the behavioural one with the basal ganglia; and the cognitive with the dorsomedial prefrontal cortex [207].

Overall, both males and females showed reduced reaction to the stimuli. Answers to the posed questions can be generally characterised by drastically shorter (lower sound duration) and slower (lower Speech Rate) speech. For the female population, a difference in voice quality (lower HNR) is obvious in both questions. Males suffering from apathy react less emotionally to the negative question as indicated by a lower variance of prosody (lower F_0 Range). Interestingly, male and female subjects with apathy show different patterns in their speech features according to the type of free speech task. For males, significant differences between apathy and non-apathy subjects can be seen in temporal features for both the negative and positive story. Females show similar patterns in the positive story, but not in the negative one. Until today, no work on gender dependent symptoms of apathy was found that could explain this pattern. Parts of this effect could be caused by the fact that men from this generation are in general less likely to talk enthusiastically about a positive event and show greater responses to threatening cues [206].

Gender differences in emotional processing and expressivity [97] as well as in emotional memory retrieval [312] could be another reason and should be further investigated, since current literature mostly focuses on exploring age as a variable. Gender differences have been observed in brain activity during emotional tasks with primarily females recalling more autobiographical memories when it's of emotional content and cues are given verbally [358]. It is possible that females in this study were more likely to be triggered to an emotional reaction when asked about a positive event and vice versa for males. Apathy might have an effect on this biased emotional memory retrieval. Hence, it can be assumed that the type of affective stimulus with which speech is being provoked might play a major role and might have to be adapted depending on a patient's gender.

Generally, when classifying between apathy and non-apathy subjects, features related to sound and pause segments seem to dominate with higher AUC results obtained for the male group. These features might have been particularly affected by the cognitive and behavioral aspects of apathy which seem to be reflected in the general amount of speech produced. Recent findings suggesting that apathetic patients have decreased visual attentional bias for social stimuli compared with non-apathetic patients[69] might apply as well for speech production since it implies engagement in social interaction. Several reasons could explain these findings drawn from related studies on depression [277] and negative symptoms in schizophrenia [80]. This may be attributed to reduced muscle tension as well as impaired neuromuscular motor or articulatory coordination [79] caused potentially by alterations in the neurotransmitter system namely low norepinephrine and/or dopamine levels [265]. Changes in affective states can impact the phonation and articulation muscular systems via the somatic and autonomic nervous systems [344]. Commonly observed psychomotor retardation in apathy can lead to small disturbances in muscle tension which in turn can affect the speaker's speech pattern and, for instance, reduce pitch variability [170].

Since patient data is always hard to acquire, the here presented sample is relatively small and future studies should strive to draw more conclusive evidence from larger datasets. Further work should examine what features in particular are predictive for apathy, how they relate to depression and how the two could be better discriminated. One potential solution could be to perform a semantic analysis of the content of speech to better differentiate apathy from depression and anxiety. Adding other additional measurements, for instance, of facial, head or body movement dynamics, by the means of video might further improve accuracy. In the field of depression, research has demonstrated more powerful detection when applying a multi-modal audio-visual data fusion approach [98].

Nevertheless, it can be concluded that automatic speech analysis could become a promising new screening and assessment tool for follow-up measurements ('digital endpoints') in clinical trials of pharmacological and other interventions that aim to monitor apathy in patients.

6.2 Detecting late-life Apathy in Dementia using Sentiment and Psycholinguistic Analysis of Emotional Language

It can be argued that ’thinking’ and ’speaking’ are intimately linked together through the sensory and motor systems which are highly affected by syndromes such as apathy or depression [189], but not always easy to detect by the human ear. Computer-aided analysis of sentiment and psycholinguistic aspects of language might have the potential to objectively identify the lack of interest and reduced emotional response characteristic to apathy.

This chapter analyses sentiment and psycholinguistic aspects of manually transcribed responses to emotional questions, to predict the presence of apathy in older adults with cognitive impairment. Results are compared to approaches utilising automatic speech recognition, in order to verify a fully automatic pipeline. Section 6.2.1 discusses related work in the fields of apathy detection, sentiment and psycholinguistic analysis. Section 6.2.2 describes the first experiment on manual transcripts. Section 6.2.3 leverages results from the previous section to verify an automatic pipeline. Section 6.2.4 concludes the study and gives an outlook of applications, possible future directions and limitations.

6.2.1 Background

Sentiment analysis (SA) is the task of classifying documents, or their parts, according to the emotions or opinions they express. In its simplest form, it distinguishes between positive and negative sentiment, optionally with a neutral class. More complex SA identifies different kinds of emotions. Applications of SA are various and range from opinion-based summaries of product reviews [235] over sentiment annotation of children’s stories in order to enrich their presentation by speech synthesis [16] to connecting the influence of (shifting) emotionality in twitter messages to their spread through the social network [209].

In general, SA applications focus on the correct classification of opinion or emotion, not the quantification thereof, although there are frameworks which also consider the intensity of the found emotion (e.g., the Google Cloud Natural Language API [2] offers a sentiment magnitude value). For apathy detection, the concrete kind of emotion is less relevant than its overall intensity, thus, many of the usual tools and techniques need to be re-evaluated for their fitness to such a task.

In her survey of SA for text-mining, [252] describes three main groups of unimodal text-based approaches to SA: lexicon-based approaches, approaches using machine learning methods and hybrid approaches. The lexicon-based approach scans the documents entries of a pre-built or specifically derived sentiment-annotated lexicon, while machine learning methods compute linguistic features from the documents that serve as input to their respective classification model. In a hybrid approach, the lexicon search may be used to derive one or more input dimensions for a more sophisticated model. In any case, feature selection and engineering are an important step in the classification process. While a comparatively simple token-based lexicon lookup can already yield reasonable results [410], the implementation of more complex natural language processing (NLP) techniques is usually able to boost the classification performance, like the correct handling of negation [179], exploiting sentence structure for better detection [253] or abstracting from single tokens to the concepts behind them [337].

Regarding the lexicon-based SA approaches, Linguistic Inquiry and Word Count (LIWC) is a psycholinguistic tool that has been used to evaluate emotional and cognitive function. It works through comparing a predetermined dictionary of hierarchally-categorised, psychometrically-validated words to transcribed speech or text, over 5 dimensions (Linguistic processes, psychological processes, personal concerns, spoken properties and punctuation) [303].

LIWC has been successfuly used in a wide range of clinical applications: Looking at mental health [77], social behaviours and cognitive disorders such as Depression [92], Dementia [401], Parkinson's Disease [364], and Schizophrenia [264]. Focusing on apathy in Parkinson's, [364] used custom and prebuilt LIWC dictionaries to study *expressive behaviour*—fabricating LIWC categories for apathy, hopefulness, protest and

[2] https://cloud.google.com/natural-language/docs/analyzing-sentiment

hopelessness. Participants were interviewed and asked to describe an enjoyable or frustrating activity and then rated on verbal, determined by both LIWC dictionaries, and nonverbal expression patterns. Participants predictably used more words in the positive emotion and hopefulness categories as well as positive facial expressions while describing an enjoyable event but produced less speech and facial expressions when describing a negative aspect of their life, exhibiting increased apathetic behaviour.

Despite the paper mentioned above, research on applications of LIWC for apathy is very sparse. However, as depression shares significant symptoms with apathy, applications of LIWC in depression may be promising. [92] used curated LIWC categories to consider the linguistic (22 specific linguistic styles) and emotional indicators (positive affect, negative affect, activation, and dominance) of depression on the popular social media platform, Twitter. They combined these with other features to achieve a 70% classification accuracy. All emotional and six of the linguistic indicators, determined by LIWC, were reported to be significantly different between groups.

Some clinical applications of LIWC prune categories to determine which categories are relevant to the clinical domain. For example [23], used LIWC-driven features in machine learning classification experiments to distinguish spontaneous speech between healthy controls and those with MCI. They achieved an accuracy of 84% by selecting category-specific features and noted that using all LIWC categories yielded a lower performance (76.2% accuracy), concluding that all LIWC categories may not be suitable for specific classification tasks.

Similarly, some LIWC categories have been found to be more prominent in indicating highly goal-oriented behaviour on twitter, presumably the inverse of the diminished goal-directed behaviour symptom in apathy. [241] performed machine learning classification tasks based on LIWC categories from the LIWC 2001 dictionary to analyse *influencers* on social media platforms. Physical states correlated negatively with influencer behaviour and were suggested to be an indicator of illness or inactivity, and a possible signifier of apathetic tendencies. Physical states are listed under the personal concerns dimension in the LIWC 2007 dictionary, meaning that this may be hypothesised as a possible indicator of apathetic behaviour. To further this point, LIWC was used to consider word usage and social engagement behaviours on Twitter, where several psycholinguistic categories correlated significantly with replies and retweets, a proxy for social engagement. Several psycholinguistic traits correlated negatively, interpreted as less likely to produce an action, with response rate (anger, anxiety) and retweet rate (physical states, tentative). These LIWC categories may indicate being apprehensive to engagement, which would be expected to be a trait for those with apathy. LIWC categories were then used to build predictive models about users engagement behaviours, where selecting significantly-correlated LIWC categories out-performed all LIWC categories in both response (4.4% AUC improvement)and retweet (12.6% AUC improvement) prediction [242].

Incorporating LIWC categories into an automatic pipeline for screening in a clinical application, [176] established the validity of an automatic pipeline for aided diagnosis of dementia and its subtypes. They utilized both ASR and LIWC features in binary classification tasks. Using three to five minute clips of spontaneous speech gathered from the semi-structured interview portion of the Western Aphasia Battery of a relatively small corpus (48 participants), they were able to achieve an accuracy of 88% when distinguishing between ADs and controls as well as ADs and those with FTLD, by combining part-of-speech tagging, LIWC categories and acoustic features. They report that 22 of the 81 LIWC features were statistically significant.

To conclude, research indicates that there might be a link between LIWC categories and symptoms of apathy and that it might be conducive to leverage this in an automatic pipeline for apathy screening.

6.2.2 Apathy Detection from Manual Transcripts

The first experiment aims to predict apathetic signals in a person based on features extracted from manual transcripts of responses to two emotional questions.

6.2.2.1 Data

Speech recordings from both the ELEMENT and MoTap projects were used. The studies were approved by Nice Ethics Committee (ELEMENT ID RCB 2017-A01896-45, MoTap ID RCB 2017-A01366-47) and were conducted according to the Declaration of Helsinki. All participants were aged 65 or older and were recruited at the Memory Clinic located at the Institute Claude Pompidou in the Nice University Hospital. Speech recordings were collected using an automated recording app on a tablet computer. Only native speakers of French were included.

Participants completed a battery of cognitive tests, the MMSE, the Apathy Inventory (AI) [328] and the Neuropsychiatric Inventory (NPI) [86]. Participants were excluded if they had any major auditory or language problems, history of head trauma, loss of consciousness, or psychotic or aberrant motor behaviour. Following the clinical assessment, patients were grouped into three categories in accordance with the DSM-5 diagnostic guide: patients without any impairment, minor impairment or major impairment. In this study, we only look at patients with either minor or major impairments, to prevent confounding of group differences by cognitive state. Patients are split into groups according to their AI score (≥ 4) and groups are matched for MMSE. Due to a strong correlation between cognitive health and apathy, a group difference in MMSE remains ($p = 0.02$). Demographic data and clinical test results by diagnostic groups are reported in Table 6.5.

It has been shown that free speech allows a greater range of induced emotional effects, particularly when asked to describe events that triggered recent emotional arousal [237]. Hence, to elicit free speech, people were asked to to perform two tasks: (1) to talk about a positive event in their live and (2) to talk about a negative event in their live. Instructions ("Pouvez-vous me raconter en une minute d'un événement positif/négatif?") for the vocal tasks were pre-recorded by one of the psychologist and played from a tablet computer ensuring standardised instruction over both experiments. The vocal tasks were recorded with the tablet computer's internal microphone. Administration and recording were controlled by the application and facilitated the assessment procedure. To increase comparability, all recordings were sampled at 22.050 kHz and encoded with 16 Bit in the *wav* format.

Afterwards, recordings were transcribed on a word level by a group of trained speech pathology students following the CHAT protocol [239]. The transcriptions were aligned with the speech signal using PRAAT [48]. The words *négatif*, *positif*, *agréable* and *désagréable* were removed from the transcripts, as they were part of the instructions and were often repeated as a form of time-filling speech while the participants were in thought.

6.2.2.2 Features

To find symptoms of apathy, such as emotional blunting and diminished goal directed behaviour, in responses to these emotional questions, we utilise methods from both

	Control	Apathy
N	31	31
Gender (% male)	35%	45%
Age	77.74 (6.02)	79.00 (7.19)
MMSE	23.45 (3.85)	20.23* (4.86)
AI	1.03 (1.20)	5.68*** (1.74)
AI-Intr	0.42 (0.67)	2.39*** (0.88)
AI-Init	0.58 (0.81)	2.55*** (1.09)
AI-Affect	0.03 (0.18)	0.74*** (1.00)
NPI-Apathy	1.00 (1.55)	5.97*** (3.29)
NPI-Depression	0.94 (1.67)	1.45 (2.49)
NPI-Anxiety	1.74 (2.63)	2.90 (3.36)

Table 6.5: Demographic data for population by gender and apathy; Mean (standard deviation); Significant difference from the control population in a Wilcoxon-Mann-Whitney test are marked with $^{*}: p < 0.05,^{**}: p < 0.01,^{***}: p < 0.001$.

Abr.: MMSE='Mini Mental State Examination', AI='Apathy Inventory', AI-Intr='AI domain Interest', AI-Init='AI domain Initiative', AI-Affect='AI domain affective', NPI='Neuropsychiatric Inventory', NPI-Apathy='NPI domain apathy', NPI-Depression='NPI domain depression', NPI-Anxiety='NPI domain anxiety' sentiment analysis and the psycholinguistic analysis tool *LIWC* [304] in the form of its French adaption [313]. In this section, the subscripts $_p$ and $_n$ refer to features calculated on the positive and negative story respectively.

Since participants are actively required to fulfil a task – answer the posed question – a first easy baseline is the number of words in the patients response (WC_p, WC_n). In this context, it is an implicit measure of engagement and should be reduced in patients showing symptoms of apathy.

6.2.2.3 Sentiment Analysis

Another form of task fulfilment – i.e. goal-directed behaviour – can be seen in the emotionality of the answers. Since participants are asked to tell a positive/negative story from their lives, measuring the sentiment expressed in the response should be indicative for engagement in the task. Furthermore, emotionally blunting – meaning the absence/reduction of emotion – should be encoded in a response's sentiment. To classify the sentiment of responses, the French Expanded Emotion Lexicon (*FEEL*) [3] is used. This resource is a supervised translation from the English word-emotion association lexicon [266] and has been used as feature in machine learning classification experiments to distinguish between responses of expert medical opinions and non-experts in French internet health forums[2]. For a total of 14.000 French words, it contains a characterisation as either positive or negative. For a given participant, we analyse the sentiment of the response by iterating over all words in the associated transcript, checking if they are present in the dictionary and if so, counting the valence category they are associated with. In the end, each sample has a count for positive (P_p, P_n) and negative words (N_p, N_n). We explicitly calculate a sentiment score for each task separately. Since we know that a positive sentiment was asked in the positive story and a negative in the negative story respectively and we assume apathy to be unrelated to direct compliance with these instructions, we can assume that negative words in the positive story and positive words in the negative story were used to amplify emotional expressions (e.g.*'very sad'*). For the positive story we calculate the sentiment as

$$S_p = P_p \cdot N_p$$

and for the negative story, the overall sentiment is defined as

$$S_n = -N_n \cdot P_n$$

We also explicitly calculate the magnitude of sentiment in both tasks. For the positive story this is defined as

$$M_p = P_p + N_p$$

and for the negative story as

$$M_n = N_n + P_n$$

The last used feature from the sentiment scores is the range of observed sentiment between the two tasks. It should represent the possible range of emotions shown by a patient and through that be sensitive to emotional blunting. It is defined as

$$R = S_p - S_n$$

6.2.2.4 LIWC

LIWC is a psycholinguistic dictionary, that categorises words into groups known as *word subcategories*. For example, words such as *fils* (son), *mari* (husband) and *mère* (mother) are all grouped into the *Family* word subcategory. One word can belong to multiple subcategories. LIWC 2007 knows a total of 68 word subcategories, which can be summed up in five so called *word categories*: (1) Linguistic Dimensions, (2) Psychological Processes, (3) Personal Concerns, (4) Spoken Categories and (5) punctuation [303]. Linguistic dimension contains all subcategories relating to linguistic features, such as *Articles* and *Prepositions*. In Psychological processes, subcategories such as *Affective*, *Anxiety* and *Anger* are grouped together. Subcategories relating to topics of daily life, such as *Religion*, *Money* and *Family* are in the Personal concerns category. Lastly, the Spoken category refers to spoken events, namely *Swear words*, *Nonfluencies* and *Fillers*. In our analysis, we excluded the punctuation category as all punctuation is removed from the transcripts during the preprocessing step. For more detailed information we refer to the LIWC 2007 user manual [303]. In our analysis, we relied on the French adaptation of LIWC 2007 provided by Piolat et al. [313].

By analysing psycholinguistic aspects of speech, we hope to identify differences in psychological processes – reduced emotion over multiple categories as a sign of emotional blunting – and possibly in linguistic variables – as has been shown to be the case in depression. Each transcript was stripped of external markings (i.e. clinician speech), tokenised and then fed into LIWC. The result is a 68 dimensional feature vector. For more detailed investigation, we split up the vector into four feature subsets corresponding to the four LIWC categories.

	SVM			LR		
	Acc	F_1	AUC	Acc	F_1	AUC
WC	0.742	0.758	0.818	0.742	0.750	0.854
Sentiment	0.774	0.781	0.826	**0.790**	**0.794**	**0.874**
Sentiment + WC	**0.790**	**0.800**	**0.847**	**0.790**	0.787	0.835

Table 6.6: Classification results for models trained on word count and sentiment features. Best performances over feature sets in a particular metric are indicated in bold.

6.2.2.5 Prediction experiment

To verify the predictive power of the extracted features, we constructed machine learning models that are capable of deciding whether a person exhibiting signs of apathy. Each person in the dataset was assigned a label according to their group (Apathy vs. Control[3]). We compare multiple different feature sets: the word count (baseline), sentiment features, LIWC categories, all LIWC features and combinations thereof. We utilise different ML algorithms, namely SVM 2.4.3.2 and LR 2.4.3.1, and compare their performance. All classifier's implementations were taken from the scikit-learn framework [301]. Because of our limited data set (n=62), we cannot keep a held-out set for neither parameter tuning nor testing. Therefore, we exploit LOO-CV for testing our ML classifiers. Sentiment and word count features are normalised using z-standardisation based on the mean and standard deviation of the training set in each iteration and LIWC features are scaled using maximum absolute scaling to preserve sparseness. Hyperparameters are optimised using an inner loop of 3-Fold CV on the training set in each outer loop of the LOO-CV. For SVMs the choice of kernel (*linear*, *rbf*), error-parameter C (10^{-4}... 10^{4}) and for the *rbf* kernel parameter γ (10^{-6}... 10^{-3}) are optimised. For LR the penalty (L_1, L_2) and the error-parameter C (10^{-4}... 10^{4}) are optimised. To analyse performance of trained ML models, we report a variety of metrics: Accuracy, F_1 Score (F_1) and AUC.

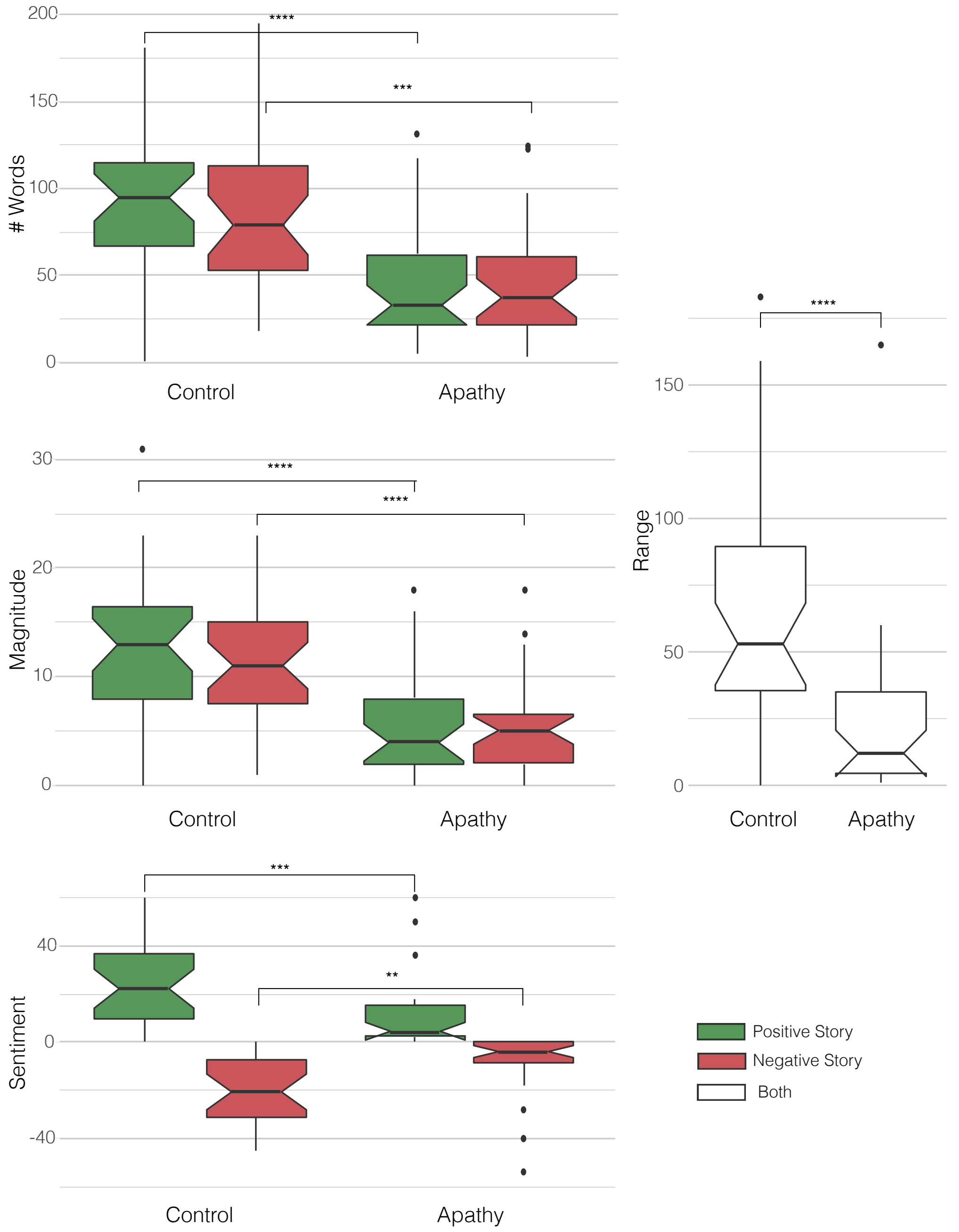


Figure 6.3: Boxplot group comparisons between control and apathy population *WC*, *S* and *M* in both positive (blue) and negative (red) story. Arches indicate group differences in a Wilcoxon test. $^* : p < 0.05$; $^{**} : p < 0.01$; $^{***} : p < 0.001$; $^{****} : p < 0.0001$

	SVM			LR		
	Acc	F_1	AUC	Acc	F_1	AUC
Linguistic	-	-	-	-	-	-
Psychological	-	-	-	0.565	-	-
Personal	0.613	0.647	0.680	0.597	0.627	0.599
Spoken	-	-	-	-	-	-
All	-	-	-	0.565	-	-

Table 6.7: Classification results for models trained on LIWC features. Best performances over feature sets in a particular metric are indicated in bold. “-“ indicates performances below the 0.5 chance baseline.

6.2.2.6 Results

First, the distribution of sentiment features and word count is reviewed to understand how people with apathy differ from controls. Figure 6.3 displays boxplots of these variable with group comparisons through non-parametric Kruskal-Wallis tests. Significant differences between Controls and people with apathy are found in the word count (Positive: $\chi^2 = 18.69$, df=1, $p < 0.001$; Negative: $\chi^2 = 14.83$, df=1, $p < 0.001$). People suffering from apathy use less words in their responses. This pattern is consistent between positive and negative story. A significant difference is also found in the magnitude (Positive: $\chi^2 = 17.42$, df=1, $p < 0.001$; Negative: $\chi^2 = 17.72$, df=1, $p < 0.001$), with people suffering from apathy showing smaller magnitudes. The sentiment of stories is also reduced in patients showing signs of apathy (Positive: $\chi^2 = 15.2$, df=1, $p < 0.001$; Negative: $\chi^2 = 14.46$, df=1, $p < 0.001$). Lastly, the range between sentiment in the positive and negative story is also greatly reduced for patients suffering from apathy ($\chi^2 = 17.68$, df=1, $p < 0.0001$).

Table 6.6 depicts the results of training ML models on either word count, sentiment features or both. Using the SVM classifier, the best results in all metrics–accuracy of 0.790, F_1 of 0.8 and AUC of 0.847–are obtained by combining both feature sources. The sentiment features alone show better performance than just the word count, with

[3] Here, Control refers to patients which do not suffer from apathy, regardless of cognitive status or other neurologic diseases they suffer from.

significantly higher accuracy (0.742 vs. 0.774) and higher F_1 (0.758 vs. 0.781). The differences between AUC are more marginal (0.818 vs. 0.826). Using LR, the best results are obtained by just using the sentiment feature set. Adding the word count to the sentiment set leads to comparable accuracy (0.790 vs. 0.790) and F_1 (0.787 vs. 0.794), but reduced AUC (0.835 vs. 0.874). Just using the word count leads to the worst accuracy (0.743) and F_1 (0.750), as well as reasonable AUC (0.854).

In Table 6.7 the results of training models on LIWC features of different word-categories are displayed. Generally, results are below the 0.5 chance baseline. For the linguistic word-category features, all models show performances below chance. The psychological category only shows an accuracy above chance (0.565) for the LR model. Features from the Personal category lead to models with above chance performances for both SVM (Acc=0.613; F_1=0.647; AUC=0.680) and LR (Acc=0.597; F_1=0.627; AUC=0.599), where SVM models show superior performance. The models trained on features from the Spoken word-category all showed performances below chance. Using all features, the LR model has an accuracy above chance (0.565).

Models trained on sentiment and word count clearly outperform LIWC features. Only the Personal word-category shows merit across all metrics and for both classification algorithms. To see if this can further improve classification performance we combine word count, sentiment and LIWC Personal features. In this experiment, features are scaled depending on their origin. Word count and Sentiment features are z-standardised and LIWC features scaled using maximum absolute scaling to preserve sparseness. The resulting classification models achieve a performance of Acc $= 0.758$, $F_1 = 0.769$, $AUC = 0.721$ for LR models and Acc $= 0.758$, $F_1 = 0.746$, $AUC = 0.812$ for SVM models. This is *not* an improvement above results previously achieved with only Sentiment and Word count features.

6.2.3 Effects of Automatic Speech Recognition

ASR has made great advances and can be considered a mature technology [414]. Here, we will examine to what extent classification results are maintained when no manual transcripts are available and ASR technology becomes imperative for analysis. This is a realistic scenario, as in a clinical setting the overhead of creating a manual transcript would be considered a big barrier for acceptance of such approaches.

6.2.3.1 Methodology

To automatically generate transcripts, commercially available ASR technology provided by Google[4] was used. Audio files are sent to the service via a REST API and a list of hypothesis transcriptions is returned. To create the final transcript we choose the most likely hypothesis. Word count and Sentiment features described in Section 6.2.2.2 are extracted from automatic transcripts. ML models are trained as described in Section 6.2.2.5. Experiments on LIWC features are not conducted on automatic transcripts, since they only showed limited merits on the manual ones.

As a general performance criterion for ASR, word error rate is calculated between the manual and automatic transcriptions. WER is a combination of the mistakes made by ASR systems in the process of recognition. Mistakes are categorised into substitutions, deletions and intrusions. Let S, D and I be the count of these errors respectively, and N be the number of tokens in the ground truth. Then WER is computed as given in Equation 4.1.

Seeing as Sentiment features play an important role in classification using manual transcripts, we will also in detail analyse how ASR errors affect them. For this, words will be split up in three categories: positive, negative and neutral. Positive and Negative words are words occurring and labeled in the FEEL dictionary (see Section 6.2.2.3), all other words are classified as being neutral. The specific error counts (S,D and I) will be analysed separately for each of the categories.

6.2.3.2 Error rates in ASR

The overall WER was 23.7%, with the majority of errors being deletions, 59.2% of the total error. Substitutions accounted for 30.5% of the total error and intrusions made up only 10.3% . Table 6.8 gives a more in depth look at the WER make up by considering the sentiment of the erroneous words with the FEEL dictionary, assigning either positive or negative valence. Words that were out of vocabulary of the FEEL dictionary were considered to be in a third category, neutral. Insertions from the manual to automatic transcripts were low overall. The number of inserted words per valence consisted of less than 3% of the overall word count in any category. Positive insertions counted 1.2% of all positive words, negative 1.2% and neutral 2.6%. Deletions accounted for the most

[4] https://cloud.google.com/speech-to-text/

	I	D	S	N	WER
Positive	15	99	63	1202	14.72%
Negative	4	39	18	341	17.88%
Neutral	357	2022	1029	13818	22.19%
Total	376	2160	1110	15361	23.74%

Table 6.8: Count of the errors from the WER for intrusions, deletions and substitutions separated by sentiment valence. I is for intrusions, D is for deletions. S is for substitutions and N is the word count in the ground truth.

likely error in the transcripts generated by ASR. Deletions from the positive subclass made up 8.2% of total positive words. From their respective subclasses, negative lost 11.4% and neutral 14.6%.

Figure 6.4 illustrates the shift in the sentiment classification of the errors from the manual to automatic transcripts for substitution errors. 86.5% of all substitutions resulted in a switch to the same category with 99% of same category substitutions going from a neutral valence to neutral (N2N). Due to the high number of N2N substitutions, 85.6% of all substitutions, it was removed from Figure 6.4 to better demonstrate the movement of errors that are represented by a sentiment valence in the FEEL dictionary. 13.5% of substitutions resulted in a category switch. 5.05% went from neutral in the manual transcript to positive in the automatic transcript, while 4.95% went from positive to neutral. 2.1% of errors went from neutral to negative and 1.3% went from negative to neutral. Less than 1% of substitutions switched valence from positive to negative or negative to positive.

6.2.3.3 ML Classification with ASR

Classification results for models trained on features extracted from ASR transcripts are reported in Table 6.9. Overall, decent performances are achieved. Using SVM as a classifier, the best $AUC = 0.864$ is achieved using only sentiment features. Accuracy and F_1 improve when adding the word count ($Acc = 0.774$, $F_1 = 0.817$). For LR, Sentiment features show the overall best performance ($Acc = 0.806$, $F_1 = 0.813$, $AUC = 0.849$). In both cases the lowest performance across all metrics is achieved

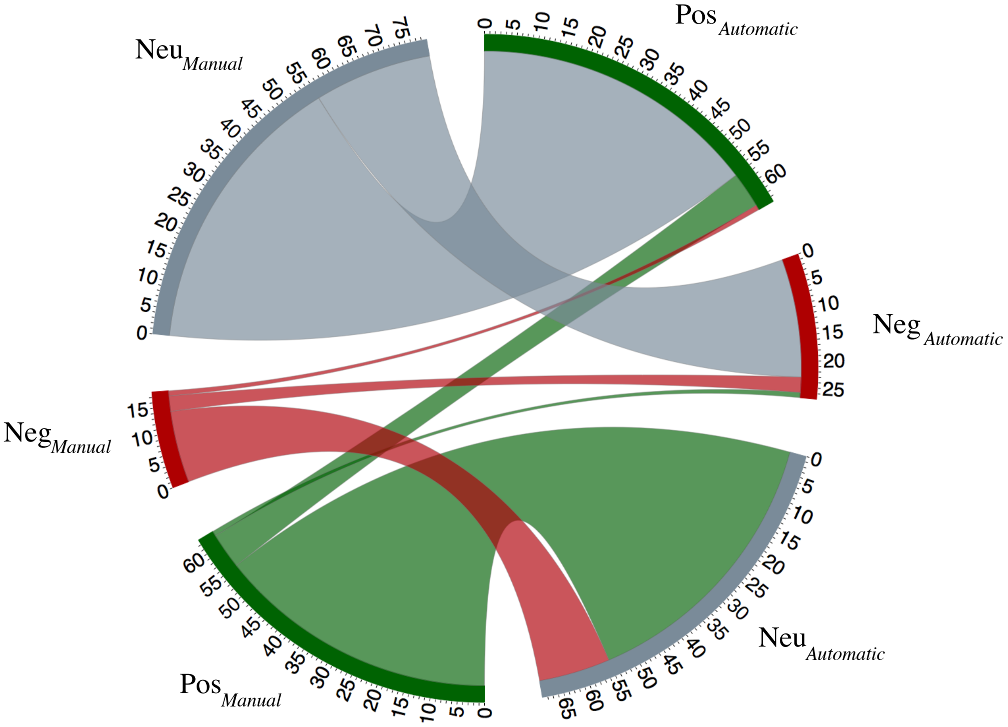


Figure 6.4: A chord diagram showing the shift in sentiment for substitution errors from the manual to automatic transcripts. The volume of the movement can be determined by the tick marks on the circumference of the circle as the raw count of the error. Green is used to represent positive, red for negative and gray for neutral. The opaque rim of the diagram is used to denote the sentiment of the category.

	SVM			LR		
	Acc	F_1	AUC	Acc	F_1	AUC
WC	0.709	0.736	0.835	0.758	0.769	0.823
Sentiment	0.774	0.774	**0.864**	**0.806**	**0.813**	**0.849**
Sentiment + WC	**0.774**	**0.817**	0.847	0.790	0.794	0.849

Table 6.9: Classification results for models trained on word count and sentiment features extracted from ASR transcripts. Best performances over feature sets in a particular metric are indicated in bold.

by only using the word count (SVM: $Acc = 0.709$, $F_1 = 0.736$, $AUC = 0.835$; LR: $Acc = 0.758$, $F_1 = 0.769$, $AUC = 0.823$).

6.2.4 Discussion

The results on manual transcripts clearly show the predictive power of linguistic sentiment features in predicting people with apathy from controls. Just looking at the word count, we already observe a steep reduction and group difference (see Figure 6.3), resulting in a strong baseline classification performance. Focusing on sentiment features, these outperform word count regardless of any metric or classifier. We observe a reduction in both magnitude and (absolute) sentiment over both tasks, which is in line with previous findings about emotional blunting in apathy [276]. Combining word count and sentiment features, an increase in predictive power is visible when using SVM as a classifier. Using LR, sentiment maintains the best performance. Both the diminished goal-directed behaviour – in form of shorter answers – as well as emotional blunting – through reduced sentiment – are detectable through these features and lead to classifiers with competitive performances.

LIWC features carry minimal, if any, information about a person's state of apathy. The classifier is not able to learn performance above the chance baseline in four out of five settings. This may be attributed to the well-known issue of data sparsity. Produced responses are very concise ($<$ 200 words) and therefore lead to sparse representation of LIWC sub-categories. Previous work has found this to cause insufficient performance [304]. Only the features from the Personal category show merit. The Personal category contains mostly sub-word-categories of Social and Semantic concepts (i.e. *Home*, *Death*, *Religion*) which are often represented by nouns or verbs. In our data, 38.9% of words in the personal category are nouns and 32.2% are verbs. The overall number of nouns and verbs used is also significantly different between the apathy and control group (Positive–Noun: $\chi^2 = 14.7$, df = 1, $p < 0.001$, Verb: $\chi^2 = 15.7$, df = 1, $p < 0.001$; Negative–Noun: $\chi^2 = 15.5$, df = 1, $p < 0.001$, Verb: $\chi^2 = 13.2$, df = 1, $p < 0.001$). This might be because in a story telling task, nouns represent actors, places and concepts, and verbs their actions, which are directly related to narrative length. Consequently, a correlation between the number of nouns and verbs used in the personal category and the overall word count is observed (Positive–Verb: $\rho = 0.46$, NOUN: $\rho = 0.66$; Negative–Verb:$\rho = 0.56$, Noun: $\rho = 0.69$). This further explains why adding the Personal word-category features to the previously best performing combination of Sentiment and Word Count, does not improve performances, as those features are implicitly counting nouns and verbs which are correlated with the word count.

Overall, a reasonable performance of an 0.874 AUC has been achieved in the best case. The next section will explore how stable these results and features are, when experiments are carried out on automatic instead of manual transcripts.

As previous research has shown [176], ASR is a viable tool for conducting automatic analysis in specific clinical settings, such as apathy detection in persons with cognitive impairment. Looking at the break down of WER in terms of sentiment, it is clear that there is no major shift in the overall sentiment of a transcript. WER is relatively low, 23.7%, considering the average age of the speakers. For comparison, WER rates in the range 26.3% to 34.1% have been reported for healthy individuals in the same age category.[154].

However, there is a surprising shift of neutral to positive valence in substitutions, and vice versa. A positive to neutral valence shift was anticipated as there are many words in the French language that would not be captured by the FEEL dictionary and it is likely that an automatic transcription error would find a lexical that is out of the bounds of the dictionary, resulting in the neutral classification. Less expected is that there is an almost equal proportion of neutral substitutions resulting in a positive valence. We looked at each of the errors in the transcripts and were not able to find a clear explanation for this phenomenon but would like to note that it makes up less than 3% of all error, and results in an equal shift in overall valence.

Classification results are stable between the manual and automatic setting. The best achieved results of $AUC = 0.874$ in the manual case is comparable to the best result of $AUC = 0.864$ in the automated case. Additionally, in both settings the best LR models are achieved when training on sentiment features, where SVM models are able to improve performance when the word count is added.

This study investigated the feasibility of using sentiment and psycholinguistic analysis of emotional language as a diagnostic screening tool for apathy.

Working on manual transcripts, sentiment and word count features both showed high baseline performances. Depending on the ML algorithm used, a combination of both feature sets lead to the best performance. Significant reductions in the number of words, the magnitude of sentiment and the overall sentiment were found for the apathetic population. This effect was consistent between the positive and the negative story. Psycholinguistic features extracted using LIWC mostly did not show any merit, with most word-categories having sub-random performances in classification. Only the personal category showed an AUC of 0.680 in the best case. This was determined to be due to its strong correlation to the word count.

ASR was introduced to fully automate the pipeline and increase its clinical feasibility. Classification performances remained largely the same with a small decrease ($AUC = 0.864$). Nevertheless, a WER of 23.7% was observed with over half of errors being deletions. This pattern was consistent when looking at sentiment word categories. A detailed analysis of substitution for these categories revealed, that a large proportion of positive and negative words were missrecognised by ASR as being a neutral word. At the same time, a roughly equal amount of neutral words were substituted with positive/negative words, keeping the overall distribution of sentiment word categories stable.

Overall, encouraging performances were achieved (best $AUC = 0.874$) when separating apathetic from non-apathetic patients using ML models. These were sustained and only dropped slightly when ASR was introduced to build a fully automatic pipeline (best $AUC = 0.864$). These results validate sentiment analysis as a potential tool for apathy detection in a clinical context. The examined robustness of results against ASR errors and the wide availability of quality ASR in multiple languages renders this approach a potential low-cost screening tool for clinical apathy.

The understanding of some limitations is vital for the correct interpretation of the presented results. First, the included population is rather small. Although, one always strives to perform an experiment on the biggest possible population, collection of clinical data is a tedious, involved and therefore an expensive process. The size of the dataset is further reduced through the process of matching demographic and clinical variables which is important to be able to draw meaningful conclusions from findings. Another problem with the population is the slight difference in cognitive health as measured by MMSE. Due to the strong correlation between affective syndromes such as apathy and cognitive decline, matching populations to have insignificant differences in MMSE would lead to a substantial loss of data. However, we do not expect slight differences in cognitive health of this magnitude to have a direct effect on any of the measured variables.

6.3 Summary

This chapter explored approaches to detect clinical apathy in people suffering from dementia based on analysis of their answers to emotional questions.

Section 6.1 specifically focused on acoustic analysis of these responses. The population of older French adults suffering from early dementia was split into males and females to account for gender differences in voice patterns. Each sub-population was matched between apathy and non-apathy patients inside a gender. Prosodic, Formant, Temporal and Source measures were automatically extracted from the audio signal, for both the positive and negative story separately. Overall, both males and females showed reduced reaction to the stimuli. Answers to the posed questions can be generally characterised by drastically shorter (lower sound duration) and slower (lower Speech Rate) speech. For the female population, a difference in voice quality (lower HNR) is obvious in both questions. Males suffering from apathy react less emotionally to the negative question as indicated by a lower variance of prosody (lower F0 Range). Interestingly, male and female subjects with apathy show different patterns in their speech features according to the type of free speech task. For males, significant differences between apathy and non-apathy subjects can be seen in temporal features for both the negative and positive story. Females show similar patterns in the positive story, but not in the negative one. Correlation analysis with diagnostic sub-scales of apathy revealed strong correlation between the sub-domains interest and initiative of the AI and temporal speech features.

The affective subdomain, which represents the emotional blunting in apathy, seemed more associated with prosodic speech features. A classifier trained on the data was able to separate the groups with an $AUC = 0.88$ for the males and $AUC = 0.77$ for the females.

Section 6.2 focused on analysis of the semantic content of the given responses. Populations were no longer split by gender and sentiment as well as psycholinguistic features were extracted from text transcripts of the responses. Reduced sentiment and emotional range was found in both positive and negative story. Trained classifiers already showed good performance when only including the word count and could be further improved by adding sentiment features. Psycholinguistic features were examined overall and in sub-categories and showed no merit. Automatic speech recognition was used to produce automatic transcripts with an overall word error rate of 23.7%. Features were again extracted from these transcripts and classifiers trained on them showed nearly the same performance as the ones trained on manual transcripts.

The next section will close this thesis by summing up the major results from each chapter and putting them into the bigger picture. Furthermore, applications of results and future work will be discussed.

Chapter 7

Conclusions and Future Work

7.1 Thesis summary

This thesis investigated the possibility to utilise speech and language analysis together with machine learning, to automatically detect people with early cognitive impairments as well as emotional disturbances in uni- and multilingual settings.

From the related literature, verbal fluency exercises were identified to be appropriate tasks to provoke speech from which early signs of cognitive impairment are apparent. A variety of new semi- and fully-automatic analysis techniques for this task were introduced and validated on data from persons with mild cognitive impairment. Classification results clearly indicated the predictive power of these analysis in discriminating people with early cognitive impairments and normally ageing adults.

Experiments using data from different languages to amend the small available resources showed success. Through acoustic analysis, data from a larger English resource of picture description tasks could be used to improve the detection of persons with mild cognitive impairment in French and Swedish data. In a second step, linguistic transcripts from the English data resource were semantically analysed and combined with a small french data set into predictive models through domain adaptation. Multilingual experiments clearly showed the benefits of adding foreign language data for detection of Alzheimer's disease. With a sufficient amount of data, a good cross-lingual performance was achieved as well.

Finally, a free speech task was used to try to automatically separate people with cognitive impairment that suffer from an affective disorder from those who do not. In acoustic as well as linguistic analysis, clear differences between the groups emerged. Acoustic patterns were correlated with fine-grained medical information to explain the relation to different symptoms of the disorder. Sentiment and psycholinguistic analysis were used to analyse text transcripts, which also resulted in models with high predictivity between the two groups. Automatic speech recognition was not found to severely worsen results.

7.2 Contributions

This work contributed methodology and knowledge to both medical and computational domains. It advances the complex issue of automatic early detection of cognitive decline, introduces novel approaches to utilise multi- and cross-lingual data resources in this domain and deals with the important problem of automatically detecting affective disorders in this patient group.

The feasibility of using natural language processing and machine learning to detect early signs of cognitive decline through analysing speech samples was shown. Novel analysis methods for a classical speech-based cognitive test, verbal fluency, were introduced and validated on a patient population. Experiments using automated speech recognition showed that this approach is also viable for use in real world face-to–face, as well as tele-medicine settings. It may be concluded, that speech is a feasible screening tool for early detection of dementia.

World first experiments on multilingual dementia detection were preformed. We showed beyond a reasonable doubt, that there are clear benefits in combining multilingual resources and that, given a large enough data set, even cross-lingual models are capable of detecting dementia. Furthermore, a new analysis method for the CTP task was introduced and validated.

This work also validated speech analysis as a feasible tool in detecting apathy in older demented adults. We showed that some characteristics of speech were directly indicative for specific symptoms observed in this disorder. Furthermore, we linked apathy to a reduction in conveyed sentiment as a reaction to an emotional question. This form of linguistic analysis was also validated to be feasible using ASR.

The research questions raised in Section 1.2, at beginning of the thesis, were addressed in the following way:

1. **Can Mild Cognitive Impairment be automatically detected from concise speech recordings?**
   The construction of automatic diagnostic models for MCI from speech samples was investigated in Chapter 3 and Chapter 4. The presented approach will focused on recordings of verbal fluency tasks, in which patients are asked to name as many words according to a given rule as possible in a given time frame (e.g., as many animals as possible in 60 seconds). Clinical performance in these test is usually assessed as the number of correct words and has been shown to be highly predictive for MCI. We introduced and validated novel and extended automatic analysis methods and showed that they improve the diagnostic ability of these tasks. Both the functionality to detect (see Section 3.1, 3.2, and 3.3) and stage dementia (see Section 3.4) was explored. In addition to validation experiments on manual transcripts, fully automatic experiments using automatic speech recognition were carried out (see Section 4.1). These approaches validated the utility of such analysis and assessment methods for real-world broad population screening applications—i.e., over the telephone (see Section 4.2).

2. **How can data resources from different languages be leveraged in multi- and cross-lingual dementia detection?**
   Multilingual analysis methods that allow increasing the productivity of models in under-resourced languages were explored in Chapter 5. The most widely available speech data in most language are picture descriptions of the Boston Cookie Theft Picture. The use of English data to improve productivity in other languages was explored (see Section 5.1). In addition to general domain adaptation methods, novel multi-lingual analysis methods were used (see Section 5.2).

3. **Can affective disorders in dementia be automatically detected based on speech recordings?**
   Methods for the detection of clinical apathy in dementia patients from speech and language were explored in Chapter 6. To this end, a subpopulation of cognitively matched patients telling positive and negative stories were analysed. Speech and signal processing (see Section 6.1), as well as sentiment and psycholinguistic language analysis (see Section 6.2), were considered as a diagnostic marker and

validated using machine learning.

7.3 Future Work

The field of using voice as a biomarker for diagnosis of dementia is still young and many areas have not been researched yet. The following topics have been identified as crucial to address in the future to further the clinical applicability of these vocal biomarkers.

7.3.1 Standardised data collection

The base for any research in this field is the collection of high quality clinical, as well as patient-generated speech and language data. Since single studies are limited by budget and time constraints to collect large amounts of data, standardisation of data collection protocols becomes a priority to share and combine data from different sources. This presents clinical, legal and ethical challenges that have to be addressed by researchers across domain barriers. To this end, colleagues and us have written a position paper on the necessary steps to establish data sharing in this community in the future [124].

7.3.2 Longitudinal data

Especially in early and pre-clinical stages of Alzheimer's disease, using a single measurement point of any diagnostic biomarker will only identify a part of the affected population. A higher sensitivity for early stages can be achieved by looking at the development of a single person over time. This paradigm can and should also be applied to voice and language data collection.

7.3.3 New protocols

This thesis focused heavily on the analysis of a single voice based cognitive task–the semantic verbal fluency. Although this task has broad applicability in the early diagnosis of dementia, as shown in Chapter 3 and 4, and has the advantage of being administrable in a short amount of time, it does not capture all cognitive impairments that can be present in early Alzheimer's patients. Future studies should analyse a broader protocol of voice based cognitive exams and combine their analysis results, while building on the results uncovered in this thesis. Memory test, which where specifically designed to measure learning and memory abilities, are a prime candidate for including into such a protocol.

7.3.4 Differentiating different dementias

An interesting topic to consider is the ability of speech and language to be used as a differential diagnostic marker between different forms of dementia. Next to Alzheimer's disease, being responsible for 50% of cases, there are other common organic causes for dementia, such as vascular pathologies and Parkinson's disease. To bring vocal biomarkers closer to real-world clinical application scenarios, we have to acknowledge the complexity of diagnosis. The Patients that walk into a memory clinic everyday are not as neatly divided into healthy controls, MCI and AD patients, as in the here-presented experiments. Exploring how diagnostic markers, coming from voice or other sources, react to these other and less common pathologies is important to enable real-world application. We have tried to address this topic to some extend, by diving into possible comorbidities in the form of affective disorders in Chapter 6. Even in this area, a lot of work is still to be done as exemplified in the next point.

7.3.5 Detection of Depression

Similar to the separation of different forms of dementia, looking into how speech and language can be used to detect depression in dementia patient is a worthwhile topic. Although there already is a large body of research on how speech can be used as a biomarker in depression as a separate condition [87], it would be important to research how speech and language biomarkers can be used to identify signs of depression in early MCI patients and how to separate the two. Depression, together with other affective disorders, is among the top risks for MCI patients to convert into dementia quicker.

7.4 Closing Remarks

Dementia has a large economic impact on our society. This work introduced and validated methods for the automatic detection of early stage dementia and related affective disorders through processing speech and spoken language. These findings open up the possibility for multiple clinical applications that could help to pervasively screen for dementia in large populations and thereby treat and slow disease progression. Technologies as the ones described in this thesis could therefore be used to prevent harm and suffering caused by these devastating diseases. Further research is needed to validate the proposed approaches is larger clinical cohorts and across different disease areas, before this technology can be applied in clinical practice.

Appendix A

Abbreviations and Definitions

This section provides a list of abbreviations used throughout the thesis. A small definition is given for each concept. It serves as a list for readers to return to, when the meaning of an abbreviation is unclear.

A	Apathy, a psychiatric affective syndrome often observed in demented patients (see Section 2.2)
AD	Alzheimer's Disease, a neurodegenerative disorders that leads to severe cognitive impairments (see Section 2.1.1.2)
ADRD	Alzheimer's Disease and related Dementias, a group of disorders, Alzheimer's and other types of dementia, that a very similar to Alzheimer's in their symptomatic
AI	Apathy inventory
aMCI	Amnestic MCI
ALS	Amyotrophic lateral sclerosis
ASR	Automatic Speech Recognition, technology used to automatically transcribe the words contained in an audio recording
BNT	Boston Naming Test
CDR	Clinical Dementia Rating Scale

CDR-SOB	Clinical Dementia Rating Scale Sum of Boxes
CT	Computer Tomography
CTP	Cookie Theft Picture, a line drawing of a kitchen scene. Often used as stimulus used to elicit free speech from patients
DB	Dementia Bank, a linguistic corpus containing English recordings of dementia patients performing the Cookie Theft picture description task. The data was collected by [240].
DSM-V	Diagnostic and Statistical Manual Version V
ESA	Explicit Semantic Analysis
fMRI	Functional Magnetic Resonance Imaging
FTLD	Frontotemporal lobe degeneration
GDS	Global Deterioration Scale
HC	Healthy Control, a healthy person included in study to serve as a point of reference
HD	Huntington's disease
LM	Language Model, a probabilistic model used to represent the probability distribution of language data in either spoken or written form (see Section 3.2.2)
LOOCV	Leave-One-Out Cross validation
LR	Logistic Regression, a class of machine learning models using a logistic loss function (see Section 2.4.3.1)
LSA	Latent Semantic Analysis
MAE	Mean absolute error
MCI	Mild Cognitive Impairment, a stage of cognitive impairment known as one of the predecessors to AD (see Section 2.1.1.3)

MCS	Mean cluster size, refers to the mean number of clustered words produced in succession during the semantic verbal fluency task (see Section 2.3.1.1)
ML	Machine Learning
MMSE	Mini Mental State Examination, a short screening test for dementia designed by [116]. The resulting score can be used as a global index of cognitive health.
MoCA	Montreal Cognitive Assessment, a short screening test used to detect early stages of dementia.
MRI	Magnetic resonance imaging
NLP	Natural language processing
NOS	Number of Switches, refers to the number of semantic category breaks in verbal fluency tasks (see Section 2.3.1.1)
PD	Parkinson’s disease
PET	Positron emission tomography
PPA	Primary progressive aphasia
PVF	Phonemic Verbal Fluency, a cognitive task in which patients are asked to name as many words starting with a certain letter as possible in a given time interval
RMSE	Root mean square error
ROC	Receiver operator curve
SA	Sentiment analysis
SCI	Subjective Cognitive Impairment, the concept refers to people who have a subjective complaint about a decline in cognitive ability that is not supported by an objective evaluation
SD	Standard deviation

SVM Support Vector Machine, a machine learning model that uses so called support vectors to construct its decision boundary (see Section 2.4.3.2)

SVF Semantic Verbal Fluency, a cognitive task in which patients are asked to name as many words belonging to a semantic category as possible in a given time interval (see Section 2.3.1)

SVR Support Vector Regression, regression model based on the same idea as SVM

TMT Trail making test

VD Vascular dementia

VF Verbal Fluency, a cognitive task in which patients are asked to name as many words under a given semantic (see SVF) or phonemic (see PVF) constraint as possible in a given time interval

Appendix B

Cognitive Domains in DSM-5

This appendix contains a detailed listing of the cognitive domains from the *Diagnostic and Statistical Manual of Mental Disorders* (DSM–5) [26] mentioned in Section 2.3.

Cognitive Domain	Sub-Domain	Description	Example Test
Complex attention	*Sustained attention*	Holding up attention over longer periods of time	Pressing an input every time a tone is heard, over long periods of time
	Selective attention	Holding up attention despite competing stimuli and/or distractors	Hearing letters and numbers, pressing an input only on numbers
	Divided attention	Attending to two competing tasks at the same time	Pressing an input when seeing a red stimulus and counting the stimuli
	Processing speed	Speed at which information can be processed	Counting from 970 to 1000
Executive function	*Planning*	Ability to plan actions required to achieve a specified goal	Describing the steps required to pack a suitcase
	Decision making	Deciding between competing alternatives	Simulated gambling task
	Working memory	Ability to, for a short period of time, hold and manipulate information	Sequential addition of heard numbers
	Feedback/error utilisation	Inferring the rules of a given scenario through continuous feedback	–

	Inhibition	Inhibiting an urge to achieve a correct outcome	Naming the color a color-word is drawn in, not its meaning
	Mental/cognitive flexibility	Ability to shift between two concepts	Mapping numbers to letters
Learning and memory	*Immediate memory span*	Ability to repeat information	Hearing a list of words and repeating it immediately
	Recent memory	Storing new information of longer periods of time	Repeating a newly learned word list twenty minutes later
	Long-term memory	Storing information over extended periods of time	Talking about past life events
Language	*Expressive language*	Describing situations/objects or recalling families of words	Naming as many animals as possible in 60 seconds
	Grammar and Syntax	Use of grammatical constructions, articles prepositions and auxiliary verbs	Describing a picture
	Receptive language	Comprehension of language	Performing tasks to verbal command
Perceptual-motor function	*Visual perception*	Understanding visual inputs	Line bisection task
	Visuoconstructional	Creation of items requiring coordination of eye and hand	Copying a complex figure
	Perceptual-motor skills	Integrating perception with purposeful movements	Inserting differently shaped blocks in a form board
	Praxis	Integrity of learned movements	Pantomime use of objects (e.g. a hammer)

	Gnosis	Perceptual integrity of awareness and recognition	Recognising faces and colors
Social cognition	*Recognition of emotions*	Recognising emotions in images or faces	Emotion recognition picture sets
	Theory of mind	Capacity to consider other persons mental sates	Asking questions to emotionally complex pictures

Table B.1: Description of cognitive domains and sub-domains according to DSM-5.

Appendix C

Literature Review

This chapter contains a collection of related literature from the space of automatic dementia detection from speech and language. All listed papers used speech or language analysis on patient populations and classified these using machine learning. The number of patients included, the pathology of the population, the language of speech samples, the task used to elicit speech, the kind of analysis (language or speech), if the experiment was fully automatic or required transcription and the classification performance are reported.

Author	Year	N	Population	Language	Task	Features	Auto	Metric	Results
Al-Hameed et al. [11]	2016	473	AD	English	PD	S	✓	ACC	94.7
Al-Hameed et al. [12]	2017	64	MCI&AD	English	PD	S	✓	MAE	3.1
Alhamai et al. [15]	2017	92	AD	English	Other	S&L		AUC	92
Alhanai et al. [13]	2018	92	AD	English	Multi-Task	S&L		AUC	76
Alúsio et al. [17]	2016	60	MCI&AD	Portuguese	Narration	L		F1	81.7
Asgari et al. [23]	2017	41	MCI	English	Interview	L		ACC	84
Balagopalan et al. [29]	2018	500+	AD	English	Multi-Task	S&L		F1	89
Bucks et al. [60]	2000	24	AD	English	Interview	L		ACC	100
Budhkar et al. [61]	2018	246	AD	English	PD	L		F1	77.5
Espinoza-Cuadros et al. [107]	2014	19	MCI	Spanish	Reading	S	✓	ACC	78.9
Fraser et al. [128]	2016	473	AD	English	PD	S&L		ACC	81
Fraser et al. [126]	2018	105	MCI	Swedish	PD	L		ACC	85
Fraser et al. [125]	2018	66	MCI	Swedish	PD	L		AUC	87
Garrard et al. [130]	2014	42	SD	English	PD	L		ACC	90
Gonzalez-Moreira et al. [140]	2014	19	MCI	Spanish	Reading	S	✓	ACC	84.2
Gonzalez-Moreira et al. [141]	2015	20	MCI	Spanish	Reading	S	✓	ACC	85.0
Gosztolya et al. [146]	2016	84	MCI	Hungarian	Story Retelling	S	✓	F1	89.1
Gosztolya et al. [147]	2018	75	MCI&AD	Hungarian	Story Retelling	S&L		F1	86
Hernández et al. [161]	2016	63	AD	Spanish	Object description	L		F1	88
Hernández et al. [162]	2018	262	MCI&AD	English	PD	S&L		AUC	76
Jarrold et al. [177]	2010	44	AD	English	Interview	L		ACC	82.6
Jarrold et al. [176]	2014	48	Multiple	English	Multi-Task	S&L	✓	ACC	80

Khodabakhsh et al. [190]	2014	54	AD	Turkish	Interview	S	✓	ACC	94
König et al. [205]	2015	64	MCI&AD	French	Multi-Task	S	✓	ACC	87
Koenig et al. [204]	2018	95	MCI&AD	French	VF	L	✓	AUC	75.8
Lehr et al. [216]	2012	72	MCI	English	Story Retelling	L	✓	AUC	80.9
Lizarduy et al. [95]	2017	225	MCI	Spanish	VF	S	✓	ACC	75
Lopez-de-Ipinã et al. [228]	2013	70	AD	Multilingual	Interview	S		ACC	up to 100
Lopez-de-Ipinã et al. [227]	2015	70	AD	Multilingual	Interview	S&L		ACC	97.7
Lopez-de-Ipinã et al. [231]	2015	225	MCI	Spanish	VF	S	✓	ACC	84.9
Lopez-de-Ipinã et al. [229]	2015	70	AD	Multilingual	Interview	S	✓	ACC	92.5
Lopez-de-Ipinã et al. [230]	2017	225	MCI&AD	Spanish	Multi-Task	S	✓	ACC	80
Luz et al. [234]	2017	398	AD	English	PD	S	✓	AUC	73.4
Martínez-Sánchez et al. [246]	2018	145	AD	Spanish	Reading	S	✓	ACC	92.4
Masrani et al. [247]	2017	309	MCI	English	PD	S&L		F1	71.2
Meilan et al. [254]	2014	66	AD	Spanish	Reading	S	✓	ACC	84.8
Mirheidari et al. [259]	2016	39	AD	English	Interview	S&L	✓	ACC	95
Mirheidari et al. [258]	2017	30	FMD	English	Interview	L		ACC	97
Mirheidari et al. [261]	2018	?	AD	English	Other	S	✓	ACC	?
Mirheidari et al. [260]	2019	30	FMD	English	Interview	S&L	✓	ACC	90
Mirzaei et al. [262]	2018	48	MCI&AD	French	Reading	S	✓	ACC	61
Orimaye et al. [288]	2014	484	AD	English	PD	L		F1	74
Orimaye et al. [287]	2016	38	MCI	English	PD	L		ACC	87.5
Orimaye et al. [286]	2017	198	AD	English	PD	L		AUC	93.3
Orimaye et al. [289]	2018	99	MCI	English	PD	L		AUC	80
Pompili et al. [315]	2018	475	AD	English	PD	L		ACC	77

Pou-Prom et al. [316]	2018	499	AD	English	Multi-Task	L		F1	82.4
Prud'hommeaux et al. [322]	2011	393	MCI	English	Story Retelling	L		AUC	82.7
Rentoumi et al. [325]	2014	36	AD	English	PD	L		ACC	75
Sadeghian et al. [336]	2017	72	AD	English	PD	S&L	✓	ACC	91.7
Satt et al. [342]	2013	89	MCI&AD	Greek	Multi-Task	S	✓	ACC	84.5
Satt et al. [341]	2014	64	MCI&AD	French	Multi-Task	S	✓	ACC	87
Sirts et al. [350]	2017	499	AD	English	Other	L		F1	84
Themistocleous et al. [368]	2018	55	MCI	Swedish	PD	S	✓	ACC	83
Thomas et al. [369]	2005	95	AD	English	Interview	L		ACC	95
Tóth et al. [374]	2015	51	MCI	Hungarian	Story Retelling	S	✓	F1	85.3
Tóth et al. [375]	2018	86	MCI	Hungarian	Story Retelling	S	✓	AUC	73.4
Vincze et al. [387]	2016	29	MCI	Hungarian	Story Retelling	S&L		F1	75.1
Wankerl et al. [392]	2017	499	AD	English	PD	L		AUC	83
Warnita et al. [393]	2018	488	AD	English	PD	S	✓	ACC	73.6
Weiner et al. [398]	2016	98	AD	German	Interview	S	✓	F1	80
Weiner et al. [399]	2016	23	AD	German	Interview	S	✓	F1	87
Weiner et al. [?]	2017	74	AD	German	Interview	S&L		UAR	62.3
Weiner et al. [401]	2018	218	AD	German	Interview	S	✓	UAR	64.5
Weiner et al. [395]	2018	98	AD	German	Interview	S&L		UAR	81.9
Weiner et al. [400]	2018	51	AD	German	Interview	S&L	✓	UAR	82.6
Yancheva et al. [408]	2015	393	AD	English	PD	S&L		MAE	2.91
Zhou et al. [417]	2016	473	AD	English	PD	L		ACC	-
Zhu et al. [419]	2018	412	AD	English	Multi-Task	S&L		ACC	78

Zhu et al.[420]	2018	246	AD	English	PD	S&L	ACC	79.9

Bibliography

[1] P. Aalten, F. R. Verhey, M. Boziki, R. Bullock, E. J. Byrne, V. Camus, M. Caputo, D. Collins, P. P. De Deyn, K. Elina, G. Frisoni, N. Girtler, C. Holmes, C. Hurt, A. Marriott, P. Mecocci, F. Nobili, P. J. Ousset, E. Reynish, E. Salmon, M. Tsolaki, B. Vellas, and P. H. Robert. Neuropsychiatric syndromes in dementia. Results from the European Alzheimer Disease Consortium: part I. *Dementia and geriatric cognitive disorders*, 24(6):457–463, 2007.

[2] A. Abdaoui, J. Azé, S. Bringay, N. Grabar, and P. Poncelet. Expertise in french health forums. *Health informatics journal*, 25(1):17–26, 2019.

[3] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

[4] A.-L. R. Adlam, S. Bozeat, R. Arnold, P. Watson, and J. R. Hodges. Semantic knowledge in mild cognitive impairment and mild alzheimer's disease. *Cortex*, 42(5):675–684, 2006.

[5] S. Ahmed, C. A. de Jager, A.-M. Haigh, and P. Garrard. Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed alzheimer's disease. *Neuropsychology*, 27(1):79–85, 2013.

[6] S. Ahmed, C. A. de Jager, A.-M. F. Haigh, and P. Garrard. Logopenic aphasia in alzheimer's disease: clinical variant or clinical feature? *Journal of Neurology, Neurosurgery and Psychiatry*, 83(11):1056–1062, 2012.

[7] S. Ahmed and P. Garrard. Spoken discourse in Alzheimer's disease: a review. *Linguistica*, 52(1):9, 2012.

[8] S. Ahmed, A.-M. F. Haigh, C. A. de Jager, and P. Garrard. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737, 2013.

[9] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski. A comparison of auc estimators in small-sample studies. In *Machine learning in systems biology*, pages 3–13. PMLR, 2009.

[10] P. Aisen, J. Touchon, R. Amariglio, S. Andrieu, R. Bateman, J. Breitner, M. Donohue, B. Dunn, R. Doody, N. Fox, S. Gauthier, M. Grundman, S. Hendrix, C. Ho, M. Isaac, R. Raman, P. Rosenberg, R. Schindler, L. Schneider, R. Sperling, P. Tariot, K. Welsh-Bohmer, M. Weiner, and B. Vellas. EU/US/CTAD Task Force: Lessons Learned from Recent and Current Alzheimer's Prevention Trials. *J Prev Alzheimers Dis*, 4(2):116–124, 2017.

[11] S. Al-Hameed, M. Benaissa, and H. Christensen. Simple and robust audio-based detection of biomarkers for alzheimer's disease. In *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 32–36, 2016.

[12] S. Al-Hameed, M. Benaissa, and H. Christensen. Detecting and predicting alzheimer's disease severity in longitudinal acoustic data. In *Proceedings of the International Conference on Bioinformatics Research and Applications 2017*, pages 57–61. ACM, 2017.

[13] T. Al Hanai, R. Au, and J. Glass. Role-specific language models for processing recorded neuropsychological exams. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 746–752, 2018.

[14] M. S. Albert, S. T. DeKosky, D. Dickson, B. Dubois, H. H. Feldman, N. C. Fox, A. Gamst, D. M. Holtzman, W. J. Jagust, R. C. Petersen, et al. The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, 7(3):270–279, 2011.

[15] T. Alhanai, R. Au, and J. Glass. Spoken language biomarkers for detecting cognitive impairment. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 409–416. IEEE, 2017.

[16] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, 2005.

[17] S. Aluísio, A. Cunha, and C. Scarton. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 109–114. Springer, 2016.

[18] Alzheimer's, Association. 2015 alzheimer's disease facts and figures. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 11(3):332, 2015.

[19] Alzheimer's Association. What is alzheimer's?, 2018.

[20] Alzheimer's Association. What is dementia?, 2018.

[21] F. Aminzadeh, A. Byszewski, F. J. Molnar, and M. Eisner. Emotional impact of dementia diagnosis: Exploring persons with dementia and caregivers perspectives. *Aging & Mental Health*, 11(3):281–290, 2007.

[22] E. Aramaki, S. Shikata, M. Miyabe, and A. Kinoshita. Vocabulary size in speech may be an early indicator of cognitive impairment. *PLOS ONE*, 11(5):1–13, 2016.

[23] M. Asgari, J. Kaye, and H. Dodge. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228, 2017.

[24] L. Ashendorf, A. L. Jefferson, M. K. O'Connor, C. Chaisson, R. C. Green, and R. A. Stern. Trail making test errors in normal aging, mild cognitive impairment, and dementia. *Archives of Clinical Neuropsychology*, 23(2):129–137, 2008.

[25] A. Association et al. 2016 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 12(4):459–509, 2016.

[26] A. P. Association, C. on Nomenclature, Statistics, et al. *Diagnostic and statistical manual: mental disorders (DSM-5)*. American Psychiatric Association, 2013.

[27] S. Auer and B. Reisberg. The GDS/FAST staging system. *International Psychogeriatrics*, 9(SUPPL. 1):167–171, 1997.

[28] S. Auriacombe, N. Lechevallier, H. Amieva, S. Harston, N. Raoux, and J.-F. Dartigues. A Longitudinal Study of Quantitative and Qualitative Features of Category Verbal Fluency in Incident Alzheimer's Disease Subjects: Results from the PAQUID Study. *Dementia and geriatric cognitive disorders*, 21(4):260–266, 2006.

[29] A. Balagopalan, J. Novikova, F. Rudzicz, and M. Ghassemi. The effect of heterogeneous data for alzheimer's disease detection from speech. *arXiv preprint arXiv:1811.12254*, 2018.

[30] C. Ballard, S. Gauthier, A. Corbett, C. Brayne, D. Aarsland, and E. Jones. Alzheimer's disease. *Lancet*, 377(9770):1019–1031, Mar. 2011.

[31] V. Bambini, G. Arcara, I. Martinelli, S. Bernini, E. Alvisi, A. Moro, S. F. Cappa, and M. Ceroni. Communication and pragmatic breakdowns in amyotrophic lateral sclerosis patients. *Brain and Language*, 153:1–12, 2016.

[32] D. M. Barch, D. Pagliaccio, and K. R. Luking. Mechanisms underlying motivational deficits in psychopathology: Similarities and differences in depression and schizophrenia. *Current topics in behavioral neurosciences*, 27:411–49, 2016.

[33] A. Barney, D. Nikolic, V. Nemes, and P. Garrard. Detecting repeated speech: A possible marker for alzheimer's disease. *Models and Analysis of Vocal Emissions for Biomedical Applications*, pages 31–32, 2013.

[34] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[35] D. Bates, M. M??chler, B. Bolker, and S. Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015.

[36] K. A. Bayles, A. W. Kaszniak, and C. K. Tomoeda. *Communication and Cognition in Normal Aging and Dementia.* College-Hill Press/Little, Brown & Co, 1987.

[37] K. A. Bayles and C. K. Tomoeda. Confrontation naming impairment in dementia. *Brain and Language*, 19(1):98–114, 1983.

[38] K. A. Bayles, C. K. Tomoeda, P. E. McKnight, N. Helm-Estabrooks, and J. N. Hawley. Verbal perseveration in individuals with alzheimer's disease. In *Seminars in Speech and Language*, volume 26, pages 335–347, 2004.

[39] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.

[40] J. K. Belanoff, J. Jurik, L. D. Schatzberg, C. DeBattista, and A. F. Schatzberg. Slowing the progression of cognitive decline in alzheimer's disease using mifepristone. *Journal of Molecular Neuroscience*, 19(1-2):201–206, 2002.

[41] K. Berg, T. Askim, S. Balandin, E. Armstrong, and M. B. Rise. Experiences of participation in goal setting for people with stroke-induced aphasia in norway. a qualitative study. *Disability and rehabilitation*, 39(11):1122–1130, 2017.

[42] F. Bernardini, A. Lunden, M. Covington, B. Broussard, B. Halpern, Y. Alolayan, A. Crisafio, L. Pauselli, P. M. Balducci, L. Capulong, L. Attademo, E. Lucarini, G. Salierno, L. Natalicchi, R. Quartesan, and M. T. Compton. Associations of acoustically measured tongue/jaw movements and portion of time speaking with negative symptom severity in patients with schizophrenia in italy and the united states. *Psychiatry Research*, 239:253–258, 2016.

[43] K. C. Berridge and T. E. Robinson. Parsing reward. *Trends in neurosciences*, 26(9):507–513, 2003.

[44] C. Bickel, J. Pantel, K. Eysenbach, and J. Schröder. Syntactic comprehension deficits in alzheimer's disease. *Brain and Language*, 71(3):432–448, 2000.

[45] J. Blitzer, R. McDonald, and F. Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

[46] L. X. Blonder, E. D. Kort, and F. A. Schmitt. Conversational discourse in patients with alzheimer's disease. *Journal of Linguistic Anthropology*, 4(1):50–71, 1994.

[47] A. Blum and T. Mitchell. Combining labeled and unlabelled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

[48] P. Boersma. Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9):341–345, 2001.

[49] M. F. Bonner, S. Ash, and M. Grossman. The New Classification of Primary Progressive Aphasia into Semantic, Logopenic, or Nonfluent/Agrammatic Variants. *Current Neurology and Neuroscience Reports*, 10(6):484–490, 2010.

[50] L. Borin, M. Forsberg, M. Hammarstedt, D. Rosen, R. Schäfer, and A. Schumacher. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University*, pages 17–18, Nov. 2016.

[51] L. Borin, M. Forsberg, and J. Roxendal. Korp – the corpus infrastructure of Språkbanken. In *The 8th international conference on Language Resources and Evaluation (LREC)*, pages 474–478, Istanbul, Turkey, 2012.

[52] V. Boschi, E. Catricala, M. Consonni, C. Chesi, A. Moro, and S. F. Cappa. Connected speech in neurodegenerative language disorders: a review. *Frontiers in Psychology*, 8:1–21, 2017.

[53] N. Botting. Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching and Therapy*, 18(1):1–21, 2002.

[54] E. Brando, R. Olmedo, and C. Solares. The application of technologies in dementia diagnosis and intervention: A literature review. 16:1–11, 5 2017.

[55] M. Brookes. Voicebox: Speech processing toolbox for Matlab. www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html, 1997.

[56] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[57] H. Brunnström and E. Englund. Cause of death in patients with dementia disorders. *European Journal of Neurology*, 16(4):488–492, 2009.

[58] T. Bschor, K.-P. Kühl, and F. M. Reischies. Spontaneous speech of patients with dementia of the alzheimer type and mild cognitive impairment. *International Psychogeriatrics*, 13(3):289–298, 2001.

[59] R. L. Buckner. Memory and executive function in aging and ad: multiple factors that cause decline and reserve factors that compensate. *Neuron*, 44(1):195–208, 2004.

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

[61] A. Budhkar and F. Rudzicz. Augmenting word2vec with latent dirichlet allocation within a clinical application. *arXiv preprint arXiv:1808.03967*, 2018.

[62] L. Caeiro, J. M. Ferro, and J. Costa. Apathy secondary to stroke: a systematic review and meta-analysis. *Cerebrovascular Diseases*, 35(1):23–39, 2013.

[63] S. Carlomagno, A. Santoro, A. Menditti, M. Pandolfi, and A. Marini. Referential communication in alzheimer's type dementia. *Cortex*, 41(4):520–534, 2005.

[64] B. Carpenter and J. Dave. Disclosing a Dementia Diagnosis: A Review of Opinion and Practice, and a Proposed Research Agenda. *The Gerontologist*, 44(2):149–158, 2004.

[65] T. C. Castanho, C. Portugal-Nunes, P. S. Moreira, L. Amorim, J. A. Palha, N. Sousa, and N. Correia Santos. Applicability of the Telephone Interview for Cognitive Status (Modified) in a community sample with low education level: association with an extensive neuropsychological battery. *Int J Geriatr Psychiatry*, 31(2):128–136, Feb. 2016.

[66] F. Cathomas, M. Hartmann, E. Seifritz, C. Pryce, and S. Kaiser. The translational study of apathy–an ecological approach. *Frontiers in Behavioral Neuroscience*, 9:241, 2015.

[67] J. Chae and A. Nenkova. Predicting the fluency of text with shallow structural features: case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–147. Association for Computational Linguistics, 2009.

[68] S. B. Chapman, H. K. Ulatowska, K. King, J. K. Johnson, and D. D. McIntire. Discourse in early alzheimer's disease versus normal advanced aging. *American Journal of Speech-Language Pathology*, 4(4):124–129, 1995.

[69] S. A. Chau, J. Chung, N. Herrmann, M. Eizenman, and K. L. Lanctôt. Apathy and attentional biases in alzheimer's disease. *Journal of Alzheimer's Disease*, 51(3):837–846, 2016.

[70] P.-C. Chung, Y.-L. Hsu, C.-Y. Wang, C.-W. Lin, J.-S. Wang, and M.-C. Pai. Gait analysis for patients with Alzheimer's disease using a triaxial accelerometer. In *2012 IEEE International Symposium on Circuits and Systems*, pages 1323–1326. IEEE, 2012.

[71] D. G. Clark, P. M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R.-M. Dukes, P. Kapur, T. P. DeRamus, et al. Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2:113–122, 2016.

[72] D. E. Clarke, R. v. Reekum, M. Simard, D. L. Streiner, M. Freedman, and D. Conn. Apathy in dementia: An examination of the psychometric properties of the apathy evaluation scale. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 19(1):57–64, 2007.

[73] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[74] A. M. Collins and M. R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247, 1969.

[75] M. T. Compton, A. Lunden, S. D. Cleary, L. Pauselli, Y. Alolayan, B. Halpern, B. Broussard, A. Crisafio, L. Capulong, P. M. Balducci, F. Bernardini, and M. A. Covington. The aprosody of schizophrenia: Computationally derived acoustic phonetic underpinnings of monotone speech. *Schizophrenia Research*, 197:392–399, 2018.

[76] M. Consonni, E. Catricala, E. Dalla Bella, V. C. Gessa, G. Lauria, and S. F. Cappa. Beyond the consensus criteria: multiple cognitive profiles in amyotrophic lateral sclerosis? *Cortex*, 81:162–167, 2016.

[77] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10. Association for Computational Linguistics, 2015.

[78] A. Costa, T. Bak, P. Caffarra, C. Caltagirone, M. Ceccaldi, F. Collette, S. Crutch, S. Della Sala, J. F. Démonet, B. Dubois, et al. The need for harmonisation and innovation of neuropsychological assessment in neurodegenerative dementias in europe: consensus document of the joint program for neurodegenerative diseases working group. *Alzheimer's research & therapy*, 9(1):27, 2017.

[79] M. A. Covington, C. He, C. Brown, L. Naçi, J. T. McClain, B. S. Fjordbak, J. Semple, and J. Brown. Schizophrenia and the structure of language: The linguist's view. *Schizophrenia Research*, 77(1):85–98, 2005.

[80] M. A. Covington, S. A. Lunden, S. L. Cristofaro, C. R. Wan, C. T. Bailey, B. Broussard, R. Fogarty, S. Johnson, S. Zhang, and M. T. Compton. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia research*, 142(1-3):93–95, 2012.

[81] J. R. Crawford and J. D. Henry. Assessment of executive dysfunction. In P. W. Halligan and D. Wade, editors, *The effectiveness of rehabilitation for cognitive deficits*, pages 233–246. Oxford University Press, 2005.

[82] S. T. Creavin, S. Wisniewski, A. H. Noel-Storr, C. M. Trevelyan, T. Hampton, D. Rayment, V. M. Thom, K. J. Nash, H. Elhamoui, R. Milligan, et al. Mini-mental state examination (mmse) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane Database of Systematic Reviews*, (1):1–160, 2016.

[83] B. Croisile, B. Ska, M.-J. Brabant, A. Duchene, Y. Lepage, G. Aimard, and M. Trillet. Comparative study of oral and written picture description in patients with alzheimer's disease. *Brain and Language*, 53(1):1–19, 1996.

[84] K. Croot, J. R. Hodges, J. Xuereb, and K. Patterson. Phonological and articulatory impairment in alzheimer's disease: a case series. *Brain and Language*, 75(2):277–309, 2000.

[85] F. Cuetos, J. C. Arango-Lasprilla, C. Uribe, C. Valencia, and F. Lopera. Linguistic changes in verbal expression: a preclinical marker of alzheimer's disease. *Journal of the International Neuropsychological Society*, 13(3):433–439, 2007.

[86] J. L. Cummings, M. Mega, K. Gray, S. Rosenberg-Thompson, D. A. Carusi, and J. Gornbein. The neuropsychiatric inventory comprehensive assessment of psychopathology in dementia. *Neurology*, 44(12):2308–2308, 1994.

[87] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[88] N. Cummins, V. Sethu, J. Epps, S. Schnieder, and J. Krajewski. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, 75:27–49, 2015.

[89] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller. Enhancing speech-based depression detection through gender dependent vowel-level formant features. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 209–214. Springer, 2017.

[90] H. Daumé III. Frustratingly easy domain adaptation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–263, 2007.

[91] R. David, E. Mulin, L. Friedman, F. L. Duff, E. Cygankiewicz, O. Deschaux, R. Garcia, J. A. Yesavage, P. H. Robert, and J. M. Zeitzer. Decreased Daytime Motor Activity Associated With Apathy in Alzheimer Disease: An Actigraphic Study. *The American Journal of Geriatric Psychiatry*, 20(9):806–814, 2012.

[92] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

[93] C. A. de Jager, A. C. Schrijnemaekers, T. E. Honey, and M. M. Budge. Detection of MCI in the clinic: evaluation of the sensitivity and specificity of a computerised test battery, the Hopkins Verbal Learning Test and the MMSE. *Age Ageing*, 38(4):455–460, July 2009.

[94] N. H. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.

[95] U. M. de Lizarduy, P. C. Salomón, P. G. Vilda, M. E. Torres, and K. L. de Ipiña. Alzumeric: A decision support system for diagnosis and monitoring of cognitive impairment. *Loquens*, 4(1):37, 2017.

[96] E. Demetriou and R. Holtzer. Mild cognitive impairments moderate the effect of time on verbal fluency performance. *Journal of the International Neuropsychological Society: JINS*, 23(1):44, 2017.

[97] Y. Deng, L. Chang, M. Yang, M. Huo, and R. Zhou. Gender differences in emotional response: Inconsistency between experience and expressivity. *PLOS ONE*, 11(6):1–12, 6 2016.

[98] H. Dibeklioğlu, Z. Hammal, Y. Yang, and J. F. Cohn. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 307–310, New York, NY, USA, 2015. ACM.

[99] H. H. Dodge, N. Mattek, M. Gregor, M. Bowman, A. Seelye, O. Ybarra, M. Asgari, and J. A. Kaye. Social Markers of Mild Cognitive Impairment: Proportion of Word Counts in Free Conversational Speech. *Current Alzheimer research*, 12(6):513–519, 2015.

[100] B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, et al. Preclinical alzheimer's disease: definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323, 2016.

[101] A. Duong, F. Giroux, A. Tardif, and B. Ska. The heterogeneity of picture-supported narratives in alzheimer's disease. *Brain and Language*, 93(2):173–184, 2005.

[102] M. Eckerström, A. I. Berg, A. Nordlund, S. Rolstad, S. Sacuiu, and A. Wallin. High Prevalence of Stress and Low Prevalence of Alzheimer Disease CSF Biomarkers in a Clinical Sample with Subjective Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, 42(1-2):93–105, 2016.

[103] A. Egerhazi, R. Berecz, E. Bartok, and I. Degrell. Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(3):746–751, 2007.

[104] J. S. Ehrlich, L. K. Obler, and L. Clark. Ideational and semantic contributions to narrative production in adults with dementia of the alzheimer's type. *Journal of Communication Disorders*, 30(2):79–99, 1997.

[105] S. A. Elgamal, E. A. Roy, and M. T. Sharratt. Age and verbal fluency: the mediating effect of speed of processing. *Canadian geriatrics journal: CGJ*, 14(3):66, 2011.

[106] S. Epelbaum, R. Genthon, E. Cavedo, M. O. Habert, F. Lamari, G. Gagliardi, S. Lista, M. Teichmann, H. Bakardjian, H. Hampel, and B. Dubois. Preclinical Alzheimer's disease: A systematic review of the cohorts underlying the concept. *Alzheimers Dement*, 13(4):454–467, 2017.

[107] F. Espinoza-Cuadros, M. A. Garcia-Zamora, D. Torres-Boza, C. A. Ferrer-Riesgo, A. Montero-Benavides, E. Gonzalez-Moreira, and L. A. Hernandez-Gómez. A spoken language database for research on moderate cognitive impairment: design and preliminary analysis. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 219–228. Springer, 2014.

[108] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[109] K. Faber-Langendoen, J. C. Morris, J. W. Knesevich, E. LaBarge, J. P. Miller, and L. Berg. Aphasia in senile dementia of the alzheimer type. *Annals of Neurology*, 23(4):365–370, 1988.

[110] M. Faurholt-Jepsen, M. Vinberg, M. Frost, S. Debel, E. Margrethe Christensen, J. E. Bardram, and L. V. Kessing. Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder. *International Journal of Methods in Psychiatric Research*, 25(4):309–323, 2016.

[111] G. Fergadiotis and H. H. Wright. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology*, 25(11):1414–1430, 2011.

[112] S. Fernaeus and O. Almkvist. Word production: Dissociation of two retrieval modes of semantic memory across time. *Journal of Clinical and Experimental Neuropsychology*, 20(2):137–143, 1998.

[113] S. E. Fernaeus, P. Östberg, Å. Hellström, and L. O. Wahlund. Cut the coda: Early fluency intervals predict diagnoses. *Cortex*, 44(2):161–169, 2008.

[114] G. G. Fillenbaum, B. Peterson, and J. C. Morris. Estimating the validity of the clinical Dementia Rating Scale: the CERAD experience. Consortium to Establish a Registry for Alzheimer's Disease. *Aging (Milano)*, 8(6):379–385, 1996.

[115] W. Fitts, L. Massimo, N. Lim, M. Grossman, and N. Dahodwala. Computerized assessment of goal-directed behavior in parkinson's disease. *Journal of clinical and experimental neuropsychology*, 38(9):1015–1025, 2016.

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

[117] K. E. Forbes, M. F. Shanks, and A. Venneri. The evolution of dysgraphia in Alzheimer's disease. *Brain Research Bulletin*, 63(1):19–24, 2004.

[118] K. E. Forbes, A. Venneri, and M. F. Shanks. Distinct patterns of spontaneous speech deterioration: an early predictor of alzheimer's disease. *Brain and Cognition*, 48(2-3):356–361, 2002.

[119] K. Forbes-McKay, M. F. Shanks, and A. Venneri. Profiling spontaneous speech decline in alzheimer's disease: A longitudinal study. *Acta Neuropsychiatrica*, 25(6):320–327, 2013.

[120] K. E. Forbes-McKay and A. Venneri. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological Sciences*, 26(4):243–254, 2005.

[121] K. S. Fowler, M. M. Saling, E. L. Conway, J. M. Semple, and W. J. Louis. Computerized neuropsychological tests in the early detection of dementia: prospective findings. *Journal of the International Neuropsychological Society*, 3(2):139–146, 1997.

[122] C. Franco, J. Demongeot, C. Villemazet, and N. Vuillerme. Behavioral Telemonitoring of the Elderly at Home: Detection of Nycthemeral Rhythms Drifts from Location Data. In *24th International Conference on Advanced Information Networking and Applications Workshops*, pages 759–766. IEEE, 2010.

[123] K. C. Fraser, N. Linz, B. Li, K. L. Fors, F. Rudzicz, A. König, J. Alexandersson, P. Robert, and D. Kokkinakis. Multilingual prediction of alzheimer's disease through domain adaptation and concept-based language modelling. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2019), June 2-7, Minneapolis,, Minnesota, USA*. o.A., 2019.

[124] K. C. Fraser, N. Linz, H. Lindsay, and A. König. The importance of sharing patient-generated clinical speech and language data. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: Reconciling Outcomes. Computational Linguistics and Clinical Psychology Workshop (CLPsych-2019), 6th, befindet sich 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), June 6, Minneapolis, MN, United States*, 2019.

[125] K. C. Fraser, K. Lundholm Fors, M. Eckström, C. Themistokleous, and D. Kokkinakis. Improving the sensitivity and specificity of MCI screening with linguistic information. In *Proceedings of the 2nd Workshop on Resources and processing of linguistic, para-linguistic and extra-linguistic data from people with various forms of cognitive/psychiatric impairments (RaPID)*, pages 19–26, 2018.

[126] K. C. Fraser, K. Lundholm Fors, and D. Kokkinakis. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language*, 53:121–139, 2018.

[127] K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60, 2014.

[128] K. C. Fraser, J. A. Meltzer, and F. Rudzicz. Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422, 2016.

[129] E. Gabrilovich and S. Markovitch. Wikipedia-based Semantic Interpretation for Natural Language Processing. *Journal of Artificial Intelligence Research*, 34(1):443–498, Mar. 2009.

[130] P. Garrard, V. Rentoumi, B. Gesierich, B. Miller, and M. L. Gorno-Tempini. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, 55:122–129, 2014.

[131] S. Gauthier, B. Reisberg, M. Zaudig, R. C. Petersen, K. Ritchie, K. Broich, S. Belleville, H. Brodaty, D. Bennett, H. Chertkow, et al. Mild cognitive impairment. *The Lancet*, 367(9518):1262–1270, 2006.

[132] S. Gauthier, P. Rosa-Neto, J. Morais, and C. Webster. *World Alzheimer Report 2021: Journey through the diagnosis of dementia*. Alzheimer's Disease International (ADI), 2021.

[133] F. Gayraud, H.-R. Lee, and M. Barkat-Defradas. Syntactic and lexical context of pauses and hesitations in the discourse of Alzheimer patients and healthy elderly subjects. *Clinical linguistics & phonetics*, 25(3):198–209, 2011.

[134] F. J. F. Gerald, B. E. Murdoch, and H. J. Chenery. Multiple sclerosis: Associated speech and language disorders. *Australian Journal of Human Communication Disorders*, 15(2):15–35, 1987.

[135] E. Giles, K. Patterson, and J. R. Hodges. Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: missing information. *Aphasiology*, 10(4):395–408, 1996.

[136] D. Giuliani, A. Ottani, D. Zaffe, M. Galantucci, F. Strinati, R. Lodi, and S. Guarini. Hydrogen sulfide slows down progression of experimental alzheimer's disease by targeting multiple pathophysiological mechanisms. *Neurobiology of learning and memory*, 104:82–91, 2013.

[137] G. Glosser and T. Deser. Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language*, 40(1):67–88, 1991.

[138] A. M. Goberman and L. W. Elmer. Acoustic analysis of clear versus conversational speech in individuals with parkinson disease. *Journal of Communication Disorders*, 38(3):215–230, 2005.

[139] R. G. Gomez and D. A. White. Using verbal fluency to detect very mild dementia of the Alzheimer type. *Archives of Clinical Neuropsychology*, 21(8):771–775, 2006.

[140] E. Gonzalez-Moreira, D. Torres-Boza, M. Garcia-Zamora, C. Ferrer, and L. Hernandez-Gomez. Prosodic speech analysis to identify mild cognitive impairment. In *VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina 29, 30 & 31 October 2014*, pages 580–583. Springer, 2015.

[141] E. Gonzalez-Moreira, D. Torres-Boza, H. A. Kairuz, C. Ferrer, M. Garcia-Zamora, F. Espinoza-Cuadros, and L. A. Hernandez-Gómez. Automatic prosodic analysis to identify mild dementia. *BioMed research international*, 2015, 2015.

[142] H. Goodglass and E. Kaplan. *Boston Diagnostic Aphasia Examination*. Lea & Febiger, 1983.

[143] P. Goodglass, B. Barresi, and E. Kaplan. Boston Diagnostic Aphasia Examination, 1983. Philadelphia: Lippincott Williams and Willkins. A Wolters Kluwer Company.

[144] M. L. Gorno-Tempini, N. F. Dronkers, K. P. Rankin, J. M. Ogar, L. Phengrasamy, H. J. Rosen, J. K. Johnson, M. W. Weiner, and B. L. Miller. Cognition and anatomy in three variants of primary progressive aphasia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 55(3):335–346, 2004.

[145] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. Rohrer, S. Black, B. F. Boeve, et al. Classification of primary progressive aphasia and its variants. *Neurology*, 76(11):1006–1014, 2011.

[146] G. Gosztolya, L. Tóth, T. Grósz, V. Vincze, I. Hoffmann, G. Szatlóczki, M. Pákáski, and J. Kálmán. Detecting mild cognitive impairment from spontaneous speech by correlation-based phonetic feature selection. In *INTERSPEECH*, pages 107–111, 2016.

[147] G. Gosztolya, V. Vincze, L. Tóth, M. Pákáski, J. Kálmán, and I. Hoffmann. Identifying mild cognitive impairment and mild alzheimer's disease based on spontaneous speech using asr and linguistic features. *Computer Speech & Language*, 53:181–197, 2019.

[148] M. F. Green, W. P. Horan, D. M. Barch, and J. M. Gold. Effort-based decision making: A novel approach for assessing motivation in schizophrenia. *Schizophrenia Bulletin*, 41(5):1035–1044, 2015.

[149] J. D. Grill and J. Karlawish. Addressing the challenges to successful recruitment and retention in Alzheimer's disease clinical trials. *Alzheimer's research & therapy*, 2(6):34, 2010.

[150] M. Grossman. Sentence processing in parkinson's disease. *Brain and Cognition*, 40(2):387–413, 1999.

[151] P. J. Gruenewald and G. R. Lockhead. The Free Recall of Category Examples. *Journal of Experimental Psychology: Human Learning and Memory*, 6:225–240, 1980.

[152] C. T. Gualtieri and L. G. Johnson. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology*, 21(7):623–643, Oct. 2006.

[153] C. I. Guinn and A. Habash. Language analysis of speakers with dementia of the alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13. Menlo Park, CA, 2012.

[154] D. Hakkani-Tür, D. Vergyri, and G. Tur. Speech-based automated cognitive status assessment. In *INTERSPEECH*, pages 258–261, 2010.

[155] K. M. Hasebroock and N. J. Serkova. Toxicity of mri and ct contrast agents. *Expert Opinion on Drug Metabolism & Toxicology*, 5(4):403–416, 2009.

[156] K. Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*, pages 187–197, 2011.

[157] Health Physics Society. Radiation exposure from medical exams and procedures, 2010.

[158] J. D. Henry and J. R. Crawford. A meta-analytic review of verbal fluency performance following focal cortical lesions. *Neuropsychology*, 18(2):284–295, 2004.

[159] J. D. Henry and J. R. Crawford. Verbal fluency deficits in parkinson's disease: A meta-analysis. *Journal of the International Neuropsychological Society*, 10(4):608–622, 2004.

[160] J. D. Henry, J. R. Crawford, and L. H. Phillips. Verbal fluency performance in dementia of the alzheimer's type: a meta-analysis. *Neuropsychologia*, 42(9):1212–1222, 2004.

[161] L. Hernández-Domínguez, E. García-Cano, S. Ratté, and G. Sierra. Detection of alzheimer's disease based on automatic analysis of common objects descriptions. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 10–15, 2016.

[162] L. Hernández-Domínguez, S. Ratté, G. Sierra-Martínez, and A. Roche-Bergua. Computer-based evaluation of alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10:260–268, 2018.

[163] E. Herrera, F. Cuetos, and R. Ribacoba. Verbal fluency in parkinson's disease patients on/off dopamine medication. *Neuropsychologia*, 50(14):3636 – 3640, 2012.

[164] D. B. Hier, K. Hagenlocker, and A. G. Shindler. Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, 25(1):117–133, 1985.

[165] T. T. Hills, M. N. Jones, and P. M. Todd. Optimal Foraging in Semantic Memory. *Psychological review*, 119(2):431–440, Apr. 2012.

[166] J. R. Hodges and K. Patterson. Is semantic memory consistently impaired early in the course of alzheimer's disease? neuroanatomical and diagnostic implications. *Neuropsychologia*, 33(4):441–459, 1995.

[167] J. R. Hodges, D. P. Salmon, and N. Butters. Semantic memory impairment in alzheimer's disease: failure of access or degraded knowledge? *Neuropsychologia*, 30(4):301–314, 1992.

[168] I. Hoffmann, D. Nemeth, C. D. Dye, M. Pákáski, T. Irinyi, and J. Kálmán. Temporal parameters of spontaneous speech in alzheimer's disease. *International journal of speech-language pathology*, 12(1):29–34, 2010.

[169] J. Hoidekr, J. V. Psutka, A. Prazák, and J. Psutka. Benefit of a class-based language model for real-time closed-captioning of TV ice-hockey commentaries. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 2064–2067, 2006.

[170] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt. On the relative importance of vocal source, system, and prosody in human depression. In *2013 IEEE International Conference on Body Sensor Networks*, pages 1–6, May 2013.

[171] C.-W. Hsu, C.-C. Chang, and C. jen Lin. A Practical Guide to Support Vector Classification, 2010.

[172] Y.-L. Hsu, P.-C. Chung, W.-H. Wang, M.-C. Pai, C.-Y. Wang, C.-W. Lin, H.-L. Wu, and J.-S. Wang. Gait and Balance Analysis for Patients With Alzheimer?s Disease Using an Inertial-Sensor-Based Wearable Instrument. *IEEE Journal Of Biomedical And Health Informatics*, 18(6):1822–1830, 2014.

[173] L. Huang, Y. Jin, Y. Gao, K.-H. Thung, D. Shen, A. D. N. Initiative, et al. Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of Aging*, 46:180–191, 2016.

[174] C. P. Hughes, L. Berg, W. L. Danziger, L. A. Coben, and R. L. Martin. A New Clinical Scale for the Staging of Dementia. *The British Journal of Psychiatry*, 140(6):566–572, 1982.

[175] M. Inoue, D. Jimbo, M. Taniguchi, and K. Urakami. Touch Panel-type Dementia Assessment Scale: a new computer-based rating scale for Alzheimer's disease. *Psychogeriatrics*, 11(1):28–33, 2011.

[176] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar. 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, 2014.

[177] W. L. Jarrold, B. Peintner, E. Yeh, R. Krasnow, H. S. Javitz, and G. E. Swan. Language analytics for assessing brain health: Cognitive impairment, depression and pre-symptomatic alzheimer's disease. In *International Conference on Brain Informatics*, pages 299–307. Springer, 2010.

[178] A. M. Jensen, H. J. Chenery, and D. A. Copland. A comparison of picture description abilities in individuals with vascular subcortical lesions and huntington's disease. *Journal of Communication Disorders*, 39(1):62–77, 2006.

[179] L. Jia, C. Yu, and W. Meng. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1827–1830, New York, NY, USA, 2009. ACM.

[180] M. N. Jones and D. J. Mewhort. Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological review*, 114(1):1–37, Jan. 2007.

[181] S. Joubert, S. M. Brambati, J. Ansado, E. J. Barbeau, O. Felician, M. Didic, J. Lacombe, R. Goldstein, C. Chayer, and M.-J. Kergoat. The cognitive and neural expression of semantic memory impairment in mild cognitive impairment and early alzheimer's disease. *Neuropsychologia*, 48(4):978–988, 2010.

[182] E. Kaplan, H. Goodglass, S. Weintraub, and O. Segal. Boston naming test. In *Psychological Corporation*, Philadelphia: Lea & Febiger., 1983.

[183] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki. The DemCare Experiments and Datasets: a Technical Report. Technical report, Centre for Research and Technology Hellas (CERTH), 2014.

[184] G. Kavé and M. Goral. Word retrieval in connected speech in alzheimer's disease: a review with meta-analyses. *Aphasiology*, 32(1):4–26, 2018.

[185] G. Kavé and Y. Levy. Morphology in picture descriptions provided by persons with alzheimer's disease. *Journal of Speech, Language and Hearing research*, 46(2):341–352, 2003.

[186] S. Kemper, M. Thompson, and J. Marquis. Longitudinal change in language production: effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, 16(4):600, 2001.

[187] D. Kempler. Language changes in dementia of the alzheimer type. *Dementia and Communication*, pages 98–114, 1995.

[188] D. Kempler, S. Curtiss, and C. Jackson. Syntactic preservation in alzheimer's disease. *Journal of Speech, Language and Hearing Research*, 30(3):343–350, 1987.

[189] D. Kernot, T. Bossomaier, and R. Bradbury. The impact of depression and apathy on sensory language. *Open Journal of Modern Linguistics*, 7(1):8–32, 2 2017.

[190] A. Khodabakhsh and C. Demiroglu. Analysis of Speech-Based Measures for Detecting and Monitoring Alzheimer's Disease. In C. Fernández-Llatas and J. M. García-Gómez, editors, *Data Mining in Clinical Medicine, Methods in Molecular Biology*, volume 1246, chapter 11, pages 159–173. Springer Science+Business Media, New York, 2015.

[191] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu. Evaluation of Linguistic and Prosodic Features for Detection of Alzheimer's Disease in Turkish Conversational Speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 9:1–15, 2015.

[192] H. S. Kirshner. Primary progressive aphasia and alzheimer's disease: brief history, recent evidence. *Current Neurology and Neuroscience Reports*, 12(6):709–714, 2012.

[193] H. S. Kirshner, W. G. Webb, and M. P. Kelly. The naming disorder of dementia. *Neuropsychologia*, 22(1):23–30, 1984.

[194] B. Klimova and K. Kuca. Alzheimer's disease: Potential preventive, non-invasive, intervention strategies in lowering the risk of cognitive decline–a review study. *Journal of Applied Biomedicine*, 13(4):257–261, 2015.

[195] A. Kluge, M. Kirschner, O. M. Hager, M. Bischof, B. Habermeyer, E. Seifritz, S. Walther, and S. Kaiser. Combining actigraphy, ecological momentary assessment and neuroimaging to study apathy in patients with schizophrenia. *Schizophrenia Research*, 195:176–182, 2018.

[196] P. Klumpp, T. Janu, T. Arias-Vergara, and J. C. V. Correa. Apkinson–A Mobile Monitoring Solution for Parkinson's Disease. *INTERSPEECH 2017–18th Annual Conference of the International Speech Communication Association*, pages 1839–1843, 2017.

[197] D. Kokkinakis, K. L. Fors, K. Fraser, and A. Nordlund. A Swedish Cookie-Theft corpus. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 2018.

[198] D. Kokkinakis, K. Lundholm Fors, E. Björkner, and A. Nordlund. Data Collection from Persons with Mild Forms of Cognitive Impairment and Healthy Controls - Infrastructure for Classification and Prediction of Dementia. In *Proceedings of the 21st Nordic Conference of Computational Linguistics*, volume 75, pages 172–182. Linköping University Electronic Press, 2017.

[199] A. König, P. Aalten, F. Verhey, G. Bensadoun, P.-D. Petit, P. Robert, and R. David. A review of current information and communication technologies: can they be used to assess apathy? *International journal of geriatric psychiatry*, 29(4):345–358, 2014.

[200] A. König, C. F. Crispim Junior, A. Derreumaux, G. Bensadoun, P.-D. Petit, F. Bremond, R. David, F. Verhey, P. Aalten, and P. Robert. Validation of an automatic video monitoring system for the detection of instrumental activities of daily living in dementia patients. *Journal of Alzheimer's Disease*, 44(2):675–685, 2015.

[201] A. König, N. Linz, J. Tröger, M. Wolters, J. Alexandersson, and P. Robert. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and geriatric cognitive disorders*, 45(3-4):198–209, 2018.

[202] A. König, N. Linz, R. Zeghari, X. Klinge, J. Tröger, J. Alexandersson, and P. Robert. Detecting apathy in older adults with cognitive disorders using automatic speech analysis. *Journal of Alzheimer's Disease*, 69(4), 2019.

[203] A. König, G. Sacco, G. Bensadoun, F. Bremond, R. David, F. Verhey, P. Aalten, P. Robert, and V. Manera. The role of information and communication technologies in clinical trials with patients with alzheimer's disease and related disorders. *Frontiers in Aging Neuroscience*, 7:110, 2015.

[204] A. König, A. Satt, A. Sorin, R. Hoory, A. Derreumaux, R. David, and P. H. Robert. Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. *Current Alzheimer Research*, 15(2):120–129, 2018.

[205] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, et al. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):112–124, 2015.

[206] M. E. Kret and B. De Gelder. A review on sex differences in processing emotional signals. *Neuropsychologia*, 50(7):1211–1221, 2012.

[207] F. Kumfor, A. Zhen, J. R. Hodges, O. Piguet, and M. Irish. Apathy in alzheimer's disease and frontotemporal dementia: Distinct clinical profiles and neural correlates. *Cortex*, 103:350–359, 2018.

[208] G. R. Kuperberg. Language in schizophrenia part 1: an introduction. *Language and Linguistics Compass*, 4(8):576–589, 2010.

[209] E. Kušen, M. Strembeck, G. Cascavilla, and M. Conti. On the influence of emotional valence shifts on the spread of information in social networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, pages 321–324, New York, NY, USA, 2017. ACM.

[210] Y.-h. Lai, H.-h. Pai, et al. To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, 22(5):465–475, 2009.

[211] K. L. Lanctôt, L. Agüera-Ortiz, H. Brodaty, P. T. Francis, Y. E. Geda, Z. Ismail, G. A. Marshall, M. E. Mortby, C. U. Onyike, P. R. Padala, et al. Apathy associated with neurocognitive disorders: recent progress and future directions. *Alzheimer's & Dementia*, 13(1):84–100, 2017.

[212] J. B. Langbaum, A. S. Fleisher, K. Chen, N. Ayutyanont, F. Lopera, Y. T. Quiroz, R. J. Caselli, P. N. Tariot, and E. M. Reiman. Ushering in the study and treatment of preclinical alzheimer disease. *Nature Reviews Neurology*, 9(7):371–381, 2013.

[213] C. Laske, H. R. Sohrabi, S. M. Frost, K. López-de Ipiña, P. Garrard, M. Buscema, J. Dauwels, S. R. Soekadar, S. Mueller, C. Linnemann, S. A. Bridenbaugh, Y. Kanagasingam, R. N. Martins, and S. E. O'Bryant. Innovative Diagnostic tools for Early Detection of Alzheimer's Disease. *Alzheimers Dement*, 11(5):561–578, 2015.

[214] K. Ledoux, T. D. Vannorsdall, E. J. Pickett, L. V. Bosley, B. Gordon, and D. J. Schretlen. Capturing additional information about the organization of entries in the lexicon from verbal fluency productions. *Journal of Clinical and Experimental Neuropsychology*, 36(2):205–220, 2014.

[215] H. Lee, F. Gayraud, F. Hirsch, and M. Barkat-Defradas. Speech dysfluencies in normal and pathological aging: A comparison between alzheimer patients and healthy elderly subjects. In *ICPhS*, pages 1174–1177, 2011.

[216] M. Lehr, E. Prud'hommeaux, I. Shafran, and B. Roark. Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment. In *INTERSPEECH 2012–13th Annual Conference of the International Speech Communication Association*, pages 1039–1042, 2012.

[217] F. V. Leslie, S. Hsieh, J. Caga, S. A. Savage, E. Mioshi, M. Hornberger, M. C. Kiernan, J. R. Hodges, and J. R. Burrell. Semantic deficits in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 16(1-2):46–53, 2015.

[218] B. Li, Y.-T. Hsu, and F. Rudzicz. Detecting dementia in Mandarin Chinese using transfer learning from a parallel corpus. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[219] G. J. Lidstone. Note on the general case of the bayes-laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8:182–192, 1920.

[220] N. Linz, K. L. Fors, H. Lindsay, M. Eckerström, J. Alexandersson, and D. Kokkinakis. Temporal analysis of the semantic verbal fluency task in persons with subjective and mild cognitive impairment. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology: Reconciling Outcomes. Computational Linguistics and Clinical Psychology Workshop (CLPsych-2019), 6th, located at 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), June 6, Minneapolis,, MN, USA*. o.A., 2019.

[221] N. Linz, X. Klinge, J. Tröger, J. Alexandersson, R. Zeghari, P. Robert, and A. König. Automatic detection of apathy using acoustic markers extracted from free emotional speech. In *Proceedings of the 2nd Workshop on AI for Ageing, Rehabilitation and Independent Assisted Living (ARIAL)*, pages 17–21, 2018.

[222] N. Linz and J. Tröger. Language modelling for the clinical semantic verbal fluency task. In D. Kokkinakis, editor, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA).

[223] N. Linz, J. Tröger, J. Alexandersson, and A. König. Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*, 2017.

[224] N. Linz, J. Tröger, J. Alexandersson, M. Wolters, A. König, and P. Robert. Predicting dementia screening and staging scores from semantic verbal fluency performance. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 719–728. IEEE, 2017.

[225] J. O. d. Lira, T. S. C. Minett, P. H. F. Bertolucci, and K. Z. Ortiz. Analysis of word number and content in discourse of patients with mild to moderate alzheimer's disease. *Dementia & Neuropsychologia*, 8(3):260–265, 2014.

[226] J. A. Lonie, K. M. Tierney, and K. P. Ebmeier. Screening for mild cognitive impairment: a systematic review. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 24(9):902–915, 2009.

[227] K. Lopez-de Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun. On automatic diagnosis of alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognitive Computation*, 7(1):44–55, 2015.

[228] K. López-de Ipiña, J. B. Alonso, C. M. Travieso, H. Egiraun, M. Ecay, A. Ezeiza, N. Barroso, and P. Martinez-Lage. Automatic analysis of emotional response based on non-linear speech modeling oriented to alzheimer disease diagnosis. In *2013 IEEE 17th International Conference on Intelligent Engineering Systems (INES)*, pages 61–64. IEEE, 2013.

[229] K. López-de Ipiña, U. Martinez-de Lizarduy, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, F. Torres, and M. Faundez-Zanuy. Automatic analysis of categorical verbal fluency for mild cognitive impartment detection: A non-linear language independent approach. In *2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 101–104. IEEE, 2015.

[230] K. López-de Ipiña, U. Martinez-de Lizarduy, P. Calvo, B. Beitia, J. García-Melero, M. Ecay-Torres, A. Estanga, and M. Faundez-Zanuy. Analysis of disfluencies for automatic detection of mild cognitive impartment: a deep learning approach. In *2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI)*, pages 1–4. IEEE, 2017.

[231] K. López-de Ipiña, J. Solé-Casals, H. Eguiraun, J. B. Alonso, C. M. Travieso, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and B. Beitia. Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, 30(1):43–60, 2015.

[232] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2011.

[233] K. Lundholm Fors, K. C. Fraser, and D. Kokkinakis. Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. In *Proceedings of the Medical Informatics Europe (MIE) Conference*, pages 705–709, 2018.

[234] S. Luz. Longitudinal monitoring and detection of alzheimer's type dementia from spontaneous speech data. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 45–46. IEEE, 2017.

[235] D. K. Ly, K. Sugiyama, Z. Lin, and M.-Y. Kan. Product review summarization from a deeper perspective. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, JCDL '11, pages 311–314, New York, NY, USA, 2011. ACM.

[236] C. A. Lynch, C. Walsh, A. Blanco, M. Moran, R. F. Coen, J. B. Walsh, and B. A. Lawlor. The Clinical Dementia Rating Sum of Box Score in Mild Dementia. *Dementia and geriatric cognitive disorders*, 21(1):40–43, 2006.

[237] P. D. MacIntyre. Motivation, anxiety and emotion in second language acquisition. *Individual differences and instructed language learning*, 2:45–68, 2002.

[238] C. Mackenzie, M. Brady, J. Norrie, and N. Poedjianto. Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology*, 21(3-4):340–354, 2007.

[239] B. MacWhinney. The CHILDES Project Part 1: The CHAT transcription format. *Department of Psychology*, page 181, 2009.

[240] B. MacWhinney, D. Fromm, M. Forbes, and A. Holland. AphasiaBank: Methods for Studying Discourse. *Aphasiology*, 25(11):1286–1307, 2011.

[241] J. Mahmud. Why do you write this? prediction of influencers from word use. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[242] J. Mahmud, J. Chen, and J. Nichols. Why are you more engaged? predicting social engagement from word use. *arXiv preprint arXiv:1402.6690*, 2014.

[243] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[244] C. A. Marczinski and A. Kertesz. Category and letter fluency in semantic dementia, primary progressive aphasia, and alzheimer's disease. *Brain and Language*, 97(3):258–265, 2006.

[245] A. Martin and L. L. Chao. Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, 11(2):194–201, 2001.

[246] F. Martínez-Sánchez, J. J. G. Meilán, J. Carro, and O. Ivanova. A prototype for the voice analysis diagnosis of alzheimer's disease. *Journal of Alzheimer's Disease*, 64(2):473–481, 2018.

[247] V. Masrani, G. Murray, T. Field, and G. Carenini. Detecting dementia through retrospective analysis of routine blog posts by bloggers with dementia. *BioNLP 2017*, pages 232–237, 2017.

[248] V. Masrani, G. Murray, T. S. Field, and G. Carenini. Domain adaptation for detecting mild cognitive impairment. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 248–259, 2017.

[249] N. Mattsson, D. Brax, and H. Zetterberg. To know or not to know: ethical issues related to early diagnosis of alzheimer's disease. *International journal of Alzheimer's disease*, 2010, 2010.

[250] G. M. McKhann, D. S. Knopman, H. Chertkow, B. T. Hyman, C. R. Jack Jr, C. H. Kawas, W. E. Klunk, W. J. Koroshetz, J. J. Manly, R. Mayeux, et al. 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, 2011.

[251] P. McNamara, L. K. Obler, R. Au, R. Durso, and M. L. Albert. Speech monitoring skills in alzheimer's disease, parkinson's disease, and normal aging. *Brain and Language*, 42(1):38–51, 1992.

[252] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[253] A. Meena and P. Tadinada. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. *Advances in Information Retrieval. ECIR 2007. Lecture Notes in Computer Science*, 4425:573–580, 2007.

[254] J. J. G. Meilán, F. Martínez-Sánchez, J. Carro, D. E. López, L. Millian-Morell, and J. M. Arana. Speech in alzheimer's disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334, 2014.

[255] M. F. Mendez and J. L. Cummings. *Dementia: a clinical approach*. Butterworth-Heinemann, 2003.

[256] J. Merrilees, G. Dowling, E. Hubbard, J. Mastick, R. Ketelle, and B. Miller. Characterization of apathy in persons with frontotemporal dementia and the impact on family caregivers. *Alzheimer disease and associated disorders*, 27(1):62–67, 2013.

[257] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

[258] B. Mirheidari, D. Blackburn, K. Harkness, T. Walker, A. Venneri, M. Reuber, and H. Christensen. Toward the automation of diagnostic conversation analysis in patients with memory complaints. *Journal of Alzheimer's Disease*, 58(2):373–387, 2017.

[259] B. Mirheidari, D. Blackburn, M. Reuber, T. Walker, and H. Christensen. Diagnosing people with dementia using automatic conversation analysis. In *Proceedings of Interspeech*, pages 1220–1224. ISCA, 2016.

[260] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen. Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79, 2019.

[261] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen. Detecting signs of dementia using word vector representations. In *INTERSPEECH*, pages 1893–1897, 2018.

[262] S. Mirzaei, M. El Yacoubi, S. Garcia-Salicetti, J. Boudy, C. Kahindo, V. Cristancho-Lacroix, H. Kerhervé, and A.-S. Rigaud. Two-stage feature selection of voice parameters for early alzheimer's disease prediction. *Irbm*, 39(6):430–435, 2018.

[263] S. Mirzaei, M. El Yacoubi, S. Garcia-Salicetti, J. Boudy, C. K. S. Muvingi, V. Cristancho-Lacroix, H. Kerhervé, A.-S. R. Monnet, E. Yacoubi, C. Kahindo, et al. Automatic speech analysis for early alzeimer's disease diagnosis. In *JETSAN 2017: 6e Journées d'Etudes sur la Télésanté*, pages 114–116, 2017.

[264] M. Mitchell, K. Hollingshead, and G. Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20, 2015.

[265] R. A. Mitchell, N. Herrmann, and K. L. Lanctôt. The role of dopamine in symptoms and treatment of apathy in alzheimer's disease. *CNS Neuroscience & Therapeutics*, 17(5):411–427, 2010.

[266] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[267] D. W. Molloy, E. Alemayehu, and R. Roberts. Reliability of a Standardized Mini-Mental State Examination compared with the traditional Mini-Mental State Examination. *American Journal of Psychiatry*, 148(1):102–105, 1991.

[268] A. U. Monsch, M. W. Bondi, N. Butters, D. P. Salmon, R. Katzman, and L. J. Thal. Comparisons of verbal fluency tasks in the detection of dementia of the alzheimer type. *Archives of Neurology*, 49(12):1253–1258, 1992.

[269] E. Morris, A. Chalkidou, A. Hammers, J. Peacock, J. Summers, and S. Keevil. Diagnostic accuracy of 18 f amyloid pet tracers for the diagnosis of alzheimer's disease: a systematic review and meta-analysis. *European journal of nuclear medicine and molecular imaging*, 43(2):374–385, 2016.

[270] J. C. Morris. Clinical Dementia Rating: A Reliable and Valid Diagnostic and Staging Measure for Dementia of the Alzheimer Type. *International Psychogeriatrics*, 9(S1):173–176, 1997.

[271] J. C. Morris, D. W. McKeel, K. Fulling, R. M. Torack, and L. Berg. Validation of clinical diagnostic criteria for Alzheimer's disease. *Annals of Neurology*, 24(1):17–22, 1988.

[272] K. D. Mueller, B. Hermann, J. Mecollari, and L. S. Turkstra. 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, 2018.

[273] K. D. Mueller, R. L. Koscik, A. LaRue, L. R. Clark, B. Hermann, S. C. Johnson, and M. A. Sager. Verbal Fluency and Early Memory Decline: Results from the Wisconsin Registry for Alzheimer's Prevention. *Archives of Clinical Neuropsychology*, 30(5):448, 2015.

[274] K. D. Mueller, R. L. Koscik, L. S. Turkstra, S. K. Riedeman, A. LaRue, L. R. Clark, B. Hermann, M. A. Sager, and S. C. Johnson. Connected language in late middle-aged adults at risk for alzheimer's disease. *Journal of Alzheimer's Disease*, 54(4):1539–1550, 2016.

[275] O. M. Mueller. Low-cost magnetic resonance imaging (mri) cryo-system, Apr. 2005. US Patent 6,879,852.

[276] E. Mulin, E. Leone, K. Dujardin, M. Delliaux, A. Leentjens, F. Nobili, B. Dessi, O. Tible, L. Agüera-Ortiz, R. S. Osorio, J. Yessavage, D. Dachevsky, F. Verhey, A. J. C. Jentoft, O. Blanc, P. Llorca, and P. H. Robert. Diagnostic criteria for apathy in clinical practice. *International Journal of Geriatric Psychiatry*, 26(2):158–165, 2011.

[277] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking. Vocal acoustic biomarkers of depression severity and treatment response. *Biological psychiatry*, 72(7):580–587, 2012.

[278] C. Munro Cullum, L. Hynan, M. Grosch, M. Parikh, and M. Weiner. Teleneuropsychology: Evidence for Video Teleconference-Based Neuropsychological Assessment. *Journal of the International Neuropsychological Society*, 20(10):1028–1033, 2014.

[279] K. J. Murphy, J. B. Rich, and A. K. Troyer. Verbal fluency patterns in amnestic mild cognitive impairment are characteristic of alzheimer's type dementia. *Journal of the International Neuropsychological Society*, 12(4):570–574, 2006.

[280] L. L. Murray. Distinguishing clinical depression from early alzheimer's disease in elderly people: Can narrative analysis help? *Aphasiology*, 24(6-8):928–939, 2010.

[281] L. L. Murray and L. P. Lenz. Productive syntax abilities in huntington's and parkinson's diseases. *Brain and Cognition*, 46(1-2):213–219, 2001.

[282] NHS, NICE, and SCIE. *Dementia: Quick Reference Guide*. National Institute for Health and Clinical Excellence, 2006.

[283] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks. Empty speech in alzheimer's disease and fluent aphasia. *Journal of Speech, Language, and Hearing Research*, 28(3):405–410, 1985.

[284] T. Nikolai, O. Bezdicek, H. Markova, H. Stepankova, J. Michalec, M. Kopecek, M. Dokoupilova, J. Hort, and M. Vyhnalek. Semantic verbal fluency impairment is detectable in patients with subjective cognitive decline. *Applied Neuropsychology:Adult*, 25(5):448–457, 2018.

[285] S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, R. Doody, and T. A. R. Texas Alzheimer's Research Consortium. Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores: a Texas Alzheimer's Research Consortium Study. *Archives of Neurology*, 65(8):1091–1095, 2008.

[286] S. O. Orimaye, J. S. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, 18(1):1–13, 2017.

[287] S. O. Orimaye, J. S.-M. Wong, and J. S. G. Fernandez. Deep-deep neural network language models for predicting mild cognitive impairment. In *Advances in Bioinformatics and Artificial Intelligence (BAI 2016): Bridging the Gap*, pages 14–20. Rheinisch-Westfaelische Technische Hochschule Aachen, 2016.

[288] S. O. Orimaye, J. S.-M. Wong, and K. J. Golden. 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, 2014.

[289] S. O. Orimaye, J. S.-M. Wong, and C. P. Wong. Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one*, 13(11):e0205636, 2018.

[290] P. Paavilainen, I. Korhonen, L. Cluitmans, J. Lötjönen, A. Särelä, and M. Partinen. Circadian activity rhythm in demented and non-demented nursing-home residents measured by telemetric actigraphy. *Journal of Sleep Research*, 14(1):61–68, 2005.

[291] A. Pachet, K. Astner, and L. Brown. Clinical utility of the mini-mental status examination when assessing decision-making capacity. *Journal of geriatric psychiatry and neurology*, 23(1):3–8, 2010.

[292] J. Pagonabarraga, J. Kulisevsky, A. Strafella, and P. Krack. Apathy in parkinson's disease: clinical features, neural substrates, diagnosis, and treatment. *The Lancet Neurology*, 14(5):518–531, May 2015.

[293] S. V. Pakhomov, L. Eberly, and D. Knopman. Characterizing Cognitive Performance in a Large Longitudinal study of Aging with Computerized Semantic Indices of Verbal Fluency. *Neuropsychologia*, 89:42–56, 2016.

[294] S. V. Pakhomov, S. E. Marino, S. Banks, and C. Bernick. Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency. *Speech Communication*, 75:14–26, 2015.

[295] S. V. Pakhomov, G. E. Smith, D. Chacon, Y. Feliciano, N. Graff-Radford, R. Caselli, and D. S. Knopman. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology*, 23(3):165–177, 2010.

[296] S. V. Pakhomov, G. E. Smith, S. Marino, A. Birnbaum, N. Graff-Radford, R. Caselli, B. Boeve, and D. S. Knopman. A computerized technique to assess language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23(2):127–144, 2010.

[297] S. Palmqvist, B. Terzis, C. Strobel, and A. Wallin. Mmse-sr: Mini mental state examination - swedish revision, version 2. 2013.

[298] P. Palta, B. Snitz, and M. Carlson. Chapter 7 - neuropsychologic assessment. In M. J. Aminoff, F. Boller, and D. F. Swaab, editors, *Neuroepidemiology*, volume 138 of *Handbook of Clinical Neurology*, pages 107–119. Elsevier, 2016.

[299] V. C. Pangman, J. Sloan, and L. Guse. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Applied Nursing Research*, 13(4):209–213, 2000.

[300] E. Papastavrou, A. Kalokerinou, S. S. Papacostas, H. Tsangari, and P. Sourtzi. Caring for a relative with dementia: family caregiver burden. *Journal of Advanced Nursing*, 58(5):446–457, 2007.

[301] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[302] S. Pekkala, M. Albert, A. Spiro Iii, and T. Erkinjuntti. Perseveration in alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 25(2):109–114, 2008.

[303] J. W. Pennebaker, R. J. Booth, and M. E. Francis. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 2007.

[304] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn. The development and psychometric properties of liwc2015. Technical report, 2015.

[305] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2015, 2015.

[306] P. Péran, O. Rascol, J.-F. Démonet, P. Celsis, J.-L. Nespoulous, B. Dubois, and D. Cardebat. Deficit of verb generation in nondemented patients with parkinson's disease. *Movement Disorders: Official Journal of the Movement Disorder Society*, 18(2):150–156, 2003.

[307] D. P. Perl. Neuropathology of alzheimer's disease. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 77(1):32–42, 2010.

[308] M. Pessiglione, L. Schmidt, B. Draganski, R. Kalisch, H. Lau, R. J. Dolan, and C. D. Frith. How the brain translates money into force: a neuroimaging study of subliminal motivation. *Science*, 316(5826):904–906, 2007.

[309] J. Peter, J. Kaiser, V. Landerer, L. Köstering, C. P. Kaller, B. Heimbach, M. Hüll, T. Bormann, and S. Klöppel. Category and design fluency in mild cognitive impairment: Performance, strategy use, and neural correlates. *Neuropsychologia*, 93:21–29, 2016.

[310] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen. Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, 56(3):303–308, 1999.

[311] S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. *LREC Conference Preprint*, 2011.

[312] M. Piefke and G. R. Fink. Recollections of one's own past: the effects of aging and gender on the neural mechanisms of episodic autobiographical memory. *Anatomy and Embryology*, 210(5):497–512, Dec. 2005.

[313] A. Piolat, R. Booth, C. Chung, M. Davids, and J. Pennebaker. La version française du dictionnaire pour le liwc: modalités de construction et exemples d'utilisation. *Psychologie Française*, 56(3):145–159, 9 2011.

[314] K. Podoll, P. Caspary, H. Lange, and J. Noth. Language functions in huntington's disease. *Brain*, 111(6):1475–1503, 1988.

[315] A. Pompili, A. Abad, D. M. de Matos, and I. P. Martins. Topic coherence analysis for the classification of alzheimer's disease. In *IberSPEECH*, pages 281–285, 2018.

[316] C. Pou-Prom and F. Rudzicz. Learning multiview embeddings for assessing dementia. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2812–2817, 2018.

[317] J. T. Povlishock and D. I. Katz. Update of neuropathology and neurological recovery after traumatic brain injury. *The Journal of head trauma rehabilitation*, 20(1):76–94, 2005.

[318] P. Prettenhofer and B. Stein. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL)*, pages 1118–1127, 2010.

[319] S. E. Price, G. J. Kinsella, B. Ong, E. Storey, E. Mullaly, M. Phillips, L. Pangnadasa-Fox, and D. Perre. Semantic verbal fluency strategies in amnestic mild cognitive impairment. *Neuropsychology*, 26(4):490–497, 2012.

[320] M. Prince, A. Comas-Herrera, M. Knapp, M. Guerchet, and M. Karagiannidou. *World Alzheimer Report 2016: Improving Healthcare for People Living with Dementia: Coverage, Quality and Costs*. Alzheimer's Disease International (ADI), 2016.

[321] E. Prud'hommeaux and B. Roark. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578, 2015.

[322] E. T. Prud'hommeaux and B. Roark. Alignment of spoken narratives for automated neuropsychological assessment. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 484–489. IEEE, 2011.

[323] N. Raoux, H. Amieva, M. L. Goff, S. Auriacombe, L. Carcaillon, L. Letenneur, and J.-F. Dartigues. Clustering and switching processes in semantic verbal fluency in the course of Alzheimer's disease subjects: Results from the PAQUID longitudinal study. *Cortex*, 44(9):1188–1196, 2008.

[324] A. M. Rapp and B. Wild. Nonliteral language in alzheimer dementia: a review. *Journal of the International Neuropsychological Society*, 17(2):207–218, 2011.

[325] V. Rentoumi, L. Raoufian, S. Ahmed, C. A. de Jager, and P. Garrard. Features and machine learning classification of connected speech samples from patients with autopsy proven alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42:3–17, 2014.

[326] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2081–2090, 2011.

[327] P. Robert, K. Lanctôt, L. Agüera-Ortiz, P. Aalten, F. Bremond, M. Defrancesco, C. Hanon, R. David, B. Dubois, K. Dujardin, et al. Is it time to revise the diagnostic criteria for apathy in brain disorders? the 2018 international consensus group. *European Psychiatry*, 54:71–76, 2018.

[328] P. H. Robert, S. Clairet, M. Benoit, J. Koutaich, C. Bertogliati, O. Tible, H. Caci, M. Borg, P. Brocker, and P. Bedoucha. The Apathy Inventory: assessment of apathy and awareness in Alzheimer's disease, Parkinson's disease and mild cognitive impairment. *International Journal of Geriatric Psychiatry*, 17(12):1099–1105, 2002.

[329] P. H. Robert, V. Lafont, I. Medecin, L. Berthet, S. Thauby, C. Baudu, and G. Darcourt. Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society*, 4(6):539–546, 1998.

[330] K. M. Robinson, M. Grossman, T. White-Devine, and M. D'esposito. Category-specific difficulty naming with verbs in alzheimer's disease. *Neurology*, 47(1):178–182, 1996.

[331] W. G. Rosen, R. C. Mohs, and K. L. Davis. A new rating scale for Alzheimer's disease. *The American journal of psychiatry*, 141(11):1356–1364, 1984.

[332] S. Rountree, W. Chan, V. Pavlik, E. Darby, S. Siddiqui, and R. Doody. Persistent treatment with cholinesterase inhibitors and/or memantine slows clinical progression of alzheimer disease. *Alzheimer's Research & Therapy*, 1(2):1–7, 2009.

[333] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1):350–367, 2011.

[334] S. R. Sabat. Language function in alzheimer's disease: A critical review of selected literature. *Language & Communication*, 14(4):331–351, 1994.

[335] P. Sachdev, D. Blacker, D. Blazer, M. Ganguli, D. V Jeste, J. Paulsen, and R. Petersen. Classifying neurocognitive disorders: The dsm-5 approach. *Nature Reviews Neurology*, 10:634–642, 9 2014.

[336] R. Sadeghian, J. D. Schaffer, and S. A. Zahorian. Speech processing approach for diagnosing dementia in an early stage. *Proc. Interspeech 2017*, pages 2705–2709, 2017.

[337] M. Sahlgren and R. Cöster. Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[338] S. A. Sajjadi, K. Patterson, M. Tomek, and P. J. Nestor. Abnormalities of connected speech in semantic dementia vs alzheimer's disease. *Aphasiology*, 26(6):847–866, 2012.

[339] D. P. Salmon, N. Butters, and A. S. Chan. The deterioration of semantic memory in alzheimer's disease. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 53(1):108–118, 1999.

[340] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert. Speech-based Automatic and Robust Detection of very Early Dementia. In *INTERSPEECH 2014–15th Annual Conference of the International Speech Communication Association*, pages 2538–2542, 2014.

[341] A. Satt, R. Hoory, A. König, P. Aalten, and P. H. Robert. Speech-based automatic and robust detection of very early dementia. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[342] A. Satt, A. Sorin, O. Toledo-Ronen, O. Barkan, I. Kompatsiaris, A. Kokonozi, and M. Tsolaki. Evaluation of speech-based protocol for detection of early-stage dementia. In *Proceedings of Interspeech*, pages 1692–1696, 2013.

[343] M. Y. Savundranayagam, M. L. Hummert, and R. J. Montgomery. Investigating the effects of communication problems on caregiver burden. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 60(1):48–55, 2005.

[344] K. R. Scherer. Vocal affect expression: A review and a model for future research. *Psychological bulletin*, 99(2):143, 1986.

[345] A. Scheuringer, R. Wittig, and B. Pletzer. Sex differences in verbal fluency: The role of strategies and instructions. *Cognitive processing*, 18(4):407–417, 2017.

[346] Z. Shao, E. Janse, K. Visser, and A. S. Meyer. What do verbal fluency tasks measure? predictors of verbal fluency performance in older adults. *Frontiers in psychology*, 5:772, 2014.

[347] B. Sheehan. Assessment scales in dementia. *Therapeutic advances in neurological disorders*, 5(6):349–358, 2012.

[348] D. Shibata, S. Wakamiya, A. Kinoshita, and E. Aramaki. Detecting Japanese patients with Alzheimer's disease based on word category frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85, 2016.

[349] S. Sindi, F. Mangialasche, and M. Kivipelto. Advances in the prevention of alzheimer's disease. *F1000Prime Reports*, 7:1–12, 2015.

[350] K. Sirts, O. Piguet, and M. Johnson. Idea density for predicting alzheimer's disease from transcribed speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 322–332, 2017.

[351] S. R. Smith, H. J. Chenery, and B. E. Murdoch. Semantic abilities in dementia of the alzheimer type. ii. grammatical semantics. *Brain and Language*, 36(4):533–542, 1989.

[352] P. J. Snyder, K. Kahle-Wrobleski, S. Brannan, D. S. Miller, R. J. Schindler, S. DeSanti, J. M. Ryan, G. Morrison, M. Grundman, J. Chandler, R. J. Caselli, M. Isaac, L. Bain, and M. C. Carrillo. Assessing Cognition and Function in Alzheimer's Disease Clinical Trials: Do we have the right Tools? *Alzheimers Dement*, 10(6):853–860, 2014.

[353] S. P. Sonty, M.-M. Mesulam, C. K. Thompson, N. A. Johnson, S. Weintraub, T. B. Parrish, and D. R. Gitelman. Primary progressive aphasia: Ppa and the language network. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 53(1):35–49, 2003.

[354] R. Speer, J. Chin, A. Lin, S. Jewett, and L. Nathan. Luminosoinsight/wordfreq: v1.7, Sept. 2017.

[355] R. Speer, J. Chin, A. Lin, L. Nathan, and S. Jewett. wordfreq: v1.5.1, 2016.

[356] R. A. Sperling, P. S. Aisen, L. A. Beckett, D. A. Bennett, S. Craft, A. M. Fagan, T. Iwatsubo, C. R. Jack Jr, J. Kaye, T. J. Montine, et al. Toward defining the preclinical stages of 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):280–292, 2011.

[357] A. St-Hilaire, C. Hudon, G. T. Vallet, L. Bherer, M. Lussier, J. F. Gagnon, M. Simard, N. Gosselin, F. Escudier, I. Rouleau, and J. Macoir. Normative data for phonemic and semantic verbal fluency test in the adult French-Quebec population and validation study in Alzheimer's disease and depression. *The Clinical Neuropsychologist*, 30(7):1126–1150, 2016.

[358] P. L. St Jacques, M. A. Conway, and R. Cabeza. Gender differences in autobiographical memory for everyday events: retrieval elicited by sensecam images versus verbal cues. *Memory*, 19(7):723–732, Oct. 2011.

[359] S. E. Starkstein, R. Jorge, R. Mizrahi, and R. G. Robinson. A prospective longitudinal study of apathy in alzheimer's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(1):8–11, 2006.

[360] F. Stella, O. V. Forlenza, J. Laks, L. P. Andrade, J. Castilho Cação, J. S. Govone, K. Medeiros, and C. G. Lyketsos. Caregiver report versus clinician impression: disagreements in rating neuropsychiatric symptoms in alzheimer's disease patients. *International journal of geriatric psychiatry*, 30(12):1230–1237, 2015.

[361] T. Suzuki, S. Murase, T. Tanaka, and T. Okazawa. New Approach for The Early Detection of Dementia by Recording In-House Activities. *Telemedicine and e-Health*, 13(1):41–44, 2007.

[362] G. Szatloczki, I. Hoffmann, V. Vincze, J. Kalman, and M. Pakaski. Speaking in alzheimer's disease, is that an early sign? importance of changes in language abilities in alzheimer's disease. *Frontiers in aging neuroscience*, 7:195, 2015.

[363] O. Täckström, R. McDonald, and J. Uszkoreit. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, pages 477–487, 2012.

[364] K. Takahashi, L. Tickle-Degnen, W. J. Coster, and N. K. Latham. Expressive behavior in parkinson's disease as a function of interview context. *American Journal of Occupational Therapy*, 64(3):484–495, 2010.

[365] V. Taler and N. A. Phillips. Language performance in alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, 30(5):501–556, 2008.

[366] I. M. Tallberg, E. Ivachova, K. Jones Tinghag, and P. Östberg. Swedish norms for word fluency tests: FAS, animals and verbs. *Scandinavian Journal of Psychology*, 49(5):479–485, 2008.

[367] I. Tarnanas, W. Schlee, M. Tsolaki, R. Müri, U. Mosimann, and T. Nef. Ecological Validity of Virtual Reality Daily Living Activities Screening for Early Dementia: Longitudinal Study. *JMIR Serious Games*, 1(1), 2013.

[368] C. Themistocleous, M. Eckerström, and D. Kokkinakis. Identification of mild cognitive impairment from speech in swedish using deep sequential neural networks. *Frontiers in neurology*, 9:975, 2018.

[369] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp. Automatic detection and rating of dementia of alzheimer type through lexical analysis of spontaneous speech. In *Mechatronics and Automation, 2005 IEEE International Conference*, volume 3, pages 1569–1574. IEEE, 2005.

[370] K.-H. Thung, C.-Y. Wee, P.-T. Yap, and D. Shen. Identification of alzheimer's disease using incomplete multimodal dataset via matrix shrinkage and completion. In *International Workshop on Machine Learning in Medical Imaging*, pages 163–170. Springer, 2013.

[371] T. N. Tombaugh, J. Kozak, and L. Rees. Normative Data Stratified by Age and Education for Two Measures of Verbal Fluency: FAS and Animal Naming. *Archives of Clinical Neuropsychology*, 14(2):167–177, 1999.

[372] T. N. Tombaugh and N. J. McIntyre. The Mini-Mental State Examination: A Comprehensive Review. *Journal of the American Geriatrics Society*, 40(9):922–935, 1992.

[373] C. K. Tomoeda, K. A. Bayles, M. W. Trosset, T. Azuma, and A. McGeagh. Cross-sectional analysis of alzheimer disease effects on oral discourse in a picture description task. *Alzheimer disease and associated disorders*, 10(4):204–215, 1996.

[374] L. Tóth, G. Gosztolya, V. Vincze, I. Hoffmann, G. Szatlóczki, E. Biró, F. Zsura, M. Pákáski, and J. Kálmán. Automatic detection of mild cognitive impairment from spontaneous speech using asr. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 1–5, 2015.

[375] L. Tóth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán. A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Current Alzheimer Research*, 15(2):130–138, 2018.

[376] J. Tröger, N. Linz, J. Alexandersson, A. König, and P. Robert. Automated speech-based screening for alzheimer's disease in a care service scenario. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 292–297. ACM, 2017.

[377] J. Tröger, N. Linz, A. König, P. Robert, and J. Alexandersson. Telephone-based dementia screening I: Automated semantic verbal fluency assessment. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare. International ICST Conference on Pervasive Computing Technologies for Healthcare (Pervasive Health-2018), May 21-24, New York, USA*. ACM, 2018.

[378] J. Tröger, N. Linz, A. König, P. Robert, J. Alexandersson, J. Peter, and J. Kray. Exploitation vs. exploration—computational temporal and semantic analysis explains semantic verbal fluency impairment in alzheimer's disease. *Neuropsychologia*, 131:53 – 61, 2019.

[379] A. K. Troyer, M. Moscovitch, and G. Winocur. Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *neuropsychology*, 11(1):138, 1997.

[380] A. K. Troyer, M. Moscovitch, G. Winocur, L. Leach, and M. Freedman. Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society*, 4(2):137–143, 1998.

[381] P. T. Trzepacz, H. Hochstetler, S. Wang, B. Walker, and A. J. Saykin. Relationship between the Montreal Cognitive Assessment and Mini-mental State Examination for assessment of mild cognitive impairment in older adults. *BMC geriatrics*, 15(1):107–115, 2015.

[382] S. Tsermentseli, P. N. Leigh, L. J. Taylor, A. Radunovic, M. Catani, and L. H. Goldstein. Syntactic processing as a marker for cognitive impairment in amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 17(1-2):69–76, 2016.

[383] L. K. Tyler and H. E. Moss. Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, 5(6):244–252, 2001.

[384] H. K. Ulatowska, L. Allard, A. Donnell, J. Bristow, S. M. Haynes, A. Flower, and A. J. North. Discourse performance in subjects with dementia of the alzheimer type. In *Neuropsychological Studies of Nonfocal Brain Damage*, pages 108–131. Springer, 1988.

[385] M. T. Ullman, S. Corkin, M. Coppola, G. Hickok, J. H. Growdon, W. J. Koroshetz, and S. Pinker. A neural dissociation within language: Evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9(2):266–276, 1997.

[386] van Dalen J, van Wanrooij LL, M. van Charante EP, B. C, van Gool WA, and R. E. Association of apathy with risk of incident dementia: A systematic review and meta-analysis. *JAMA Psychiatry*, 2018.

[387] V. Vincze, G. Gosztolya, L. Tóth, I. Hoffmann, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán. 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)*, volume 2, pages 181–187, 2016.

[388] P. Vitali, J. Abutalebi, M. Tettamanti, J. Rowe, P. Scifo, F. Fazio, S. F. Cappa, and D. Perani. Generating animal and tool names: An fMRI study of effective connectivity. *Brain and Language*, 93(1):32–45, 2005.

[389] E. Vuorinen, M. Laine, and J. Rinne. Common pattern of language impairment in vascular dementia and in alzheimer disease. *Alzheimer Disease & Associated Disorders*, 14(2):81–86, 2000.

[390] A. Wallin, A. Nordlund, M. Jonsson, K. Lind, Å. Edman, M. Göthlin, J. Stålhammar, M. Eckerström, S. Kern, A. Börjesson-Hanson, M. Carlsson, E. Olsson, H. Zetterberg, K. Blennow, J. Svensson, A. Öhrfelt, M. Bjerke, S. Rolstad, and C. Eckerström. The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of cerebral blood flow and metabolism: official journal of the International Society of Cerebral Blood Flow and Metabolism*, 36(1):114–31, 2016.

[391] X. Wan. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, pages 235–243, 2009.

[392] S. Wankerl, E. Nöth, and S. Evert. An n-gram based approach to the automatic diagnosis of alzheimer's disease from spoken language. In *INTERSPEECH*, pages 3162–3166, 2017.

[393] T. Warnita, T. Warnita, N. Inoue, and K. Shinoda. Detecting alzheimer's disease using gated convolutional neural network from audio data. *Proc. Interspeech 2018*, pages 1706–1710, 2018.

[394] D. Wechsler. Wechsler abbreviated intelligence scale. In *Psychological Corporation*, San Antonio, TX, USA, 1999.

[395] J. Weiner, M. Angrick, S. Umesh, and T. Schultz. Investigating the effect of audio duration on dementia detection using acoustic features. In *INTERSPEECH*, pages 2324–2328, 2018.

[396] J. Weiner, M. Engelbart, and T. Schultz. Manual and Automatic Transcriptions in Dementia Detection from Speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3117–3121, 2017.

[397] J. Weiner, C. Herff, and T. Schultz. Speech-based detection of Alzheimer's disease in conversational German. In *Proceedings of Interspeech*, pages 1938–1942, 2016.

[398] J. Weiner, C. Herff, and T. Schultz. Speech-based detection of alzheimer's disease in conversational german. In *INTERSPEECH*, pages 1938–1942, 2016.

[399] J. Weiner and T. Schultz. Detection of intra-personal development of cognitive impairment from conversational speech. In *Speech Communication; 12. ITG Symposium*, pages 1–5. VDE, 2016.

[400] J. Weiner and T. Schultz. Automatic screening for transition into dementia using speech. In *Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

[401] J. Weiner and T. Schultz. Selecting Features for Automatic Screening for Dementia based on Speech. In *20th International Conference on Speech and Computer SPECOM 2018*, 2018.

[402] K. Wild, D. Howieson, F. Webbe, A. Seelye, and J. Kaye. The status of computerized cognitive testing in aging: a systematic review. *Alzheimers Dement*, 4(6):428–437, Nov. 2008.

[403] M. Wolters. Give me your data, and i will diagnose you. In *Data Power Conference 2017, Ottawa, CA*, 2017.

[404] M. K. Wolters, N. Kim, J. H. Kim, S. E. MacPherson, and J. C. Park. Prosodic and Linguistic Analysis of Semantic Fluency data: A Window into Speech Production and Cognition. In *INTERSPEECH 2016–17th Annual Conference of the International Speech Communication Association*, pages 2085–2089, 2016.

[405] D. L. Woods, J. M. Wyma, T. J. Herron, and E. W. Yund. Computerized Analysis of Verbal Fluency: Normative Data and the Effects of Repeated Testing, Simulated Malingering, and Traumatic Brain Injury. *PLOS ONE*, 11(12):1–37, 2016.

[406] World Health Organization. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. World Health Organization, 1992.

[407] A. Worthington and R. L. Wood. Apathy following traumatic brain injury: A review. *Neuropsychologia*, 118:40–47, 2018.

[408] M. Yancheva, K. Fraser, and F. Rudzicz. Using linguistic features longitudinally to predict clinical scores for alzheimer's disease and related dementias. In *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, pages 134–139, 2015.

[409] M. Yancheva and F. Rudzicz. Vector-space topic models for detecting alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2337–2346, 2016.

[410] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 133–136, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[411] H. Yazbek, S. Raffard, J. Del-Monte, F. Pupier, A. Larue, J.-P. Boulenger, M.-C. Gély-Nargeot, and D. Capdevielle. L'apathie dans la schizophrénie: une revue clinique et critique de la question. *L'Encéphale*, 40(3):231–239, 2014.

[412] C. A. Yeager and L. Hyer. Apathy in dementia: relations with depression, functional competence, and quality of life. *Psychological reports*, 102(3):718–722, 2008.

[413] B. Yu, T. F. Quatieri, J. R. Williamson, and J. C. Mundt. Cognitive impairment prediction in the elderly based on vocal biomarkers. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 3734–3738, 2015.

[414] D. Yu and L. Deng. *Automatic Speech Recognition*. Springer, 2016.

[415] G. S. Yuen, F. M. Gunning-Dixon, M. J. Hoptman, B. AbdelMalak, A. R. McGovern, J. K. Seirup, and G. S. Alexopoulos. The salience network in the apathy of late-life depression. *International journal of geriatric psychiatry*, 29(11):1116–1124, 2014.

[416] C. Zadikoff, S. H. Fox, D. F. TangWai, T. Thomsen, R. M. de Bie, P. Wadia, J. Miyasaki, S. DuffCanning, A. E. Lang, and C. Marras. A comparison of the mini mental state exam to the montreal cognitive assessment in identifying cognitive deficits in parkinson's disease. *Movement disorders*, 23(2):297–299, 2008.

[417] L. Zhou, K. C. Fraser, and F. Rudzicz. Speech recognition in alzheimer's disease and in its assessment. In *Interspeech*, pages 1948–1952, 2016.

[418] X. Zhu, H.-I. Suk, L. Wang, S.-W. Lee, D. Shen, A. D. N. Initiative, et al. A novel relational regularization feature selection method for joint regression and classification in ad diagnosis. *Medical Image Analysis*, 38:205–214, 2017.

[419] Z. Zhu, J. Novikova, and F. Rudzicz. Detecting cognitive impairments by agreeing on interpretations of linguistic features. *arXiv preprint arXiv:1808.06570*, 2018.

[420] Z. Zhu, J. Novikova, and F. Rudzicz. Isolating effects of age with fair representation learning when assessing dementia. 2018.