

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



AutoSpeech: Automatic Speech Analysis of Verbal Fluency for Older Adults

João António Fernandes da Costa

Mestrado Integrado em Bioengenharia

Supervisor: Liliana Ferreira

Co-Supervisor: Aníbal Ferreira

July 16, 2019

AutoSpeech: Automatic Speech Analysis of Verbal Fluency for Older Adults

João António Fernandes da Costa

Mestrado Integrado em Bioengenharia

July 16, 2019

Resumo

Os testes de Fluência Verbal são exames flexíveis e de fácil administração, frequentemente usados na área da Neuropsicologia para avaliar funções executivas e de linguagem. Estes testes consistem na produção do maior número de palavras possível que pertençam a uma determinada categoria, num intervalo de tempo de 1 minuto.

Existem dois tipos principais de testes de fluência verbal usados: fonémicos (em que são produzidas palavras que possuem uma determinada letra inicial, por exemplo, a letra *P*), e semânticos (em que as palavras produzidas pertencem a um determinado grupo semântico, por exemplo, animais). A má performance em testes de fluência verbal está associada a défices cognitivos, sendo que este teste pode ser usado para distinguir diferentes tipos de lesões cerebrais, perturbações mentais, e distúrbios neurocognitivos (como a demência).

Testes de fluência verbal têm dois fatores principais a influenciar o resultado final obtido. Os examinados têm de rapidamente identificar possíveis subcategorias que possam ser exploradas, enquanto inibem a produção de palavras incorretas e enquanto realizam uma procura por novas subcategorias a explorar - isto é comumente apelidado de troca de *clusters*/grupos, e está intrinsecamente associada às funções executivas. Para além disto, o examinado tem de invocar da memória palavras para produzir que pertencem a uma determinada subcategoria identificada - isto determina o tamanho dos *clusters* produzidos, uma tarefa principalmente associada a funções de linguagem e memória.

O aumento da incidência de distúrbios neurocognitivos e a falta de especialistas disponíveis para realizar triagem e diagnóstico motiva o aparecimento de métodos de automação de testes de fluência verbal.

Nesta dissertação é proposto e implementado um sistema de análise automático de testes de Fluência Verbal (letra *P* e animais), em Português Europeu. Este sistema usa os serviços de transcrição automática do Google Cloud Speech-to-Text, e técnicas modernas de Processamento de Linguagem Natural para extrair automaticamente *clusters* temporais, fonéticos e semânticos produzidos.

Este sistema é testado e usado para determinar o impacto de diferentes níveis de atividade física de idosos na performance dos testes de Fluência Verbal ($N = 57$). Verificou-se que idosos com um nível mais baixo de atividade física tendem a ter uma performance inferior a outros idosos mais ativos. No entanto, outros fatores que afetam este resultado, como estado cognitivo geral dos participantes e a sua idade, não permitem tirar conclusões definitivas sobre a influência de atividade física na cognição de idosos.

Palavras-Chave: Testes de Fluência Verbal, Processamento de Linguagem Natural, Reconhecimento de Fala, Funções Cognitivas, Funções Executivas

Abstract

Verbal Fluency tests are quick and flexible tests used in the area of Neuropsychology to evaluate executive and language functions of subjects. These tests consist in the production of as many words as possible within a 1 minute time constraint, with respect to a given category.

There are two main types of verbal fluency tests administered: phonemic (where words that begin with a given letter are produced, e.g., the letter *P*), and semantic (where words that belong to the same semantic group are produced, e.g., animals). Verbal fluency performance is shown to be related to cognitive impairment and can be used to distinguish different types of brain lesions, mental disorders, and neurocognitive disorders (such as dementia).

Verbal fluency tasks have two main factors influencing the overall performance obtained. Subjects must be able to identify a possible subcategory that can be explored for the production of exemplars, while inhibiting incorrect words, and searching for new possible subgroups to use afterward - this is known as cluster switching, which is more related to executive functions. Besides this, a subject has to effectively produce from memory words belonging to a given subcategory - this determines cluster size, a task mainly associated with language and memory functions.

The increase of incidence of neurocognitive disorders and the lack of available specialists to perform screening examinations is a motivation for the automation of these tasks.

An automatic pipeline for the analysis of both phonemic (letter *P*) and semantic (animal category) verbal fluency tests is here proposed, for European Portuguese. This pipeline uses Google Cloud Speech-to-Text online service to perform Automatic Speech Recognition, and Natural Language Processing techniques to automatically evaluate temporal, phonemic, and semantic clusters produced.

This pipeline is tested and used to determine the impact of physical activity on cognitive function and verbal fluency performance on elderly subjects ($N = 57$). Subjects with a lower level of physical activity tend to have lower cognitive and verbal fluency performance than more active subjects. However, due to other factors that affect this result, such as age and general cognitive status of subjects, no definitive conclusions can be taken regarding the influence of physical activity on elderly cognition.

Keywords: Verbal Fluency Tests, Natural Language Processing, Automatic Speech Recognition, Cognitive Functioning, Executive Functioning

Para a minha Mãe

Agradecimentos

Demorou, mas chegou a hora.

Agradeço aos meus orientadores, prof. Liliana Ferreira, prof. Aníbal Ferreira, e à Ana Pereira, pela ajuda prestada durante a realização deste trabalho.

Deixo aqui os meus profundos agradecimentos aos psicólogos Ricardo Araújo e Marta Martins, que sem a sua dedicação incessante para ajudar com a recolha de dados, não me seria possível estar a escrever isto neste momento. Junto a estes agradecimentos os meus agradecimentos à Sílvia Rêgo, pela sua constante disponibilidade e ajuda com todo este processo temeroso de recolha de dados.

Agradeço também a todos os idosos que participaram no estudo aqui desenvolvido. A sua colaboração, disponibilidade, alegria e simpatia também alegraram quem lhes pedia para dizer o maior número de animais num minuto.

Um enorme obrigado ao Dr. Pedro Amorim, à Dra. Maria João Susano e ao Dr. Francisco Seixas, do Hospital Santo António, pelo enorme interesse que apresentaram neste projeto, e pela imediata disponibilidade que tiveram em reunir, conversar, e perceber melhor o que estava a fazer. A ajuda que deram a este trabalho é enorme e têm toda a minha gratidão!

E claro, agradecer à Dra. Mariana Dias! Que "jantares de amigos" com inesperadas coincidências da vida continuem a acontecer pelos anos fora.

... e a ti. Se estás neste momento a ler isto, então é porque eu assim o quis. Se assim o quis, é porque queria que lesses esta parte, porque é para ti. Obrigado, por tudo!

João Costa

“I don’t like anything here at all.” said Frodo, “step or stone, breath or bone. Earth, air and water all seem accursed. But so our path is laid.”

“Yes, that’s so,” said Sam, “And we shouldn’t be here at all, if we’d known more about it before we started. But I suppose it’s often that way. The brave things in the old tales and songs, Mr. Frodo, adventures, as I used to call them. I used to think that they were things the wonderful folk of the stories went out and looked for, because they wanted them, because they were exciting and life was a bit dull, a kind of a sport, as you might say. But that’s not the way of it with the tales that really mattered, or the ones that stay in the mind. Folk seem to have been just landed in them, usually their paths were laid that way, as you put it. But I expect they had lots of chances, like us, of turning back, only they didn’t. And if they had, we shouldn’t know, because they’d have been forgotten. We hear about those as just went on, and not all to a good end, mind you; at least not to what folk inside a story and not outside it call a good end. You know, coming home, and finding things all right, though not quite the same; like old Mr Bilbo. But those aren’t always the best tales to hear, though they may be the best tales to get landed in! I wonder what sort of a tale we’ve fallen into?”

J.R.R. Tolkien - The Two Towers

Contents

Abbreviations and Symbols	xix
1 Introduction	1
1.1 Context	1
1.1.1 Dementia in Portugal	2
1.1.2 Diagnosis of Neurocognitive Disorders	3
1.2 Motivation	3
1.3 Main Objectives	4
1.4 Main Contributions	5
1.5 Document Structure	5
2 Literature Review	7
2.1 Human Behaviour	7
2.1.1 Cognitive Assessment Batteries	9
2.2 Verbal Fluency Tests	10
2.2.1 Subcategories of Verbal Fluency Tests	11
2.2.2 Scoring Methods for Verbal Fluency	12
2.2.3 Functional Brain Imaging during Verbal Fluency Tests	15
2.2.4 Manual Administration of Verbal Fluency Tests	16
2.3 The Impact of Ageing and Pathology on Verbal Fluency Tests	17
2.3.1 Ageing	17
2.3.2 Neurocognitive Disorders	19
2.4 Speech-to-Text Systems	22
2.4.1 Performance of State-of-the-Art Speech-to-Text Systems	22
2.4.2 Speech Recognition Online Services	23
2.5 Natural Language Processing	23
2.5.1 Word Stemming and Lemmatisation	24
2.5.2 Word Ontologies and WordNet	25
2.5.3 Word Embeddings	27
2.6 Automatic Administration of Verbal Fluency Tests	29
3 Implemented Pipeline for the Automatic Analysis of Verbal Fluency Tests	33
3.1 Audio Processing	35
3.1.1 Google Cloud Storage	36
3.2 Speech-to-Text	36
3.3 Non-Test Word Removal	37
3.3.1 Non-Test Word Removal for Phonemic Tests	37
3.3.2 Non-Test Word Removal for Semantic Tests	37

3.4	Phonemic Transcription	40
3.5	Word Embeddings Dictionary	42
3.6	Incorrect Word Removal	42
3.6.1	Incorrect Word Removal for Phonemic Tests	43
3.6.2	Incorrect Word Removal for Semantic Tests	43
3.7	Feature Extraction	45
3.7.1	Temporal Features	45
3.7.2	Semantic Features	47
3.7.3	Phonemic Features	48
4	AutoSpeech Study Design	53
4.1	Main Objectives	53
4.2	Participant Recruitment	53
4.2.1	Recruitment Strategies	54
4.2.2	Inclusion and Exclusion Factors	54
4.2.3	Informed Consent	55
4.3	Test Protocol	55
4.3.1	Test Material	56
4.4	Risk Management	57
4.5	Benefits for the Participant	57
4.6	Confidentiality	57
4.6.1	Data Acquisition Methods	57
4.7	Ethical Considerations Regarding Test Results	58
5	Data Characterisation	61
5.1	Centro Hospitalar Universitário do Porto Data	61
5.1.1	Data Collection Method	62
5.1.2	Population Characterisation	62
5.2	AutoSpeech Study Data	62
5.2.1	Data Collection Methods	63
5.2.2	Population Characterisation	63
5.3	Manual Feature Extraction	64
5.3.1	Semantic Switching and Clustering Features	65
5.3.2	Phonemic Switching and Clustering Features	66
6	Results and Discussion	69
6.1	Speech-to-Text Performance	70
6.1.1	Speaker Model	73
6.1.2	Test Type	74
6.1.3	Recording Environment	74
6.2	Automatic Feature Extraction Performance on Manual Transcripts	75
6.2.1	Number of Correct Words	75
6.2.2	Embeddings Cosine Similarity as a Measure of Semantic Similarity	76
6.2.3	Switching and Clustering Features	77
6.3	Automatic Feature Extraction Performance on Automatic Transcripts	80
6.3.1	Number of Correct Words	81
6.3.2	Switching and Clustering Features	83
6.4	Impact of Physical Activity of Elderly on Cognition and Verbal Fluency	86
6.4.1	Discrimination of Manual Features According to Physical Activity Levels	89

6.4.2	Discrimination of Automatic Features According to Physical Activity Levels	92
6.5	Final Remarks	94
7	Conclusions	97
7.1	Pipeline Improvements	98
7.2	Future Work	100
A	Statement of Gabinete Coordenador da Investigação of CHUP	103
B	Ethics Committee of Universidade do Porto Study Approval	107
C	Informed Consent to Participants, Portuguese Version	111
D	Montreal Cognitive Assessment, Portuguese Version	115
E	Animal Fluency, Portuguese Script	117
F	Sociodemographic Questionnaire, Portuguese Version	119
G	Short International Physical Activity Questionnaire, Portuguese Version	121
	References	125

List of Figures

1.1	Developed Regions Population Pyramids from 1950 and 2017	1
1.2	Longitudinal perspective of the number of cases of Dementia in Portugal, according to age group	2
2.1	Semantic Fluency vs. Phonemic Fluency in whole brain analysis	16
2.2	Graph representation of the synset <i>dog.n.01</i> and its hypernym relations	26
2.3	Different types of relations between noun synsets in WordNet	27
3.1	Schema of the implemented pipeline for automatic VF test analysis, from speech signal data	34
3.2	Frequency Profile of the employed Equalisation (Bass Boost) step for audio processing	35
3.3	Schema of the implemented pipeline for automatic VF test incorrect word removal, from Speech-to-Text transcripts of audio signals	38
3.4	Hypernym tree illustration of several animal names	40
3.5	Identified temporal clusters of an automatic PVF test transcript	46
3.6	Semantic Similarity Matrices	51
3.7	Phonemic Similarity Matrices	52
4.1	Depiction of a data collection interview from the AutoSpeech Study	55
5.1	Distribution of collected subjects, according to location	64
5.2	Age and Education distribution of collected subjects	65
6.1	Diagram of processing and analysis performed on collected data	70
6.2	Verbal Fluency Word Error Rate according to Test Type, Location, and Speaker Model	72
6.3	Number of Correct and Incorrect Words in Automatic Speech Recognition Transcripts	73
6.4	Relative Error of the Number of Correct Words Automatically Extracted from Manual Transcripts	76
6.5	Comparison of pairing frequency of animal names with their embeddings' cosine similarity	78
6.6	Spearman's ρ between Automatically and Manually Extracted Features from Manual Transcripts	79
6.7	Number of Correct Words Error Rate according to Test Type and Location	82
6.8	Spearman's ρ Correlation Matrix for Phonemic Verbal Fluency Automatic Features	84
6.9	Spearman's ρ Correlation Matrix for Semantic Verbal Fluency Automatic Features	85
6.10	Distribution of Subjects By Level of Physical Activity and Data Collection Location	86

6.11 Age and Education Distribution, According to Physical Activity Levels	87
6.12 Montreal Cognitive Assessment scores According to Physical Activity Level . . .	88
6.13 Verbal Fluency Word Error Rate According to Physical Activity Level	90
6.14 Manual Verbal Fluency Feature Distributions According to Physical Activity Levels	91
6.15 Automatic Verbal Fluency Feature Distributions According to Physical Activity Level	93

List of Tables

2.1	Normalised scores for the MoCA for the Portuguese population	10
2.2	Impact of Pathologies and Ageing on Verbal Fluency	18
2.3	Overview of Automatic Speech Recognition APIs	23
2.4	Summary of reviewed state of the art studies	31
4.1	Tests and Questionnaires used in the designed exploratory study	58
5.1	Overview of Age distribution of collected subjects, according to gender and collection location	66
5.2	Overview of Years of Education distribution of collected subjects, according to gender and collection location	67

Abbreviations and Symbols

#	Number
AACD	Age-Associated Cognitive Decline
AD	Alzheimer's Disease
ADHD	Attention Deficit/Hyperactivity Disorder
API	Application Programming Interface
ASR	Automatic Speech Recognition
CHUP	Centro Hospitalar Universitário do Porto
DAT	Dementia of the Alzheimer Type
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DPD	Dementia with Parkinson's Disease
ERPI	Estrutura Residencial de Pessoas Idosas
ESA	Explicit Semantic Analysis
HD	Huntington's Disease
IPA	International Phonetic Alphabet
IPAQ	International Physical Activities Questionnaire
LSA	Latent Semantic Analysis
MCI	Mild Cognitive Impairment
MET	Metabolic Equivalent of Task
MMSE	Mini-Mental State Examination
MoCA	Montreal Cognitive Assessment
NCD	Neurocognitive Disorders
NDP	Non-demented Parkinson's Disease
NLP	Natural Language Processing
PD	Parkinson's Disease
PF	Pairing Frequency
POI	Phonemic Organisation Index
PVF	Phonemic Verbal Fluency
SOI	Semantic Organisation Index
SVF	Semantic Verbal Fluency
SVM	Support Vector Machine
VF	Verbal Fluency
VFWER	Verbal Fluency Word Error Rate
WER	Word Error Rate

Chapter 1

Introduction

1.1 Context

Developed countries have had a newly emerging issue to tackle over the last decades: their populations are getting older [118]. The increase in life expectancy and the improvement of health care in the past century allow for longer, better lives, as can be seen in Figure 1.1. However, this can be a curse in disguise: longer lives often lead to elderly health issues, with the threat of dementias such as Alzheimer’s Disease increasing, robbing patients of their independence.

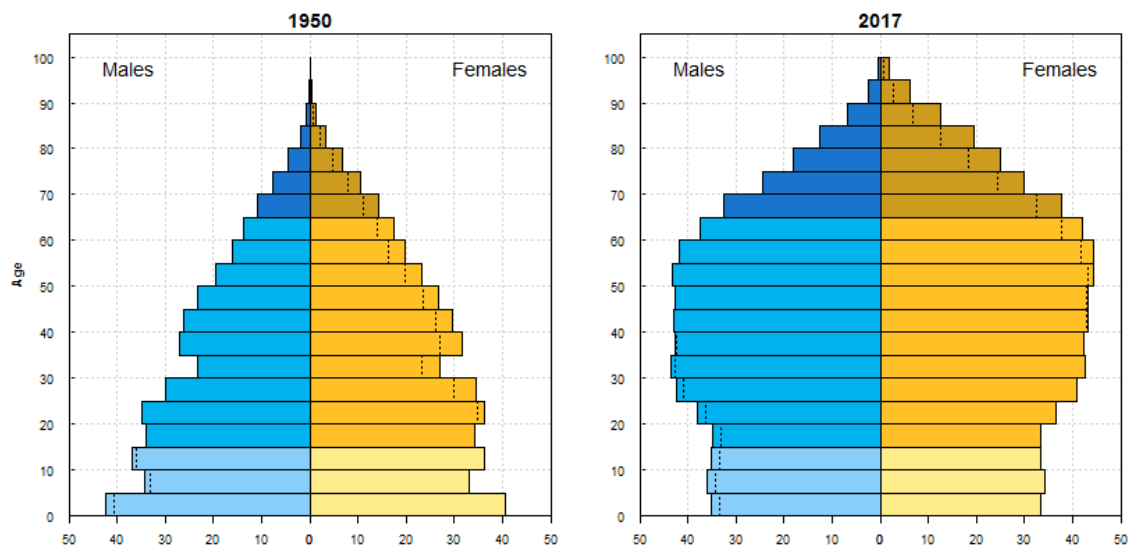


Figure 1.1: Developed Regions Population Pyramids from 1950 and 2017, from United Nations [118]

Age-associated Neurocognitive Disorders (NCD), commonly referred as dementias [3], are a grave health issue worldwide. As the World Health Organization [122] points out, dementia is currently an urgent worldwide public health issue, with over 50 million subjects diagnosed, with 10 million new yearly cases.

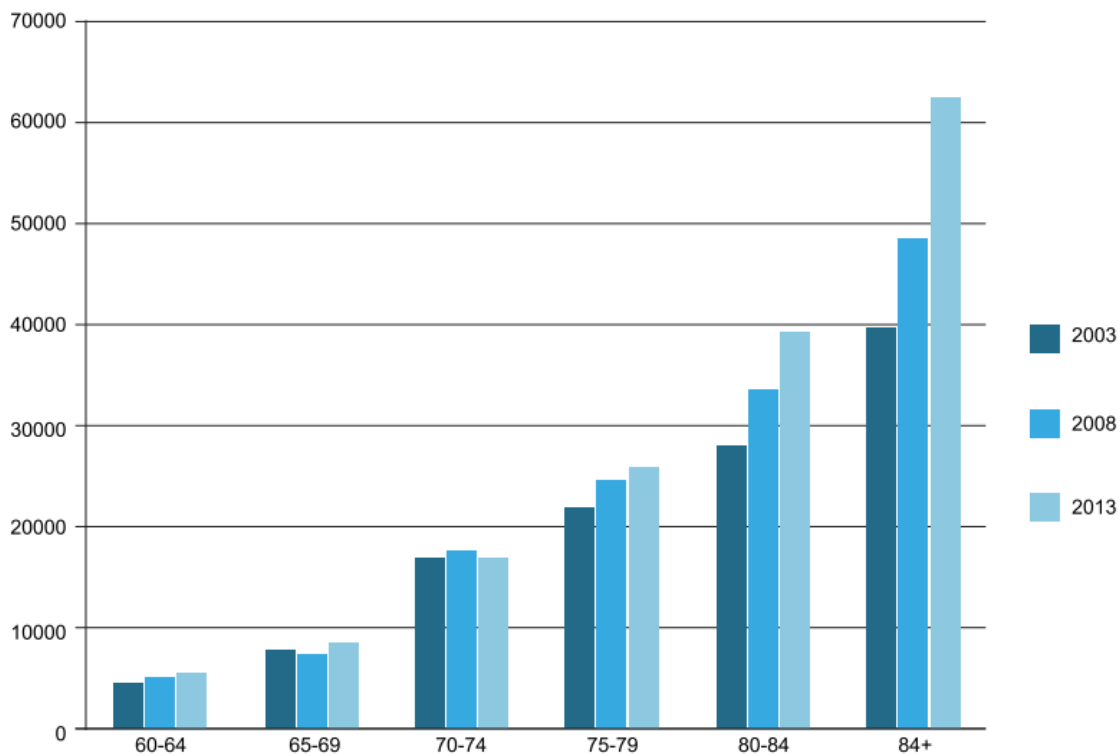


Figure 1.2: Longitudinal perspective of the number of cases of Dementia in Portugal, according to age group, from Santana et al. [106]

Dementia is a chronic and progressive syndrome that affects elderly subjects, with the deterioration of cognitive functions, such as memory and orientation, greatly impacting a patient's independence [122], being a major cause of disability. Dementias have a considerable impact on carers and patients alike, in social, economic, physical, and psychological terms. Dementia-associated costs represent approximately 1% of the gross world product: in 2010, dementia-related costs were estimated to be \$604.000M. [7].

The most common cause of dementia in the elderly is Alzheimer's Disease, which is responsible for over 60% of NCD cases [122]. Other causes for dementia are Parkinson's Disease, vascular accidents, and other dementias associated with neurodegenerative disorders (for instance Huntington's Disease).

1.1.1 Dementia in Portugal

There are almost 200.000 registered cases of dementia in Portugal, with incidence increasing over the last years [7]. Figure 1.2 shows the evolution of the number of cases of dementia in Portugal in recent years: the number of cases is increasing, and dementia incidence increases as subjects get older.

Costs associated with dementia in Portugal are estimated to be around €32M per year, in medication alone. 6% of the elderly Portuguese population is affected by some form of dementia, with 60% of elderly aged 85 or older having dementia [106].

Portugal is the 4th country in the OECD (Organisation for Economic Co-operation and Development) with the highest ratio of cases of dementia [8]. Furthermore, Santana et al. [106] indicate that dementias are particularly sub-diagnosed in Portugal, with primary health care professionals with little to no knowledge regarding NCDs. Even more concerning, there only exist 12 psychiatrists per 100.000 Portuguese inhabitants, and only 5 neurologists, a much lower ratio compared to other OECD members. It is estimated that 50% of dementia cases remain undiagnosed [8].

1.1.2 Diagnosis of Neurocognitive Disorders

Neurocognitive disorders, such as Alzheimer's Disease or Parkinson's Disease, are complex, multi-factorial pathologies, and no exam exists that can single-handedly diagnose a disorder involving neurological impairment [3, 6].

The cycle of care of dementia usually starts with memory and cognition complaints of patients, carers, or close family members. These complaints are often shared with primary health care professionals, that may suggest an in-depth consultation with an NCD specialist, such as a psychiatrist or a neurologist [3, 6].

Specialists evaluate the possibility of dementia according to several parameters so that other possible conditions can be rejected:

- Medical history;
- Physical examination, namely laboratory testing (blood and urine tests), to rule out possible medication side effects or metabolic disorders;
- Mental state examination, to evaluate cognitive functions, which can be done using standard neuropsychological batteries, such as the Mini-Mental State Examination (MMSE) [36], or the Montreal Cognitive Assessment (MoCA) [80];
- Brain imaging, using magnetic resonance (MRI), computerised tomography (CT), or positron emission techniques (PET).

Only after thorough physical and neurological examination can a dementia diagnosis be performed. These tests also help in determining the nature of the NCD at hand [3].

After a diagnosis is given, patients are regularly checked regarding the progression of the disorder and are usually prescribed with anti-dementia medication, to minimise the symptoms of these disorders. Unfortunately, in most cases, progression can be quick and sudden, often leading to a complete loss of independence, requiring constant caregiver assistance, admission into elderly care homes, or even hospitalisation [3].

1.2 Motivation

Of the commonly administered neuropsychological tests to assess cognitive impairment for the diagnosis of NCDs, verbal fluency (VF) tests stand out for being very quick and easy to administer.

VF tests are also common in neuropsychological test batteries for mental state examination: the MoCA is one such example of a battery which includes a VF test [80].

VF tests consist in the production of words respecting a given rule in a constrained time interval. Of the several possible categories, phonemic VF (production of words starting with a given letter), and semantic VF (production of words belonging to the same semantic group, e.g., production of animal names) are the most commonly used ones. VF test performance is shown to correlate to executive and language functions, being a good measure of impairment in the frontal and temporal lobes [69].

To score a VF test, the test examiner registers all words produced by the testee. The final score is the total number of correct words produced, ignoring proper nouns and repetitions [69].

Recent developments in the scoring procedures of VF tests allow for the extraction of more fine-grained metrics that can distinguish the different contributions of executive functions (self-control, inhibition, set switching) and language memory (retrieval and production of exemplars) [114].

The computation of these new metrics, however, is very laborious and time-consuming to be performed in a clinical setting, being mostly used exclusively in research. Common VF test assessments rely only on the traditional score of the total number of words produced because it is relatively quick to obtain. Nonetheless, the administration of these tests requires for the examiner to manually annotate all produced words, which can also be strenuous, particularly considering an overburdened national health system, with very few specialists available to perform these evaluations [8].

As such, recent efforts for the automation of VF test scoring have been developed [24, 61, 62, 85, 86, 87, 112, 121], both to ease and quicken test transcription and to allow the extraction of more informative metrics.

1.3 Main Objectives

The principal objective of this dissertation is the creation of a fully automated computational pipeline for the transcription and scoring of Portuguese VF tests.

Another main objective is to perform an exploratory study to assess the existence of correlation of cognitive abilities of Portuguese elderly subject with their different levels of physical activity, using automatically extracted features from VF tests.

For these goals, a Python 3 based solution for the automatic transcription of recorded VF test audio data is constructed, relying on external online services for Automatic Speech Recognition, such as Google Cloud [45] Speech-to-Text API.

After transcription, Natural Language Processing (NLP) techniques are applied to the obtained words to assess semantic and phonemic similarity and relatedness between them for computation of switching and clustering metrics [114], the golden standard metrics for VF tests. Furthermore, novel semantic and phonemic features based on NLP techniques [86, 87, 104, 112, 121]

are extracted. In addition, speech signal related features, such as the calculation of time intervals between produced words, will be studied in tandem with the features described above [62].

To validate the constructed pipeline, and to evaluate the influence of age, physical activity, and cognitive function in the performance of VF tests, a population sample of elderly adults is collected. Each subject in the studied sample is characterised according to:

- Cognitive ability;
- Demographic factors (age, gender, level of education);
- Level of physical activity.

1.4 Main Contributions

The main contribution of this dissertation work is the creation of the first fully automated VF test analysis pipeline in European Portuguese, that can automatically transcribe, analyse, and extract relevant features from these tests.

The created pipeline extracts semantic, phonemic, and temporal features from automatic transcripts of VF tests, which is also the first time temporal and semantic similarity features using NLP techniques are utilised in Portuguese VF tests. In addition, it is the first time all these three types of features are extracted simultaneously from both phonemic and semantic VF tests.

A novel database of VF test performance in healthy and cognitive impaired elderly subjects is also created with the development of this work. Subjects are thoroughly characterised, with respect to demographic factors (such as age and level of education), and cognitive impairments (by performing a mental state examination, namely the Montreal Cognitive Assessment [80]), but also with respect to the level of physical activity performed regularly.

This database is used to evaluate the influence of physical activity, age, and education on the performance of VF tests, both phonemic and semantic, with particular emphasis on the automatic features extracted with the developed pipeline. These features are known to correlate to verbal and executive functions [114], so an overall brain function profile can be obtained, and populations with different levels of physical activity can be compared.

1.5 Document Structure

Chapter 2 provides an overview of the current consensus and trends present in literature, mainly regarding the administration and applicability of Verbal Fluency Tests, and what Automatic Speech Recognition and Natural Language Processing tools currently exist that can be used to automatise test administration.

Chapter 3 presents the methods implemented in this work for the automation of VF tests, detailing what procedures are performed to process audio, perform automatic transcription of tests, and automatically extract meaningful features from them.

Chapter 4 defines the exploratory study of this work, which aims to evaluate the influence of physical activity on the cognitive health of a sample of elderly Portuguese subjects. The design of test protocols and participant recruitment is presented and explained here. Several ethical and legal questions are also addressed here, namely risk management, confidentiality, and informed consent.

Chapter 5 characterises the data gathered in the aforementioned study, describing participant's age, level of education, data collection location, among other influencing factors.

Chapter 6 presents and discusses the gathered results of the previously described study in four main topics:

- Section 6.1 studies the overall performance of the utilised Speech-to-Text system, compared with manual transcriptions of VF tests;
- Section 6.2 examines the correlation between automatically extracted features from manual transcripts of VF tests, comparing these features with the current gold standard of VF test analysis;
- Section 6.3 determines the correlation of automatically extracted features from automatic test transcripts provided after Automatic Speech Recognition with manually extracted ones, thus evaluating if the automatic VF test analysis pipeline extracts valid and reliable metrics;
- Section 6.4 analyses the impact of physical activity on the performance of VF tests on elderly subjects, by comparing obtained automatic features of groups determined by physical activity levels.

Chapter 7 exposes the main conclusions taken from the production of this work, what were its main challenges and areas to improve, and possible future work to expand on it.

Chapter 2

Literature Review

This chapter will present reviewed material of required prior knowledge regarding the subject of the dissertation. Fully understanding what a Verbal Fluency Test is, how it is administered, what it measures, and how its results are affected by certain pathologies is essential to interpret its outputs and how they can be used to distinguish individuals in a given population. Furthermore, since the design and execution of a pipeline for automatic analysis of verbal fluency tests is a prime objective of the dissertation, a thorough understanding of each component is paramount for the success of the designed algorithms, namely Speech-to-Text systems and Natural Language Processing techniques.

2.1 Human Behaviour

Human behaviour is a complex interplay of cerebral processes, in response to internal and external stimuli, that encapsulates cognition, emotion, and control. Behaviour and its expression are the main target of study and analysis of Psychology and Neuropsychology, since it can provide insight into the functions of the brain. Verbal Fluency tests are one of the tools commonly used by psychologists, neuropsychologists, and neurologists to assess behaviour and brain functions that affect it [69].

Behaviour is commonly subdivided into its three main dimensions, regarding distinct processes, tasks, and abilities, but intricately interconnected: Cognitive Functions, Emotionality, and Executive Functions.

Cognitive Functions Cognitive functions concern information handling, that is, its reception (perceptive abilities), storage and retrieval (memory), organisation (thinking), and communication (expression). These functions are the most predominantly studied, not only due to their correlation with neuroanatomical systems but also to the fact that impairment in these functions generates prominent and evident symptoms. Furthermore, the structured nature of medical examination lessens the possibility of evaluation of emotional or executive/control deficits [69].

The majority of neuropsychological tests involve the evaluation of cognitive functions, for instance, perception, memory, language skills, drawing, spatial orientation, and reasoning. Lezak et al. [69] provides an extensive list of currently used tests to evaluate cognitive functioning.

Executive Functions Whereas cognitive functions consider the overall ability of a subject to achieve certain tasks, executive functions pertain to a subject's willingness to execute said tasks. That is, these functions are related to the existence of independent behaviour, the ability to create and follow self-guided plans, and to employ strategies to initiate, monitor, and retroactively control activities [69].

Executive function impairment tends to be expressed as global alterations in the behaviour of a subject. Cognitive impairment may only affect a particular set of functions, while others remain relatively unaltered; executive impairment has a broad impact on the behaviour of a subject, affecting cognitive functions because subjects no longer possess adequate approach or planning strategies to tackle certain tasks. A subject who has sustained an extensive cognitive loss, but with relatively intact executive functions, can still manage to have an independent life, being able to self-care, be productive, and maintain relationships with others. When executive functions are impaired, a general *anergia* (inertia, lack of motivation, inability to initiate activity) is present, with defective self-control, emotional and personality alterations, carelessness, and difficulties in keeping focused attention.

Executive functions are somewhat elusive due to their intrinsic difficulty of dissociation from other cognitive tasks and their paradoxical nature of testing (a structured protocol and scoring method has to be constructed for a task where subjects are expected to generate their own structure and strategy to undertake it). Possible executive components are still debated [59].

Lezak et al. [69] has an extensive list of executive function tasks. Note that Verbal Fluency is considered an executive ability and can be used to test this dimension of behaviour.

Executive functions are deeply related to frontal lobe functioning, as can be seen with cases of brain damage to this area (particularly the prefrontal region) and respective consequences on planning, decision making, and self-control, leading to a set of symptoms commonly referred as *executive dysfunction*. Damage to these areas does not directly correlate to cognitive difficulties, but has a high impact on the overall quality of life of patients, since their executive functions become impaired, becoming highly dependent on others [50, 59, 69].

Emotionality Emotions and personality are a mainstay of a subject's behaviour. A stable emotional background is essential for a subject's well-being and can be often transformed when exposed to brain injury or disorder, whether from the damage itself or from the added emotional baggage it may carry [69].

Brain lesions are commonly associated with emotional and personality changes in patients. Depression is often present, due to radical changes in lifestyle, loss of function and consequent frustration with dealing with a newly found impaired state. Some display emotional lability, i.e., general instability with exaggerated mood swings, lack of control, and strong expressive episodes.

Particular brain dysfunctions may first present strong emotional changes, thus emotional functioning and personal adjustment assessments are valuable as diagnosing tools. Besides this, the emotional state of a subject must always be taken into consideration when performing neuropsychological assessments, since it can affect overall test performance.

2.1.1 Cognitive Assessment Batteries

Verbal Fluency Tests are often utilised by neuropsychologists in conjunction with other types of test, as part of neuropsychological test batteries. These batteries are designed to evaluate several cognitive and executive functions, providing an overall view of the mental state of a subject. In the particular case of neurocognitive disorders, battery results are crucial for the diagnosis of diseases such as Alzheimer's, Parkinson's, etc., along with other tests, such as brain imaging techniques [3].

These test batteries are commonly administered in the scope of mental status examinations, performed as relatively structured interviews with verbally intact patients to assess their level of cognitive and executive functions, by examining notable behaviours and capabilities regarding appearance, orientation, speech, thought process, attention/concentration and memory, cognitive functioning, emotional state, preoccupations, and insight. These interviews are the main source of data to determine a patient's ability for self-care or regarding legal issues.

A particular use of interest for mental status examinations is for dementia screening and rating. Through these test batteries, a professional can accurately and systematically score the capabilities of a subject and compare them to the standard normal values for its population: a significant decrease of performance may indicate the impact of a progressive dementing disease. Thus, mental status examinations have received growing interest as screening tools for the ever more growing population of demented elderly, particularly with the possibility of their automatic administration.

Two particular prominent sets of tests for mental status examinations are the Mini-Mental State Examination (MMSE) [36], and the Montreal Cognitive Assessment (MoCA) [80], both relatively quick and easy to administer, and frequently used to screen dementias and other cognitive impairments [69].

2.1.1.1 The Montreal Cognitive Assessment (MoCA)

The Montreal Cognitive Assessment (MoCA), developed by Nasreddine et al. [80], is a more thorough mental state examination, being slightly longer than the MMSE, usually lasting 10 to 15 minutes. MoCA includes evaluating items for orientation in space and time, sustained attention, memory and recall, confrontational naming, among others.

Particularly relevant for this work, the MoCA includes a Verbal Fluency test, specifically a phonemic verbal fluency task: in English using the letter *F* [80], in Portuguese the letter *P* [38]. Refer to Section 2.2 for more information on phonemic verbal fluency tests.

Table 2.1: Normalised scores for the MoCA for the Portuguese population, with respect to age and educational level, from Freitas et al. [39, Table 5]. Scores are presented as the average $\pm$ standard deviation

Age	Educational level (years)				All education
	Primary (1-4)	Middle (5-9)	High (10-12)	University (>12)	
25-49	23.55 $\pm$ 2.56	26.42 $\pm$ 2.18	27.39 $\pm$ 1.86	28.83 $\pm$ 1.38	26.98 $\pm$ 2.55
50-64	21.78 $\pm$ 2.86	25.58 $\pm$ 2.25	26.61 $\pm$ 2.28	27.51 $\pm$ 2.13	24.46 $\pm$ 3.43
≥ 65	21.27 $\pm$ 3.37	24.60 $\pm$ 2.87	25.11 $\pm$ 1.94	26.35 $\pm$ 1.87	22.71 $\pm$ 3.60
All ages	21.71 $\pm$ 3.18	25.65 $\pm$ 2.50	26.77 $\pm$ 2.15	28.04 $\pm$ 1.94	24.70 $\pm$ 3.67

The MoCA has shown to be a good predictor for the presence of mild cognitive impairment, being validated for use in detecting Alzheimer’s Disease, Parkinson’s Disease, vascular cognitive impairment, and substance abuse disorders [60, 80, 109].

The maximum score for the MoCA is 30 points. Age and education have shown to have an impact on average performance [69], thus normative studies exist to remove the influence of these demographic variables. Freitas et al. [39] performs a normative study of MoCA scores for the Portuguese population, with age and education having a large influence in the obtained results on a population sample of healthy subjects. Table 2.1 shows the normalised scores of the MoCA for the Portuguese population.

2.2 Verbal Fluency Tests

Verbal Fluency (VF) tests aim to assess all associated mental processes that must be in place for fluent speech to occur. These are complex and varied, ranging from more executive aspects (such as self-regulation and flexible thinking) to more language-related processes (such as language processing and production speed), with required short and long term memory involvement [69].

VF tests evaluate executive aspects of verbal behaviour and how (and how quickly) subjects organise their thoughts to generate words from a given category, with necessary self-regulation strategies in place to inhibit incorrect attempts, and with efficient semantic search strategies to find subcategories and items within these subcategories [69].

Standard VF tests are administered individually, with a quick oral introduction of the test and associated rules: subjects are asked produce as many words as possible from a given category, within a time limit of sixty seconds, avoiding repetitions and proper nouns. The examiner records all words produced, including errors and repetitions, for posterior analysis [69].

These tasks, although superficially similar, combine complex neural pathways present in distinct cerebral regions, which are activated differently according to the type of test performed. Alteration of performance in one or more of VF test subcategories can be used as indicator of frontal or temporal brain injuries [11, 115], but also for brain dysfunctions, such as Attention-Deficit/Hyperactivity Disorder (ADHD) [5, 55], chronic depression [10], and neurodegenerative

diseases, such as Parkinson's Disease (PD), Alzheimer's Disease (AD), Mild Cognitive Impairment (MCI) [94, 97, 116], and Huntington's Disease (HD) [54].

VF testing proves to be a useful tool in the neuropsychologist's belt. Ease of administration, with a quick quantitative and qualitative assessment of a subject's executive functions, and working and verbal memory, make these tests highly valuable. As Regard et al. [99] state:

"Fluency Tasks require little time and effort to administer and score and may prove useful in assessing adults and children with neurological disorders"

2.2.1 Subcategories of Verbal Fluency Tests

VF tests have several modalities, differing in the type and relation of words that are asked to be produced by the subject. Two main modalities are presented, corresponding to the most used in neuropsychological assessments, and the most informative regarding brain function [69]:

Phonemic VF Tests The subject is asked to produce words that start with a given letter (e.g. *"Please say words that start with the letter F"*);

Semantic VF Tests The subject is asked to produce words belonging to a given semantic category (e.g. *"Please say names of animals"*);

Phonemic VF tests are considered the harder task, being mostly dependent of effective executive capabilities and frontal lobe functioning, whereas Semantic VF tests have more words produced, relying more on verbal memory and temporal lobe functioning [69].

2.2.1.1 Phonemic Verbal Fluency Tests

Phonemic Verbal Fluency Tests (or simply, phonemic fluency tests, or PVF tests) are a subcategory of VF tests, where the main task is the spontaneous production of words that belong to the same lexical category (i.e., share the same initial letter), within a time limit (usually sixty seconds) [69].

Phonemic Verbal Fluency Tests are also named in the literature Phonological Fluency Tests, Controlled Oral Word Association Tests (COWAT), Lexical Verbal Fluency Tests, or simply Letter Fluency Tests [69].

Considering the large number of words present in any language's lexicon that can be used to fulfil this task, and the variability of commonly used words for each initial letter, one of the key factors for the performance of PVF tests is the initial letter that is chosen for this task. Borkowski et al. [16] performed a normative study, classifying each letter in the English alphabet according to its difficulty for the PVF task, as Easy, Moderate or Hard. Commonly used letters for the English Language are *F*, *A* and *S*, all of easy difficulty. In some cases, PVF tests are simply named the *FAS* Test, due to this fact.

For the Portuguese language, commonly used letters for normative studies are *M*, *R*, and *P*, for both European and Brazilian studies, as used by Cavaco et al. [21], Rodrigues et al. [101]. These letters are used due to their varied difficulty levels for the PVF task.

VF tests are generally quick and easy to administer to any subject. However, for the particular case of phonemic testing, one must take into careful consideration the level of education of the subject. As shown by Reis and Castro-Caldas [100], phonemic fluency performance is severely hampered for illiterate subjects, due to deficient phonological processing capability. This does not occur for subjects who have learned to read and write, possibly because this complex learning process creates new neural pathways that enhance phonological analysis through the more abstract written representation of words.

2.2.1.2 Semantic Verbal Fluency Tests

Semantic Verbal Fluency Tests (or simply, semantic fluency tests, or SVF tests) are a subcategory of VF tests, where the main task is the spontaneous production of nouns that belong to the same semantic category, within a time limit (usually sixty seconds) [69].

SVF tests are also named in the literature Category Fluency Tests. Similarly to PVF testing, the chosen category to evaluate fluency will affect the overall performance of subjects. Frequently used categories are names of animals, names of fruits and vegetables, names of supermarket items, among others. The most commonly used category is animals, so much that SVF tests often appear in literature simply referred to as Animal Fluency Tests [69].

Similarly to the English language, European Portuguese SVF tests often employ the animal naming test to evaluate performance, as seen used in the works of Cavaco et al. [21], Nogueira et al. [82]. Brazilian Portuguese tests also employ this category the most, for instance in the work of Brucki and Rocha [18].

Semantic tests can be applied to a certain extent to illiterate subjects. Reis and Castro-Caldas [100] show that semantic fluency performance of illiterate subjects is comparable to more educated subjects. However a difference is still observable, with illiterate subjects performing worse than literate subjects, but unlike phonemic fluency, they do not appear to be severely hampered on this task.

2.2.2 Scoring Methods for Verbal Fluency

The application of VF tests allows for a quantitative comparison of performance between subjects or between distinct populations, through measurable characteristics.

2.2.2.1 Total Number of Correct Words

The most commonly used score for this objective is the total number of correct words produced (also known as the number of words, or in some cases, raw score), ignoring errors, repetitions, and proper nouns. To allow posterior analysis of the performance of a subject on a given test, all words are recorded by the examiner, even if they are incorrect or repeated, being ignored afterwards.

General correct word criteria are further described for each type of VF test, defining what counts as an admissible exemplar, although each study may present slightly distinct rules.

In PVF, nouns with suffix variation (such as *friend*, *friends*, *friendly*, or other gender or number variation of the same name or concept) are to be avoided, and only the first exemplar is counted as a correct word, whilst the others are ignored [21, 111].

For SVF, subcategory labels (such as *bird*) are not counted if a specific member is also produced (for instance, *sparrow*, or any other species of bird). Gender (e.g. *hen* and *rooster* and age variants (e.g. *cat* and *kitten*) of animal names are considered the same animal, and thus only counted once [113]. Similarly, Portuguese studies, such as Cavaco et al. [21], only count gender (*vaca*, *boi*) and age (*bezerro* or *vitelo* instead of *vaca/boi*) variants once, and do not count superordinate categories if members of it have also been given (e.g. *pássaro*). Regional designations of animals are considered correct (for instance "*chico*" instead of *porco*, a regional designation in Barcelos).

2.2.2.2 Switching and Clustering Metrics

Troyer et al. [114] suggest a two-factor impact in VF test, based on qualitative evidence from previous VF studies: subjects tend to produce words in relatively isolated clusters, where words produced share semantic or phonemic characteristics. Production speed tends to be faster intra-cluster than inter-cluster, with notable "breaks" during the test when the subject switches to a newly found cluster to deplete.

Whiteside et al. [120] discuss and corroborate this multimodal nature of VF testing, supporting this model with evidence from patients with frontal or temporal lobe lesions [115], neurodegenerative diseases [54, 97, 116], and with brain imaging techniques [10, 27, 41], showing two distinct components of VF:

- Effective search over possible subcategories, and quickly switch between them, correlated with brain executive function and the frontal lobe;
- Search and exhaust word stores for a given phonemic or semantic subcategory, correlated with brain language function and the temporal lobe;

Subcategory search is a complex and dynamic task, where a subject's brain has to employ effective, but flexible search tactics, being able to rapidly shift between different subcategory candidates, while inhibiting and discarding incorrect ones, i.e., quickly and efficiently switching to different semantic or phonemic clusters. These processes are often related to the general executive function of the frontal lobe of the brain [120].

With identified semantic or phonemic clusters, the next major task is the search, selection, and production of words within these subgroups, by depleting existing word stores. This is more dependent on existing verbal memory and word storage neural circuitry, tasks related to the semantic and memory functions of the temporal lobe [120].

To quantitatively score these two factors involved in the execution of PVF and SVF tests, Troyer et al. [114] developed two new metrics that translate these switching and clustering phenomena:

Mean Cluster Size average number of words for a produced cluster, starting from the second contained word (1-word cluster has size 0, 2-word cluster has size 1, etc.), including errors and repetitions;

Number of Switches total number of transitions between clusters, considering 1-word clusters, including errors and repetitions.

Troyer et al. [114] performed PVF studies using the letters *F*, *A*, and *S*, and defined phonemic clusters as being constructed from words sharing one of the following phonemic characteristics:

- First two letters (e.g., *art* and *arm*);
- Rhyming (e.g., *sand* and *stand*);
- First and last sounds, i.e., words differing only by a vowel sound, regardless of spelling (e.g., *sat*, *sought*, *soot*, etc.);
- Homophony, i.e., words with the same pronunciation, but distinct spelling (e.g., *some* and *sum*)

Troyer et al. [114] also evaluated switching and clustering phenomena on SVF tests using the animal category. Clusters were derived *a posteriori* from the actual patterns of words generated by test subjects, being an extensive list of animal groups, depending on their habitat region (e.g., African animals, Australian animals, aquatic animals, etc.), human use (e.g., pets, beasts of burden, etc.) or zoological group (e.g., birds, insects, felines, fish, etc.), with corresponding members. Defined clusters are not mutually exclusive, with some exemplars belonging to two or more clusters (for instance *cat* is both a pet and a feline).

With the introduction of these two new metrics, VF tests can now be used to effectively measure both executive and language function of a subject. The total number of correct words, although informative, cannot distinguish performance perturbances due to language or executive function impairment [114].

Mayr [74] points out the limitations and shortcomings of the metrics suggested by Troyer et al. [114], namely that the developed switching metric is ambiguous, not wholly isolating the executive aspect of VF. Considering that a time difference exists between intra-cluster and inter-cluster word retrieval, the number of switches does, in fact, depend on the difficulty of accessing a new cluster, i.e., depends on the inter-cluster time. However, the number of switches will also decrease in the case of subjects with difficulty in producing new intra-cluster words (i.e., longer intra-cluster retrieval times), since with more spent time exhausting a cluster, less time remains in the test to access new clusters. Thus, the switching score is dependent on both intra- and inter-cluster times, and can't differentiate between the two. Mayr [74] suggests detailed timing protocols to evaluate the variability of retrieval times and further improve switching and clustering scores, as is done by Fitzgerald [35].

Raoux et al. [97] also use Troyer scores [114], and critique several limitations regarding scoring procedures, namely scoring rules failing to capture all word associations performed by subjects. For instance, during their studies, Raoux et al. [97] identified subjects that formed phonemic clusters during SVF tests by grouping animals by their initial sounds (e.g., *fox* and *ferret*), and semantic clusters during PVF tests. Additionally, cultural variability is discernible by analysing word associations performed: some patients associated animals commonly referred in popular fables and tales. Thus, a qualitative appreciation of results must be performed, to ensure all possible associations are captured using Troyer et al. [114] scoring rules and adequate corrections to capture cultural or regional associations.

Nonetheless, Troyer et al. [114] scores are still used to assess executive and language functions, for both classical studies [97, 115, 116, 117, 126], and automated procedures for VF testing, with some adaptations for automatic calculation [24, 61, 62, 85, 86, 87, 112, 121].

2.2.3 Functional Brain Imaging during Verbal Fluency Tests

Besides neuropsychological assessments, functional brain imaging is an important tool to understand the anatomy and function of different brain regions, making it possible to observe which areas are most activated during tests that specifically evaluate distinct behavioural processes (such as cognitive reasoning, executive functions, etc.). For each type of test, the consequently activated regions are thus hypothesized to control the underlying behavioural process [69].

For Verbal Fluency, several functional imaging studies exist that evaluate brain activity during the administration of these tests, namely PVF and SVF. Several studies that indicate VF has language areas involved, namely the left inferior frontal gyrus [27], left posterior temporal lobe and left inferior parietal lobe [2]. Executive function associated areas [59], such as the anterior cingulate cortex, insula, cerebellum and hippocampus [107] are also shown to be active during VF tests, suggesting VF involves both a language processing factor, and an executive function factor [120]. Generally, the frontal (executive function) and temporal (language processing) regions of the brain are activated during VF tests [88], but in a distinct manner, depending on the test type.

Glikmann-Johnston et al. [41] perform a functional brain imaging study using functional Magnetic Resonance Imaging (fMRI) techniques, to assess the different activated regions of the brain for PVF and SVF. Figure 2.1 represents the different activated areas for PVF (in blue) and SVF (in orange) in healthy subjects.

Birn et al. [13] found PVF to be more dependent on frontal regions (associated with executive functions), and SVF with temporal regions (associated with language processing and word storage). This can also be seen in Figure 2.1, with PVF generating higher levels of activation in frontal regions (such as the middle frontal gyrus) than SVF.

Functional image studies further corroborate the VF dichotomy into two distinct processes: subcategory search is associated with executive functions and the frontal lobe, whereas subcategory word retrieval with language skills of the temporal lobe. For PVF, executive processing plays a more important role, due to the more abstract representation associated with phonological search, requiring more effective search and suppression tactics [55].

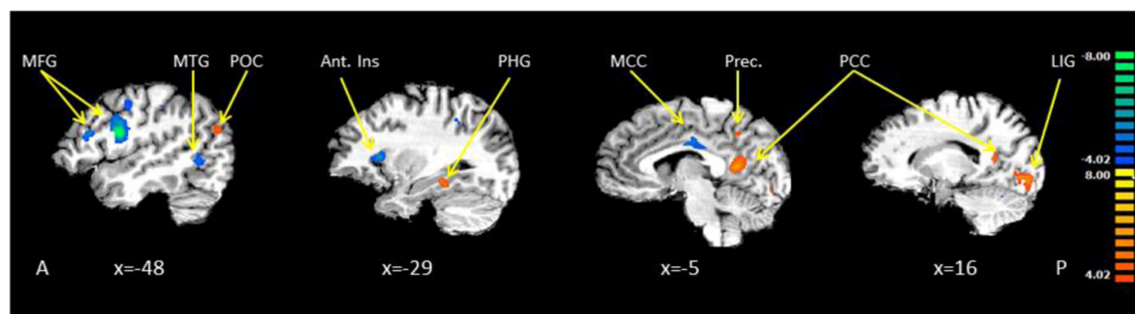


Figure 2.1: Results of semantic fluency vs. phonemic fluency in a whole brain analysis using grey matter mask with balanced blocks, from Glikmann-Johnston et al. [41]. Regions with higher activations during semantic fluency are shown in orange. Regions with higher activation during phonemic fluency are shown in blue.

Legend:

MFG: middle frontal gyrus	MTG: middle temporal gyrus	POC: parieto-occipital gyrus
Ant. Ins.: anterior insula	PHG: parahippocampal gyrus	MCC: middle cingulate cortex
Prec.: precuneus	PCC: posterior cingulate cortex	LIG: lingual gyrus
A: anterior	P: posterior	

2.2.4 Manual Administration of Verbal Fluency Tests

VF tests are commonly administered in an interview setting with a patient, in a calm and private manner, usually as part of a test battery to assess mental status. This traditional method of administration is also designated as pencil-and-paper administration since it only requires these materials to be performed. Most of the currently used neuropsychological tests are performed in pencil-and-paper fashion; the same thing happens for mental state batteries, such as the MMSE and MoCA, with each having a fillable form with the tests to be performed and respective instructions [69].

The administration of a VF test begins with the tester verbally explaining the purpose of the test to the subject, and what rules must be followed. For instance, Borkowski et al. [16] state the following to administer a PVF test:

“I shall say a letter of the alphabet, then I want you to give me as many words that begin with that letter as you can and as quickly as you can. For instance, if I say, ‘B’, you might give me, ‘bag, battle, bed’. Do not use words which begin with a capital letter like names of brands, people, or places such as ‘Boston or Bob’. Also, do not give numbers, or the same word but with a different ending, for example, ‘eat, eating.’ Do you have any questions? Begin when I say the letter. The first letter is –”

Similar instructions are provided for all types of VF test, to ensure the testee fully understands what rules must be followed before starting exemplar production.

An example of a traditional administration of a PVF, included in the MoCA battery, is ACT on Alzheimer’s [1, at 8 minutes, 55 seconds] video. Rules to follow are first stated and it is ensured the testee understands them fully. After starting the test, the tester records on paper all produced words, including repetitions and errors. A final score is calculated after the end of the test, by counting all correct words produced [69]. Note that the tester also controls elapsed test time to

know when to stop the test. Overall, one can consider the final product of a VF test to be the transcript of uttered words, which can be used to calculate features, such as the total number of correct words produced.

With the advent of new extractable features [114], VF test transcripts become more complex to analyse: each word must be associated with a given cluster (or clusters), identifying what words belong to the same group and when the testee has performed a cluster switch. Troyer et al. [114, Appendix] provides the protocol to follow to extract clustering and switching metrics, including the semantic clusters used to categorise exemplars in the animal SVF test. However, due to the increase in processing time and complexity, these scores are mostly exclusive to neuropsychological research, not being generally used in a clinical setting [87].

2.3 The Impact of Ageing and Pathology on Verbal Fluency Tests

VF tests, being good estimators of executive and language functions of a subject, present characteristic alterations when in the presence of pathologies that affect these areas of human behaviour.

A varied number of health conditions exist that affect these specific areas of cognition of the human brain and impact the overall performance of VF tests, most notably:

- Brain lesions, particularly frontal and temporal lesions;
- Mental disorders, such as Attention Deficit Hyperactivity Disorder (ADHD) and Depression;
- Neurocognitive disorders, such as Mild Cognitive Impairment (MCI), Alzheimer's, Parkinson's, and Huntington's Disease.

Table 2.2 synthesises the overall impact of age and brain pathologies on PVF and SVF, and associated metrics (mean cluster size, number of switches, and total words produced), where data is available.

Considering the main pathologies affecting the elderly population, a special emphasis is given here to the impact of neurocognitive disorders on the performance of VF tests, particularly Alzheimer's and Parkinson's Disease, discussed in Subsection 2.3.2. In addition, ageing also proves to be an important factor affecting the overall performance of VF tests, as discussed below in Subsection 2.3.1.

2.3.1 Ageing

Age increase is associated with a higher incidence of neurocognitive disorders such as Dementia, Alzheimer's Disease, Parkinson's Disease, etc. However, non-malignant cognitive difficulties are also evident, where older subjects tend to have more problems, such as forgetfulness of minor events, as first discussed by Kral [63], but with a very distinct evolution from other memory dysfunctions. This type of benign cognitive deficit is usually classified as Age-Associated Cognitive

Table 2.2: Impact of Pathologies and Ageing on Verbal Fluency.

- indicates a decreased performance in that metric, **0** indicates no alteration of performance, and **NA** indicates no data is available to assess impact.

Data gathered from:

Ageing: Cavaco et al. [21], Clark et al. [25], Nogueira et al. [82], Tombaugh et al. [111], Troyer [113];

Frontal and Temporal Lesions: Henry and Crawford [51], Troyer et al. [115];

ADHD Andreou and Trott [5], Hurks et al. [55], Tucha et al. [117];

Depression: Audenaert et al. [10];

MCI: Lopez et al. [71], Nutter-Upham et al. [83], Peter et al. [92];

Alzheimer's: Buckner [19], Henry et al. [52], Piatt et al. [94], Raoux et al. [97], Rodríguez-Aranda et al. [102], Troyer et al. [116];

Parkinson's: Piatt et al. [94], Rodrigues et al. [101], Troyer et al. [116];

Huntington's: Ho et al. [54].

	Phonemic			Semantic		
	Cluster	Switches	Total	Cluster	Switches	Total
Ageing	0	0	0	0	-	-
Frontal Lesions	0	-	-	0	-	-
Temporal Lesions	0	0	-	-	0	-
ADHD	0	-	-	0	-	0
Depression	NA	NA	-	NA	NA	-
MCI	0	-	-	0	-	-
Alzheimer's	-	0	-	-	-	-
Parkinson's	-	-	-	0	-	-
Huntington's	0	-	-	0	-	-

Decline (AACD), described as a self-reported slight impairment of cognitive functions, but not of a sufficient degree for a diagnosis of mild cognitive impairment or dementia [68].

Salthouse [105] show a decrease of memory, reasoning, spatial visualisation, and speed metrics prior to age 60, in a cross-sectional study. Longitudinal studies performed by the same authors also confirm this pattern of cognitive decline as subjects get older. Their findings show age-related cognitive decline starts after reaching early adulthood, and steadily decreases with age, although these effects may not be detectable without prior cognitive testing.

Several contributing factors for cognitive decline with ageing can be identified, such as genetic predisposition, cardiovascular disease, and biological factors (namely, inflammation processes and neurotransmitter depletion) [30]. Particularly, active and healthy lifestyles, with adequate nutrition and physical activity, are associated with reduced risk of cognitive impairment and reduced incidence of dementias [66], being an important protective factor. These lifestyles are associated with a lower risk of cardiovascular diseases, which are known to impact cognition in elderly subjects [30].

Frontostriatal damage is thought to be the main impacting factor on executive functioning in AACD. White matter damage in this anterior region of the brain, due to small infarcts and small vessel damage, is present in most AACD cases, as well as frontal grey matter atrophy, striatal

volume loss and neurotransmitter depletion (such as dopamine) [19, 59].

Normative studies show a correlation between age and performance in VF tests: the effects of age on phonemic fluency are less pronounced than on semantic fluency. Elderly subjects are associated with a decrease in performance, namely total words produced [21, 25, 82, 111] and decrease in the number of cluster switches performed [113], particularly for SVF testing. Nogueira et al. [82] results show that SVF is the most sensitive to age effects, followed by PVF.

Presented causes for AACD line up with the observed influence of ageing on VF: frontostriatal damage mainly impacts executive functioning of the brain, an essential part of VF, particularly for switching efficiently, which is shown to be affected for SVF tasks with the increase of age.

2.3.2 Neurocognitive Disorders

Neurocognitive disorders (NCD) are a broad type of mental disorders frequently observed on elderly people that progressively impact cognitive capabilities and impact daily functioning [3]. Commonly referred to as dementia, NCDs have varied etiologies associated, such as (but not limited to) the ones discussed below.

Verbal Fluency tests, as standards for the measurement of executive function and language capabilities, prove to be extremely useful for the diagnosis of NCDs. Many of these disorders impact frontal and/or temporal areas of the brain, with gradual consequences to executive and language functions. Different NCDs have clear and distinct evolution and impacted regions: VF results can be important decisive factors to perform a correct diagnosis, being able to distinguish different NCDs [126]; however, they are rarely considered alone, being part of commonly administered neuropsychology test batteries [3, 69].

2.3.2.1 Mild Cognitive Impairment

Mild Cognitive Impairment (MCI) is a diagnosis given to patients that show greater than expected cognitive impairment (considering their specific age and age-associated cognitive decline), but not sufficiently severe to warrant a major NCD diagnosis. Research of NCD has recently been shifting to MCI since it is in this phase that possible treatments would be most effective [69].

MCI can be divided into two distinct subtypes: amnesic MCI, thought to be a very early sign of Alzheimer's Disease, and nonamnesic MCI, related to other NCDs. Furthermore, MCI can be classified as a single domain or multiple domain impairing, depending on what cognitive domains and functions it affects [69].

Longitudinal studies of MCI subjects which progressively evolve to an AD diagnosis have evaluated how these subjects are distinct from healthy individuals. Lopez et al. [71] observe on MCI patients an initial memory impairment, but remaining cognitive functions are comparable to healthy controls. Nutter-Upham et al. [83], Peter et al. [92] point out a subtle reduction of performance of both PVF and SVF on MCI patients, due to slightly impaired executive processes and switching difficulty. These difficulties are further exacerbated with the diagnosis of AD when semantic stores become progressively depleted.

From this studies, it is possible to conclude that the application of VF tests can prove as a decisive factor for the early diagnosis of NCD: first signs of MCI can be assessed with the application of VF tests. Since MCI is many times a prodromal condition to severe future disorders (such as Alzheimer's), early detection of these fluency deficits allows for earlier treatment, mitigating future effects of future NCDs [69].

2.3.2.2 Alzheimer's Disease

Alzheimer's Disease (AD) is a prevalent disease that is a common cause of NCD, in this case commonly referred to as Dementia of the Alzheimer type (DAT), classified as a cortical dementia. AD is characterised by the progressive deposition of insoluble extracellular amyloid and intraneuronal changes, with progressive consequences on cognitive and executive functioning, mainly affecting the grey matter in the cerebral cortex [17]. Cognitive functions that usually deteriorate first are episodic memory, mental tracking, and verbal fluency, gradually leading to mental deterioration and complete loss of independence [69].

DAT and its effects on VF are commonly studied, whether comparing to control subjects [52, 94], age-associated cognitive decline [19, 102], or to other dementias [116].

Troyer et al. [116] perform a comparative study of DAT and Parkinson's Disease with associated dementia (DPD) and without (NPD), for both PVF (FAS test) and SVF (animals) tests, utilizing their designed switching and clustering metrics, evaluating executive (frontal lobe) and verbal memory (temporal lobe) processes. DAT patients produce a reduced number of words for both fluency tasks compared to the control group, with lower mean cluster sizes. Switching only appears to be affected for SVF, with fewer switches performed for this particular fluency test.

Considering PVF, DAT subjects having a diminished mean cluster size suggests an impairment in lexical-phonemic word stores. The reduced cluster size produced by DAT patients for SVF is consistent with an impoverished semantic memory and word store: these patients have more difficulty generating exemplars from each specific subcategory considered. This is also confirmed when DAT subjects are asked to generate words from a particular subcategory (i.e., *pets*, or *jungle animals*), where they generate, with difficulty, a reduced number of exemplars [116]

Switching performance is only impaired on SVF, having no significant difference for PVF. This indicates that general executive functions, namely subcategory search processes and shifting strategies, are still intact for DAT patients. However, for SVF, semantic memory deficits result in an increased difficulty distinguishing different subcategories, and thus difficulty switching to them: semantic associations can be so severely impaired in DAT that this impairment shows on both clustering and switching metrics [115], demonstrating a critical flaw in these scores, as discussed by Mayr [74].

Raoux et al. [97] perform a longitudinal study of several elderly subjects, including AD patients, over the course of 9 years, evaluating semantic fluency with the animal naming test. The progression of AD and its impact on VF can be studied since the performance of SVF was recorded for several years prior to diagnosis.

Interestingly, in Raoux et al. [97], future AD cases showed decreased switching on SVF tests, with relatively intact cluster sizes. Furthermore, exemplar generation given a subcategorical cue is not affected in the AD prodromal phase (before diagnosis). This indicates the early course of AD is marked by impairment of executive processes, leading to reduced switching in VF tests. After diagnosis, cluster size does reduce, indicating a reduction of exploration of semantic fields, similar to Troyer et al. [116].

2.3.2.3 Parkinson's Disease

Parkinson's Disease (PD) is a progressive NCD, caused by the depletion of dopamine on the basal ganglia [65], with notable motor disorders, starting with resting tremors, muscle rigidity, and generalised slow movement. This disorder can further affect mental capabilities, imposing cognitive impairments in several areas, including VF [69, 98]. Subcortical dementias caused by PD are commonly referred to as dementia with PD (DPD); early stage PD may not present dementia symptoms, referred to as non-demented PD (NPD).

Basal ganglia dysfunction may not be the cause of impaired cognition present in PD, but rather the disrupted frontal lobe connections due to dopamine depletion. PD patients typically lack spontaneity, show slower response times, and delayed mental shifting [98], characteristic symptoms of impaired executive functions.

Troyer et al. [116] evaluate the influence of PD on verbal fluency, both on PVF (FAS test) and SVF (animal category), utilising switching and clustering metrics to evaluate executive and language functions of these patients, respectively. Non-demented PD patients did not show VF impairment when compared to the control group. DPD subjects have PVF and SVF impaired, with a decreased number of correct words generated. Decreased switching was observed on both tasks. Mean cluster size is reduced for PVF, and normal cluster size is registered for SVF.

The maintained cluster size for SVF suggests semantic stores are not affected by PD, with patients being able to successfully generate words belonging to a specific subcategory. Moreover, when prompted with a given subcategory, PD subjects do not show impairment, unlike patients of Alzheimer's Disease, as discussed before. For phonemic fluency, cluster size is reduced, possibly due to the less salient subcategories intrinsic to this task: clustering on semantic fluency is a more automatic approach to the resolution of the problem, whereas clustering on phonemic fluency may be a deliberate strategy employed. Thus, generating words following a phonemic clustering strategy may not be evident for DPD patients, which show impairments in strategic problem-solving [114, 116].

Switching performance is shown to be significantly reduced for both PVF and SVF, confirming DPD patients have greater difficulty in search and retrieval processes, associated with a general impairment in executive functions [98, 116], further corroborated by the aforementioned frontal lobe connection disruption due to dopamine depletion [98].

2.4 Speech-to-Text Systems

Automatic Speech Recognition (ASR), or Speech-to-Text, is nowadays a very sought after field by major corporations such as Microsoft, Google, IBM, and Apple. The development of new computerised personal assistants, such as *Google Assistant* and Apple's *Siri*, which rely on Speech Recognition to translate what is spoken to them to text, has made this area more evident over the years. Recent advancements of Deep Learning techniques, such as Deep Neural Networks (DNN) have also boosted the performance of ASR algorithms to very reliable standards [53].

The development of an ASR system is a complex endeavour which requires thousands of annotated training data and immense computational power to provide a usable translator [53]. As such, it is beyond the scope of this dissertation. Moreover, there are several options of APIs (application programming interfaces) available in the market, allowing the use of highly trained and complex models such as the ones described above to translate acoustic data into text.

2.4.1 Performance of State-of-the-Art Speech-to-Text Systems

A frequently used metric to assess the performance of Speech-to-Text systems is the Word Error Rate, or WER. WER is defined on Equation 2.1, where S is the number of substitutions performed, D the number of deletions, I the number of insertions, N the number of spoken words in the original reference, and C is the number of correctly guessed words.

$$WER = \frac{S+D+I}{N} = \frac{S+D+I}{S+D+C} \quad (2.1)$$

Xiong et al. [124] describe the currently used Speech-to-Text system for Microsoft's conversational speaker recognition system, based on Convolutional Neural Networks (CNN), with Bidirectional Long Short-Term Memory cells (BLSTM) for acoustic modelling, achieving a 5.1% WER on the Switchboard 2000 dataset [42].

Chiu et al. [23] details the currently used Speech-to-Text system for Google Voice, which is composed of three main modules: the listener, attender, and speller modules, in what is known as a LAS system. The listener module transforms the input acoustic data to higher dimensional data; the attender module selects which features provided by the listener module can be used to determine the next output symbol/letter; and the speller module takes into account the current output of the attender module, but also its previous outputs and previously predicted symbols, to output the most probable current symbol, thus outputting the most probable word spoken. The LAS system can achieve a 5.6% WER on a voice search dataset.

The direct comparison of these two methods is not easy or direct, bearing in mind each system was trained and tested on distinct datasets. However, 5% WER is an excellent overall metric for accurate speech recognition, rivalling the performance of human transcribers [124].

Table 2.3: Overview of Automatic Speech Recognition APIs

	PT?	Python?	Open-source?	Online?	Free?	Cost
Google Cloud Speech-to-Text [45]	✓	✓	×	✓	×	60min free, \$0.006/15sec
Microsoft Azure Speech to Text API [76]	✓	✓	×	✓	×	5h free, \$1/h
IBM Watson Speech to Text [56]	*	✓	×	✓	×	100min free, \$0.02/min
CMU Sphinx [26]	×	C	✓	×	✓	-
Kaldi ASR [96]	×	C++	✓	×	✓	-

*: Only Brazilian Portuguese

2.4.2 Speech Recognition Online Services

There are several online services available for Speech-to-Text, with corresponding APIs that can be used in the development of this dissertation. Table 2.3 provides an overview of available APIs, along with their costs and features.

Since VF tests will be administered in Portuguese and answered by elderly members of the Portuguese population, it is essential that these APIs have Portuguese language support, preferentially European Portuguese. Besides this, it is advisable that they should have the least amount of costs associated, which can vary from API to API. Finally, since the main programming tasks of this project will be performed in Python 3, it is desirable to use APIs for Python 3.

The selected API can be used in the future developed programming application to perform ASR of administered VF tests, not being necessary to implement an interface for this system/service from the ground up.

2.5 Natural Language Processing

Natural Language Processing, or NLP, is a sub-field of computer science and linguistics that deals with the processing of naturally written language and the extraction of information from bodies of text. NLP techniques started by being rule based, with individualised conditions for each particular case present in a given language, creating a set of heuristics for processing text. More recently, NLP bases itself on language models that rely on statistical inference over large text corpora. With the advent of Deep Learning NLP models, the tools of this area are being ever more robust and powerful to deal with the immense variability of written text, languages, and contexts [12].

NLP has a plethora of possible uses when dealing with bodies of text, being also known as Text Mining for these applications. Document classification is a major area of interest, utilising commonly present words in documents, articles, emails, etc., to identify what type of document

it is; a very prominent example is the use of email classification for the detection and filtering of spam messages. Sentiment analysis is also an interesting application, being able to identify positive or negative sentiment in text; with this, companies can extract overall client sentiment in relation to a given product in the market, for instance [12].

Written text is a very diverse and flexible media to communicate ideas and information. Text is commonly digitally encoded as having a distinct sequence of bits for each possible symbol. Different encodings have been developed over the years, from the 7 bits of ASCII [22] to the 1 to 4 bytes used in UTF-8 encoding for Unicode characters [125]. However, the underlying representation of text and words is not informative, being only a table of correspondence between a bit representation and a given symbol.

As such, the main way of processing text data is to directly apply transformations to it, for instance segmenting whole bodies of text into its constituent phrases, and phrases into its constituent words. Words can also be processed to obtain its root, without prefixes or suffixes (such as gender or number variation).

The main tool available to perform NLP in Python is the Natural Language Toolkit, or NLTK [81], containing a large number of text processing functions, such as tokenization and stemming, including for the Portuguese Language.

For this task of automatic VF test analysis, NLP mainly contemplates normalisation of inflection of gender, size, and number, and identifying semantic and phonemic relationships between words.

2.5.1 Word Stemming and Lemmatisation

Words can be said or written in several different inflections. Considering names, the words that describe them can have variations due to gender, size, and number. For instance, the words *boi*, *vaca*, *vaquinhas*, *vitelo* all refer to the same animal, but also transmit information regarding gender, number, size, and even age.

For NLP it can be relevant to normalise words to their underlying concepts, to better identify relations between them and to ease analysis. On the other hand, the use of inflected forms of certain words (e.g. diminutives) can be associated with the general sentiment of a certain phrase; the determination of inflected forms can be useful for sentiment analysis, among other NLP tasks [12].

Inflection normalisation can be achieved in two ways: word stemming and word lemmatisation.

2.5.1.1 Word Stemming

Word stemming aims to extract the root forms (or stems) of inflected or derived words. The identification of words stems allows the identification of words that are morphological related, but with different suffixation. The words stems extracted by stemming algorithms/stemmers may not always be semantically intelligible, but can ease feature extraction of NLP models [12].

For the Portuguese language, the *Removedor de Sufixos da Língua Portuguesa* (RSLP) Stemmer [84] is a particular example of a stemming algorithm developed for the Portuguese language, and implemented in the NLTK Project [81] Python package.

Considering the previous example words, the application of the RSLP Stemmer originates the stems *boi*, *vac-*, *vaqu-*, and *vitel-*. The RSLP Stemmer applies a set of predetermined rules for suffix removal, which results in the identification of word stems based on word morphology. In this example, all words are inflections of the same name, but different stems are extracted for each.

Nonetheless, these algorithms are extremely important for the normalisation of gender, number, and suffixation.

2.5.1.2 Word Lemmatisation

Word lemmatisation aims to group all inflected forms of a word/concept onto a single lemma, i.e. the dictionary or canonical form of a word [12]. As such, lemmatisation output is always the underlying concept of inflected words, being an intelligible word with meaning, unlike stemming.

For the case of the Portuguese language, no effective lemmatisation algorithms are commonly available, for instance in the NLTK Project [81] package, although list based alternatives are available, such as michmech [75] repository of lemmatisation lists.

Considering the previous example words, the application of a lemmatiser converts all given words to their lemma *boi*, the underlying animal species that all word examples refer to.

2.5.2 Word Ontologies and WordNet

A major component needed for NLP algorithms and models to be as effective as possible is a way of representing all possible words in a given language, and how these words relate to each other, as a lexical database. To properly understand how words in a given document are related allows for complex text analysis tasks to be performed, such as part-of-speech tagging.

Ontologies are a representation of words of a given field, with their associated definition and possible relations with other words or concepts of said field. Word ontologies aim to organise words of a given language according to their definition and relation with other concepts. These ontologies are often visualised as connected graphs or networks, with words/concepts as vertices of said graph, with several different types of connections between them, illustrating the different possible relations between words [34].

A commonly used word ontology for NLP, available in the NLTK Project [81] Python package, is WordNet [77, 78]. Although not initially designed as a word ontology of English, but rather as a lexical database of English based on Psycholinguistics concepts, it is now commonly used for this effect in many NLP applications [12, 77].

WordNet allows for the identification and quantification of semantic and lexical relations between words, being a structured dataset of English words organised in a hierarchical fashion. WordNet divides words into four large groups: nouns, verbs, adjectives, and adverbs, each with distinguishing hierarchical organisation and relations [77]. The utility of using WordNet in this

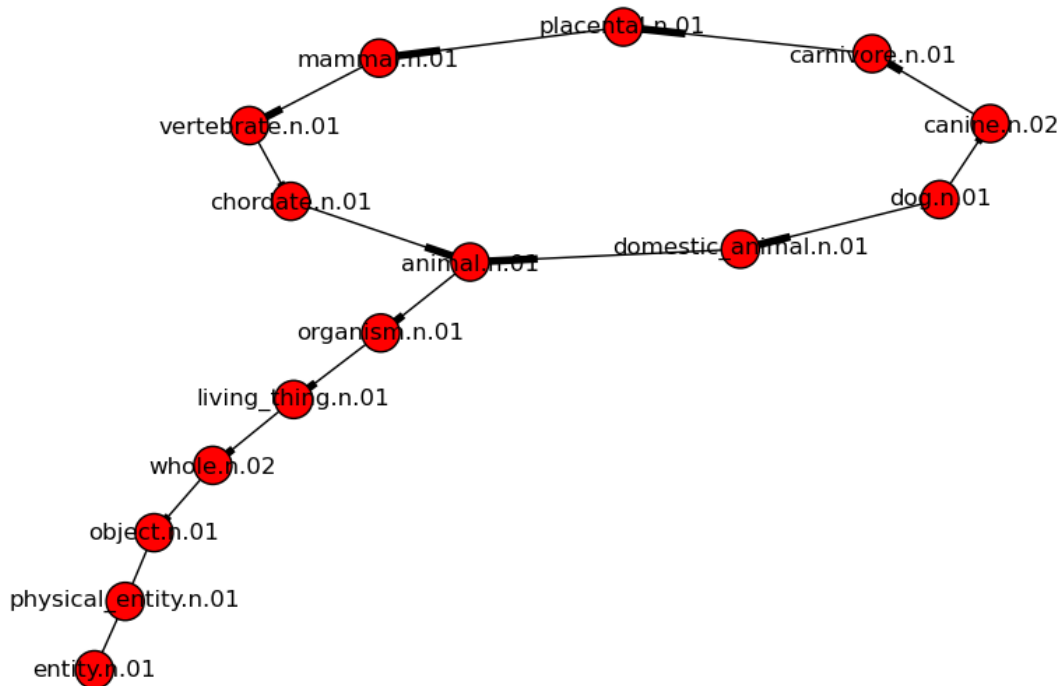


Figure 2.2: Graph representation of the synset *dog.n.01* and its hypernym relations

dissertation is based mostly on the hierarchical structure of nouns, their definitions, and relations it provides.

Several similar lexical databases exist for other languages, also commonly referred as wordnets. Portuguese wordnets exist, such as de Paiva and Rademaker [29], Gonalo Oliveira and Gomes [44], Simoes and Guinovart [108]. For this dissertation, de Paiva and Rademaker [29] is used, which is an open-access wordnet accessible through the Multilingual WordNet [15] interface available on the NLTK Project [81] package.

2.5.2.1 WordNet Structure and Relations

As described by Miller et al. [77], Miller [78], a concept in WordNet is represented as a *synset*, or synonym set. Each synset comprises of a list of word lemmas that belong to the same lexical category and are roughly synonymous. Words can appear in several synsets if they are polysemous, i.e., if they have more than one possible meaning. Each synset also has a definition associated, similar to a dictionary.

Since nouns in WordNet are organised in a lexical inheritance system [77], all are connected to a base synset of *entity*, through several "is a" connections to more and more general groups and concepts. Figure 2.2 shows this connection path for the synset *dog*.

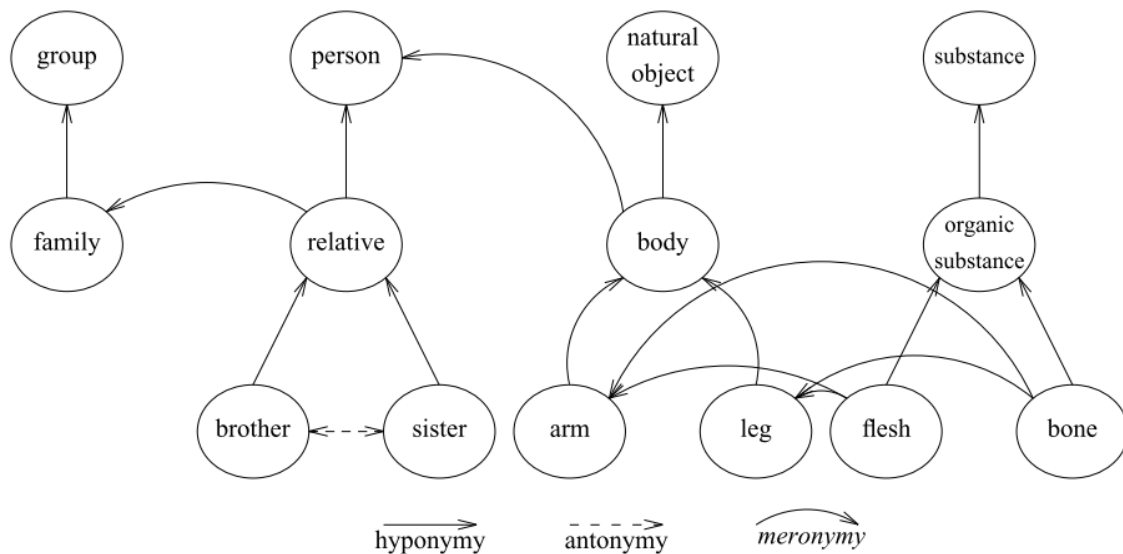


Figure 2.3: Different types of relations between noun synsets in WordNet, from Miller et al. [77].

Four main relations are present in WordNet:

Synonymy where two words have identical meanings and share the same synset, such as *cão* and *cachorro*;

Antonymy where two words have opposite meanings, such as *hot* and *cold*;

Hyponymy/Hypernymy where synsets have a unidirectional "is a type of" relation between them, i.e., one is included in the semantic field of the other. The hyponym is the word corresponding to a specific instance of the hypernym word. For instance, *pássaro* is a hypernym of *canário*, and *canário* a hyponym of *pássaro*;

Holonymy/Meronymy where synsets have a unidirectional "is a part of" relation between them. The meronym is the word corresponding to the part, and holonym to the whole. WordNet further specifies what type of holonymy/meronymy exists between synsets, depending on the relation between the synsets: part, member, or substance holonymy/meronymy. Some examples are:

Part *pata* is a component part of *cão*;

Member *cão* is a member of *matilha*;

Substance *carbono* is the stuff that *carvão* is made from.

Figure 2.3 shows some examples of these different relations between some English synsets.

2.5.3 Word Embeddings

Word embeddings are a major area of research in NLP, being the set of techniques to model language and extract features from text corpora by mapping words to vectors of real numbers with high dimensions.

This process of embedding words to highly dimensional vectors is performed in such a way so that semantically similar words, i.e. words that appear in the same context, have similar resulting vectors. Furthermore, the distance between pairs of word embeddings that share a common transformation (for instance *king-queen* and *man-woman*) is kept similar for these. Due to this, the similarity between corresponding word embeddings can give a sense of the level of relationship between words [43].

Several approaches for word embeddings can be considered, divided as two groups: classical methods (that rely on document co-occurrence matrices and related transformations), and Deep Learning methods (that rely on current Deep Learning techniques on large text corpora to produce word embeddings).

2.5.3.1 Word Embeddings from Document Co-Occurrence Matrices

Word embeddings can be extracted from document co-occurrence matrices of text corpora, which quantify the frequency of terms in each document present in a selected corpus. By applying statistical transformations to this matrix, such as calculating term frequency - inverse document frequency (*tf-idf*) [73] values for each term and document, and Singular Value Decomposition (SVD), terms can be embedded into a vector space that can transmit semantic similarity between them. Two major methods for this are Latent Semantic Analysis and Explicit Semantic Analysis.

Latent Semantic Analysis Latent Semantic Analysis, or LSA, is a NLP technique used to determine the relationship between pairs of words, by constructing a semantic space where terms that are commonly associated are closer together [31].

LSA is performed by calculating the *tf-idf* co-occurrence matrix of a text corpus [73] and applying SVD to perform a low-rank approximation of the document-term matrix [31], embedding each word into a linearly independent vector. The similarity between terms can be determined using the cosine similarity metric. Terms which appear frequently together will have a similarity score close to 1. This similarity score can also be utilised to calculate a similarity measure between different documents.

LSA can be performed with a large text corpus, containing thousands of documents, thus obtaining a general similarity score between words: terms that have a high similarity are terms that tend to be simultaneously present in a large number of documents, therefore they are probably semantically related [64].

Explicit Semantic Analysis Explicit Semantic Analysis, or ESA, is an adaptation of LSA introduced by Gabrilovich and Markovitch [40], which makes use of predetermined natural concepts defined by Wikipedia articles. ESA is *explicit* since it directly maps concepts grounded in human cognition to a feature space, instead of latent and indirect concepts provided by LSA, which only considers statistical co-occurrence of terms.

The algorithm of ESA developed by Gabrilovich and Markovitch [40] converts each Wikipedia concept (i.e., each Wikipedia article) to a vector attribute composed of the *tf-idf* values of the

words that occur in said article. An inverted index is created, which maps each word to the concepts/articles in which it appears with the corresponding *tf-idf* score.

As such, an embedding that maps words to concepts is generated, with each word being associated with a N -dimensional vector representation, where N is the total number of concepts (i.e., the total number of analysed Wikipedia articles). Words can be compared for semantic relatedness with the cosine similarity metric.

ESA has achieved better results than LSA (trained with WordNet) for the computation of semantic relations between words [40]. This method proves to be more robust to determine word semantic similarity because it relies on information collected and organised by humans, whereas LSA only utilises statistical co-occurrence of words to determine their possible connection.

2.5.3.2 Word Embeddings from Deep Learning Models

The application of Deep Learning algorithms and very large text corpora for the production of word embedding models is an exciting area of research in NLP, with several alternatives available. Some examples of models of word embeddings utilising Deep Learning techniques are *word2vec* [43], *GloVe* [91], *fastText* [14], *ELMo* [93], and *BERT* [32].

Instead of depending on statistical analysis of text corpora to determine a high-dimensional vector for each word present, these Deep Learning methods consider an embeddings table, where each word has a corresponding embedding. The values of this embeddings table are parameters that are optimised as part of a Deep Learning model that performs text classification, for instance. By performing model optimisation using gradient descent algorithms, these parameters are optimised to reach peak performance on text classification tasks.

This optimised embeddings table can then be used as a source of word embeddings for other NLP tasks, such as sentiment analysis, semantic similarity determination, among others.

2.6 Automatic Administration of Verbal Fluency Tests

Recent developments in ASR technologies and NLP techniques for semantic relatedness measures have been used to tackle the impracticability of extracting clustering and switching scores from VF transcripts, as well as mitigate some limitations of Troyer et al. [114] proposed scores, as discussed by Mayr [74].

Table 2.4 presents a summary of the reviewed state of the art studies, chronologically detailing the progression of research in this area, from the original idea of word production as a cognitive function predictor [16], to fully automated algorithms that can predict risk of dementia over telephone conversations [61].

Troyer et al. [114] proposed metrics allow for a deeper insight on the executive and language functions of subjects, allowing to distinguish between different NCDs [116], as discussed previously. However, the new proposed metrics significantly increase the level of processing of VF transcripts, which leads to increased processing time. The extraction of these values cannot be done *in situ* since the transcript needs to be analysed so thoroughly. With these scores, VF tests

can no longer be a simple pencil-and-paper test, able to provide an immediate result, and are mainly applied in neuropsychological research, and not in a clinical setting. The determination of phonemic and semantic similarity between words in a VF test is the step that requires the most processing time.

Fortunately, with the advent of new NLP techniques that can calculate semantic similarity between words in an automatic fashion [40, 43, 89], these tasks can be automatised. These techniques are used in several works related to VF tests, such as König et al. [61], Pakhomov and Hemmy [86], Pakhomov et al. [87], Woods et al. [121], to create automatic analysis tools of manual transcripts of these tests, based on the semantic similarity of words that can be calculated with these methods.

Computer friendly dictionaries of phonemic representations of words, like the one used by *CMUDict* [119], have also been developed, that can be used to assess phonemic similarity. For instance, Ryan et al. [104] utilises this dictionary to define phonemic similarity metrics that can determine phonemic clusters of PVF tests, analogous to those manually extracted by Troyer et al. [114].

König et al. [62] offers an analysis purely based on word production timings, obtained through audio processing of recorded VF tests. This analysis is completely language-independent, and can be used to obtain insightful information on the temporal patterns of how words are produced by subjects [121].

Finally, an impressive boost of applicability of these techniques comes with the introduction of powerful, Deep Learning based, online, and accessible ASR systems [45, 56, 76]. With these it is no longer necessary to perform manual transcription of tests, allowing for fully automatic systems of VF test analysis, such as the one described by König et al. [61]. Fully automatic systems can be applied in large scales, and are able to extract meaningful features from short verbal tests, which can be used as screening tests for NCDs, such as dementia. As described by König et al. [61], telephone-based dementia screenings are now a real possibility of application of fully automatic VF test analysis pipelines.

Table 2.4: Summary of reviewed state of the art studies

Study	Language	VF Test	ASR?	NLP	Main Ideas
Borkowski et al. [16]	English	PVF	x	x	Total number of correct words produced is a predictor of brain damage
Troyer et al. [114]	English	PVF, SVF	x	x	Switching relates to executive functioning; Clustering relates to language memory These two metrics can be used to distinguish NCD
Pakhomov et al. [87]	English	SVF	x	LSA + WordNet	Automated semantic similarity scores based on LSA are akin to human defined similarity
Ryan et al. [104]	English	PVF	x	CMUDict	Automated phonetic similarity scores based on phonemic transcriptions of words, based on a phonemic dictionary
Pakhomov and Hemmy [86]	English	SVF	x	LSA + Wikipedia	Automated semantic similarity scores can be adapted to compute Troyer switching and clustering metrics
König et al. [62]	French	SVF	x	x	Language-independent vocal features, based on timing of word production
Woods et al. [121]	English	PVF, SVF	x	ESA	ESA is more robust than LSA, relies on concepts rather than word co-occurrence
König et al. [61]	French	SVF	Google	word2vec	SVM classification to distinguish healthy, MCI, and dementia subjects; Automatic transcription of tests with ASR; Automatic semantic-related feature extraction

Chapter 3

Implemented Pipeline for the Automatic Analysis of Verbal Fluency Tests

As discussed previously, VF tests are a powerful and quick-to-use tool for the assessment of executive and language functions of subjects, which can be of particular importance when dealing with elderly subjects and the threat of dementias.

Being tests that rely solely on the speech production of the subject at hand, audio recording VF tests is a natural evolutionary step for the analysis of these procedures, with the ever more increasing ubiquitousness of everyday devices that allow for audio recording, such as *smartphones*.

With the advent of more powerful Speech-to-Text services, such as Google Cloud [45] Speech-to-Text, automatic transcription of recorded VF tests becomes a real possibility, which can facilitate the analysis of these tests by professionals. Moreover, new NLP techniques also allow for the automatic analysis of the extracted text, allowing for a much more quick and efficient administration of these tests, where the health professional administering is not required to quickly write down all produced words, but rather just record what the subject is saying.

Being a main objective of this dissertation work, this chapter details in-depth the implemented pipeline and algorithms used for the automation of VF tests administered to Portuguese subjects, by using as input audio recorded from these tests. Several fully-automatic pipelines already exist in literature, as detailed in Section 2.6, but the one here suggested is the first of its kind for European Portuguese.

A schematic of the proposed pipeline is shown in Figure 3.1, delineating the main blocks required for the automatic extraction of clinically relevant features from VF tests. These features can then be used for supervised and unsupervised Machine Learning methods to better characterise populations studied with these tests.

The pipeline for automatic analysis of VF tests proposed here considers both PVF and SVF tests as possible inputs. In particular, the steps detailed below consider PVF tests to be performed with words starting with the letter *P*, a commonly used letter in Portuguese tests and studies [21], and the letter used in the PVF present in the Portuguese version of the MoCA [80, 38] (see Appendix D for the Portuguese version of this neuropsychological test battery). SVF tests are applied

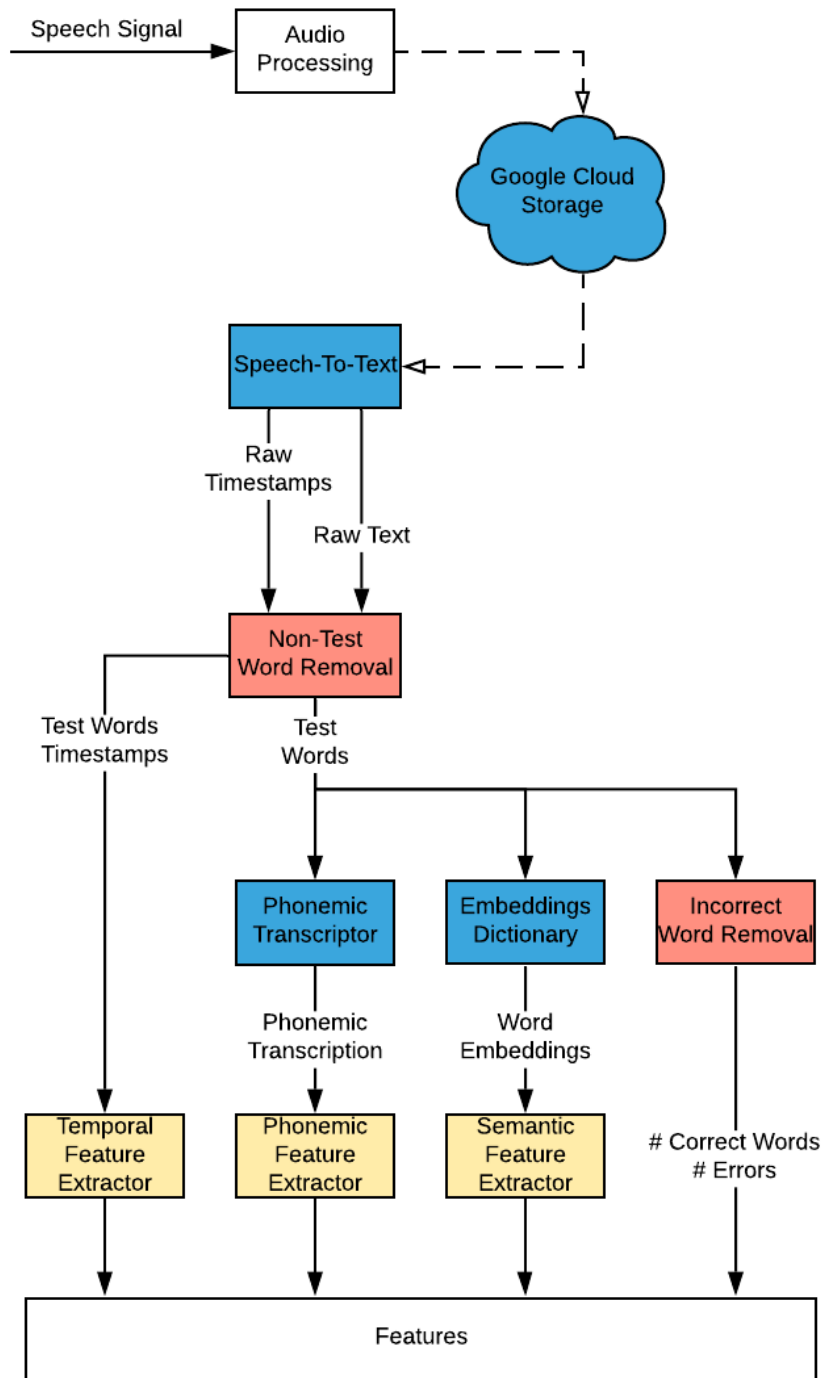


Figure 3.1: Schema of the implemented pipeline for automatic VF test analysis, from speech signal data. Blocks in blue represent pipeline steps that depend on external resources, such as online services. Blocks in yellow are feature extraction steps. Blocks in red represent word filtering steps performed with the aid of a Portuguese WordNet [29].

with the animal category, the most commonly used category for these tests, including in Portuguese studies [21] and clinical applications. Nonetheless, the pipeline can be adjusted to analyse PVF tests with different letters and SVF tests with different semantic categories.

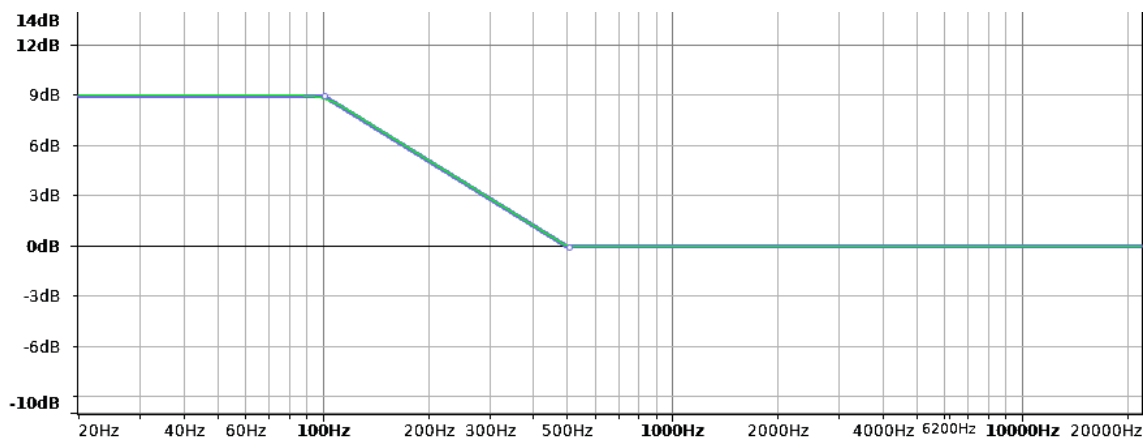


Figure 3.2: Frequency Profile of the employed Equalisation (Bass Boost) step for audio processing

Several of the described blocks rely on external information, such as online services and websites, or freely available downloaded data. Below each of these blocks is described in-depth to fully understand how these steps were implemented in Python 3.6.

3.1 Audio Processing

The collected recordings are first cut to remove the initial and final portions of audio, where the examiner can be heard saying for the subject to start the test, to indicate the test's conclusion, or to clarify any rules that may have been misunderstood.

The collected audio signals are highly dependent on the subject performing the test, as well as the environment in which it is performed. To normalise and enhance speech levels on the gathered audio signals, all are subjected to a 5-step pipeline, developed as a macro for Audacity, a free, open-source software program to handle and edit audio [9]:

1. Normalisation (default parameters);
2. Equalisation (Bass Boost) (Figure 3.2 shows the frequency profile used);
3. Dynamic Range Compression (default parameters);
4. Normalisation (default parameters);
5. Export to *.wav*, at $44.1kHz$ and 16-bit depth.

The normalisation steps ensure audio amplitude is equippable for all collected audio files, thus ensuring subjects with different voice volumes will have sound-waves with similar amplitudes.

Since normalisation can also amplify background noise, equalisation and compression are applied to try to mitigate this effect. Equalisation using the Bass Boost frequency profile will enhance the lower frequencies, up to $500Hz$, as seen on Figure 3.2, which are commonly associated with speech, thus enhancing these components over background noise at higher frequencies. Dynamic Range Compression reduces the dynamic range of the sound signal, and reduces the volume

of frequencies according to their energy, thus allowing the reduction of background noise when compared to speech.

The audio signals are again normalised to ensure no wave peaks that may have been produced with the application of the previous effects remain. Finally, the now processed audio signal is exported and saved as a *.wav* file, which is an uncompressed audio file format, with sampling frequency and bit depth similar to CD quality, which is adequate for use with Speech-to-Text services and maintains file sizes around 5 megabytes for 1 minute recordings.

3.1.1 Google Cloud Storage

The *.wav* audio files are transferred to a Google Cloud [46] Storage bucket associated with the Speech-to-Text API of this project, enabling communication between the two services. Sound files longer than 1 minute are required by the Google Cloud Speech-to-Text API to be provided via Google Cloud Storage, thus all collected audio files are stored in the latter. The use of cloud services to store audio data also facilitates usability and portability of the developed tool, since sound files are not required to be stored in a local machine to be transcribed.

3.2 Speech-to-Text

Speech-to-Text is the crucial initial step that allows for the conversion of audio signals to a list of uttered words.

This block relies on the Google Cloud [45] Speech-to-Text online service. From the speech recognition APIs analysed in Section 2.4, Google's solution proved to be the best one, since it is compatible with Portuguese, has a Python API, and is one of the most commonly used APIs for the automation of VF tests [61].

These files are read from the Google Cloud Storage bucket associated with the project, and transmitted to the Google Cloud Speech-to-Text API client, which then converts the audio signal to a list of uttered words. The client is setup to detect the European Portuguese language, with the default configurations for speech detection and transcription, and with the option to return word offsets.

Two speech models are available for use for transcription: the *default* speech model, applicable for most common situations, and the *command-and-search* speech model, used for voice commands or voice search. Both these models are used for comparison of their performances.

After a waiting period of approximately 5 seconds, the client returns a response containing all uttered words, with confidence values, and associated start and end times.

When creating a new Google Cloud account, a \$300, 1-year grace period is given, alongside the 60 free minutes per month. This grace quota allows the use of this API to carry out this dissertation, but since each request spends approximately 1 minute of available time, it is vital that API requests are kept to a bare minimum, to ensure minimum expenses using it. To do this, each resulting Speech-to-Text transcription is saved to a *.csv* file, with words transcribed and their corresponding timestamps.

The Speech-to-Text API can and will transcribe hesitations, vocal pauses, and filler words, such as "ahhh", "hmm", or "ok". These words are filtered from the transcription result, since they do not provide any meaningful information of the VF test performance. Each word is compared to a list of common Portuguese stopwords, present in NLTK Project [81] package: if it is on the list, it is removed from the transcription. One careful consideration is taken here: if the word starts with the letter *P* it is not removed, since for PVF tests stopwords starting with *P* are valid words and must be considered in the final transcription.

3.3 Non-Test Word Removal

After speech transcription, the resulting text provided by the Google Cloud Speech-to-Text API is further analysed to automatically extract relevant features using NLP techniques.

However, since the application of VF tests is not rigid, the speech present in the recordings of these tests does not always correspond to actual attempts of exemplar production, but can include subject remarks, non-test monologue (e.g. "*Não me lembro de mais nenhuma palavra...*"), or rules clarifications by the VF test administrator. These phrases can and will appear in the automatic transcript provided by the Speech-to-Text service, making it necessary for these to be filtered out to only keep words relevant to the test at hand.

To achieve this, simple decisions rules were implemented to assess if a word in a given test transcription is a test exemplar or not, based on the fundamental rules of each test type. The overall implemented pipeline for word removal is shown in Figure 3.3, with non-test word removal being the upper section. Since PVF and SVF have a distinct set of rules, non-test word removal processes depend on the considered test type.

The filtered transcriptions are then further processed for feature extraction. Following Troyer et al. [114], errors and repetitions are considered for feature extraction, since these mistakes can provide insight to the way the subject's brain processes the VF test task; therefore the test word transcript is used for these tasks. To determine the total number of correct words produced, the test word transcript is further filtered to remove errors, as described in Section 3.6.

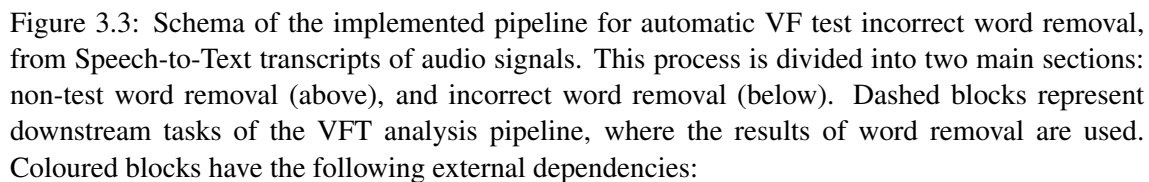
3.3.1 Non-Test Word Removal for Phonemic Tests

Since the main rule of PVF tests is the production of words that start with a given letter, words that do not start with said letter are filtered out of the automatic transcript of the test.

In the case of the collected PVF tests for this work, the letter *P* was used, thus, words whose first character is not *P* are discarded from the automatic test transcript.

3.3.2 Non-Test Word Removal for Semantic Tests

The considered category for SVF tests was the animal category, since it is the most commonly used for this VF test type. As such, to remove non-test words implies the classification of each word in the transcript as being an animal or not an animal.



(available through NLTK Project [81]);

Blue - [Infopédia Online Dictionary](#) [95];

Pink - RSLP Word Stemmer [84] (available through NLTK Project [81]).

To do this, the Multilingual WordNet is used [15], particularly the Portuguese WordNet available within [29]. The NLTK Project [81] package provides an interface for this WordNet for

Python.

This WordNet has entries for several animals, accompanied by their corresponding lexical relations with other concepts. For this particular case, hypernymity relations are very useful, since the concept of *animal* is a hypernym of all animal instances.

However, since the Portuguese WordNet used does not always contain word inflection variations (such as gender and number), attempting to search for these can result in no results found. For instance, searching for *ornitorrincos* returns no synsets, although one exists for *ornitorrinco*. Consequently, inflection normalisation is necessary before attempting animal classification.

3.3.2.1 Inflection Normalisation

Since word stemming does not produce intelligible results that may be used for WordNet synset search, but only split roots of words, stemming is not an option for inflection normalisation.

Instead, a lemmatisation algorithm is necessary to normalise gender, number, and size variation of produced words. For the particular case of the Portuguese language, no effective lemmatisers are available in the NLTK Project [81] package.

As such, to perform lemmatisation, michmech [75] lemmatisation list repository is used, namely the Portuguese lemmatisation list, with 850,264 lemma-token pairs. This file is read and stored in a Python dictionary for easy access, associating each token with its corresponding root lemma.

Most of the pairs present in the Portuguese lemmatisation list are associated with verbs and different verbal conjugations, but animal names and corresponding number, gender, and size inflections can also be found, allowing for the normalisation of all transcript words before non-animal word removal. For instance, the words *boi*, *vaca*, *vaquinha*, *vaquinhas*, are all converted to *boi*, thus normalising gender, number, and even size denominations.

3.3.2.2 Animal Classification and Non-animal Removal

To determine if a given word is the name of an animal, the Portuguese WordNet is searched for synsets that have this word as a lemma, effectively outputting all possible synsets that a given word corresponds to. The full hypernym path of each synset is computed, i.e., the collection of synsets that have a "is a" relationship with the given synset, resulting from the hypernyms of a synset, the hypernyms of these hypernyms, and onward, in a recursive fashion. If any of these paths contains the *animal* synset, the word is classified as being an animal. Words that are not classified as an animal are excluded from the transcript, retaining only animal names that were produced in the scope of the SVF test.

For instance, consider the words *ornitorrinco*, *periquito*, and *garfo*. The corresponding hypernym path is depicted in Figure 3.4: since the hypernym path of the synsets associated with *ornitorrinco* and *periquito* contain the *animal* synset, these are classified as animals and would be kept by the non-test word filter; on the other hand *garfo*, not having *animal* in its hypernym path, would be discarded.

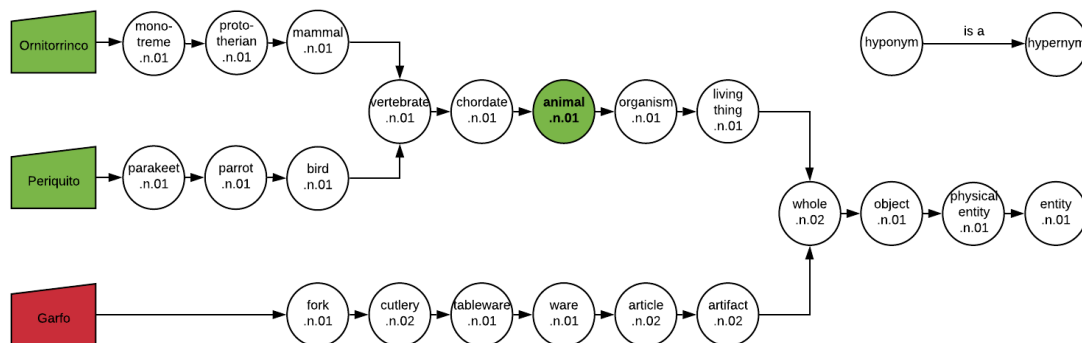


Figure 3.4: Hypernym tree illustration of *ornitorrinco*, *periquito*, and *garfo*. WordNet synsets are represented by circles, with synsets increasing in generality from left to right. Green input words are classified as animals and are kept, whereas red input words are discarded.

To correctly identify words that do not directly reference an animal, such as words related to food like *vitela*, all holonym synsets are also analysed in the same fashion as above. For these cases, their holonyms will be the synsets associated with the animal *per se*, and not just their food items, allowing for a correct classification as animals.

3.4 Phonemic Transcription

A major motif present and described in the performance of VF tests of subjects is the association of words according to how they sound: similarly sounding words tend to be produced in quick succession, such as rhyming words, words starting with the same sound, or words with one sound replaced. Phonemic search of possible word candidates is an important factor that affects performance of VF tests, especially PVF, since this is the most prevalent search strategy for tasks of this nature [114].

Since letters in different words can represent different sounds, how a word is spelled brings little to no information regarding how it sounds. As such, phonemic transcription of the words produced during VF tests is required to automatically evaluate how similarly sounding they are, by decomposing each word into its constituent sounds.

Automatic phonemic transcription is a difficult and challenging computational task, which is beyond the scope of this dissertation. Instead, a dictionary-based approach was opted, where each word in the speech-to-text transcript is searched on an online dictionary that contains the phonemic transcription of said word.

The chosen website was [Infopédia](#) [95], since it is a large European Portuguese dictionary that contains phonemic transcriptions for most words referenced in it: in particular, all animals and *P*-words uttered during all collected VF tests have a phonemic transcription available on Infopédia.

Phonemic transcriptions on the Infopédia website follow the International Phonetic Alphabet (IPA) [57], with each character corresponding to a specific phoneme (unit of sound in a word). For instance, the word "*gato*" has /gatu/ as its transcription, and /purkijudɐˈĩdʒɐ/ is the transcription

of "*porquinho-da-índia*". By directly comparing the phonemic transcription of two words, i.e., comparing the symbols that make these transcriptions, it is possible to determine their phonemic similarity, which will be a main focus of phonemic feature extraction.

To perform phonemic transcription using Python 3.6, an algorithm is developed that parses the *HTML* code of the web-page corresponding to each word said. Consider, for instance, the word "*coelho*". The phonemic transcription is obtained as such:

1. A request is made to the web-page <https://www.infopedia.pt/dicionarios/lingua-portuguesa/coelho>;
2. If successful, i.e., if the status code of the request is 200, the phonemic transcription of the word is found in the *HTML* code using regular expressions.

Some words, such as "*coelho*" have more than one possible phonemic transcription, due to regional or dialect variability. For this word, two alternatives are presented in the website:

- /kwɐ(j)ɫu/, with an open first vowel;
- /kweɫu/, with a closed first vowel.

Each of these alternatives can be captured by the use of regular expressions on the *HTML* code given by the request.

Note however the /j/ sound present in the first alternative: this indicates a possible semi-vowel sound that may or may not be present in that particular pronunciation of "*coelho*", depending on how long a subject makes that sound transition.

If a phonemic transcription has parenthesis, indicating the possibility of the existence of a given phoneme in a given word, two transcriptions are derived: one with the sound present, and one without.

As such, there are effectively three possible phonemic transcriptions for this word:

- /kwɐjɫu/, with an open first vowel and long semi-vowel transition;
- /kwɐɫu/, with an open first vowel, but no semi-vowel sound;
- /kweɫu/, with a closed first vowel.

Each word is thus associated with a list of possible phonemic transcriptions. In most words, this is just a length 1 list with the common phonemic transcription. If the searched word does not have a phonemic translation available, or if the web-page does not exist or is not available at the time, the algorithm associates an empty list to the word.

This procedure is performed for each word in the filtered VF test transcript, thus generating a list of lists of phonemic transcriptions, one for each word. This can be done in a quick fashion, only taking approximately 5 seconds to request phonemic transcriptions for all words in a given VF test. Nonetheless, to avoid unnecessary requests to Infopédia, phonemic transcriptions for each VF test are stored on a *.csv* file.

3.5 Word Embeddings Dictionary

Another major driving force for the production of exemplars during a VF test, particularly with SVF, is the semantic relatedness between words: words tend to be produced within defined semantic clusters, as described by Troyer et al. [114].

The measurement of semantic similarity between words is a challenging NLP task which is required for the determination of semantic clusters present in a given VF test. Some techniques for the determination of semantic similarity, such as Latent or Explicit Semantic Analysis, are described in Subsection 2.5.3, since these have already been used in this same panorama of VF tests [86, 87, 121].

In general, to achieve this result, words must be converted to a highly dimensional embedding space, where each word is represented by a highly dimensional vector. To determine semantic relatedness between words, a similarity metric between the corresponding embedding vectors is calculated, usually the cosine similarity measure, as in Equation 3.1, with $\vec{t}_i$ and $\vec{t}_j$ being the word embeddings of the i -th and j -th word (w_i and w_j), respectively:

$$\langle w_i, w_j \rangle^{sem} = \cos \theta = \frac{\vec{t}_i \cdot \vec{t}_j}{||\vec{t}_i|| \cdot ||\vec{t}_j||} \quad (3.1)$$

For this dissertation, the implementation of a state-of-the-art algorithm with equal or better performance than the ones referenced in Subsection 2.5.3 is a daunting task, requiring an immense processing power, hundreds of hours of training, and very large text corpora as input.

Instead, pre-trained Portuguese word embeddings from *fastText* are used [47], which contain 2 million Portuguese word embeddings in 300-dimensional vectors. These embeddings are downloaded from the *fastText* website [48] as a text file and are used to create a Python dictionary, with each word as the key to its corresponding embedding, to facilitate and quicken access to these vectors.

After Speech-to-Text transcription of the recorded VF test and non-test word removal, the list of transcribed words is converted to the corresponding list of word embeddings, using the aforementioned embeddings dictionary as a conversion tool.

3.6 Incorrect Word Removal

The number of correctly produced words during a VF test is the oldest and most commonly used feature extracted for these tests, being the most studied metric with respect to the impact of brain lesions, mental disorders, and NCDs [16].

Besides filtering non-test words, it is necessary to identify which test words do not comply with the specific rules of VF tests, particularly for PVF and SVF. Proper nouns, repetitions, different conjugations of a verb, and small alterations to words, such as suffixation, are to be ignored to obtain the number of correct words produced. Figure 3.3 represents the major steps for incorrect word removal, on the lower section of the image.

3.6.1 Incorrect Word Removal for Phonemic Tests

The particular rules for PVF administration and scoring [21] state that proper nouns (names of people and places) are to be discarded. In addition, for repetitions and words representing the same object/action/concept with slight alterations to their root (e.g. variation of number, gender, or size), only the first response is credited.

To only credit correctly produced *P*-words, three major steps are applied: repetition removal, proper noun removal, and suffix variation removal.

3.6.1.1 Repetition Removal

To remove repeated words produced, each word in the transcription word list is successively compared with the words following it: if a matching pair is found, the second instance is removed from the list.

3.6.1.2 Proper Noun Removal

Names of people and places are easily transcribed by the Google Cloud Speech-to-Text API and can appear on automatic transcriptions if the subject performing the PVF test forgets to inhibit the production of proper nouns.

It was noted during the development of the pipeline that, for the phonemic transcription of words using Infopédia, proper nouns, such as *Paulo*, *Pedro*, and *Portugal*, do not have an entry in this online dictionary, along with other similar nouns: when queried for these words, Infopédia returned a 404 webpage error.

Taking advantage of this, a simple algorithm was developed to identify proper nouns: if a given word had no Infopédia entry, i.e., returned a 404 error when queried for it, it is most likely a proper noun and was discarded from the word transcription list.

3.6.1.3 Suffix Variation Removal

Unlike the previously described method for inflection normalisation for non-animal removal, suffix variation removal does not require the conversion of tokens to their respective lemmas, being only required to identify the root of each word.

As such, the NLTK Project [81] implementation of the *Removedor de Sufixos da Lingua Portuguesa* (RSLP) Stemmer [84] is used to stem words to their lexical roots. Similarly to repetition removal, each word root is successively compared with the word roots following it in the test transcript: if a match is found, the second instance is removed from the transcript list.

3.6.2 Incorrect Word Removal for Semantic Tests

Rules for scoring SVF tests, particularly animal naming tests, are described in Cavaco et al. [21]. Regional designations of animals are to be credited. Only the first exemplar of a given animal species is credited, including gender and size variations in their naming: for instance *boi*, *vaca*, and

vitelo are to be scored as a single correct word. General, superordinate categories of animals are to be credited only when no specific animals of that category are produced: if a subject produces the word *pássaro*, and doesn't name a particular bird, then this word is credited, otherwise it is discarded.

To remove incorrectly produced words and to only credit words that respect the rules described above, three filters are applied: repetition removal, synonym removal, and hypernym removal.

3.6.2.1 Repetition Removal

Similar to PVF tests, to remove repeated words produced, each word in the transcription word list is successively compared with the words following it: if a matching pair is found, the second instance is removed from the list.

3.6.2.2 Synonym Removal

Some words, particularly animal names, can refer to the same concept or animal species, but have distinct morphology and word origin. These synonyms must be accounted for, by only crediting the first instance of a group of synonyms, and discard the rest.

Some different animal designations, such as gender variations, are already normalised upstream in the inflection normalisation step using lemmatisation lists. Nonetheless, some synonymical designations still remain, particularly for the case of the younger offspring of some species, and for food items related to some animals. For instance, it is of interest to identify *boi* and *vitela* as synonyms, since subjects produce these words while thinking in the same animal species.

To identify synonyms present in a test transcript, each word is successively compared with the following words to determine if they are synonyms: in the positive case, the second word (in the order of the transcription) is removed.

To determine if two words are synonyms, the Multilingual WordNet [15, 29, 81] is once again used. First, the sets of synsets associated with each word are determined. If there exist shared synsets, the words are considered synonyms. Alongside this, the holonym sets of these synsets are also considered and compared: if two words share some synset or have shared holonym synsets, they are considered synonyms.

3.6.2.3 Hypernym Removal

Superordinate animal categories can be thought of as hypernyms of specific elements from these categories. Thus, to determine the possible existence of hypernym designations of a given animal in a test transcript, all possible word pairs are analysed to determine if there exists a hyponym-hypernym relationship between words. If such a pair is found, then the hypernym of that pair is removed from the transcript word list.

To determine if there exists a hyponym-hypernym relationship between a pair of words, the Multilingual WordNet [15], particularly the Portuguese WordNet within [29], is once again used, through the NLTK Project [81] package implementation.

For each pair of words analysed, the lowest common hypernym synset between them is determined. Consider for instance the words and respective hypernyms represented in Figure 3.4: the lowest common hypernym between *periquito* and *ornitorrinco* is *vertebrate*.

In the cases where one word is the hypernym of another, the lowest common hypernym will be the synset corresponding to that hypernym word.

Consider now the words *periquito* and *pássaro*: the lowest shared common hypernym will be the synset associated with the word *pássaro*, i.e., the bird synset. By performing the intersection between the list of lowest common hypernyms and the synsets associated with each word, hypernym relations can be identified: if the intersection is non-empty, then one of the words is the hypernym of the other, and is discarded from the test transcript for scoring.

3.7 Feature Extraction

After removal of uttered words that are not produced in the scope of the performed VF test, the resulting list of test words (and associated production times) are analysed for feature extraction.

The number of correct words produced, and the number of errors/repetitions a user makes when performing the test are very important features, and are obtained after incorrect word removal, as described previously.

In addition, as observed and analysed by Troyer et al. [114], words tend to be produced in clusters organised according to semantic and phonemic relatedness between words. The analysis of these clusters can bring insight to a subject's executive and language functions, easing the screening, detection, and distinction of different types of brain lesions, NCDs, and mental disorders [69]. As such, extracting features related to these clustering phenomena is an important step of the automatic VF analysis pipeline.

Word production timing is also insightful to determine how the human brain tackles exemplar production, inhibition, and subcategory searching and switching. As proposed by Mayr [74], detailed timing protocols can provide new information to VF test analysis, and can mitigate some intrinsic problems with the switching and clustering scores of Troyer et al. [114]. Therefore, temporal feature extraction is also performed, based on the word production start timestamps provided by the used Speech-to-Text system.

In total, 20 features are extracted from a VF test transcript: total number of correct words, number of errors, 6 temporal features, 6 semantic features, and 6 phonemic features.

3.7.1 Temporal Features

König et al. [62] perform temporal analysis of VF tests, namely the timing of exemplar production, through the analysis of corresponding audio signals.

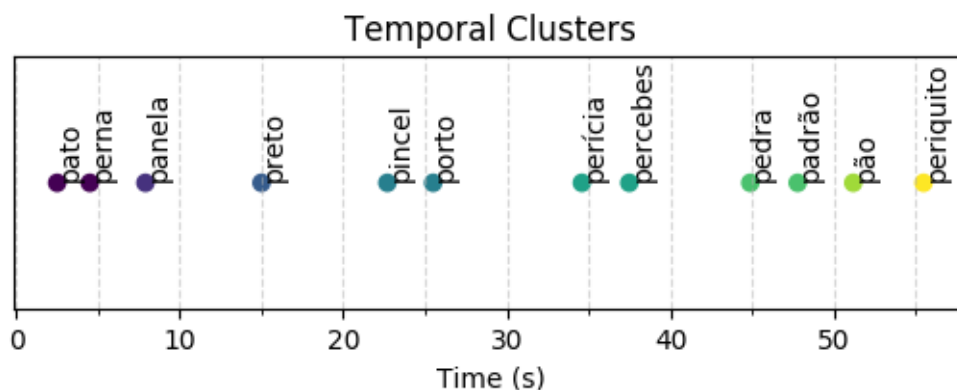


Figure 3.5: Identified temporal clusters of an automatic PVF test transcript, using the DBSCAN clustering algorithm. Points with the same colour represent words clustered in the same temporal cluster.

The temporal features here automatically extracted are based on the metrics used by König et al. [62], Woods et al. [121], as well as on the concepts of switching and clustering [114] applied to exemplar production.

Unlike König et al. [62] where the production timing is determined by analysing the recorded audio signal of VF test administration, or Woods et al. [121] where production timing is obtained according to the timing of keyboard presses, for the analysis here proposed, word timings are obtained using Google Cloud [46] Speech-to-Text, that provides transcriptions and word timestamps.

To determine temporal clusters, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm [123] is applied to the list of word timestamps, with the implementation provided by the *scikit-learn* Python package [90].

The DBSCAN algorithm allows clustering of data points according to their distances; in this case, time between word production. A threshold of 3 seconds is used as the maximum time difference between exemplar production for them to be considered part of the same cluster. To identify isolated words as distinct temporal clusters, the minimum number of points in a cluster is set to 1. Figure 3.5 shows the resulting temporal clusters obtained for a given PVF test.

From the identified word timestamps and respective temporal clusters, several metrics are obtained, which are used as temporal features of a given test [62, 121]:

Mean Temporal Cluster Size average number of words per temporal cluster;

Maximum Temporal Cluster Size size of the largest temporal cluster present in the test;

Number of Temporal Cluster Switches total number of switches between temporal clusters, i.e. the total number of temporal clusters $- 1$;

Mean Intra-Cluster Time average time interval between produced words that belong to the same clusters, ignoring clusters of size 0;

Mean Inter-Cluster Time average time difference between the last word of a cluster and the first word of the following cluster;

Temporal Decline Percentage percentage of words produced during the first half of the VF test.

Similar to Troyer et al. [114], cluster size is defined as the number of words belonging to a cluster -1 : for instance, a 1-word cluster has size 0, 2-word clusters have size 1, and so on.

3.7.2 Semantic Features

Semantic analysis of produced exemplars is performed based on their corresponding *fastText* Portuguese word embeddings [47], which are obtained as described in Section 3.5.

From these embeddings, the cosine similarity (Equation 3.1) is determined for all possible pairs of words, generating a cosine similarity matrix for a given VF test, according to Equation 3.2, with S^{sem} as the semantic similarity matrix, and w_i and w_j as the i -th and j -th words of the test transcript, respectively. A representation of this matrix for a PVF and SVF test is shown in Figure 3.6.

$$S_{ij}^{sem} = \langle w_i, w_j \rangle^{sem} \quad (3.2)$$

The cosine similarity matrix is then used to determine semantic clusters. Three different grouping strategies are considered:

Chain where words are aggregated in clusters if their cosine similarity with the last word of the cluster is above the pairwise threshold $\delta_{pairwise}^{sem}$, as described in Equation 3.3 [70], where N represents the total number of words in the transcript and w_C the last word of the cluster C ;

$$w_n \in C \iff \langle w_n, w_C \rangle^{sem} > \delta_{pairwise}^{sem}, \quad (3.3a)$$

$$\delta_{pairwise}^{sem} = \frac{1}{N-1} \sum_{n=1}^N \langle w_n, w_{n-1} \rangle^{sem} \quad (3.3b)$$

Cluster where words are aggregated in clusters if their cosine similarity with the average cluster embedding is above the global threshold δ_{global}^{sem} , as described in Equation 3.4 [70], with $\vec{\mu}_C$ as the average embedding vector of cluster C ;

$$w_n \in C \iff \langle w_n, \vec{\mu}_C \rangle^{sem} > \delta_{global}^{sem}, \quad (3.4a)$$

$$\delta_{global}^{sem} = \frac{2}{N(N-1)} \sum_{i>j} \langle w_i, w_j \rangle^{sem}, \quad (3.4b)$$

$$\vec{\mu}_C = \frac{1}{|C|} \sum_{\vec{x} \in C} \vec{x} \quad (3.4c)$$

DBSCAN where the DBSCAN clustering algorithm [123] is applied, using the cosine similarity matrix as a precomputed distance matrix, with maximum distance between embeddings of 0.5 to be considered neighbours.

From the identified clusters from these models, several metrics are extracted, which are used as semantic features of a given test [70, 121]:

Mean Semantic Cluster Size average number of words per semantic cluster;

Maximum Semantic Cluster Size size of the largest semantic cluster present in the test;

Number of Semantic Cluster Switches total number of switches between semantic clusters, i.e., the total number of semantic clusters—1;

Average Pairwise Semantic Similarity average embedding similarity between sequential words pairs of a transcript, $\delta_{pairwise}^{sem}$ in Equation 3.3;

Average Global Semantic Similarity average embedding similarity between all possible word pairs of a transcript, δ_{global}^{sem} in Equation 3.4;

Semantic Organisation Index the ratio between $\delta_{pairwise}^{sem}$ and δ_{global}^{sem} , which transmits the level of semantic organisation of exemplars: a value of 1 indicates exemplars are produced in a seemingly random fashion, whereas higher values indicate a stronger influence of semantic word retrieval.

The mean semantic cluster size and number of semantic clusters switches are an automatic analogue of the metrics proposed by Troyer et al. [114]. Cluster size is defined again as the number of words in a given cluster—1.

For example, using the clustering grouping method on the SVF test transcript "*cão gato coelho porco cavalo vaca boi ovelha cabra leão rinoceronte zebra elefante crocodilo peixe lagarto cabra crocodilo*" results in the following semantic clusters:

- *cão gato coelho porco cavalo*;
- *vaca boi ovelha cabra*;
- *leão rinoceronte*;
- *zebra*;
- *elefante crocodilo peixe lagarto*;
- *cabra*;
- *crocodilo*.

3.7.3 Phonemic Features

Phonemic analysis of VF test transcripts is performed using the phonemic transcriptions of each word, obtained as described in Section 3.4.

From these phonemic representations using the IPA alphabet, phonemic similarity between words is calculated using the Levenshtein distance [67] between corresponding phonemic transcriptions [104]. The Levenshtein distance is an edit distance for strings that indicates the number of alterations necessary to transform one string into another, by using insertions, substitutions and

deletions. When applied to phonemic transcriptions, the Levenshtein distance indicates the number of edits in phonemes to transform the sound of one word into another: similar sounding words have a low edit distance, whereas dissimilar words have a high edit distance associated.

The Levenshtein distance algorithm utilised in the automatic VF pipeline was the implementation available in the *Levenshtein* package [49].

For a given pair of words i and j of a test transcript, their phonemic similarity is calculated as in Equation 3.5, with P_i being the set of possible phonemic transcriptions of word i , P_j the set of possible phonemic transcriptions for word j , $\text{Lev}(p_i, p_j)$ the Levenshtein distance between phonemic transcripts p_i and p_j , and $|p|$ representing the length (number of characters) in a phonemic transcript.

$$\langle w_i, w_j \rangle^{pho} = \max_{p_i \in P_i, p_j \in P_j} \left(1 - \frac{\text{Lev}(p_i, p_j)}{\max(|p_i|, |p_j|)} \right) \quad (3.5)$$

Since the Levenshtein distance between phonemic transcripts is a positive integer, it is first normalised by dividing it by the length of the largest phonemic transcription of the pair, thus containing it in the $[0, 1]$ interval. Because the Levenshtein distance is a measure of edit distance, and not of similarity between a pair of string, it is converted to a similarity metric by subtracting it to 1. Finally, since there are several possible phonemic transcripts for each word in the compared pair, the value picked for phonemic similarity is the largest possible one, i.e., the one that originates from the most similar options for phonemic transcription.

From this phonemic similarity metric defined above, a phonemic similarity matrix for a given test transcript is calculated, according to Equation 3.6. Examples of phonemic similarity matrices are shown in Figure 3.7.

$$S_{ij}^{pho} = \langle w_i, w_j \rangle^{pho} \quad (3.6)$$

This phonemic matrix is then used to determine phonemic clusters. The same three grouping strategies of chaining (analogous to Equation 3.3), clustering (Equation 3.7), and DBSCAN are used, in a similar way as in the determination of semantic clusters, but utilising the phonemic similarity metric defined in Equation 3.5.

For the clustering method, a slightly different approach from Equation 3.4 was used, since there are no "phonemic embeddings". As such, clusters are determined following Equation 3.7, where a word will be assigned to a cluster C if the resulting cluster words (i.e., words already belonging to cluster C , plus the new word w_i) have an higher average phonemic similarity δ_C^{pho} than the average global phonemic similarity δ_{global}^{pho} .

$$w_n \in C \iff \delta_C^{pho} > \delta_{global}^{pho}, \quad (3.7a)$$

$$\delta_{global}^{pho} = \frac{2}{N(N-1)} \sum_{i>j} \langle w_i, w_j \rangle^{pho}, \quad (3.7b)$$

$$\delta_C^{pho} = \frac{1}{|C|+1} \sum_{w_i, w_j \in \{C \cup w_n\}} \langle w_i, w_j \rangle^{pho} \quad (3.7c)$$

From the identified clusters from these models, several metrics are extracted, which are used as phonemic features of a given test [104, 121]:

Mean Phonemic Cluster Size average number of words per phonemic cluster;

Maximum Phonemic Cluster Size size of the largest phonemic cluster present in the test;

Number of Phonemic Cluster Switches total number of switches between phonemic clusters, i.e., the total number of phonemic clusters—1;

Average Pairwise Phonemic Similarity average phonemic similarity between sequential words pairs of a transcript, $\delta_{pairwise}^{pho}$ in Equation 3.3;

Average Global Phonemic Similarity average phonemic similarity between all possible word pairs of a transcript, δ_{global}^{pho} in Equation 3.4;

Phonemic Organisation Index the ratio between $\delta_{pairwise}^{pho}$ and δ_{global}^{pho} , which transmits the level of semantic organisation of exemplars: a value of 1 indicates exemplars are produced in a seemingly random fashion, whereas higher values indicate a stronger influence of phonemic word retrieval.

The mean phonemic cluster size and number of phonemic clusters switches are an automatic analogue of the metrics proposed by Troyer et al. [114].

For example, using the clustering grouping method on the PVF test transcript "*palheiro pestana pisco pitada porco pé praia país panela parvo prensa penso pisco palhota*" results in the following semantic clusters:

- *palheiro*;
- *pestana pisco pitada porco*;
- *pé*;
- *praia*;
- *país panela*;
- *parvo*;
- *prensa penso pisco palhota*.

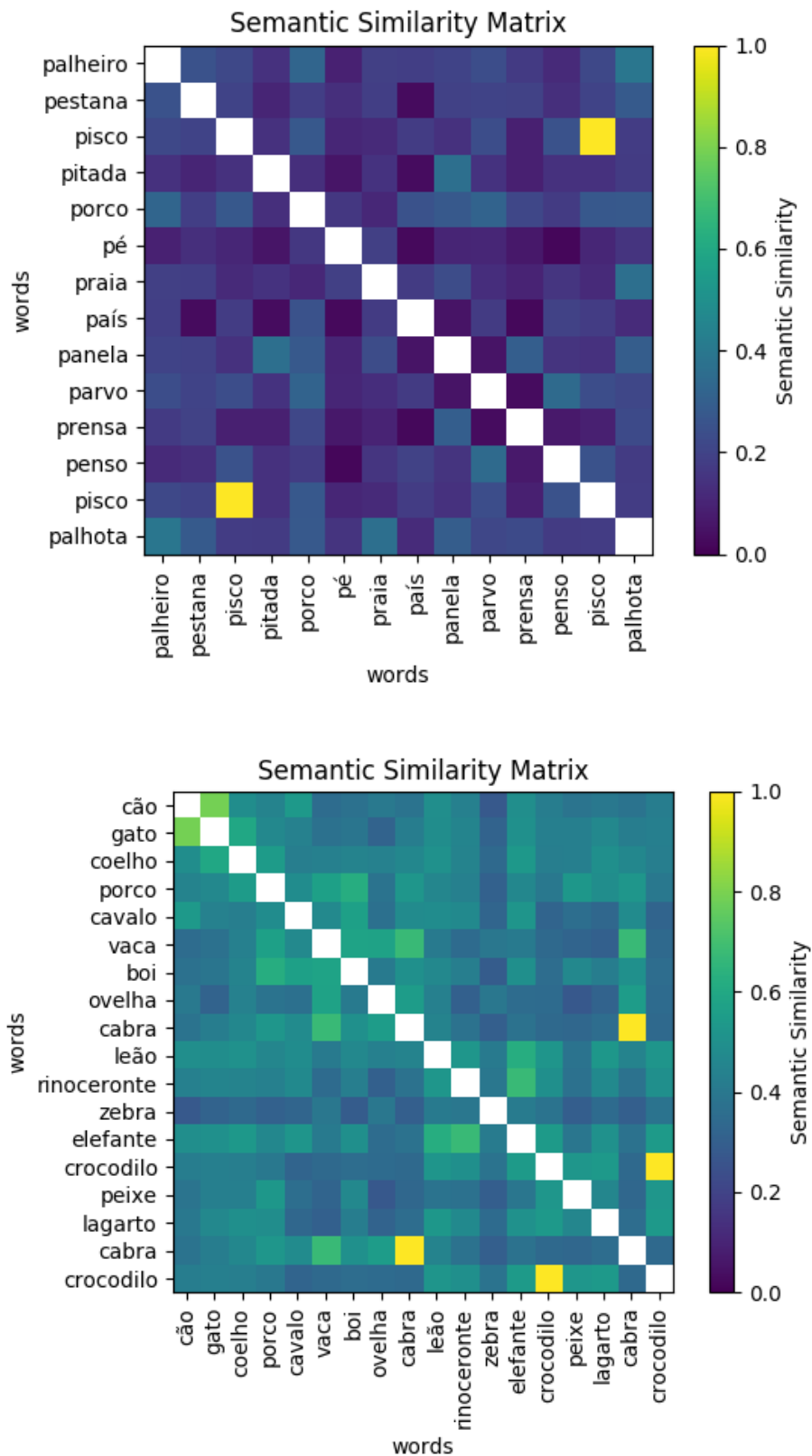


Figure 3.6: Semantic Similarity Matrices, as calculated in Equation 3.2, for a PVF (above) and SVF (below) test transcript.

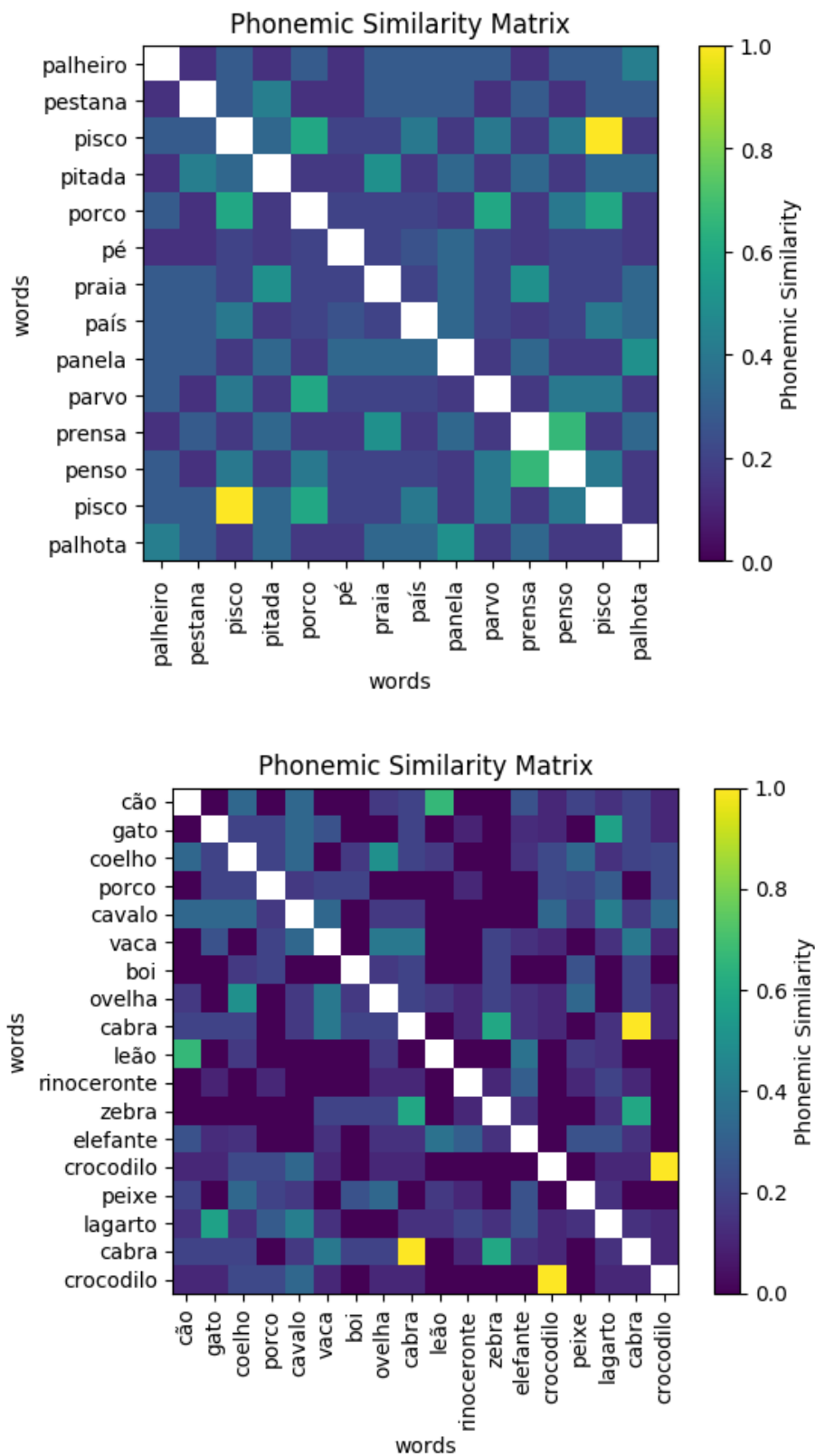


Figure 3.7: Phonemic Similarity Matrices, as calculated in Equation 3.6, for a PVF (above) and SVF (below) test transcript.

Chapter 4

AutoSpeech Study Design

In conjunction with the implemented automatic pipeline for the analysis of VF tests, an introductory and exploratory study involving elderly subjects was performed. This chapter details the main objectives of this study, along with its test protocol. In addition, legal and ethical concerns regarding confidentiality and informed consent are here discussed.

An approval request for the realisation of this study was made to the Ethics Committee of Universidade do Porto (CEUP), with the information provided in the following sections. Appendix B provides a copy of the response report of CEUP to this request, where this study is approved for realisation.

4.1 Main Objectives

The main goal of this exploratory study was to evaluate the relationship between the level of physical activity of healthy elderly Portuguese subjects with their performance on VF tests, by comparing averaged declared time performing different physical activities with VF test metrics extracted with an automatic pipeline.

Moreover, the data collected during this study was used to assess the validity of the features extracted automatically by the pipeline proposed in Chapter 3, by comparing them with traditional, manually extracted features, namely total word count, and switching and clustering metrics from Troyer et al. [114].

In addition, automatic and manual transcriptions of VF tests were used to assess the overall performance of the selected ASR system for this type of setting in European Portuguese, thus determining if this system is trustworthy enough for these types of applications.

4.2 Participant Recruitment

To achieve statistically relevant results, it was desirable to include in this study at least 40 participants that respect both inclusion and exclusion factors.

4.2.1 Recruitment Strategies

Participants were recruited in Porto by Fraunhofer Portugal, integrated in the COLABORAR project [37].

No financial compensation was awarded, except for the refund of the exact value of necessary trips for participation in this study in the facilities of Fraunhofer.

Volunteer's motivation for participation was their willingness and interest to contribute for research studies with a technological basis.

4.2.2 Inclusion and Exclusion Factors

Since VF test performance can be influenced by a multitude of factors and conditions, inclusion and exclusion factors are here defined to restrict the used population sample to one where the effects of physical activity on the performance of these tests can be studied. Inter-subject variability factors, such as age and level of education, were recorded for each participant, as discussed in Section 4.3.

4.2.2.1 Inclusion Factors

- 65 years old or older;
- Healthy (self declared).

4.2.2.2 Exclusion Factors

- Neurocognitive disorder diagnosis (Alzheimer's, Parkinson's, etc.);
- Mild cognitive impairment diagnosis;
- Not a Portuguese native speaker;
- Presence of hearing of language impairments;
- History of head trauma, brain damage, or stroke;
- Loss of consciousness;
- Addiction (including alcoholism);
- Psychotic or abhorrent motor behaviour;
- Intake of medication that influences psychomotor capabilities.

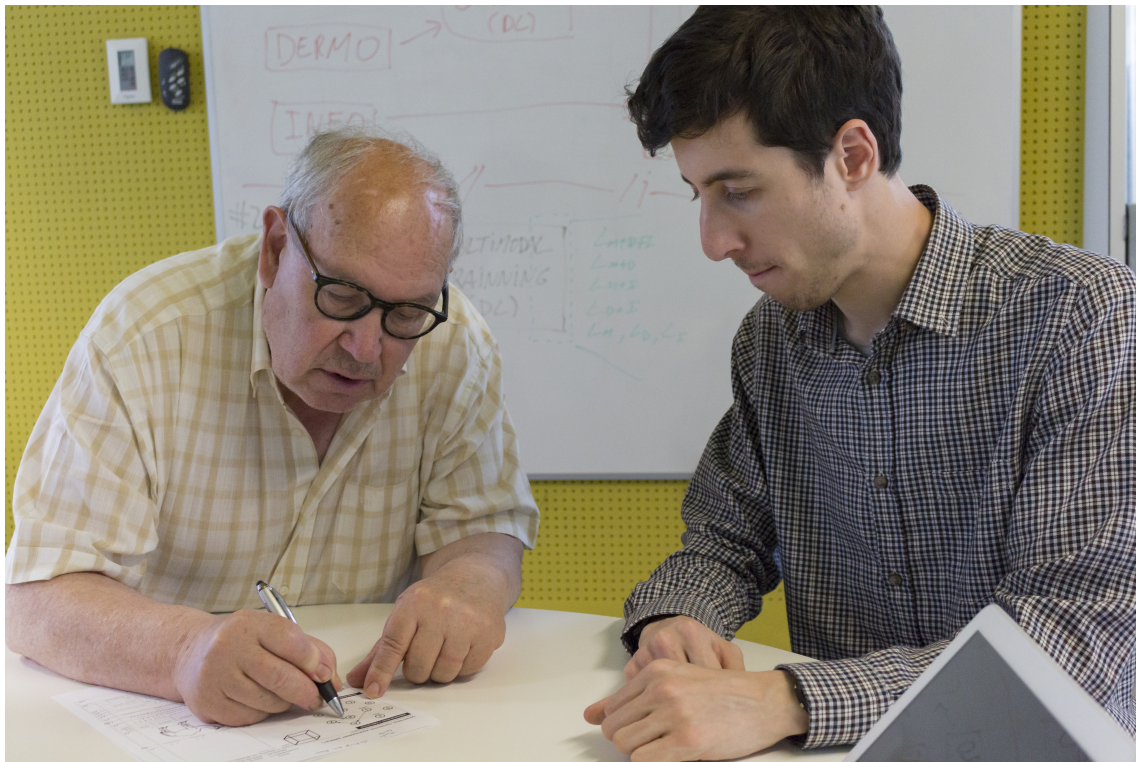


Figure 4.1: Depiction of a data collection interview from the AutoSpeech Study. The participant can be seen performing a cognitive test of the MoCA.

4.2.3 Informed Consent

Informed consent (Appendix C) was obtained for each participant (or respective representative) for the participation in the study.

Participants had access to written information detailing the objectives, risks, and benefits of the study (included in the same document of informed consent, Appendix C, leaving one of the copies for the participant), and could ask for any clarification by soliciting verbal information to the researcher.

Each participant had total freedom to decide their participation in the study, and could cease participation at any moment, with no drawbacks or consequences.

4.3 Test Protocol

This study only required one phase, in which participant data was collected in a semi-structured interview setting with a psychologist. Interviews lasted 20 to 30 minutes with each participant. Figure 4.1 illustrates the data collection procedure.

Interviews were performed in the facilities of Fraunhofer-AICOS, or in the elderly care homes that participants frequent (if travelling to the facilities was not a possibility).

Participants were healthy elderly Portuguese subjects, 65 years old or older, with no diagnosis of NCDs or other conditions that may affect performance of the performed tests.

The interview was divided into three main stages:

1. MoCA administration [80];
2. Sociodemographic Questionnaire administration;
3. Short-IPAQ administration [28].

The MoCA does not have a semantic VF test, so after the conclusion of this battery, a quick animal naming test was performed (see Appendix E for the used Portuguese script).

The MoCA was the first questionnaire to be administered since it encompassed the majority of the interview, and required total concentration and attention from the participants, thus being ideal to perform it first. The following questionnaires did not require as much concentration, and therefore could be performed after a strenuous cognitive effort such as the MoCA.

Since cognitive ability and VF test performance can be influenced by sociodemographic factors, particularly age and level of education [38], this information was also collected for each participant using a self-designed sociodemographic questionnaire.

Finally, to determine if the level of physical activity has any correlation with the performance of the MoCA and VF tests, the International Physical Activity Questionnaire (IPAQ), short version, was administered. This questionnaire estimates the overall level of activity of a subject from their self-declared times spent performing calm, mild, or vigorous physical activities over the last 7 day period. Two metrics can be extracted from the short-IPAQ, by following the criteria delineated in IPAQ Research [58]: the Metabolic Equivalent of Task (MET) minutes per week, or *MET – min/week* of a subject, and a classification of physical activity into one of three categories: low, moderate, and high physical activity. The MET can be thought of as a ratio of energy expenditure, between energy expended by a subject while performing a certain task to the energy expended by the same subject while at rest. Tasks with a higher MET indicate more vigorous physical activities. The measure of *MET – min/week* indicates the overall energy expended during a week: a subject with a higher value of *MET – min/week* expends more energy on physical activities, and thus can be said to have a higher volume of physical activity per week.

4.3.1 Test Material

To perform the aforementioned tests, the following material was necessary:

- Pen;
- *Smartphone* to record audio of VF tests;
- Printed MoCA, Portuguese version (Appendix D);
- Printed Sociodemographic Questionnaire (Appendix F);
- Printed short-IPAQ, Portuguese version (Appendix G);

4.4 Risk Management

This exploratory study involved human beings, particularly elderly individuals, aged 65 or older. No damages or inconveniences for the participant were predicted during the duration of the study.

The participant could interrupt the interview whenever deemed necessary, with no consequences.

Transport related expenses were refunded. For participants that travelled to the facilities of Fraunhofer AICOS, a personal accidents insurance was provided for them, to safeguard any possible incident during travel.

4.5 Benefits for the Participant

The result of the Montreal Cognitive assessment, applied by a psychologist, was made available to the study's participants, if they so desired, in the form of a report with detailed conclusions regarding this battery.

4.6 Confidentiality

Data confidentiality is of utmost importance when dealing with studies involving humans, particularly when such studies require the collection of personal and sensitive data, such as overall cognitive ability.

As such, data confidentiality was ensured during the development of this study, by anonymising all the collected information for each participant: each subject of the study was associated with a unique identifying code during the administration of the several written questionnaires. No identifying information, such as name, date of birth, or home address, was collected. Audio collected from VF tests was also only referred by the corresponding identifying code of each participant.

The collected data was only used for research purposes by the researchers directly involved with the project. Data files can only be accessed by these researchers, through user-name and password. Written questionnaires were archived in private cabinets, to which only the researchers have access.

4.6.1 Data Acquisition Methods

The psychologist performed data collection with the participant, through the use of several questionnaires, detailed in Table 4.1.

Several participant data was collected, namely cognitive health data (using the MoCA, Appendix D), sociodemographic variables (using the designed Sociodemographic Questionnaire, Appendix F), and physical activity information (using the short-IPAQ, Appendix G).

Alongside this, the audio of each VF test administered, i.e., the PVF (letter *P*) test present in the MoCA and the SVF (animal fluency), was recorded using a *smartphone*.

Table 4.1: Tests and Questionnaires used in the designed exploratory study

Stages	Parameters	Test/Questionnaire
Stage 1: MoCA	Cognitive Functions	MoCA (Appendix D)
	Semantic Verbal Fluency	Animal Fluency (Appendix E)
Stage 2: Sociodemographic data	Sex	Sociodemographic Questionnaire (Appendix F)
	Age	
	Level of Education	
	Marital Status	
Stage 3: IPAQ	Lives Alone	short-IPAQ (Appendix G)
	Physical Activity	

Inclusion and exclusion factor were also collected for each participant, and if a participant wished to receive the final MoCA report composed by the psychologist, their living address and/or email address was collected as well.

All the collected audio and questionnaire data was be anonymised during the duration of the study, and is destroyed in its end.

4.7 Ethical Considerations Regarding Test Results

Taking into account the intended population to study and the tests and questionnaires to be applied, some ethical questions arise with the development of this study.

Older subjects have an increased incidence of neurocognitive disorders [7], being a particularly sensitive population regarding these diseases. Moreover, the development of these disorders on subjects may have only subtle effects and symptoms on their early stages, and the rapid diagnosis of dementia onset can slow down its progression, with adequate medication treatment.

Considering that the MoCA administered in the scope of this study to elderly subjects can be used as a dementia and MCI screening tool [80], the results of this test battery on the recruited subjects can be clinically valuable to them. In addition, consider for instance a participant of this study, who self-declares to be healthy physically and mentally, whose performance on the MoCA is below the defined normal threshold, indicating for instance the possibility of MCI or even dementia: how to properly transmit this information to the subject? Should the subject be redirected to more specialised medical consultations, to rule out or confirm a possible NCD diagnosis?

These ethical questions arose during the development of this study, and were addressed with the redaction of individual reports detailing the performance of each subject regarding the several cognitive tasks evaluated with the MoCA. These reports are written by the psychologist that administers the questionnaires of this study, with comments and annotations regarding a subject's performance and behaviour during testing.

If a subject's performance is below the threshold for normal scores of the MoCA, even after taking into account age, level of education, and demographic factors [39], the report addresses this fact, and suggests possible further consultations by psychologists or neurologists to perform further examinations, to assess possible causes for decreased cognitive abilities.

Chapter 5

Data Characterisation

To evaluate the performance and feasibility of an automatic pipeline for the analysis of Portuguese VF tests, recorded using a *smartphone*, two main sources of audio samples are used, collected in different contexts and scopes. In this chapter these two sources of data are characterised according to their collection methods, as well as subject information available from each.

5.1 Centro Hospitalar Universitário do Porto Data

Amorim and Susano [4] performed a large clinical trial (ClinicalTrials.gov identifier (NCT number): NCT03171766) to assess the incidence of delirium in postoperative patients subjected to elective surgery in Centro Hospitalar Universitário do Porto (CHUP), with the main objective of determining the main predictors of postoperative delirium. As a part of patient characterisation of cognitive ability, a SVF test is administered to these subjects, using the animal category. To avoid performing *in situ* manual transcription of the subject’s responses, which is a challenging task to the examiner, responses were recorded using a *smartphone*. These recordings were later manually analysed to count the total number of correct words produced by patients.

This study was approved by the competent bodies *Departamento de Ensino Formação e Investigação*, *Comissão de Ética*, and *Conselho de Administração* of CHUP, approval number 253/16 (216-DEFI/ 205-CES). These recordings were gently provided by the principal investigators to be used as validation samples of the proposed automatic VF test analysis pipeline of this dissertation. Data sharing was approved by *Gabinete Coordenador de Investigação* of CHUP (Appendix A).

The provided recordings are anonymous (subjects are identified numerically), with no patient information provided: the age and level of education of each participant is not known, thus a qualitative analysis of their performance based on normalised scores [21] is not possible.

Results obtained with this dataset are described and discussed in-depth in Chapter 6, with a major focus on the performance of the Speech-to-Text system used (Section 6.1), and of the employed feature extraction methods, compared with manually evaluated scores (Section 6.2 and 6.3).

5.1.1 Data Collection Method

SVF tests were recorded in a semi-structured interview with patients, utilising a *HUAWEI nova smartphone* as a recording device, with a sampling rate of $48kHz$ and stereo channels.

Patients were given 1 minute to produce animal exemplars, however, for some subjects that revealed an increased difficulty in this task, tests were halted if the subject insisted he/she couldn't remember additional animal names.

Tests were recorded in two distinct environments, the hospital's infirmary and the surgical clinic (where patients are admitted for elective surgeries). Both these locations are not private in nature, with varied levels of background noise from other patients, doctors, and televisions. Some subjects perform the SVF test accompanied by a partner or relative, and occasionally hints and suggestions are provided by them. The examiner also frequently intervenes during test administration to promote exemplar production.

Manual transcriptions of tests were not provided alongside test recordings, so all animal names produced in the recordings are manually annotated for each subject by the author of this work. Each test recording is heard and manually annotated, and words that are not understood on the first run are repeated until they are.

To calculate the total number of correct words produced for each subject, the scoring rules for SVF tests delineated by Cavaco et al. [21] are utilised.

For each SVF test recording, approximately 5 minutes were necessary to fully transcribe all named animals and calculate the SVF score of total number of correct words produced.

5.1.2 Population Characterisation

In total, 164 recordings from Portuguese subjects were provided by CHUP, each containing the audio recording of the SVF test response of a single subject, with the animal category. From these 164 subjects, 107 (65% of the dataset) are male and 57 (35%) are female. All subjects have over 65 years.

5.2 AutoSpeech Study Data

Following the study described in Chapter 4, a dataset was created from the gathered recordings of PVF and SVF tests performed by each subject, as well as sociodemographic, cognitive, and physical activity information collected to characterise each participant of the study.

These recordings are used in Chapter 6 to assess the performance of the ASR system used (Section 6.1), the effectiveness of automatic feature extraction compared with manual calculation of relevant scores (Section 6.2 and 6.3), and to evaluate the influence of physical activity on cognitive health and VF performance on elderly subjects (Section 6.4).

5.2.1 Data Collection Methods

As described in Chapter 4, PVF and SVF tests were recorded during a semi-structured interview, performed by a psychologist, incorporated as cognitive tests of the MoCA. Tests were performed in succession, with 1 minute given to subjects to produce exemplars for each test sub-type.

Tests were recorded using two smartphones *Samsung SM-J320F* and an *iPhone 6S Plus*, with recording apps that recorded mono channel audio at 44.1kHz, and in similar conditions, with the device set on a table, with its microphone pointed at the subject.

Manual transcriptions are performed *in situ* by the psychologist performing the interview, who wrote down all exemplar production attempts. The total number of correct words was calculated for both tests following the rules delineated by Cavaco et al. [21] from the recorded manual transcripts.

Data collection interviews were performed by one of two psychologists (to aid and quicken data collection), but PVF, SVF, and MoCA scoring were always performed by the same professional.

Recording environments and respective background noise were dependent on the location of data collection. Participants were recruited in three major locations:

Location A an elderly Day Care centre in Porto associated with a non-profit social welfare association, with some subjects institutionalised in the scope of the *Estrutura Residencial Para Pessoas Idosas* (ERPI) programme;

Location B a Senior Social Centre in Porto associated with a private social welfare institution;

Location C Fraunhofer AICOS.

Location A recordings were performed in private meeting rooms or nurse offices, enabling a relatively quiet and calm environment for audio retrieval, although some background noise may be heard due to outside traffic.

Location B recordings were performed in a large meal room, in a common shared space with the remaining centre users. Although separated by a large corridor, the recording environment is not wholly isolated from the commotion and dialogue of the centre users, so background noise levels tend to be higher in these recordings, with distinguishable dialogues between non-test subjects.

Location C recordings were performed in private meetings rooms, enabling a quiet environment for audio recordings. Some instances have some background noise of muffled conversations due to the proximity to other meetings rooms, although these are never dominant in the recordings.

5.2.2 Population Characterisation

Data was collected from a total of 60 elderly subjects, over the period of one month. Three of the interviewed subjects were excluded from this study since they revealed during the interview

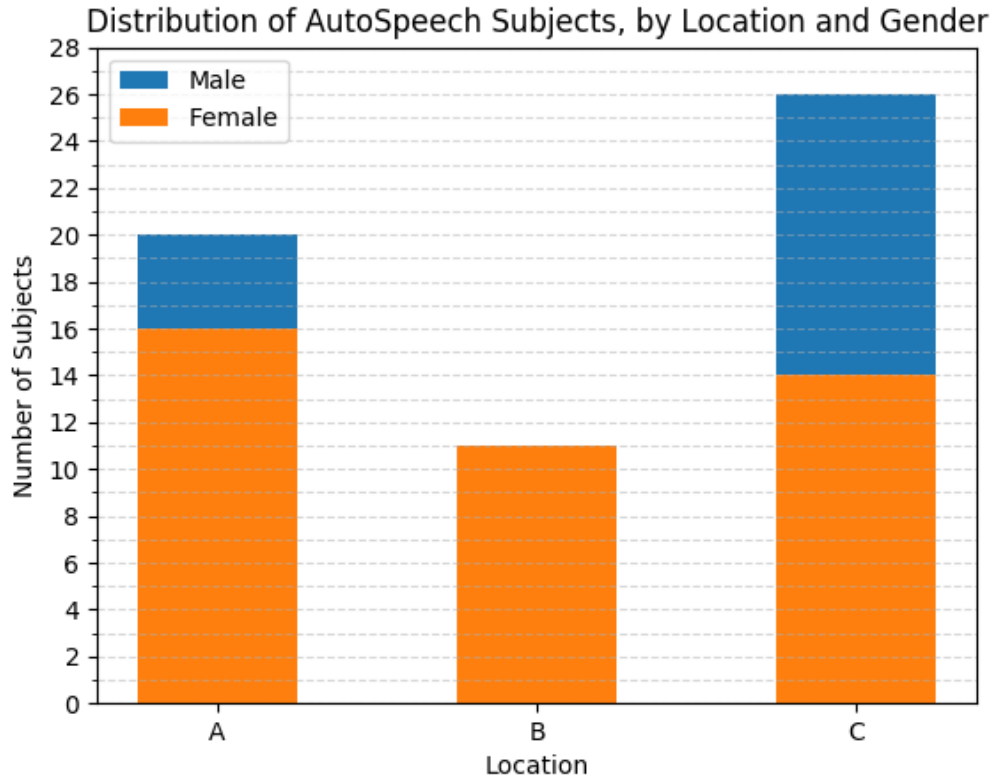


Figure 5.1: Bar Graph of the distribution of collected subjects for the AutoSpeech study, according to location and gender.

to have had conditions that were considered as exclusion factors. From the remaining 57 subjects, 16 were male and 41 were female. All subjects were from the Porto Region.

The distribution of subjects according to data collection location is shown in Figure 5.1.

The age and years of education distribution of the collected population is shown in Figure 5.2, depending on data collection location.

Table 5.1 provides an overview of average, median, and standard deviation of age and level of education according to gender and collection location.

5.3 Manual Feature Extraction

The total number of correct words produced during the administration of a VF test is a relatively easy metric to calculate, by following the rules described in Cavaco et al. [21], and can be done in less than a minute for a single VF test. As such, this is performed alongside transcription, for the case of the CHUP data, and *in situ*, for the case of the AutoSpeech study.

However, manual switching and clustering analysis following the rules delineated by Troyer et al. [114] for large datasets proves to be a daunting, time-consuming task, needing manual assignment of each exemplar produced in a test to a set of predetermined categories.

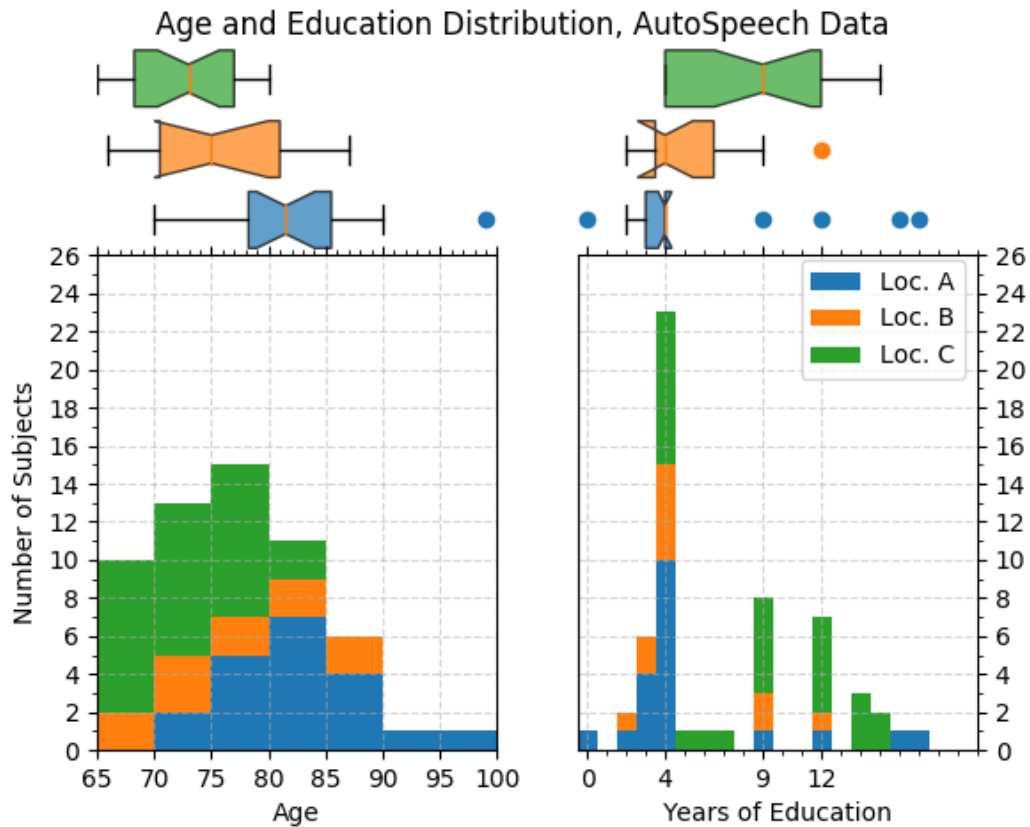


Figure 5.2: Histograms of Age and Years of Education distribution of collected subjects for the AutoSpeech study, according to collection location. Age is captured in 5-year intervals.

On top of each histogram, the box-and-whisker diagram of the population of each location is shown: for each diagram, the box represents the interquartile range (IQR) of the data, the band inside the box the median, and the whiskers extend to $1.5 \cdot IQR$. Outlier data are shown as dots beyond the whiskers. The notches in each box represent the 95% confidence interval for the value of the median.

5.3.1 Semantic Switching and Clustering Features

For the extraction of semantic switching and clustering metrics on SVF tests, a Python script was designed to automatically extract mean semantic cluster size and number of semantic switches based on the work of Troyer et al. [114] and their rules for semantic clustering, based on predetermined animal semantic groups.

For this, all unique animal names produced by all subjects over all collected data were gathered and assigned to one or more animal semantic categories, adapted from the categories available in Troyer et al. [114, Appendix], effectively creating a dataset of animals and respective categories, which is saved as a .csv file. The considered groups were:

Habitat African, Australian, Arctic, Farm, Portuguese, Aquatic;

Human Use Beasts of Burden, Food, Fur, Pets;

Table 5.1: Overview of Age distribution of collected subjects for the AutoSpeech study, according to collection location and gender. For each combination, the average $\pm$ standard deviation, median (Med), and number of subjects (N) are shown.

Location Gender	A	B	C	Total
Male	78.00 $\pm$ 6.60 Med=77.0 N=4	-	73.00 $\pm$ 4.53 Med=73.0 N=12	74.25 $\pm$ 5.56 Med=74.0 N=16
Female	82.56 $\pm$ 6.19 Med=82.0 N=16	75.64 $\pm$ 6.91 Med=75.0 N=11	72.71 $\pm$ 4.68 Med=73.0 N=14	77.34 $\pm$ 7.35 Med=77.0 N=41
Total	81.65 $\pm$ 6.54 Med=81.5 N=20	75.64 $\pm$ 6.91 Med=75.0 N=11	72.85 $\pm$ 4.61 Med=73.0 N=26	76.47 $\pm$ 7.03 Med=76.0 N=57

Zoological Category Bird, Bovine, Canine, Deer, Feline, Fish, Insect, Insectivore, Primate, Rabbit, Reptile/Amphibian, Rodent, Weasel.

From this multilabel classification, each manual transcription of SVF tests provided can be analysed to calculate switching and clustering metrics. Given a manual transcription of a test, each animal name in it is labelled according to its entry on the animal category dataset described above. After performing this for all names in the transcript, the largest contiguous groups of animals that share at least one semantic category are identified, thus emulating the output of manual clustering analysis as described in Troyer et al. [114, Appendix]. From these identified groups, the total number of switches between them and the mean cluster size can be easily calculated.

Semantic Features on PVF tests Since no specific semantic category is given during the administration of PVF tests, produced exemplars are varied and cannot easily be categorised in a limited set of semantic categories, unlike the animal category test.

As such, Ho et al. [54] proposed methodology of semantic analysis (based on the work of [114]) for letter fluency tests is manually applied to the collected PVF test transcripts, by manually determining semantically related words that may be considered as part of a semantic group.

Semantic clusters on PVF tests do not follow strict categorisation rules, but rather an understanding of what are the possible underlying semantic groups a subject is considering when performing the test. Although this makes for a somewhat scorer-dependent metric, inter-rater reliability is very high for these extracted values [114].

5.3.2 Phonemic Switching and Clustering Features

Following the rules described in Troyer et al. [114, Appendix] for the determination of phonemic clusters of VF tests (referred in Subsection 2.2.2), several of these can be automatised to ease feature extraction. Similarly to manual semantic feature extraction, a Python script is created for

Table 5.2: Overview of Years of Education distribution of collected subjects for the AutoSpeech study, according to collection location and gender. For each combination, the average $\pm$ standard deviation, median (Med), and number of subjects (N) are shown.

Location Gender	A	B	C	Total
Male	5.50 $\pm$ 3.84 Med=4.0 N=4	-	10.00 $\pm$ 3.65 Med=12.0 N=12	8.88 $\pm$ 4.18 Med=10.5 N=16
Female	5.38 $\pm$ 4.53 Med=4.0 N=16	5.27 $\pm$ 3.05 Med=4.0 N=11	7.64 $\pm$ 3.99 Med=6.5 N=14	6.12 $\pm$ 4.14 Med=4.0 N=41
Total	5.40 $\pm$ 4.40 Med=4.0 N=20	5.27 $\pm$ 3.05 Med=4.0 N=11	8.73 $\pm$ 4.01 Med=9.0 N=26	6.89 $\pm$ 4.33 Med=4.0 N=57

the semi-automatic extraction of phonemic features. The following methods are designed to be independent of phonemic transcriptions, since that is a part of the automated pipeline.

One of the criteria for grouping words together as phonemic clusters is if they share the first two letters. The determination of first letter similarities between words can be easily achieved by comparing the first two *chars* of their corresponding strings.

Rhyming is another criteria for phonemic clustering, and although it should require the comparison of phonemic transcripts of words to effectively compare the way they sound, this can be avoided. Considering rhyming words as words that share the last three letters is usually enough to identify these pairs, since in Portuguese rhyming words that don't share the last three letters are rare and unlikely to be produced as VF test exemplars. Nonetheless, a manual check is performed to ensure all rhyming words in a test transcript are correctly identified as such.

Homonym or homophone words, due to their identical spelling, cannot be distinguished in an automatic way. The same applies for words that differ by only one sound/phoneme, which may not be evident in their spelling. As such, for these cases, manual annotation of these words is necessary to correctly extract the correct metrics related to phonemic switching and clustering.

These procedures can be applied for both PVF and SVF tests, in a similar fashion. However, a special emphasis is given on the extraction of features for PVF tests, since these are more propitious for the production of homonyms, homophones, words differing in one sound, and unorthodox rhymes. In the case of SVF, for the animal category, homonyms and homophone words do not occur.

Chapter 6

Results and Discussion

The SVF test recordings provided by CHUP and the collected data in the scope of the study described in Chapter 4 allow for an extensive analysis of all the major steps of the proposed pipeline for automatic transcription, processing, and analysis of VF tests.

Figure 6.1 shows how collected data was handled and processed to evaluate the implemented pipeline, namely the Speech-to-Text service used, and the feature extraction steps based on phonemic representations and word embeddings. Moreover, it details what features and data were calculated and analysed together to assess the impact of physical activity on VF performance and cognitive ability of elderly subjects.

Note that since the CHUP data provided only consists of SVF test recordings, with no insightful patient information, the extracted features can only be compared with their corresponding manual metrics, namely the number of correct words produced, mean cluster size, and number of switches. The automatic transcription performed can also be compared to its analogous manual transcript to determine the effectiveness of the Speech-to-Text service used. As such, this data is only used in Sections 6.1, 6.2, and 6.3.

The data gathered in the scope of the AutoSpeech study described in Chapter 4, similar to the data provided by CHUP, is also used to validate the proposed pipeline. In addition, this data is exclusively used to assess the impact of physical activity on VF test performance and cognitive ability of elderly subjects that participated in this study (Section 6.4).

Recordings have three distinct processing paths, based on the type of speech transcription employed and how features are extracted from said transcriptions:

- A fully automatic analysis pathway (leftmost pathway on Figure 6.1), where tests are automatically transcribed using Google Cloud [45] Speech-to-Text, and features extracted using the algorithms described in Section 3.7;
- A manual analysis pathway (middle path), where recordings are manually transcribed, with manual scoring of the number of correct words, and semi-automatic switching and clustering feature extraction, as described in Section 5.3;

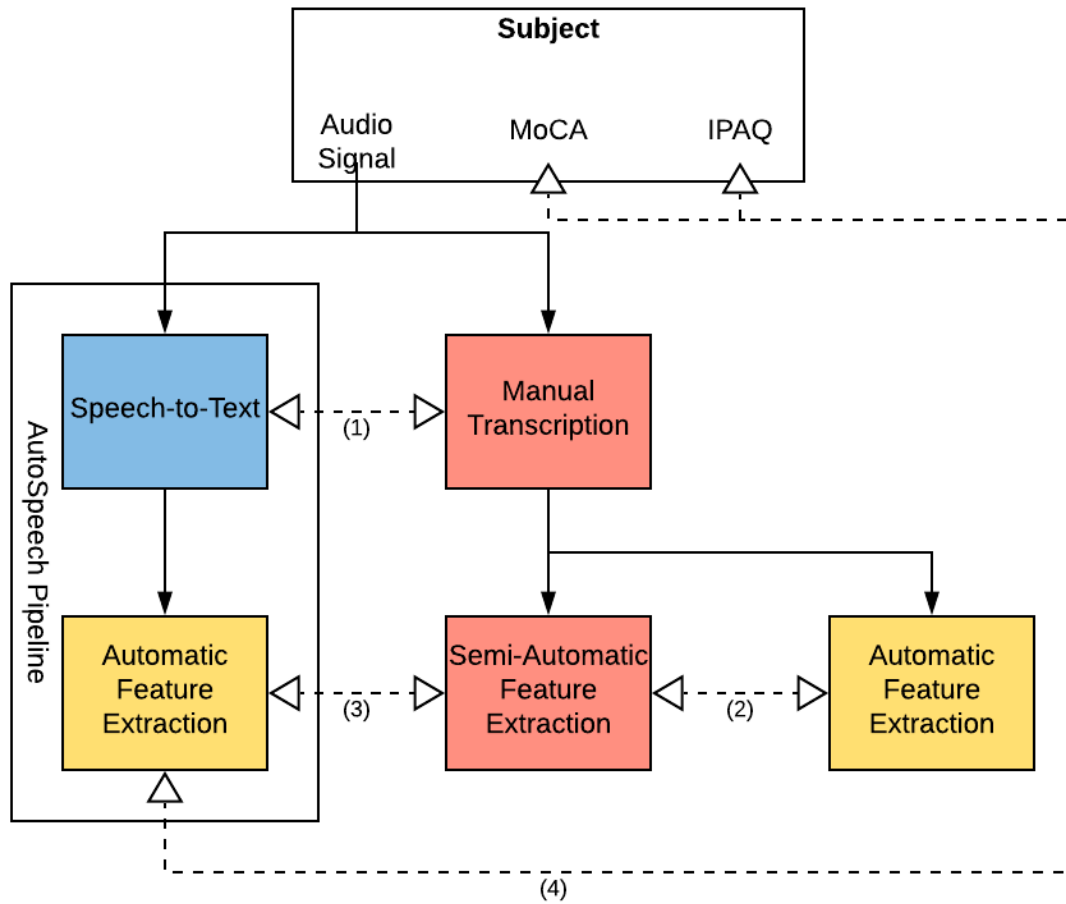


Figure 6.1: Diagram of processing and analysis performed on collected data. Red blocks represent steps in which manual processing is required, whereas yellow blocks indicate processes where analysis is performed automatically. The blue block represents a process which is solely dependent on online services. Dashed lines represent comparison of results between different processes. Four comparisons are performed:

- (6.1) - Speech-to-Text performance (Automatic Vs. Manual Transcription);
- (6.2) - Automatic Feature Extraction performance on Manual Transcripts;
- (6.3) - Automatic Feature Extraction performance on Automatic Transcripts;
- (6.4) - Impact of Physical Activity on Cognition and Verbal Fluency.

- A pathway with manual transcription of VF tests, but with automatic extraction of features (rightmost path), namely number of correct words, and switching and clustering scores.

6.1 Speech-to-Text Performance

To evaluate the performance of ASR systems, the WER is commonly used (Equation 2.1). However, in this case, since the main focus of the recording of VF tests is the capture of words produced as tests words, and there already exist manual transcripts of the tests (performed *in situ* by the psychologist administering the questionnaires of the study), a slightly different metric was used. The

Verbal Fluency Word Error Rate, or VFWER, is shown in Equation 6.1, and translates the overall performance of the ASR when considering only the test exemplars performed [61]. Using the VFWER, it is no longer necessary to perform another manual transcription of audio with all recorded utterances, being only necessary to compare the automatic and manual transcripts of the test.

In Equation 6.1, S_{VFT} represents the number of substitutions of VF test words, D_{VFT} the number of deletions of VF test words, I_{VFT} the number of insertions of VF test words (words that show in the automatic transcript but were not manually considered part of the test), C_{VFT} the number of correctly transcribed test words, and N_{VFT} the total number of words of the VF test. Note that VF test words include possible errors and repetitions.

$$VFWER = \frac{S_{VFT} + D_{VFT} + I_{VFT}}{N_{VFT}} = \frac{S_{VFT} + D_{VFT} + I_{VFT}}{S_{VFT} + D_{VFT} + C_{VFT}} \quad (6.1)$$

The VFWER of each test is automatically calculated by comparing the transcript provided by Google Speech-to-Text, after non-test word removal, with the manual transcript performed by the psychologist, using an adapted version of the algorithm provided by Romanowski [103].

In reality, the VFWER metric will be influenced by not only the overall performance of the ASR service, but also by the effectiveness of the non-test word filtering methods described in Section 3.3: for instance words incorrectly filtered out will be credited as deletions from the manual transcription, and words that are not filtered out but should have as insertions. Nonetheless, the major impacting factor of VFWER is the overall performance of the ASR system used.

The overall distributions of VFWER for the two used datasets are shown in Figure 6.2. The data of the AutoSpeech study is further divided into test type and collection location. For all data, the results are shown for the two used speaker models for ASR: *default* and *command and search*, as detailed in Section 3.2.

The ratio of errors performed during automatic transcription, for each type of test, is shown in Figure 6.3, considering the *default* Speaker Model for transcription.

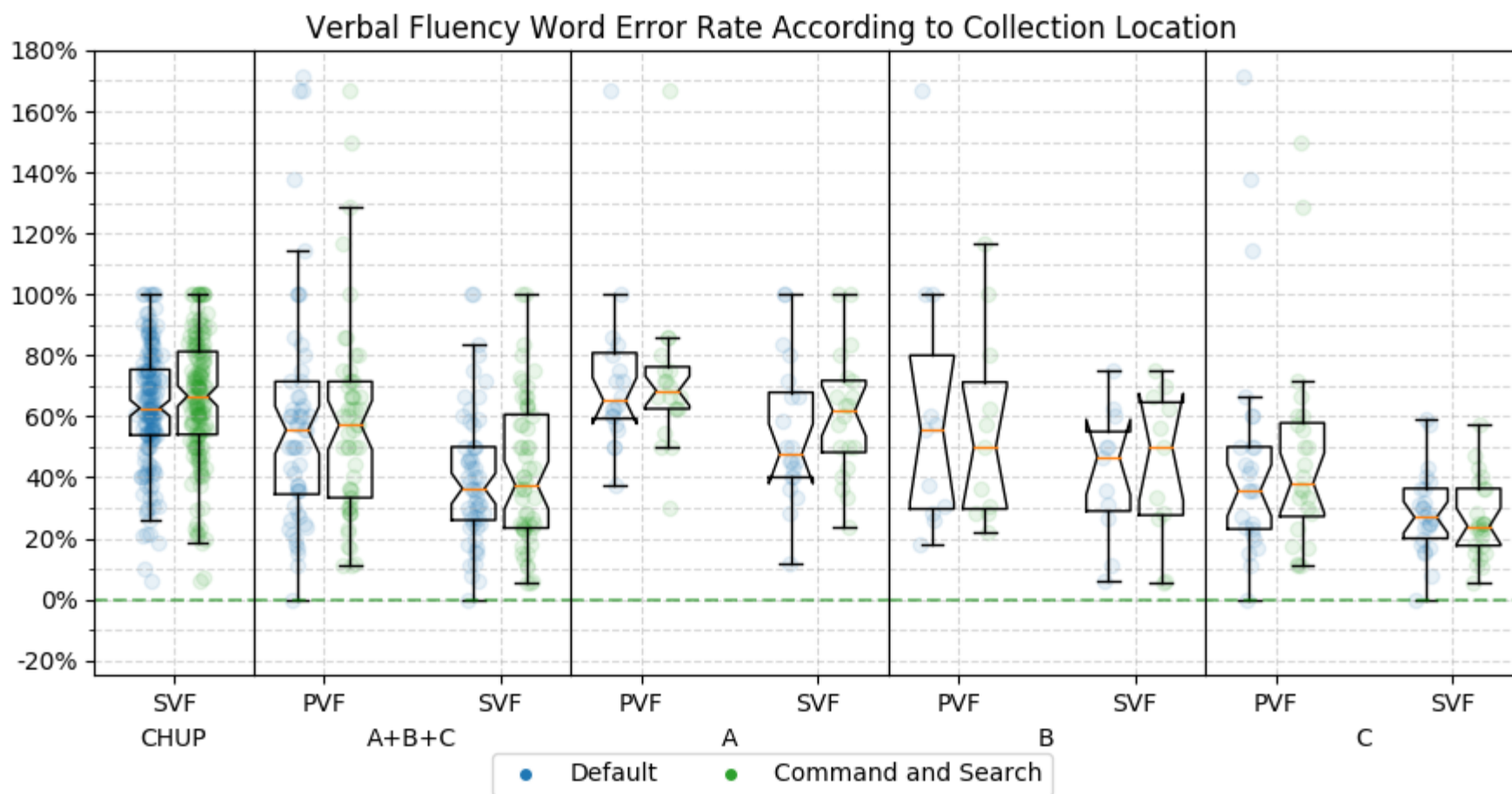


Figure 6.2: Verbal Fluency Word Error Rate according to Test Type, Location, and Speaker Model. Outliers are not shown to ease visualisation.

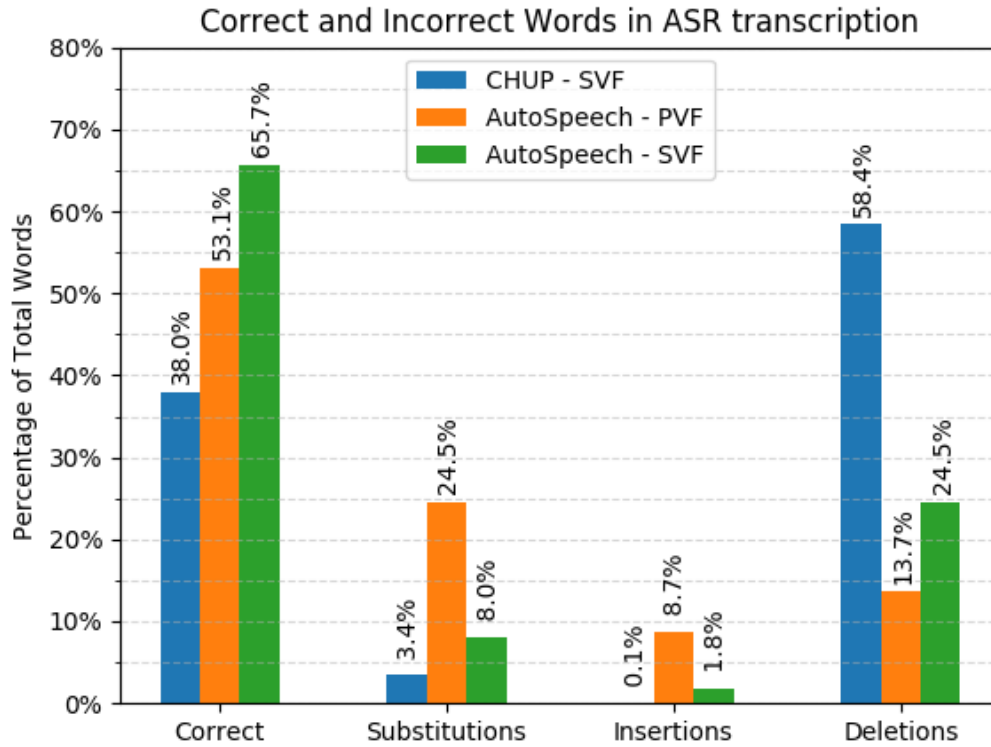


Figure 6.3: Number of Correct and Incorrect Words in Automatic Speech Recognition Transcripts. The *default* Speech Model is used to perform automatic transcription.

6.1.1 Speaker Model

Both available speaker models, *default* and *command and search*, for Speech-to-Text were compared. Since the *command and search model* is specifically trained to deal with short utterances, single words, and voiced commands, it was thought to be the more adequate model for the transcription of VF tests, since these are not composed by grammatically correct or natural phrases, but rather being a list of isolated words.

A paired *t*-test was performed to evaluate which one of the Speaker Models had a higher number of correct words transcribed: if was found that, considering all SVF transcripts (from, the *default* Speaker Model had a lower mean of VFWER ($t = -4.08, p < 0.0001$). For the phonemic tests, the *command and search* model appeared to have the best performance, but results were not statistically significant ($t = 1.14, p = 0.13$).

Since the *default* Speaker Model had lower VFWER overall, this is the model used for the ensuing data analysis.

6.1.2 Test Type

Test type has a very distinct influence in the overall results of automatic transcription and non-test word filtering.

Observing Figure 6.2, it can be seen that PVF tests tend to have a higher VFWER, when compared to SVF in the same location. Furthermore, PVF tests are the only test types where the VFWER can go beyond 100%. By examining Equation 6.1, this can only occur when there is a large number of inserted words in the automated transcript. Indeed, in Figure 6.3, PVF tests have the highest overall percentage of insertions.

A paired t -test was performed to evaluate if there exists a statistical significant difference between the VFWER of PVF and SVF tests of the same subject. These pairs of tests are assumed to be performed in the same conditions, since they are administered in quick succession. It was found that SVF tests have lower VFWER than PVF tests ($t = -1.79, p < 0.05$).

Since PVF tests were only restricted to words starting with the letter *P*, many words that were captured in the automatic transcription that were not produced in the scope of the test that also start with *P* were credited. For instance, a subject that says "*Pois, palavras com P é complicado*", does not produce the words *pois* and *palavras* with the intention of being test exemplars, but, if properly transcribed by the ASR system, will be credited nonetheless by the automatic pipeline.

In addition, approximately 25% of PVF test words are substituted by a similar one, being the test type with the highest percentage of substitutions. This is most likely due to the ASR system incorrectly transcribing *P*-starting words as similar sounding words, many of them also starting with *P*. Since they still start with the required test letter, they are not filtered out from the automatic transcript, thus being credited as substitutions in the calculation of the VFWER. In the case of SVF, it is unlikely that the ASR system erroneously transcribes an animal name as another animal name. Due to the more strict filtering rules for this type of VF test, this leads to an increased number in deletions, as can be seen in Figure 6.3.

6.1.3 Recording Environment

Welsh t -tests were performed to compare the influence of recording location on the VFWER of PVF and SVF tests. Statistically significant differences were found between SVF tests performed in: CHUP and A ($t = 1.78, p < 0.05$), CHUP and B ($t = 2.95, p < 0.01$), CHUP and C ($t = 12.22, p < 0.0001$), A and C ($t = 3.83, p < 0.001$), and B and C ($t = 2.36, p < 0.05$). No statistically significant differences were found for the VFWER of PVF tests in locations A, B, and C. From these results, it can be said that location C provided the lowest average VFWER of all considered locations, and tests performed in CHUP the highest.

Recording environment plays a definitive role in the quality of automatic transcriptions obtained from the ASR service used. Locations where there is a lot of background noise, such as the data collected in CHUP, tend to have higher levels of VFWER. Quiet recording environments,

such as location C, prove to be better for automatic transcription quality. CHUP data has the highest percentage of deletions: almost 60% of words said by subjects in the CHUP dataset are not correctly transcribed, and are filtered out.

Although some noise reduction is performed in Google Cloud [45] Speech-to-Text service, it is recommended to avoid having multiple people talking at the same time, and to perform audio recordings with the microphone as close as possible to the subject. The overall performance of an ASR system is thus explained by the level of background noise present in the recorded audio signals. Confounding noise, such as televisions, other people talking nearby, echoes, and reverberations affect the quality of the transcription, making words produced by a VF test subject less perceptible to the ASR algorithm, increasing VFWER.

Another possible factor that may affect VFWER is subject diction and accent. All subjects were from the region of Porto, and the accent of certain uttered words was not favourable for their adequate transcription by the ASR service. For instance, the word *leão* was commonly transcribed as *leon*, being afterwards filtered out from the transcript, since it was not considered an animal. Furthermore, a subject's diction tends to worsen with age progression, which can also difficult automatic transcription.

6.2 Automatic Feature Extraction Performance on Manual Transcripts

One of the main advantages of automating VF tests is the possibility of quickly determining switching and clustering metrics. In the proposed pipeline, this is performed by using semantic and phonemic similarity measures between words and grouping them together accordingly, as described in Section 3.7.

To evaluate the effectiveness of the developed algorithms for the determination of switching and clustering metrics, the results of these on manual transcripts are compared to the corresponding metrics extracted manually, following the rules in [114, Appendix]. Automatic feature extraction is done on manual transcripts to fully avoid any possible Speech-to-Text transcriptions errors that may influence the results.

Three distinct grouping strategies are considered for feature extraction: Chaining, Clustering, and DBSCAN. To assess their efficacy, features automatically extracted from manual manuscripts of tests were compared to their analogous semi-automatic features, using Spearman's ρ as a measure of correlation.

6.2.1 Number of Correct Words

The total number of correct words automatically extracted from manual transcripts were compared with the number of correct words manually determined by the psychologist.

The relative error rate (Equation 6.2) of the number of correct words automatically extracted, when compared with its manual counterpart, was determined for both PVF and SVF tests, to assess

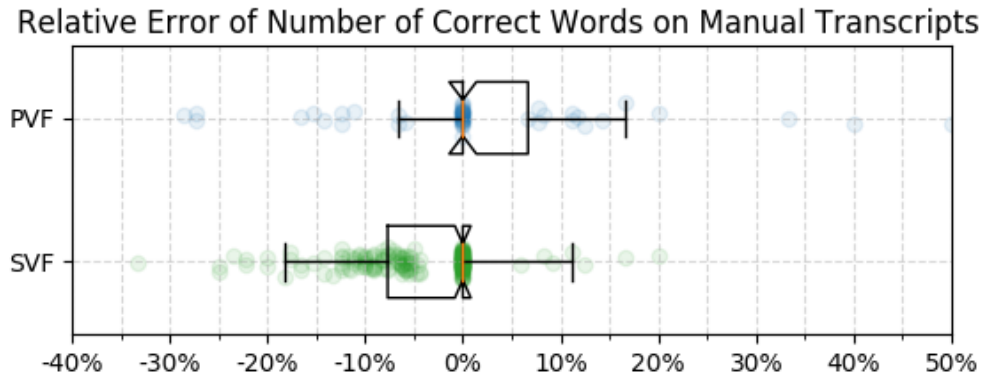


Figure 6.4: Relative error of the number of correct words automatically extracted from manual transcripts, from AutoSpeech Study and CHUP data.

the level of accuracy of the implemented word filtering algorithms. Figure 6.4 shows the relative error distribution, for both AutoSpeech Study and CHUP data.

In Equation 6.2, x represents the measured value, and x_o the true value: in this particular case, x corresponds to the number of correct words determine automatically, and x_o the number of correct words as determined by the psychologist scoring the VF tests. The relative error can be positive or negative: positive values indicate an overestimation of correct words, and negative values an underestimation.

$$\text{Relative Error Rate}(\%) = \frac{x - x_o}{x_o} \cdot 100\% \quad (6.2)$$

For PVF tests, a strong correlation was obtained ($\rho = 0.97, p < 0.0001$) between the automatic and manually determined number of correctly produced *P*-words. For SVF tests, a very strong correlation was obtained ($\rho = 0.98, p < 0.0001$) between the automatic and manually determined number of correct exemplars produced in the animal category. These values are also shown in Figure 6.6.

These strong correlations and low error rates indicate the proposed word filtering algorithms of non-test word removal and incorrect word removal, described in Sections 3.3 and 3.6, are effective. The proposed non-test word filtering algorithm can correctly identify what words respect the rules of the test at hand. Moreover, the incorrect word filtering algorithm used can identify similar words that are not to be credited, following the rules delineated by Cavaco et al. [21], such as superordinate animal categories, and variations on the same word root.

6.2.2 Embeddings Cosine Similarity as a Measure of Semantic Similarity

A question that arose during the development of this dissertation was the validity of considering the cosine similarity metric between the embeddings of two words as a *de facto* measure of their semantic similarity. Since the *fastText* embeddings were calculated based on written Portuguese Wikipedia documents [47], the relations found may not effectively encapsulate the brain's semantic organisation and categorisation of words. Furthermore, since VF tests are an oral expression

of language, and these embeddings rely instead on the written language, these two modes of communication may not be entirely similar in terms of their word relations.

To evaluate if two words with similar embeddings are semantically similar, and thus, if they are more quickly accessed by a subject's brain when performing a PVF or SVF, all uttered pairs of words produced in succession on the CHUP data were considered and their Pairing Frequency was compared against their embeddings' cosine similarity (Equation 3.1). Pairing Frequency, or PF, is a measure of how often two words tend to be paired together in a VF test dataset, as described in Equation 6.3, with A_1 and A_2 being two words present in the dataset of tests D .

$$PF(A_1, A_2) = \frac{|t \in D : \{A_1, A_2\} \subset t|}{|t \in D : (A_1 \in t) \vee (A_2 \in t)|} \quad (6.3)$$

In other words, the PF of two words is the ratio between the number of tests where these two words appear paired (one after the other) to the total number of tests where the words are produced. Higher values indicate a higher tendency of a pair of words to be produced in rapid succession: a word pair with a PF of 1 indicates those words are always produced in a sequence. Note that for the sake of pairing frequency calculation, no distinction is made regarding the order of a pair production, i.e., *cão-gato* is considered the same pair as *gato-cão*.

The resulting scatter graph is shown in Figure 6.5.

It can be seen that the great majority of produced pairs have a low PF, with a great number of points on the left side of the scatter plot. However, discarding these points, it can be seen that higher cosine similarities are associated with higher pairing frequencies (Spearman's $\rho = 0.43, p < 0.0001$), particularly for the case of the pair *cão-gato*. These words are ubiquitous over the whole dataset, with more than 50% of subjects producing these words in quick succession, and have the highest cosine similarity of all considered pairs.

Note that the pairing frequency only takes into account words that are produced one after the other in VF tests, which can be misleading. Many subjects produced closely related groups with a large number of elements, but only sequential words are considered for PF. Increasing the analysis to a larger word neighbourhood increases the scope and size that these larger groups can be detected, but may also introduce noise to the data, since associations between words not in the same semantic group would increase.

Nonetheless, the results shown in Figure 6.5 indicate that, in general, the higher the embedding similarity between words, the more likely they are to be paired together in a VF test, which indicates they are probably more semantically related, with the human brain connecting the two more quickly. The selected strategy of using *fastText* word embeddings seems to be effective for capturing semantic relatedness between pairs of words.

6.2.3 Switching and Clustering Features

The number of switches, mean cluster size, and maximum cluster size for both phonemic and semantic clusters, as defined by Troyer et al. [114], were compared with their automatically extracted counterparts, based on the methods described in Section 3.7.

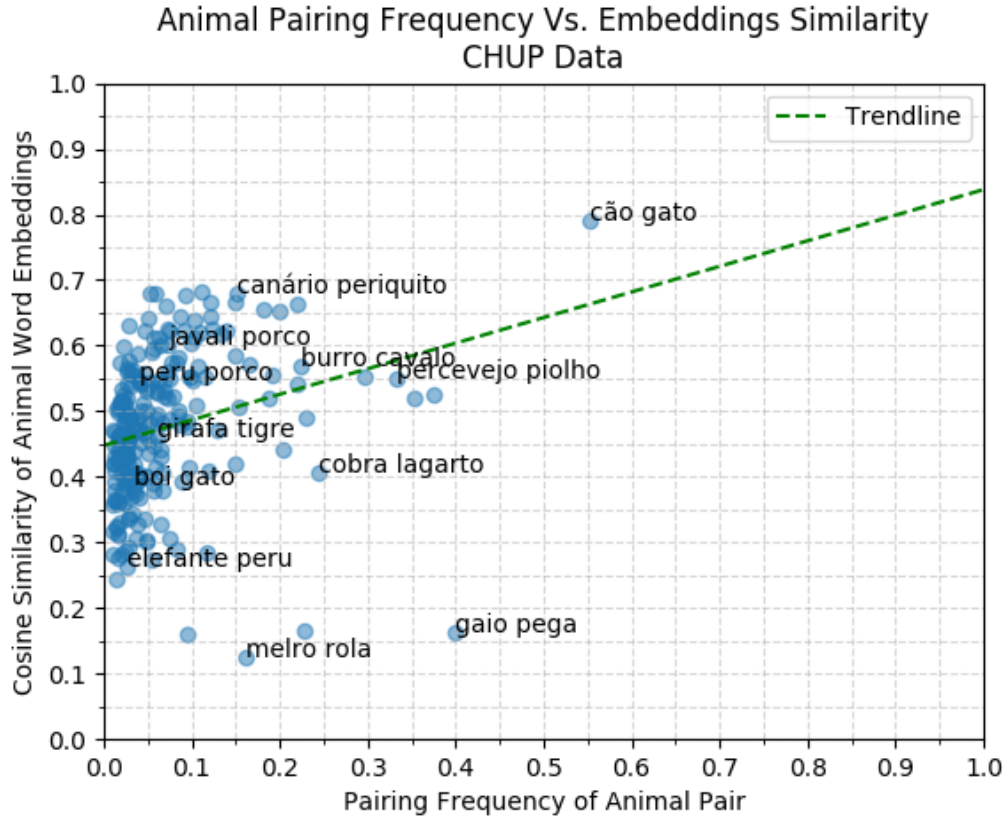


Figure 6.5: Comparison of pairing frequency (Equation 6.3) of animal names with their embeddings' cosine similarity, for SVF tests of the CHUP dataset.

Temporal feature comparison was not performed, since manual transcription of VF tests did not contemplate annotation of word production times.

The three available grouping models (chaining, clustering, and DBSCAN) were compared for both phonemic and semantic feature extraction, on PVF (AutoSpeech Study data) and SVF (both AutoSpeech and CHUP data) tests. Figure 6.6 presents the calculated Pearson's correlation coefficients, with associated significance values.

It can be seen that all grouping models have distinct performances in extracting phonemic and semantic features from both PVF and SVF tests. The total number of switches presents a consistent high ρ for all possible feature types and grouping models, indicating the automatically extracted values are similar to those calculated manually.

For phonemic and semantic features, it is advisable to perform extraction using the same grouping model for both, to maintain generality and reduce pipeline complexity. Comparing the results obtained, the clustering grouping strategy was chosen since it presented the best correlations overall, considering all extracted features, feature types, and analysed tests. DBSCAN was the second best choice, since it presents very high ρ for most of the features extracted; however, for the case of semantic features of SVF tests, performance is underwhelming, particularly considering mean and maximum cluster size values with low correlation coefficients. The chaining

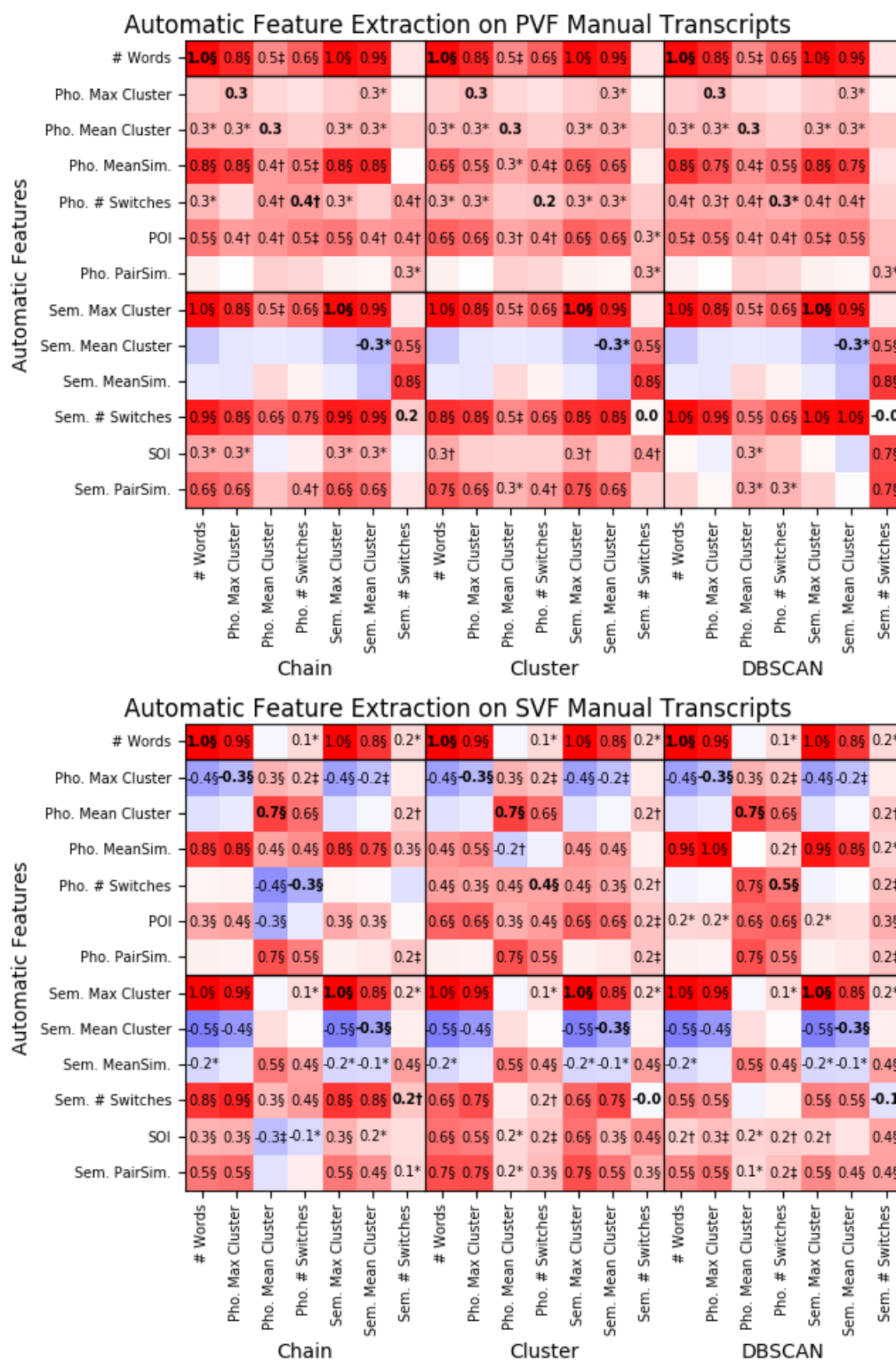


Figure 6.6: Spearman's ρ between automatically and manually extracted features from manual transcripts of PVF (top) and SVF (bottom) tests. Rows indicate automatically extracted features (phonemic and semantic) and columns manually extracted features. Automatic features are extracted using the following grouping models: chaining (left), clustering (middle), and DBSCAN (right). ρ values with no statistical significance are omitted to ease visualisation. Bold number indicate the correlation for pairs of analogous manual-automatic features. Red cells indicate positive correlation, blue cells negative correlation, and white cells no correlation.

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$, § $p < 0.0001$

method, although apparently promising during development, shows to be the weakest option, with low correlation for semantic features, and even negative values of ρ for phonemic features on SVF tests.

Overall, correlation results are somewhat underwhelming, considering that extraction is performed on manual transcripts, thus eliminating the influence of incorrect word transcription. Even so, the determined correlation coefficients are lower than in state-of-the-art systems, such as the one described by König et al. [61], with values for ρ over 0.8 for number of switches and mean cluster size on automatic SVF transcripts.

Nonetheless, visual analysis of the obtained phonemic and semantic clusters show these algorithms are effective in determining and grouping together semantically related and similar sounding words, for both PVF and SVF tests. In addition, Troyer et al. [114] scoring rules can be somewhat limiting for both phonemic and semantic features, relying too much on precise grouping rules, that may not always credit *de facto* semantic or phonemic associations because they do not respect these rules.

For instance, the words "*penso*" and "*prensa*", being effectively phonemically similar, would not be clustered together following Troyer et al. [114, Appendix] rule, since they do not rhyme, nor do they start with the same two letters. Semantic connections may not always be captured using predetermined animal categories as well: "*lebre*" and "*tartaruga*" are two commonly associated animals due to both being common fable characters, but would not be clustered together considered the aforementioned rules.

The use of more generalised algorithms for the determination of phonemic and semantic clusters, such as the use of word embeddings or phonemic transcriptions comparisons, bypasses the creation of a rule set, thus generalising and approximating the computer notion of phonemically/semantically similar words to human-like levels [86, 97, 121].

6.3 Automatic Feature Extraction Performance on Automatic Transcripts

Considering the previous results, the final pipeline for automatic analysis of VF tests uses the *default* Speaker Model for automatic transcription of recordings, and the clustering grouping model for automatic extraction of phonemic and semantic features.

Automatic analysis of a single VF test, from transcription to feature extraction, lasts approximately 30 seconds. This is a significant improvement compared to manual feature extraction, which tends to last 5 minutes per recording, from transcription, to correct word count, to the determination of switching and clustering metrics.

Automatically extracted features were compared to their corresponding manual analogues, to determine correlation and error rate. Since transcription quality was shown to be influenced by the recording location, tests are divided into their respective data collection locations.

6.3.1 Number of Correct Words

Since the number of correct words is the most important feature extracted from VF tests [16], being the target of Portuguese VF normative studies [21], it is ideal for the automatic analogue of this feature to be as similar as possible to the manually determined one, together with adequate positive correlation.

The correct word error rate (Equation 6.2) is determined for each PVF and SVF test, according to collection location. The resulting relative error distributions can be seen in Figure 6.7.

It can be seen that the relative error rate is mostly negative for all locations, indicating the automatically calculated value is normally lower than the real value. Many uses of VF tests rely on thresholding the total number of correct words produced, such as in the MoCA [80], so the underestimation of this value can be troublesome, leading to incorrect results.

Errors are of a lesser magnitude for location C, most likely due to better automatic transcriptions, with lower VFWER, as can be seen in Figure 6.2: the lower number of deletions and substitutions on automatic transcripts allows for the correct identification and counting of all produced exemplars.

SVF tests tend to have higher absolute correct word error rates than PVF tests, on the same recording conditions. As discussed in Section 6.1, SVF tests have a higher VFWER due to stricter filtering rules, in conjunction with low probability of substitutions generating a valid animal name for this test.

Spearman's ρ were calculated to assess the level of correlation between automatic and manual number of correct words, depending on recording location and test type.

For PVF tests, location A recordings did not provide statistically relevant correlation for this metric ($\rho = 0.17, p = 0.47$), whereas location B ($\rho = 0.96, p < 0.0001$) and location C ($\rho = 0.80, p < 0.0001$) did, with large positive values for ρ . SVF tests all showed statistically significant positive correlations: data collected in CHUP ($\rho = 0.58, p < 0.0001$), location A ($\rho = 0.87, p < 0.0001$), B ($\rho = 0.97, p < 0.0001$), and location C ($\rho = 0.84, p < 0.0001$) all showed that automatic determined counts of correct words are in fact correlated with their true values. These values are also present in Figure 6.8 and 6.9.

Positive correlations were achieved for the automatic count of correct words, but values of ρ are better for quieter collection locations, with low transcription VFWER. This further corroborates the necessity of performing audio recordings in quiet environments to ensure the best possible transcription.

Nonetheless, even though it may not result in exactly the same value obtained manually, these good correlation values indicate that automatic counts of correct words are very reliable to distinguish subjects with different performances, when recorded in similar conditions.

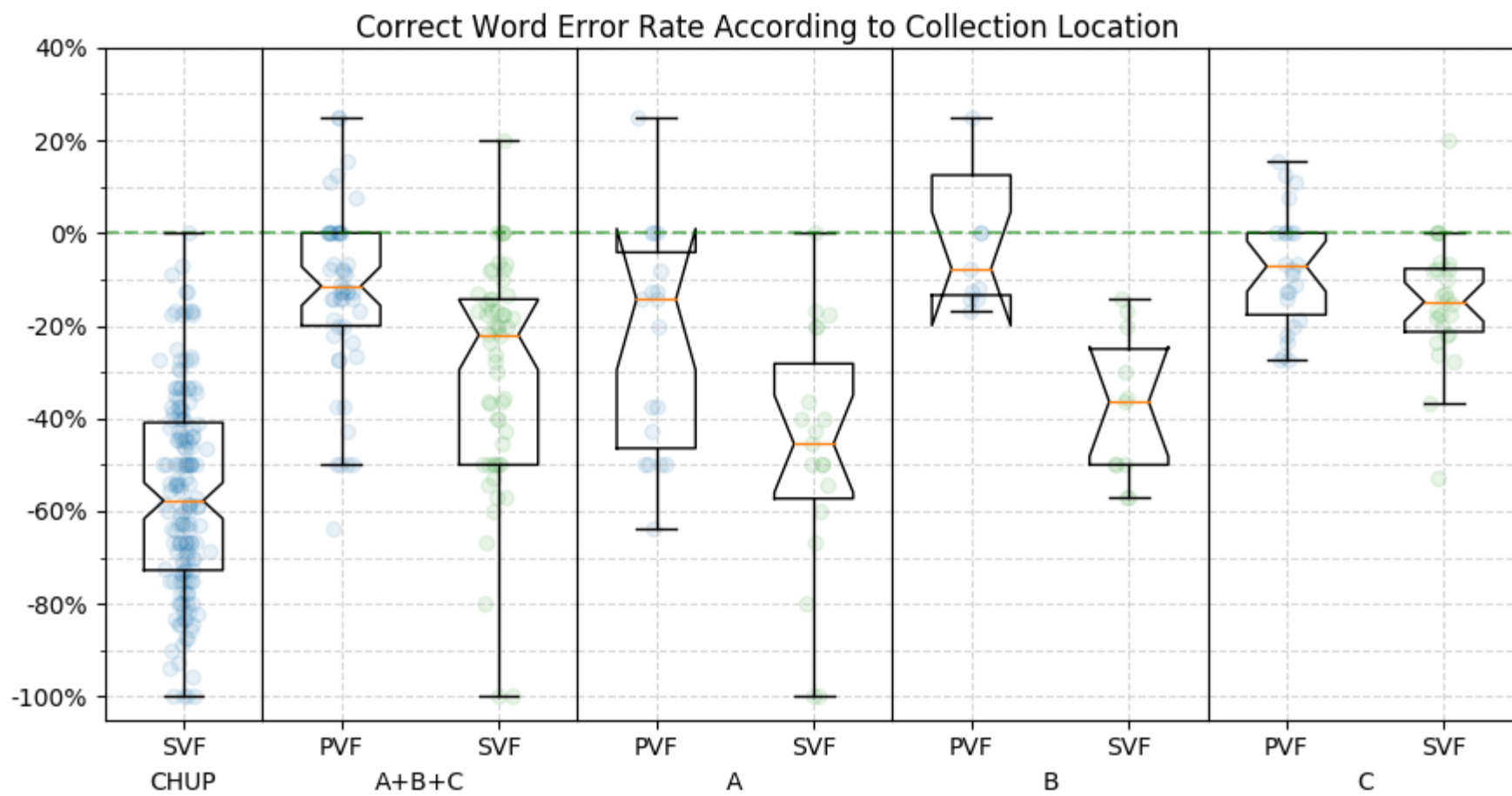


Figure 6.7: Number of Correct Words Error Rate according to Test Type and Location.

6.3.2 Switching and Clustering Features

Switching and Clustering metrics for temporal, phonemic, and semantic features were compared with their manual analogues. For phonemic and semantic feature extraction, the clustering grouping model was used.

Figures 6.8 and 6.9 present correlation matrices between automatic and manual switching and clustering features, for PVF and SVF tests, respectively.

For PVF tests, statistically significant correlation was only found for the maximum semantic cluster size, on location C. All other features did not have a clear enough correlation with their manual analogues, for the number of subjects of each location.

For SVF tests, statistically significant positive correlations were found for most of the manual-automatic feature pairs, particularly for tests provided by CHUP, location A, and location B. Obtained correlations were lower for CHUP data, most likely due to variation introduced by incorrect automatic transcription.

The p values obtained for each location not only depend on the level of VFWER observed for each location, but also on the way location subjects tackled each type of VF test. Since each data collection location is different on several parameters, such as age, level of education, and overall level of physical activity, subjects may adopt distinct strategies for exemplar production. These differences are shown and discussed in Section 6.4.

Overall, correlation of manual and automatic switching and clustering features is inferior to the correlation observed for the total number of words in a VF test, many times not being statistically significant to extract relevant conclusions from data. As discussed in Section 6.2, this may arise due to the slightly different approaches used for manual phonemic and semantic clustering, based on strict rule sets, when compared to the more generalised automatic approaches by word embeddings and phonemic transcript similarities.

Nonetheless, automatic transcription and feature extraction performance is sufficiently reliable to create a new VF analysis framework, with the results of the implemented pipeline. This automatic framework, where each subject can be characterised by 40 features (20 for each VF test type), is used in Section 6.4 to evaluate the different performances and test strategies used by subjects, depending on their age, education, and overall level of physical activity.

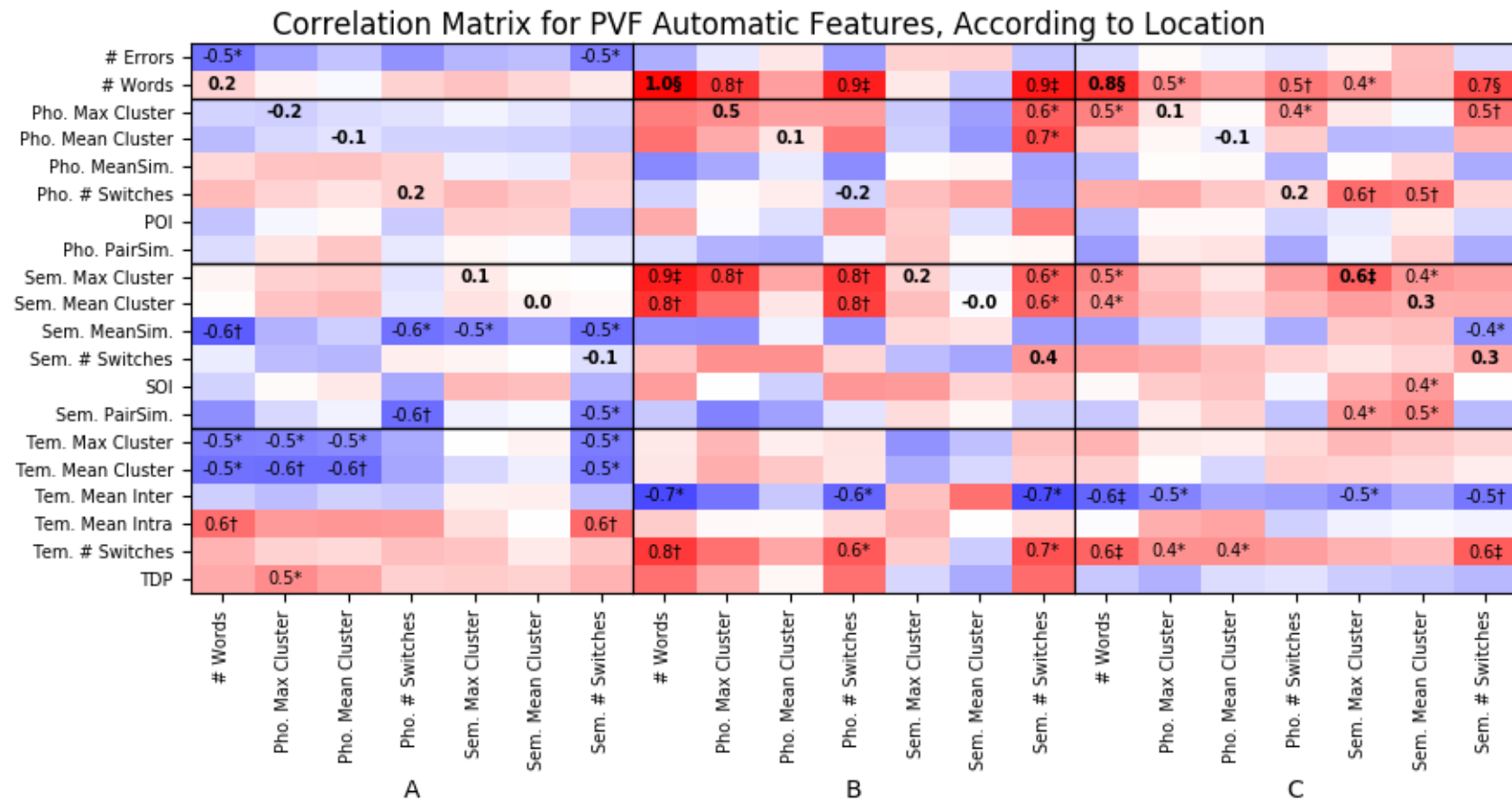


Figure 6.8: Spearman's ρ between automatically and manually extracted features from automatic transcripts of PVF tests. Rows indicate automatically extracted features (number of errors, number of correct words, phonemic, semantic, and temporal features) and columns manually extracted features. ρ values with no statistical significance are omitted to ease visualisation. Bold number indicate the correlation for pairs of analogous manual-automatic features. Red cells indicate positive correlation, blue cells negative correlation, and white cells no correlation.

* $p < 0.05$, $^{\dagger} p < 0.01$, $^{\ddagger} p < 0.001$, $^{\S} p < 0.0001$

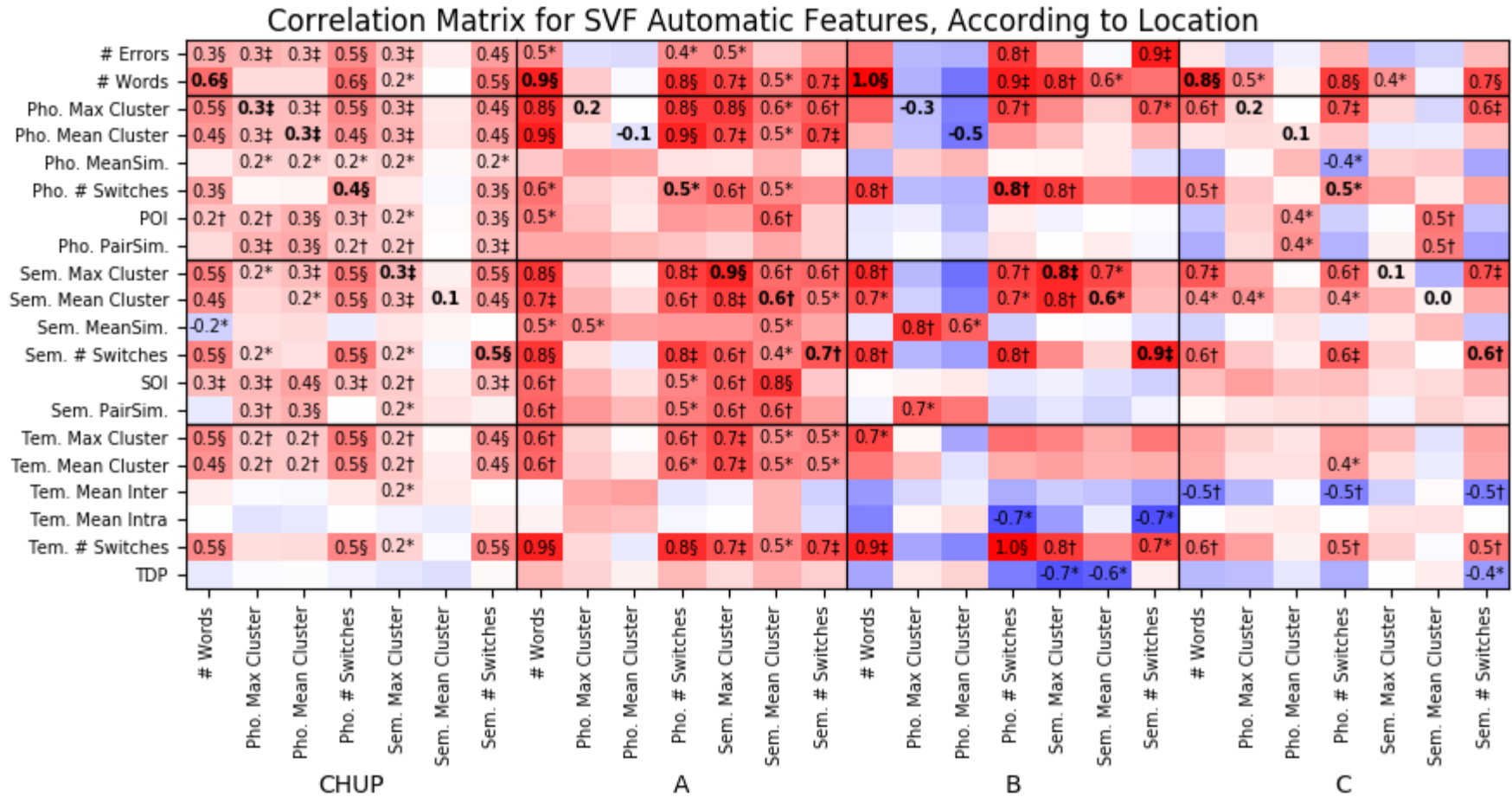


Figure 6.9: Spearman's ρ between automatically and manually extracted features from automatic transcripts of SVF tests. Rows indicate automatically extracted features (number of errors, number of correct words, phonemic, semantic, and temporal features) and columns manually extracted features. ρ values with no statistical significance are omitted to ease visualisation. Bold number indicate the correlation for pairs of analogous manual-automatic features. Red cells indicate positive correlation, blue cells negative correlation, and white cells no correlation.

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$, § $p < 0.0001$

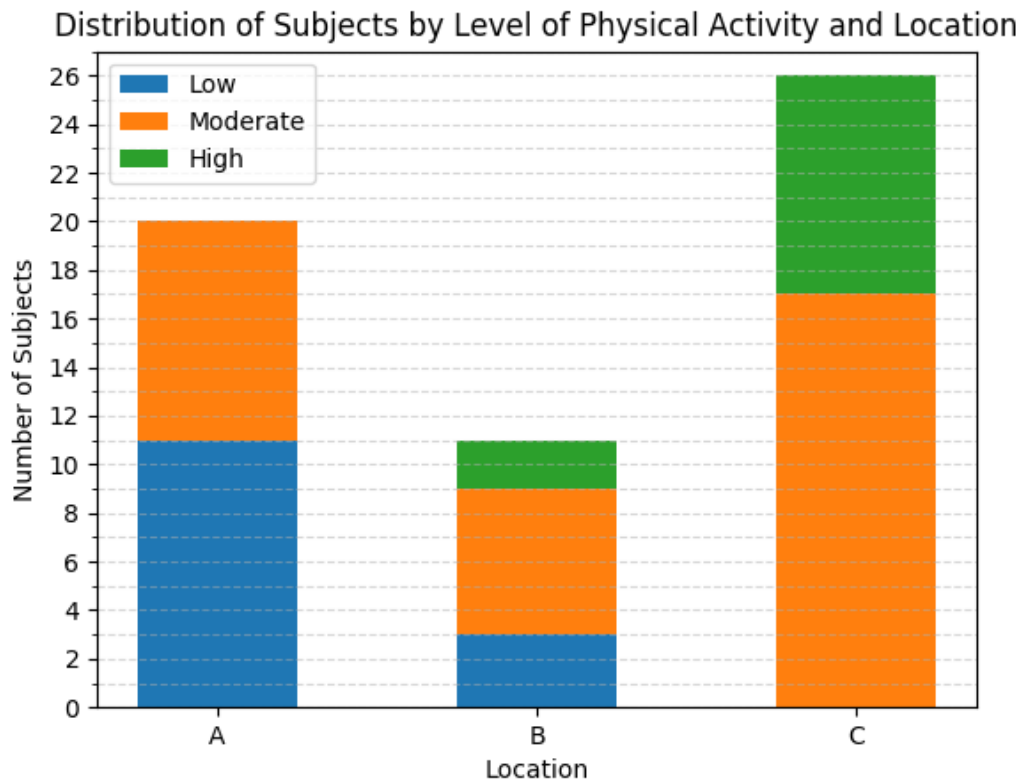


Figure 6.10: Distribution of subjects according to their level of physical activity and data collection location.

6.4 Impact of Physical Activity of Elderly on Cognition and Verbal Fluency

The automatic VF analysis pipeline was used to extract 40 relevant features, namely number of correct words and errors produced, and temporal, phonemic, and semantic switching and clustering features, from both PVF and SVF tests performed by subjects of the AutoSpeech study described in Chapter 4.

Subjects are characterised by their age, level of education (in years), and obtained overall physical activity level, ranging from low, moderate, or high, obtained by processing the answer to the short-IPAQ questionnaire, following the rules delineated by IPAQ Research [58]. The total universe of 57 subjects is thus divided into three subpopulations, depending of the level of physical activity determined: low physical activity ($N = 14$), moderate physical activity ($N = 32$), and high physical activity ($N = 11$). Figure 6.10 shows the distribution of subjects by level of physical activity and data collection location.

As could be expected, subjects from location A were overall less active than location B and C. Subjects from A, an elderly Day Care centre, did present an apparent higher level of dependency, with less physical and social activity, particularly for institutionalised subjects in the scope

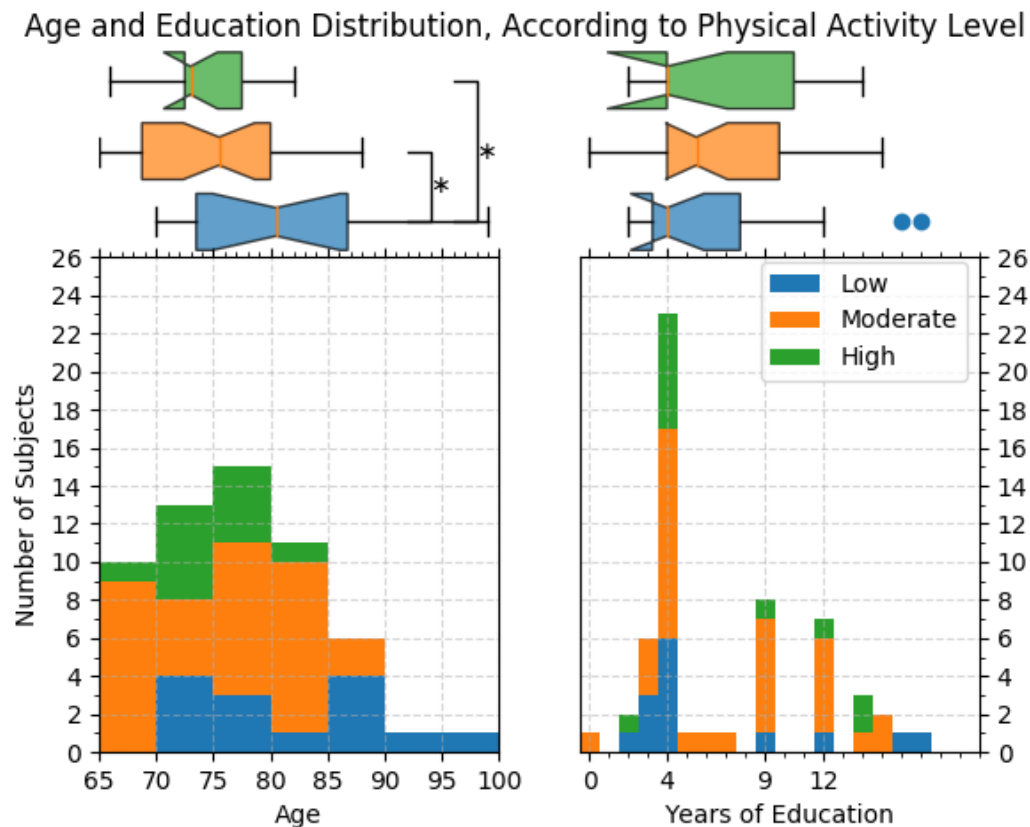


Figure 6.11: Age and education distribution according to physical activity levels. Age is grouped in 5-year intervals. Brackets indicate statistical significance of the Mann-Whitney U test between populations of different levels of physical activity.

* $p < 0.05$

of the ERPI programme. Subjects from location B were generally slightly more active than subjects from location A, being more independent on their daily lives, but some still registered low levels of activity, where their daily scope of activities would be limited to being at home or being in the Senior Social Centre. Subjects whose data was collected in location C were overall still quite active, participating in several activities and hobbies, with a deeper connection to modern technology, and performing frequent physical exercise over the course of one week.

Since age and level of educations are relevant subject factors that impact VF performance, for both total number of words [21, 82, 111] and switching and clustering features [113], each subpopulation is characterised and compared according to these factors. Figure 6.11 shows the resulting histogram distributions for age and education.

To perform comparisons between the obtained subpopulations according to level of physical activity, the Mann-Whitney U statistical test is used [72], which is a non-parametric test to evaluate the null hypothesis that a distribution is not stochastically greater than another independent distribution. This is similar to Student's t -test, but does not assume distribution normality, which is often the case for the distributions evaluated here.

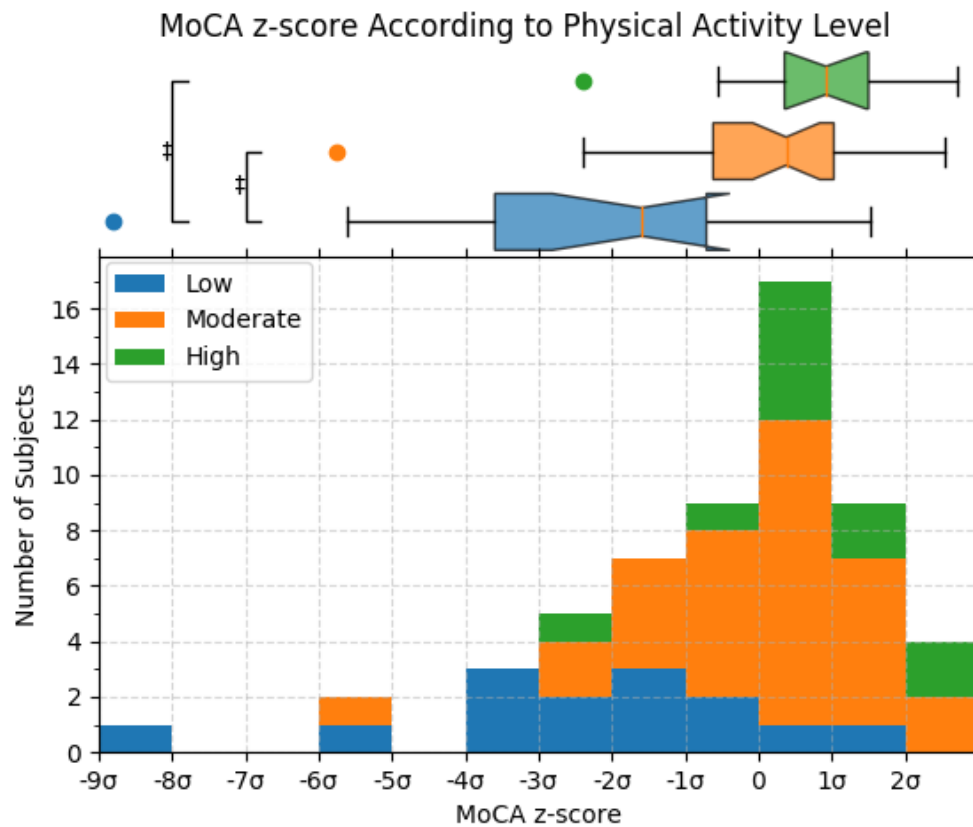


Figure 6.12: MoCA z -scores according to physical activity level. MoCA z -scores are calculated with Freitas et al. [39, Table 5]. Brackets indicate statistical significance of the Mann-Whitney U test between MoCA z -scores of populations of different levels of physical activity.

$^{\ddagger}p < 0.001$

No statistically significant difference was observed for the level of education for each subpopulation, but the low physical activity subpopulation was shown to be older than the moderate ($U = 138.0, p < 0.05$) and high ($U = 42.5, p < 0.05$) activity subpopulations: the decrease in physical ability associated with ageing, and the increased incidence of debilitating physical complications (such as back problems, herniated disks, etc.) makes physical activity more difficult to perform [110].

MoCA total scores were also compared between the different subpopulations. Since this test also has dependency on age and level of education, total scores are normalised to z -scores according to normative data for the Portuguese population in Freitas et al. [39, Table 5]. The resulting distributions can be seen in Figure 6.12

No statistically significant difference can be observed for MoCA scores between moderate and high activity subpopulations. The low activity subpopulation tends to have a lower z -score than both moderate ($U = 89.5, p < 0.001$) and high ($U = 20.0, p < 0.001$) activity subpopulations. Subjects with lower level of physical activity appear to be more cognitive declined than other populations, even when age and education factors are taken into account. However, several other

factors other than physical activity can influence these results.

Age can still impact these results, since the normative study of Freitas [38] only considers one age group for elderly subjects over the age of 65. This group is then further subdivided according to years of education. Since the age range for this study is large, with subject age ranging from 65 to 99 years, categorisation of MoCA results into a single age group may be insufficient to correctly normalise scores and adjust cognitive ability for older subjects. As such, age effects may still be noticeable for the obtained z -scores, considering that the low physical activity subpopulation was already determined to be generally older than the other groups.

Subjects from the low activity subpopulation were also mostly from location A, which were more dependent elderly subjects, relying more on the care provided by social assistants, being more sedentary and with lower social and cognitive activity. It is possible that subjects from this location (and others) may already be developing NCDs, without these being formally diagnosed, since performance in the MoCA was very low for some. Impaired cognitive ability is more frequent as age progresses, and so is the incidence of NCDs [66, 68, 71, 122], so it is probable that this subpopulation performance may be hampered not only by age-associated cognitive decline, but also from yet undiagnosed dementia cases.

Since automatic transcription performance of VF tests is very dependant on the overall quality of the recorded audio signal, which needs to have as little background noise as possible, VFWER of VF tests for each population were determined. VFWER can impact negatively features extracted downstream, so it is yet another factor that may impact specific subpopulation performances. Figure 6.13 shows the resulting distributions of VFWER for PVF and SVF tests, according to calculated physical activity levels.

The VFWER of VF tests of subjects with low physical activity are shown to be statistically larger than both moderate (PVF: $U = 331.0, p < 0.01$; SVF: $U = 320.0, p < 0.01$) and high (PVF: $U = 119.0, p < 0.01$; SVF: $U = 130.0, p < 0.001$) activity subpopulations. No significant difference was found for VFWER between moderate and high physical activity subjects (PVF: $U = 192.5, p = 0.33$; SVF: $U = 206.5, p = 0.20$).

VFWER for the low activity subpopulation is considerably higher than the other groups, which will greatly impact extracted features from subjects of this group, when compared to the others. Features are underestimated when compared to their manual counterparts, which further influences the results observed for this group. The cause of such a dramatic difference in VFWER when compared to other groups is uncertain, since most of the recordings of this group happened in location A, where background noises were not prevalent. A possible cause may be subject diction, pronunciation and accent, whose effects were more noticeable with older subjects, and can hamper ASR systems.

6.4.1 Discrimination of Manual Features According to Physical Activity Levels

To avoid the influence of transcription errors on the discrimination of these three subpopulations, manual features are extracted from manual transcripts and compared against each other, to determine statistically significant differences between the three groups.

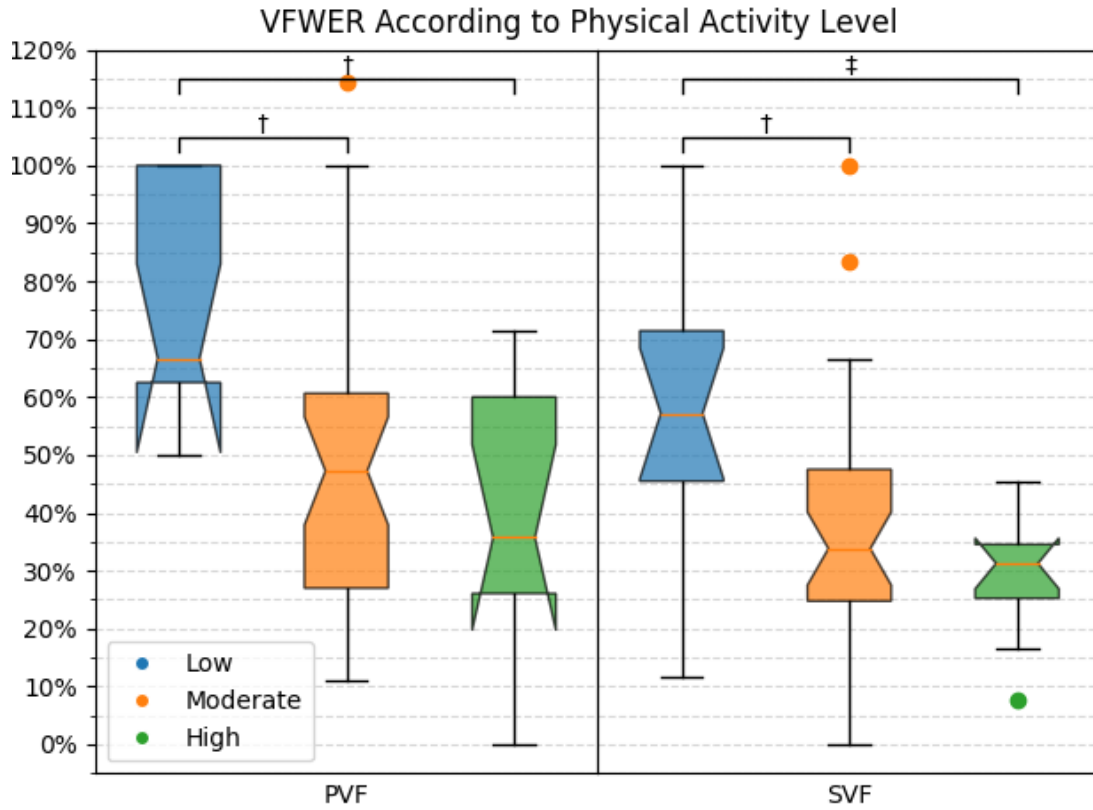


Figure 6.13: VFWER according to physical activity level. Brackets indicate statistical significance of the Mann-Whitney U test between VFWER of populations of different levels of physical activity.

† $p < 0.01$, ‡ $p < 0.001$

Manual VF feature distributions, for PVF and SVF tests, for the three physical activity level subpopulations, are compared in Figure 6.14. Feature values are shown in their z -score, normalised to the mean and standard deviation of the whole studied population. Seven features are compared for each test type: total number of correct words produced, mean and maximum phonemic cluster size, number of phonemic switches, mean and maximum semantic cluster size, and number of semantic switches. These features are extracted in a semi-automatic fashion, based on the criteria defined by [114, Appendix], as described in Section 5.3.

The three subpopulations present different distributions of total number of correct words produced, for both PVF and SVF. This metric for SVF tests is more discriminatory, with all groups being statistically different from each other. For the case of PVF, performance of this value on moderate and high activity subjects is not significantly different.

For switching and clustering features, subjects with low physical activity tend to have lesser and smaller phonemic and semantic clusters on PVF tests, but distribution of these features on moderate and high activity subpopulations is not discriminatory.

On SVF tests, semantic cluster size is highly discriminatory of these three subpopulations,

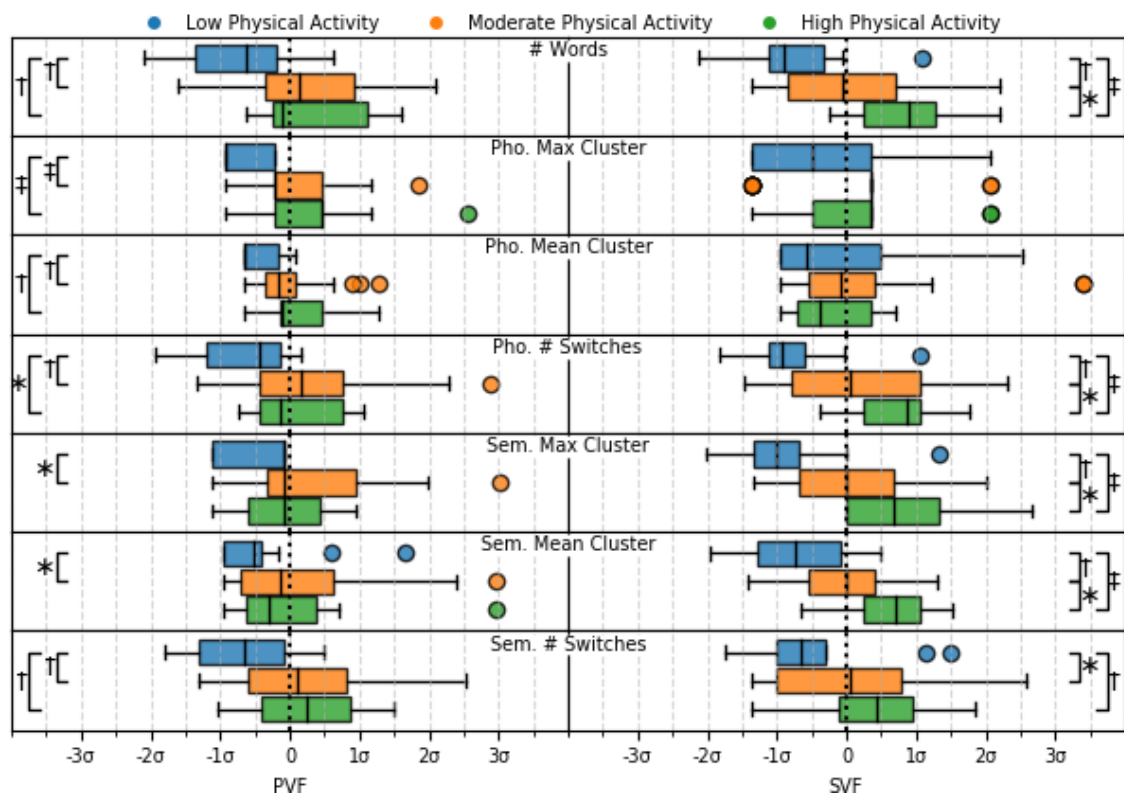


Figure 6.14: Manual VF feature distributions according to physical activity level, extracted from manual test transcripts. Feature values for PVF (left) and SVF (right) are normalised according to the mean and standard deviation of the whole study population. Brackets indicate statistical significance of the Mann-Whitney U test between features distributions of populations of different levels of physical activity.

* $p < 0.05$, $^{\dagger} p < 0.01$, $^{\ddagger} p < 0.001$

with low activity subjects having smaller clusters, and high activity subjects having large semantic clusters. Phonemic cluster size is not relevant to distinguish these groups, but phonemic switching is larger for higher activity levels.

These distributions indicate that switching and clustering are both necessary processes that need to occur simultaneously to achieve a good VF test performance. For low activity subjects, these two elements of VF appear to be reduced, with smaller and fewer clusters produced, resulting in a lower number of produced exemplars. This could be due to age and cognitive function effects observed on this population on VF tests.

SVF tests are more discriminatory for these populations, possibly due to a stronger influence of switching and clustering processes. PVF tests have low cluster size and a high count of switches, indicating exemplars tend to be produced on a one-by-one basis, since this test goes against natural brain lexical word storage [55], making it feel more "artificial". On SVF tests, switching and clustering effects become more noticeable, as subjects naturally group responses together: the size of obtained groups indicates the overall quality of verbal functions and semantic storage, and the number of switches between groups the effectiveness of searching strategies and executives

functions for controlled responses and inhibition [114].

6.4.2 Discrimination of Automatic Features According to Physical Activity Levels

All 40 automatically extracted features from PVF and SVF tests of each subject are compared between the identified subpopulations to assess if physical activity influences VF performance, and how. These features are automatically extracted, using the fully implemented pipeline for transcription, word filtering, and feature extraction.

Figure 6.15 shows the resulting features, for PVF and SVF tests, in a similar manner as Figure 6.14.

Comparing these results to manually obtained features, it can be seen that analogue features present the same approximate distributions, and can be used to distinguish these three subpopulations. This indicates that, although automatic transcription may introduce errors on the total number of words (and consequently affecting the metrics calculated for temporal, phonemic, and semantic switching and clustering), obtained features are still quite valid to distinguish different subjects and populations, since they are highly correlated with their manual counterparts, as seen previously in Section 6.2.

Once again, SVF presents a higher number of features that can be used to distinguish the three populations, with varying levels of significance. For PVF, no single feature exists that has significant difference between all three groups. Nonetheless, several PVF features have interesting distributions that can be used to distinguish one group from the other two.

The size of semantic clusters on PVF tests can be used to differentiate the low activity subpopulation, as well as the number of phonemic switches performed, as was previously seen in Figure 6.14 with their manual analogues. In addition to these, several new automatic features can also distinguish one group from the other two, namely the SOI, mean inter temporal cluster time, and number of temporal switches.

Since the SOI is the ratio between the pairwise and global semantic similarity of words, higher values indicate a stronger tendency to group words together in semantically similar contexts. It can be seen that the low activity population has considerably lower SOI, indicating that the other populations are more effective when tackling PVF because subjects are effective at finding semantically related words that start with the same letter, boosting their overall performance. In this case, the decreased value for the low activity subpopulation can indicate an increased difficulty in gathering semantically related concepts, perhaps due to more depleted verbal memory and word storage, or perhaps due to slower word searching and validation processes when considering a specific semantic subcategory and determining if a given member can be used as a PVF test exemplar.

Temporal features are an interesting addition for the analysis of PVF tests, particularly mean inter cluster time and temporal switches. Mean inter cluster time can be used to differentiate the high activity subpopulation, since this group has a significantly lower valued distribution. Mean inter cluster time is a feature that predominately measures executive functions and the effectiveness of subcategory search processes: subjects that quickly remember new semantic/phonemic clusters

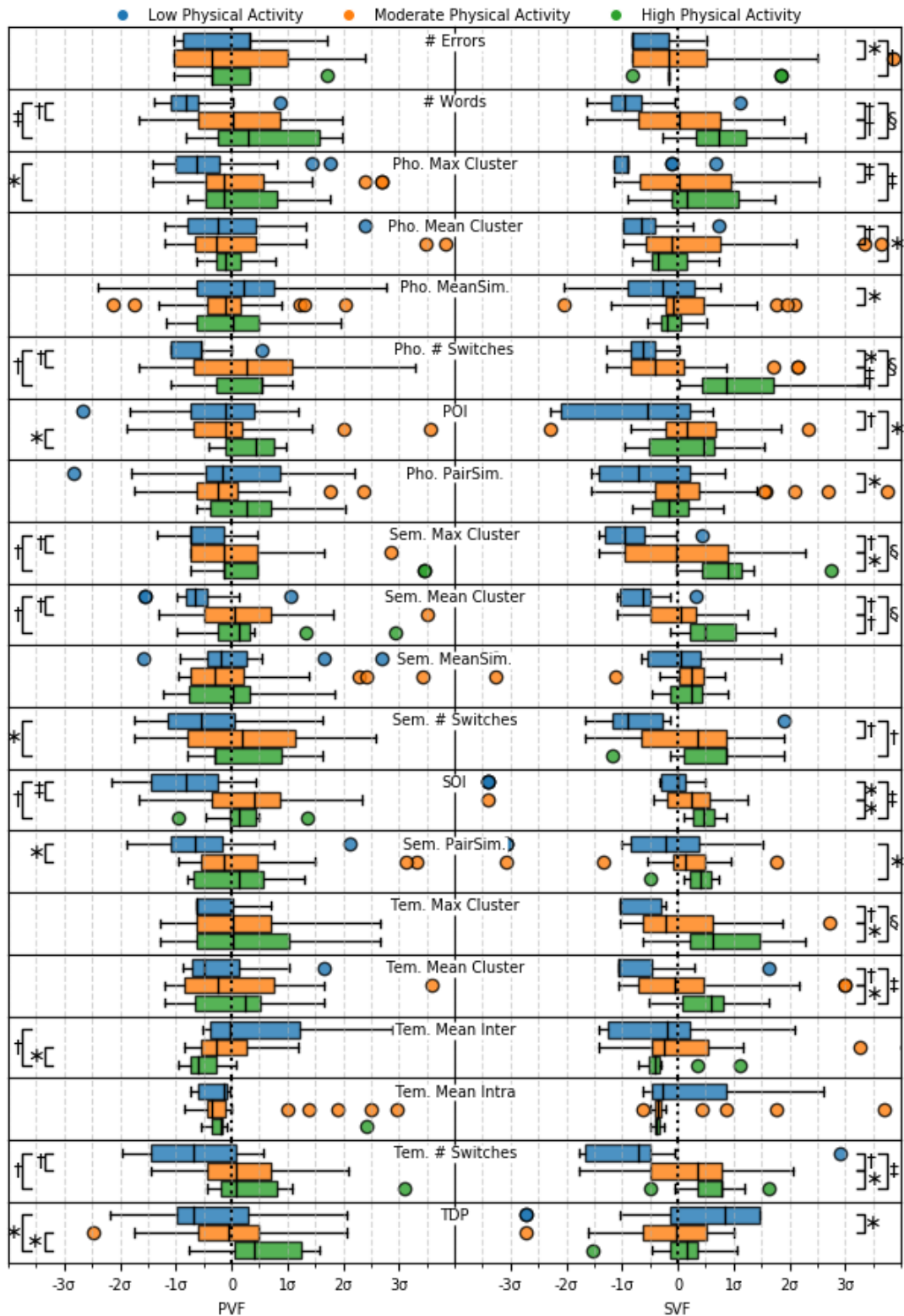


Figure 6.15: Automatic VF feature distributions according to physical activity level, extracted from automatic test transcripts. Feature values for PVF (left) and SVF (right) are normalised according to the mean and standard deviation of the whole study population. Brackets indicate statistical significance of the Mann-Whitney U test between features distributions of populations of different levels of physical activity.

* $p < 0.05$, $^{\dagger} p < 0.01$, $^{\ddagger} p < 0.001$, $^{\S} p < 0.0001$

to explore have a lower average time interval between them, thus allowing for more clusters to be explored and more words to be produced. Conversely, the number of temporal switches can also be used to determine the effectiveness of subcategory search, where a higher number of switches indicates more subcategories are found by a subject, ready to be used to produce exemplars, up to the exhaustion of their semantic stocks. Lower activity subjects tend to have a lower number of temporal switches, thus indicating this impairment in their executive functions.

For SVF, besides the total number of words produced, the number of phonemic switches, semantic cluster size, SOI, temporal cluster size, and number of temporal switches are features that can be used to distinguish all subpopulations from one another. For this test, phonemic groupings are infrequent, so the number of phonemic switches tends to be highly associated with the total number of words produced, since each new word produced can be considered a new phonemic cluster.

As discussed previously, semantic cluster size (mean and maximum) are a more prevalent feature for SVF since these tests are much more natural to the brain's semantic memory structure, which remembers concepts based on their associations with other similar, semantically related concepts [55]. The size of these clusters indicates the overall availability of semantic stocks and verbal memory. The SOI behaves in much the same way, indicating the ease/difficulty of subjects in grouping together semantically related exemplars.

Similarly to PVF tests, the analysis of temporal features obtained in SVF tests can be used to distinguish all subpopulations. Low activity subjects not only tend to produce less temporal clusters, but the clusters produced are generally smaller. This indicates a two-fold impact on VF performance due to impairments in both executive and verbal memory functions: subjects not only have increased difficulty in accessing new clusters to explore, but the number of exemplars produced for each is significantly reduced.

6.5 Final Remarks

The results obtained in this chapter aimed to study two things: the validity of using the implemented fully automatic pipeline for analysis of VF tests, and to assess the impact of physical activity on VF performance.

For the first question, VF manual features were compared to their automatic analogues, using the best found parameters for the pipeline (*default* Speaker Model and clustering grouping strategy for phonemic and semantic feature extraction). The automatic extraction of features is shown to be valid using manual transcripts from VF tests and comparing with manually determined ones. Positive correlations were found for these pairs, thus showing the obtained number and size of semantic and phonemic clusters is similar for automatic and manual methods. In reality, the automatic methods here described can be considered more reasonable and well founded, since they rely on generalised models for the determination of semantic and phonemic similarity, based on word embeddings cosine similarities and phonemic transcription edit distances, instead of predetermined rule sets, that may not account for all edge and novel cases.

The automatic transcription of VF tests implies a loss in accuracy of several crucial features, particularly in the total number of correct words determined, since deletions, substitutions, and misunderstandings of the ASR system used are common, and can lead to a considerable VFWER, particularly for noisy recording environments. Nonetheless, the determined values are consistent with their manual correlates, with significant positive correlations between them. With this, it is reasonable to say that, if two subjects are recorded in similar background noise conditions, with no major diction problems that may affect ASR performance, VF test analysis and comparisons between the two can be done with these automatically extracted features.

The impact of physical activity on VF was assessed using manual and automatic features. Conclusions taken from manual features analysis are also evident on extracted automatic features, once again showing the validity of using these as a framework for VF performance analysis.

Overall, subjects with lower physical activity showed a reduced production of clusters, with a reduced size. This can be an indicator of impairment in the two main dimensions of VF test analysis: executive functions related to self control, response inhibition, and effective search strategies are reduced in effectiveness, thus subjects generate a lower number of distinct phonemic or semantic groups. Furthermore, produced groups have a lower member count for this subpopulation, possibly indicated depleted semantic stocks and verbal memory problems.

In addition, SVF tests and its corresponding automatically extracted features proved to be more effective in distinguishing all three study subpopulations, with more significant differences, particularly in number and size of semantic and temporal clusters found.

Notwithstanding, the analysis of these results has to be performed carefully, since the low physical activity subpopulation also corresponds to a subpopulation of older, more cognitive impaired subjects, and with a higher level of VFWER, when compared to the two other groups. These factors can influence the obtained results negatively, impacting overall found values for this subpopulation and exaggerating differences and impairments found.

Furthermore, the association of a lower physical activity level to a lower VF test performance does not imply causation: it is not valid to assume that lower physical activity levels lead to a decreased performance in VF tests. VF tests are affected by a plethora of personal, social, environmental, neurological, and psychological variables, all at play here, indissociable from the found subpopulations. Higher levels of physical and social activity, coupled with healthy, active lifestyle are associated with decreased age associated cognitive decline and incidence of dementia [66], but this cannot be proved for this corss-sectional study of small dimensions. To adequately evaluate this, longitudinal studies have to be performed to assess the overall impact of physical activity on cognitive ability, including VF performance.

Chapter 7

Conclusions

In this work, a novel fully automatic pipeline for the analysis of PVF and SVF tests in European Portuguese is suggested and detailed. This pipeline has as its base the online ASR system provided by Google Cloud [45] to perform automatic transcription of VF test recordings. Classic VF features are extracted, such as the total number of correct words produced (which implies word filtering according to test type) and switching and clustering features related to phonemic, semantic, and temporal clusters in subject's responses.

This pipeline was characterised and validated using two sets of data, from CHUP [4], and from a designed study with healthy elderly subjects. Feature extraction was shown to be valid, with high correlations obtained to manual feature analogues. Automatic transcription was performed, proving to be sensitive to background noise present in recordings, which leads to the deletion of several correctly produced exemplars. This elimination of some words in the automatic transcript leads to understated values for automatically calculated features. Nonetheless, due to high positive correlation to manually determined features, comparison of VF performance of different subjects can still be performed in this new automatic framework, with more informative features (particularly temporal features) enhancing the different contributions of executive and verbal functions for the realisation of these tests.

In addition, the used semantic and phonemic switching and clustering metrics, based on more general methods of similarity calculation (word embeddings and phonemic transcription edit distances), not only removes the necessity of creating a particular rule set for manual determination of clusters, but also ensures that less obvious edge cases can be identified and credited.

These features are used to great effect in the distinction of VF performance of elderly subjects with different levels of physical activity. It was found that the subpopulation with lower levels of physical activity showed to have lower overall values for number of words produced, and number and size of clusters (particularly semantic and temporal). However, no concrete conclusion regarding the effect of regular physical activity has on VF performance can be taken, since this subpopulation was also shown to be older and more cognitive impaired than the rest of the subpopulations of the study. Larger and more detailed studies, particularly longitudinal studies, need to be performed to correctly assess the impact of activity on cognitive performance, particularly

on VF performance.

This pipeline for analysis of VF tests presents itself as an exciting new tool for VF test analysis, not only reducing workload and processing time, but also providing health professionals with a larger number of more informative features, in a short amount of time. Even though ASR introduces transcription errors, if recorded in similar conditions, extracted features can still be compared to assess the different underlying brain pathways for VF test execution.

The use of this new method of VF test analysis, with much lower processing times and more informative metrics extracted, can be used in the future in clinical settings to ease VF test administration by health professionals. Since these metrics have predictive power and can be used as screening tests for dementias [33, 61, 112], these quick 1-minute tests, can become a very powerful, easy to implement, and quick to apply tool to aid diagnosis on a more and more aged population, with NCDs being an evermore growing concern on the quality of life of our elderly.

7.1 Pipeline Improvements

The developed pipeline is shown to be effective in being a new framework of analysis and comparison of VF test performance. Nonetheless, during its development, several obstacles were faced where the pipeline struggles. These possible future improvements can greatly enhance this tool, becoming more and more reliable for clinical use.

Automatic Speech Recognition ASR is the keystone of the implemented pipeline, whose performance dictates the overall significance and quality of extracted features. As such, several improvements can be made to ensure maximum effectiveness of ASR.

Background noise was shown to highly impact the overall performance of the ASR service to extract test related words. A possible way to mitigate the impact of background noise is to record VF tests using a unidirectional microphone, thus avoiding the interference of other possible sources of noise in the recording. However, this goes against initial motivation of creating a simple to use system for VF test analysis, that does not require complex recording systems, but rather just a *smartphone*.

As such, to ensure good results, it is vital to administer VF tests in a quiet, private environment.

Furthermore, an available feature of the used ASR system is the possibility of providing hint words and have access to a list of possible transcriptions alternatives, each with a corresponding confidence value. In the future, these alternatives can also be analysed, being used for instance in SVF test transcripts, of one of the provided alternatives for a specific word is an animal name.

Furthermore, the automatic transcription frequently struggled in deciphering the regional accent of some subjects, suggesting perhaps an inadequate model is used that does not take into account these slight phonetic variations. Since the ASR model cannot be directly adjusted according to a specific population accent, perhaps a way to ameliorate this is to create a new ASR system that can be trained with these specific populations, making it more robust to these variations. This is, however, a very complex task.

Since Google Cloud [45] Speech-to-Text online services takes advantage of supplied data to further improve transcription, it is likely in time that these phonetic variations will be resolved, with ASR models getting better and better at transcribing European Portuguese and its varied regional accents and variations.

Out-of-Test Speech A major weakness of the implemented pipeline resides on its difficulty dealing with out-of-test speech, such as a subject's interjections and vocalised thoughts, or the rules clarifications performed by the tester. These can appear as unwanted intrusions in the automatic test transcript, and may be, in some cases, incorrectly credited.

Speaker diarisation algorithms, that not only automatically recognise what is being said, but also *who* says what, are possible solutions to separate interventions from subjects and examiners. Google Cloud [45] Speech-to-Text already has speaker diarisation functionality, but it is still in its beta version, and is only recommended for telephone recordings. Speaker diarisation was nonetheless attempted with this ASR service, but no meaningful results were obtained. These algorithms can be introduced in the future to distinguish an examiner's intervention from exemplar production.

Performing automatic non-test word filtering as described in the implemented pipeline is a crude method to filter non-test interventions: words that are not said in the scope of the test can be credited, while the opposite scenario can also occur. Restricting subjects to only produce test exemplars during a VF test administration is also an inadequate way of tackling the issue, since subjects will naturally think aloud, monologue, etc., and restricting this may even hamper the performance of a subject.

Possible new NLP approaches to this out-of-test speech problem may be applied in the future, such as part-of-speech tagging for the identification of subject's interjections, complete phrases, etc.

Animal Classification The use of a Portuguese WordNet [29] for the classification of words as animals proved to be a rather successful strategy for this task. However, the implemented solution struggles with several, more uncommon, animal names. For many of these, the Portuguese WordNet does not have an associated synset, resulting in its elimination.

These edge cases were treated as exceptions, being hard coded as animals, with a list of misclassified animal names being updated during the development of this work. In the future, it is preferable to not have this list, but rather a more effective method of determining animals, perhaps using a different, more extensive Portuguese word ontology.

Phonemic Transcription The implemented method of phonemic transcription relies on an external online dictionary with phonemic transcriptions available. The provided transcriptions are very good, with several options for different word pronunciations. However, currently there is no method of determining which of the pronunciations was actually said by the subject performing

the VF test. In the future, the addition of this may improve overall phonemic cluster determination, based on how words were actually said by the subject, and not on how they can be said in general.

7.2 Future Work

With the developed pipeline, future studies can be considered, to further validate this new tool, and to use it in great extent to evaluate several neuropsychological conditions.

Unsupervised Learning for Population Characterisation This 40-feature framework can be used in the future to characterise populations based on their relevant feature values. Unsupervised Learning techniques can be used to group together subjects into populations of similar VF performance. These populations can be associated with a risk of NCD, considering the distinct VF performance evolution profiles for several common dementias, such as Alzheimer's or Parkinson's Disease. Subjects grouped into clusters of high risk of developing a NCD can be quickly identified and screened, allowing for faster treatment of these conditions.

Supervised Learning for Dementia Screening Several different types of dementias have distinct VF performance profiles, particularly when considering switching and clustering features. The new framework developed in this work can be used as a feature extractor for Supervised Learning problems, using it alongside clinical studies, with healthy control groups, and demented patients with a confirmed diagnosis.

From this, a classifier can be built based on the 40 extracted features from VF tests. If trained adequately, such a classifier would be a very powerful dementia screening tool, being based on a test that is quick to administer (only takes one minute), quick to analyse (with the use of automated pipelines), and quick to determine the probability of subjects having a certain NCD.

Automation of Other/New Cognitive Tests Alongside pipeline improvements in the future, other possible research areas can be interesting areas of application of this framework, or other frameworks based on it.

Verbal cognitive tests are very frequent in Neurology and Neuropsychology, so their automation using ASR and NLP is a very promising goal. With fine tuned pipelines for the extraction of relevant features from several verbal cognitive tests, it is possible to evaluate in a quick and easy manner even more cognitive functions and abilities, besides verbal fluency.

In addition, the creation of automatic pipelines for analysis can be the starting point for the development of new and better cognitive tests. Tests that are more reliant on non-structured conversation and natural responses of a subject are hard to analyse by hand, since there is a lot of subject and examiner variability at hand. Possible automatic analysis processes can remove examiner variability altogether, and can use NLP techniques to determine important parts of a subject's speech patterns.

One such example of a cognitive test that can be automatised to great extent would be the picture description test, where a subject is asked to describe a picture (for instance, a landscape), with as much detail as possible, in one minute. Manual analysis of such a test would be almost inconceivable since natural speech is so diverse. Strict rule sets can be implemented to quantify performance (e.g., attribute one point per referenced landmark), but then tests become too strict and unforgiving, and do not take into account the vocabulary richness of a subject's discourse, or how fast a subject describes the picture. With the automation of such a test, it would be possible to determine vocabulary richness, word timing, and many other clinical meaningful features, with the use of ASR and NLP techniques, with consistent reliability.

The automation of verbal cognitive tests proves to be an exciting topic, with the application of new solutions of ASR and NLP to effectively analyse and extract more information from these tests. The use of new automatic frameworks for the analysis of these tests, including (but not limited to) VF tests, can make their application more reliable, quick, and easy, making them more informative and ubiquitous in a clinical setting. Since these are powerful tools for the NCD screening, their automation and application can lead to faster identification of at-risk subjects, allowing for better and more powerful preemptive action on these disorders, that have such a deep, incalculable impact on our elderly population and society.

Appendix A

Statement of Gabinete Coordenador da Investigação of CHUP

The following pages contain a copy of the statement of the *Gabinete Coordenador da Investigação* of CHUP regarding the use of VF test recordings for analysis using the developed algorithms in this dissertation, as an addendum to the research study performed by Amorim and Susano [4].

The provided statement indicates the favourable opinion of *Gabinete Coordenador da Investigação* of CHUP regarding the realisation of data analysis using the developed pipeline.

PRE-OPERATIVE NEUROCOGNITIVE DISORDER AND LOW NEAR-INFRARED SPECTROMETRY IS ASSOCIATED WITH POSTOPERATIVE DELIRIUM – ADENDA Nº 3

253/16 (216-DEFI/ 205-CES) Trabalho Académico de Investigação no âmbito do Programa Doutoral em Medicina pela Faculdade de Medicina da Universidade do Porto

Data de Receção dano Gabinete Coordenador da Investigação (GCI): 29/05/2019

Data de envio de email a solicitar esclarecimentos ao investigador: 06/06/2019

Data de receção dos esclarecimentos do investigador: 14/06/2019

Estudante de Doutoramento (nº 201309682) e Investigadora Principal

- Maria João Susano – Médica, Interna em Formação Específica, Serviço de Anestesiologia do CHP.

Equipa de Investigação

- Pedro Amorim – Médico Anestesta, Chefe de Serviço, Serviço de Anestesiologia do CHP.
- Humberto Machado – Médico Anestesta, Diretor do Serviço de Anestesiologia do CHP.
- Enfermeiras responsáveis pela admissão dos doentes à enfermaria e inseridas na equipa de investigação (ponto 2.6 – metodologia) – Não identificadas
- Mariana Pimentel Carvalho Tedim Dias - Estudante do 5º ano do MIM

Orientadores de Doutoramento:

- Fernando Jose Pereira Alves Abelha (Orientador)
- Pedro de Pinho e Costa Amorim (Co-Orientador)
- Deborah J. Culley (Co-Orientadora)

ADENDA nº 3:

- É submetida uma adenda ao estudo intitulado “*Pre-Operative Neurocognitive Disorder and Low Near-Infrared Spectrometry is associated with Postoperative Delirium*” autorizado pelo CHP em 26/04/2017 [253-16 (216-DEFI/205-CES)] em fase de conclusão no Serviço de Anestesiologia do CHP.
- Em 30/11/2017, foi submetida uma adenda ao estudo com algumas alterações metodológicas, nomeadamente aumento da amostra para 210 participantes (ao invés dos 110 previstos no projeto inicial); acréscimo de alguns instrumentos de recolha de dados, nomeadamente testes à avaliação pré-operatória (*Brief Pain Inventory Short-Form, 5-Item FRAIL scale, atividades da Vida Diária, Atividades Independentes da Vida Diária e o World Health Organization Disability Assessment Schedule 2.0 (WHODAS)*), o teste de fluência verbal como teste de rastreio de disfunção cognitiva no pré-operatório e a *Numeric Rating Scale para avaliar a presença da dor no pré- e pós-operatório*. Foram também acrescentados alguns *outcomes* secundários. A adenda solicitada teve parecer favorável em 29/03/2018, depreendendo-se que o estudo decorreria até à inclusão dos 210 participantes.
- Em 14/12/2018 foi submetida uma segunda adenda ao estudo, em que foi solicitada a “Recolha de sinal de Eletroencefalograma (EEG), no pré-operatório recorrendo ao equipamento EMOTIV Insight 5 Channel Mobile EEG”. A adenda teve parecer favorável em 20/03/2019.
- Agora é submetida uma nova adenda em que é solicitada uma análise secundária das palavras do teste de fluência animal, em que será comparado o método manual com um novo *software* (o *AutoSpeech* – ferramenta computacional desenvolvida para automatizar a administração e análise de testes de fluência verbal).

QUESTÕES CLARIFICADAS PELOS INVESTIGADORES:

Transcreve-se em seguida parte dos esclarecimentos enviados pelos investigadores:

- a) Tendo em conta que a fase de recrutamento e recolha de dados já terminou, as respostas dos doentes ao teste de fluência verbal manual foram gravadas em áudio? A administração manual do teste de fluência verbal implica habitualmente a gravação em áudio das respostas?

Os investigadores esclarecem que o teste de fluência verbal é “...um teste em que solicitamos ao doente que no tempo exacto e cronometrado de um minuto, nomeie o maior número de animais que consiga, explicando que não são contabilizados o macho e a fêmea de um mesmo animal, por exemplo “cão e cadela”, mas que podem nomear qualquer animal.”

Inicialmente, a contabilização do número de animais era efetuada “...fazendo um traço num papel por cada animal nomeado, enquanto com recurso a uma das funções de um telemóvel, cronometrávamos um minuto. Rapidamente nos apercebemos que nos era difícil detectar se se ocorria a repetição de um mesmo animal ou da nomeação do macho ou fêmea de um animal já nomeado, pois essa detecção ficava dependente da nossa memória, já que não era possível ir escrevendo o nome de cada animal, por impossibilidade de o fazer no devido tempo. Assim, e já que estávamos a usar o telemóvel para cronometrar o tempo, ocorreu-nos usar a funcionalidade de gravação dos telemóveis, para simultaneamente controlar com rigor o período de um minuto e gravar as respostas. Assim, após os casos iniciais, começamos a gravar, num telemóvel, cada resposta de um minuto. Conscientes da necessidade de respeitar anonimato, tivemos o cuidado de não associar qualquer identificação a cada gravação, não registando sequer o número de sequência do doente incluído no estudo, idade ou fosse que dado fosse.

Após cada teste a respectiva gravação era de imediato ouvida, registando-se na base de dados, o número de animais nomeados por aquele doente. A gravação tinha por vezes que ser ouvida mais do que uma vez para permitir detectar repetições e nomeação de macho e fêmea de um mesmo animal, o que implicava tomar algumas notas.”

(...) “Entretanto verificamos ter na nossa posse cerca de 150 gravações de um minuto, como disse, gravações sem qualquer dado respeitante ao sujeito gravado.” (...)

APRECIÇÃO e PARECER FINAL DO DEFI/ Gabinete Coordenador da Investigação

No âmbito do estudo de investigação aprovado pelo CHUP em 26/04/2017, intitulado “Pre-Operative Neurocognitive Disorder and Low Near-Infrared Spectrometry is associated with Postoperative Delirium” 253-16 (216-DEFI/ 205-CE) e subseqüentes adendas, é submetida uma 3ª adenda ao estudo a solicitar a análise secundária das palavras do teste de fluência verbal, em que será comparado o método manual com um novo *software* (o *AutoSpeech*).

Tendo em conta os esclarecimentos acima recebidos e, não sendo referidas alterações aos objetivos ou outras modificações à metodologia do estudo, do ponto de vista científico o **Gabinete Coordenador da Investigação** não tem nada a opor a esta 3ª adenda ao estudo - **PARECER FAVORÁVEL**

Não estando previstas implicações éticas para além das previamente autorizadas, o processo segue para autorização pelo DEFI.

Centro Hospitalar Universitário do Porto, 17 de junho de 2019

Isabel Fonseca, Responsável pela área da análise científica dos estudos de investigação
DEFI/ Gabinete Coordenador da Investigação

Appendix B

Ethics Committee of Universidade do Porto Study Approval

The following pages contain a copy of the report by the Ethics Committee of Universidade do Porto (CEUP) for the realisation of the AutoSpeech study, detailed in Chapter [4](#), after submission of a statement request for approval of this study.

The provided report indicates the favourable opinion of CEUP regarding the realisation of the study.



COMISSÃO DE ÉTICA

REPORT Nº 72/CEUP/2019

Report by the Ethics Committee of Universidade do Porto

AutoSpeech – Automatic Speech Analysis of Verbal Fluency For Older Adults

SUBMITTED BY:

Ana Cristina Pereira

Associação Fraunhofer Portugal Research - Fraunhofer Portugal Research Center for
Assistive Information and Communication Solutions (Fraunhofer AICOS)

The Rapporteur: Prof. Doutor Fernando Monteiro

12th June 2019

Associação Fraunhofer Portugal Research - Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (Fraunhofer AICOS) has submitted to this ethics committee a project named "*AutoSpeech – Automatic Speech Analysis of Verbal Fluency For Older Adults*". The project started on the 4th February 2019 and shall last for 6 months.

The Project aims at developing a novel technology allowing for the provision and automatic evaluation of verbal fluency tests, allowing also for an easier use by health technicians. Besides, the project involves the development of a data base of audio recordings of verbal fluency tests applied on elderly individuals with several types of scholar background and of physical activity, not only to validate the technology but also to identify the influence of physical activity in the cognitive capabilities of the elderly persons.

The study relevance regards the fact that the low performance in verbal fluency tests might be associated to cognitive deficiencies, and this test possibly might be used to distinguish between different types of brain lesions, mental disturbances and neurodegenerative diseases (as dementia)

Socio-demographic data, cognitive performance data, daily physical activity habits, and audio recordings of verbal fluency tests(phonetical and semantic) shall be collected from elderly people aged over 65, healthy, without any diagnosed neurocognitive problems with different levels of physical activity.

The audio recordings of verbal fluency tests shall allow an exploratory study of novel metrics and characteristics to be automatically extracted, applying methodologies and techniques of computer learning (namely *speech-to-text* and *natural language processing*). Such extracted values provide increased information and will facilitate the establishment of correlations with the level of physical activity in elderly people. They allow also to substitute some characteristics of those tests that are difficult to obtain manually.

The Project is carried out by Ana Pereira (MSc.)(PI)from Fraunhofer, and other researchers are participating:

Liliana Ferreira (PhD.) Fraunhofer-AICOS, Aníbal Ferreira (PhD.) Faculdade de Engenharia da Universidade do Porto (FEUP) Ana Pereira (MSc.) Fraunhofer-AICOS, João Costa (BSc.) Fraunhofer-AICOS Ricardo Araújo (BSc.) Faculdade de Psicologia e de Ciências da Educação da Universidade do Porto (FPCEUP)).

Tests Protocol

Data collection from the participants will occur during 15 to 20 minutes interviews with a psychologist (Ricardo Araújo). The interviews shall take place at Fraunhofer-AICOS, or where possible if the subjects cannot be interviewed at Fraunhofer-AICOS.

Participants are selected in Porto, by Fraunhofer Portugal, in the framework of COLABORAR network , being over 65 years old, healthy, without any diagnosed neurocognitive problems,

and without listening and comprehension disabilities that might affect the performance of the tests (information self-declared by each participant).

The interviews consist of applying three questionnaires found as annexes to the request submitted to the Ethics Committee:

1. *Montreal Cognitive Assessment (MoCA) (Anexo 3)*
2. *Sociodemographic questionnaire (Anexo 2);*
3. *International Physical Activity Questionnaire, short version (Annex 5)*

All the texts are translated into Portuguese..

The inclusion and exclusion factors to apply in the participants selection are clearly defined.

The informed consent is clear and all the required conditions to guarantee the data protection and confidentiality are well formulated. There exists the option of making the analysis of the respective participation available to each participant, after a formal demand from the respective participant.

No damage or inconvenience is foreseen for any participant.

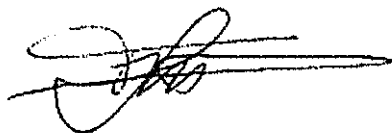
Concluding remark:

The study is well designed and has obvious reasons from the scientific and social point of view to be carried out. Confidentiality aspects are well preserved and in no case there seems to be any risk of inadequate extraction of data or of their inappropriate use.

Therefore, CEUP decided to give a favorable opinion to the realization of this project.

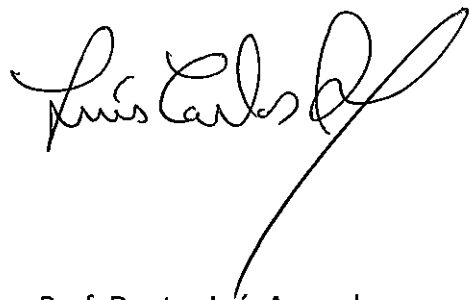
Universidade do Porto, 12th June 2019

The Rapporteur



Prof. Doutor Fernando Jorge Monteiro

The President of CEUP



Prof. Doutor Luís Amaral

Appendix C

Informed Consent to Participants, Portuguese Version

The following pages contain a copy of the consent form used to inform test participants of the AutoSpeech study of the procedures that were performed, as well as what data was collected for each.

Informed consents were printed in duplicates, so that one copy remained with the subject.

CONSENTIMENTO PARA PARTICIPAÇÃO EM INVESTIGAÇÃO

A *Associação Fraunhofer Portugal Research* pretende realizar um estudo de investigação sobre a automação de administração e análise de testes de fluência verbal.

Para este estudo, propomos a sua participação em entrevistas individuais com um psicólogo, nas quais lhe será pedido que responda a questões sociodemográficas, o seu nível de atividade física, e que realize um conjunto de pequenos testes cognitivos.

Objetivo do estudo

O estudo pretende criar uma base de dados áudio da execução de testes de fluência verbal para:

- Validar os algoritmos criados, comparando-os com a avaliação manual destes testes;
- Inferir o impacto que diferentes níveis de atividade física têm nas capacidades cognitivas de seniores portugueses, nomeadamente em funções executivas e de linguagem, que o teste de fluência verbal avalia.

Materiais usados

Neste estudo não será necessária a utilização de qualquer material ou dispositivo por parte do participante, sendo apenas necessário responder a um conjunto de questões. Algumas questões irão requerer o uso de esferográfica para desenhar formas geométricas.

Os questionários realizados serão aplicados pelo psicólogo e as respostas dadas serão registadas com recurso a papel e esferográfica.

Será utilizado um *smartphone* (pelo psicólogo) para gravar o áudio dos testes de fluência verbal aplicados.

Procedimentos

Este estudo apenas possui uma etapa, em que se procede à recolha de dados dos participantes em ambiente de entrevista com o psicólogo. A entrevista dura cerca de 15 a 20 minutos.

A entrevista é dividida em três fases:

1. Aplicação do *Montreal Cognitive Assessment* (MoCA), um conjunto de testes cognitivos.
2. Questionário sócio-demográfico;
3. Questionário referente a atividades físicas realizadas – International Physical Activity Questionnaire.

Montreal Cognitive Assessment

Este questionário é composto por um conjunto de testes avaliadores de funções cognitivas, frequentemente usado para a deteção e diagnóstico de doenças do foro neurocognitivo, como a demência. Os testes pretendem avaliar capacidades visuospaciais, executivas, de nomeação, de memória, de atenção, de linguagem, de abstração, e de orientação. Alguns destes testes irão envolver desenho de figuras geométricas, operações aritméticas simples, memorização de palavras, pensamento lógico, entre outras tarefas simples.

Questionário Sócio-Demográfico

Durante este questionário iremos recolher dados sócio-demográficos referentes a si, nomeadamente género, data de nascimento, nível de escolaridade, estado civil, e onde e com quem mora.

CONSENTIMENTO PARA PARTICIPAÇÃO EM INVESTIGAÇÃO

Questionário de Atividade Física– International Physical Activity Questionnaire

Este questionário é composto por um conjunto de perguntas que tenciona avaliar o seu nível de atividade física que efetuou durante os últimos sete dias. O questionário é de auto-preenchimento.

Fatores de Exclusão

A fim de assegurar que não há nenhum impedimento relacionado com a condição mental do participante e que o mesmo pode participar neste estudo, as condições seguintes enunciadas constituem fatores de exclusão de participação:

- Diagnóstico de doenças neurocognitivas (Alzheimer's, Parkinson's, etc.);
- Diagnóstico de Défice Cognitivo Ligeiro;
- Não ser falante nativo português;
- Presença de problemas graves de audição ou de linguagem;
- Histórico de traumatismo craniano;
- Perda de consciência;
- Adição (incluindo alcoolismo);
- Comportamento motor aberrante ou psicótico;
- Toma de medicação que influencia capacidades psicomotoras;

Os seus dados pessoais serão analisados pelos investigadores e destruídos no final do estudo. Os dados recolhidos são confidenciais e anonimizados. A *Associação Fraunhofer Portugal Research* tomará todas as medidas necessárias à salvaguarda e proteção dos dados recolhidos por forma a evitar que venham a ser acedidos por terceiros não autorizados.

Os seus direitos no âmbito da proteção de dados serão sempre assegurados, tendo em consideração as normas específicas para o desenvolvimento da atividade de investigação. Qualquer comunicação para este efeito deverá ser efetuada para o seguinte e-mail: ana.pereira@fraunhofer.pt. Informações adicionais poderão ser prestadas pelo Investigador Responsável abaixo identificado.

Gostaríamos de contar com a sua participação. A participação não envolve qualquer prejuízo ou dano material e não haverá lugar a qualquer pagamento, excepto o reembolso das despesas de transporte, mediante entrega dos respetivos recibos. A participação não terá custos para o participante.

A sua participação é voluntária, podendo em qualquer altura cessá-la sem qualquer tipo de consequência. Também poderá pedir a retificação ou destruição da informação recolhida a qualquer momento.

Agradecemos muito o seu contributo, fundamental para a nossa investigação!

Os resultados da avaliação do teste MoCA convencional a realizar pelo psicólogo que colabora neste estudo ser-lhe-ão transmitidos, se assim o desejar. Para o efeito, ser-lhe-á pedido o seu contacto de morada ou de e-mail.

CONSENTIMENTO PARA PARTICIPAÇÃO EM INVESTIGAÇÃO

O participante:

*Declaro ter lido e compreendido este documento, bem como as informações verbais fornecidas e aceito participar nesta investigação. Declaro ainda não ter nenhuma das condições listadas acima que constituem factores de exclusão deste estudo. Permito a utilização dos dados que forneço de forma voluntária, confiando que apenas serão utilizados para investigação e com as garantias de confidencialidade e anonimato que me são dadas pelo investigador. Autorizo a comunicação de dados de forma **anónima** a outras entidades que estabeleçam parceria com a Associação Fraunhofer Portugal Research para fins académicos e de investigação científica, nomeadamente a Faculdade de Engenharia da Universidade do Porto. Autorizo ainda a comunicação de alguns dados pessoais (nome e idade) para ativação de um seguro de acidentes pessoais para as atividades de resposta aos questionários, em cumprimento dos procedimentos internos da Associação Fraunhofer Portugal Research.*

Pretendo receber os resultados do teste MoCA:

- ☐ Não
- ☐ Sim, por correio para a minha morada
- ☐ Sim, por e-mail

Nome do participante: _____

Assinatura do participante: _____

Data ____ / ____ / ____

Investigador responsável pelo projecto:

Nome: Ana Pereira

Telefone: 220 430 342

E-mail: ana.pereira@fraunhofer.pt

Assinatura: _____

Appendix D

Montreal Cognitive Assessment, Portuguese Version

The following page contains the Portuguese version of the Montreal Cognitive Assessment (MoCA), originally created by Nasreddine et al. [80].

The Portuguese version here illustrated was obtained from the MoCA website [79]. On the website, detailed administration and evaluation procedures can also be found and freely downloaded, including a Portuguese version of these.

MONTREAL COGNITIVE ASSESSMENT (MOCA®)

Versão 8.1 Portuguese (Portugal)

Nome:

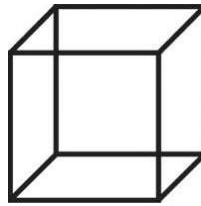
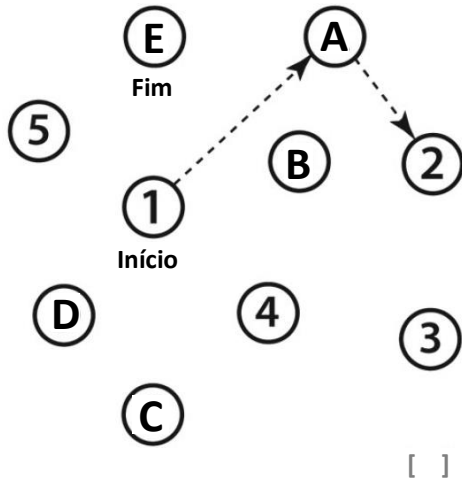
Escolaridade:

Data de Nascimento:

Género:

DATA DE AVALIAÇÃO:

VISUOESPACIAL / EXECUTIVA



Copiar o cubo

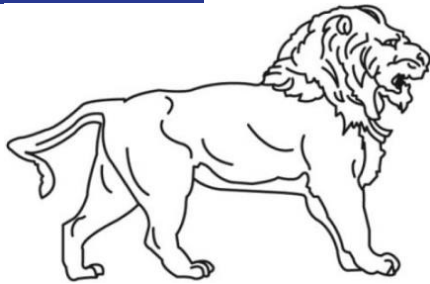
Desenhar um RELÓGIO (onze e dez)
(3 pontos)

PONTUAÇÃO

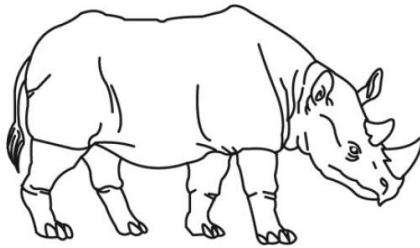
[] [] []
Contorno Números Ponteiros

___/5

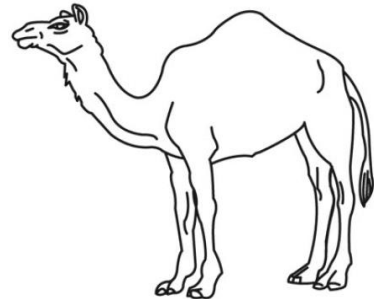
NOMEAÇÃO



[]



[]



[]

___/3

MEMÓRIA

Leia a lista de palavras. O sujeito deve repeti-la. Realize dois ensaios. Solicite a evocação da lista 5 minutos mais tarde

	BOCA	LINHO	IGREJA	CRAVO	AZUL
1º ENSAIO					
2º ENSAIO					

SEM PONTUAÇÃO

ATENÇÃO

Leia a sequência de números
(1 número/segundo)

O sujeito deve repetir a sequência em sentido direto. [] 2 1 8 5 4
O sujeito deve repetir a sequência em sentido inverso. [] 7 4 2

___/2

Leia a série de letras (1 letra/segundo). O sujeito deve bater com a mão cada vez que for dita a letra A. Não se atribuem pontos se ≥ 2 erros

[] F B A C M N A A J K L B A F A K D E A A A J A M O F A A B

___/1

Subtrair de 7 em 7 começando em 100. [] 93 [] 86 [] 79 [] 72 [] 65

4 ou 5 subtrações corretas: 3 pontos; 2 ou 3 corretas: 2 pontos; 1 correta: 1 ponto; 0 corretas: 0 pontos.

___/3

LINGUAGEM

Repetir: Eu só sei que hoje devemos ajudar o João. []
O gato esconde-se sempre que os cães entram na sala []

___/2

Fluência verbal / Dizer o maior número possível de palavras que comecem pela letra "P" [] ____ (N ≥ 11 palavras)

___/1

ABSTRAÇÃO

Semelhança p. ex. entre banana e laranja = fruta [] comboio - bicicleta [] relógio - régua

___/2

EVOCAÇÃO DIFERIDA

(IM)	Deve recordar as palavras SEM PISTAS	BOCA	LINHO	IGREJA	CRAVO	AZUL	Pontuação apenas para evocação SEM PISTAS
X3		[]	[]	[]	[]	[]	
X2	Pista de categoria						
X1	Pista de escolha múltipla						

IM = ____ / 15

___/5

ORIENTAÇÃO

[] Dia do mês [] Mês [] Ano [] Dia de semana [] Lugar [] Localidade

___/6

© Z. Nasreddine MD

www.mocatest.org

IM: /15

Administrado por: _____

(Normal ≥ 26/30)

É necessária formação e certificação para garantir a exatidão

Adicionar 1 ponto se ≤ 12º ano de escolaridade

TOTAL

___/30

Appendix E

Animal Fluency, Portuguese Script

Semantic Verbal Fluency tests are often used with the animal category to assess language and executive functions, commonly referred as simply Animal Fluency tests.

To ensure small inter-rater variability, and to guarantee each participant performs the Animal Fluency test in as similar conditions as possible, a Portuguese script is presented below, for the test administrator to follow. This script is based on the script of Cavaco et al. [21, Supplementary Material]

"Agora vou-lhe pedir que me diga o maior número de animais que conseguir. Não deve repetir a mesma espécie animal com pequenas variações, como por exemplo "cão, cadela". Tem 1 minuto para me dizer o maior número de animais diferentes. Pronto? Começar!"

[contar 60 segundos]

"Pare! Muito obrigado"

Appendix F

Sociodemographic Questionnaire, Portuguese Version

The following sociodemographic questionnaire was developed to obtain subject's characteristics that have an impact in the overall performance in VF tests, such as age, and level of education. Furthermore, more particular questions regarding with whom each subject lives are also present, to attempt to assess the possible level of daily social interaction each subject has.

Sujeito # ____/____ Local: _____ Data: ____/____/____

Informação Demográfica

Gênero: _____ Idade: _____

Habilitações Literárias: _____ Anos de Educação: _____

Estado civil: ☐ Solteiro(a) ☐ Casado(a) ☐ Divorciado(a) ☐ Viúvo(a)
Mora sozinho: ☐ Sim ☐ Não

Appendix G

Short International Physical Activity Questionnaire, Portuguese Version

The Short International Physical Activity Questionnaire, or IPAQ-short, is an internationally available and validated questionnaire to assess the level of activity of subjects within the past seven days [28]. The following pages contain the Portuguese short version of this questionnaire.

The particular version presented here was obtained from Campaniço and Sardinha [20].

Estamos interessados em conhecer os níveis de actividade física habitual dos Portugueses. As suas respostas vão ajudar-nos a compreender o quanto activos somos. As questões referem-se ao tempo que dispende na actividade física numa semana. Este questionário inclui questões acerca de actividades que faz no trabalho, para se deslocar de um lado para outro, actividades referentes à casa ou ao jardim e actividades que efectua no seu tempo livre para entretenimento, exercício ou desporto. As suas respostas são importantes. Por favor responda a todas as questões mesmo que não se considere uma pessoa activa.

Obrigado pela sua participação

Ao responder às seguintes questões considere o seguinte:

Actividade física vigorosa refere-se a actividades que requerem muito esforço físico e tornam a respiração muito mais intensa que o normal.

Actividade física moderada refere-se a actividades que requerem esforço físico moderado e torna a respiração um pouco mais intensa que o normal.

Ao responder às questões considere apenas as actividades físicas que realize durante pelo menos 10 minutos seguidos.

1a Habitualmente, por semana, quantos dias faz actividades físicas **vigorosas** como levantar e/ou transportar objectos pesados, cavar, ginástica aeróbica ou andar de bicicleta a uma velocidade acelerada?

____ dias por semana
____ Nenhum (passe para a questão **2a**)

1b Quanto tempo costuma fazer actividade física vigorosa por dia?

____ horas ____ minutos

2a Normalmente, por semana, quantos dias faz actividade física **moderada** como levantar e/ou transportar objectos leves, andar de bicicleta a uma velocidade moderada ou jogar ténis? Não inclua o andar/caminhar.

____ dias por semana
____ Nenhum (passe para a questão **3a**)

2b Quanto tempo costuma fazer actividade física moderada por dia?

____ horas ____ minutos

3a Habitualmente, por semana, quantos dias **caminha** durante pelo menos 10 minutos seguidos? Inclua caminhadas para o trabalho e para casa, para se deslocar de um lado para outro e qualquer outra caminhada que possa fazer somente para recreação, desporto ou lazer.

____ dias por semana
____ Nenhum (passe para a questão **4a**)

3b Quanto tempo costuma caminhar por dia?

____ horas ____ minutos

3c A que passo costuma caminhar?

____ Passo **vigoroso**, que torna a sua respiração muito mais intensa que o normal;

____ Passo **moderado**, que torna a sua respiração um pouco mais intensa que o normal;

____ Passo **lento**, que não causa qualquer alteração na sua respiração;

As últimas questões referem-se ao tempo que está sentado diariamente no trabalho, em casa, no percurso para o trabalho e durante os tempos livres. Estas questões incluem o tempo em que está sentado numa secretária, a visitar amigos, a ler ou sentado/deitado a ver televisão.

4a Quanto tempo costuma estar sentado num **dia de semana**?

____ horas ____ minutos

4b Quanto tempo costuma estar sentado num **dia de fim-de-semana**?

____ horas ____ minutos

References

- [1] ACT on Alzheimer's (2018). Montreal Cognitive Assessment (MoCA): Administration - YouTube.
- [2] Allen, M. D. and Fong, A. K. (2008). Clinical application of standardized cognitive assessment using fMRI. II. verbal fluency. *Behavioural Neurology*, 20(3-4):141–152.
- [3] American Psychiatric Association (2017). *Diagnostic and Statistical Manual of Mental Disorders (DSM)*, volume 44. American Psychiatric Pub.
- [4] Amorim, P. and Susano, M. J. (2017). Clinical Trial: Pre-operative Neurocognitive Disorder and Low Near-infrared Spectrometry is Associated With Postoperative Delirium.
- [5] Andreou, G. and Trott, K. (2013). Verbal fluency in adults diagnosed with attention-deficit hyperactivity disorder (ADHD) in childhood. *ADHD Attention Deficit and Hyperactivity Disorders*, 5(4):343–351.
- [6] Associação Alzheimer Portugal (2019a). Exames utilizados no Diagnóstico de Demência.
- [7] Associação Alzheimer Portugal (2019b). Fact-Sheet.
- [8] Associação Alzheimer Portugal (2019c). Portugal é o quarto país com mais demência, mas a resposta é fraca.
- [9] Audacity Team (2019). Audacity ® | Free, open source, cross-platform audio software for multi-track recording and editing.
- [10] Audenaert, K., Goethals, I., Van Laere, K., Lahorte, P., Brans, B., Versijpt, J., Vervae, M., Beelaert, L., Van Heeringen, K., and Dierckx, R. (2002). SPECT neuropsychological activation procedure with the Verbal Fluency Test in attempted suicide patients.
- [11] Batty, R., Francis, A., Thomas, N., Hopwood, M., Ponsford, J., Johnston, L., and Rossell, S. (2015). Verbal fluency, clustering, and switching in patients with psychosis following traumatic brain injury (PFTBI). *Psychiatry Research*, 227(2-3):152–159.
- [12] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*.
- [13] Birn, R. M., Kenworthy, L., Case, L., Caravella, R., Jones, T. B., Bandettini, P. A., and Martin, A. (2010). Neural systems supporting lexical search guided by letter and semantic category cues: A self-paced overt response fMRI study of verbal fluency. *NeuroImage*, 49(1):1099–1107.
- [14] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching Word Vectors with Subword Information.

- [15] Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1352–1362.
- [16] Borkowski, J. G., Benton, A. L., and Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia*, 5(2):135–140.
- [17] Braak, H. and Braak, E. (1991). Neuropathological staging of Alzheimer-related changes. *Acta neuropathologica*, 82(4):239–59.
- [18] Brucki, S. M. and Rocha, M. S. (2004). Category fluency test: Effects of age, gender and education on total scores, clustering and switching in Brazilian Portuguese-speaking subjects. *Brazilian Journal of Medical and Biological Research*, 37(12):1771–1777.
- [19] Buckner, R. L. (2004). Memory and executive function in Aging and AD: Multiple factors that cause decline and reserve factors that compensate.
- [20] Campaniço, M. and Sardinha, L. (2016). *Validade Simultânea do Questionário Internacional de Actividade Física Através da Medição Objectiva da Actividade Física por Actigrafia Proporcional*. PhD thesis.
- [21] Cavaco, S., Gonçalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., Fernandes, J., and Teixeira-Pinto, A. (2013). Semantic fluency and phonemic fluency: Regression-based norms for the portuguese population. *Archives of Clinical Neuropsychology*.
- [22] Cerf, V. (1969). ASCII format for network interchange. Technical report.
- [23] Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. (2018). State-of-the-Art Speech Recognition with Sequence-to-Sequence Models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2018-April, pages 4774–4778.
- [24] Clark, D. G., McLaughlin, P. M., Woo, E., Hwang, K., Hurtz, S., Ramirez, L., Eastman, J., Dukes, R. M., Kapur, P., DeRamus, T. P., and Apostolova, L. G. (2016). Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment. *Alzheimer’s and Dementia: Diagnosis, Assessment and Disease Monitoring*, 2:113–122.
- [25] Clark, L. J., Gatz, M., Zheng, L., Chen, Y.-L., McCleary, C., and Mack, W. J. (2009). Longitudinal verbal fluency in normal aging, preclinical, and prevalent Alzheimer’s disease. *American journal of Alzheimer’s disease and other dementias*, 24(6):461–8.
- [26] CMUSphinx (2019). CMUSphinx Open Source Speech Recognition.
- [27] Costafreda, S. G., Fu, C. H., Lee, L., Everitt, B., Brammer, M. J., and David, A. S. (2006). A systematic review and quantitative appraisal of fMRI studies of verbal fluency: Role of the left inferior frontal gyrus.
- [28] Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., Pratt, M., Ekelund, U. L. F., Yngve, A., Sallis, J. F., and others (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise*, 35(8):1381–1395.

- [29] de Paiva, V. and Rademaker, A. (2012). Revisiting a Brazilian WordNet. *GWC 2012 6th International Global Wordnet Conference Proceedings*, pages 100–104.
- [30] Deary, I. J., Corley, J., Gow, A. J., Harris, S. E., Houlihan, L. M., Marioni, R. E., Penke, L., Rafnsson, S. B., and Starr, J. M. (2009). Age-associated cognitive decline.
- [31] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.
- [32] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
- [33] Farzanfar, D., Statucka, M., and Cohn, M. (2018). Automated Indices of Clustering and Switching of Semantic Verbal Fluency in Parkinson’s Disease. *Journal of the International Neuropsychological Society*, 24(10):1047–1056.
- [34] Fensel, D. (2001). Ontologies : Silver Bullet for Knowledge. *Springer*, page 138.
- [35] Fitzgerald, J. M. (1983). A developmental study of recall from natural categories. *Developmental Psychology*, 19(1):9–14.
- [36] Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state. A grading the cognitive state of patients for the clinician. *J Psychiatr Res*, 12:189–198.
- [37] Fraunhofer AICOS (2019). COLABORAR.
- [38] Freitas, S. (2011). *Envelhecimento e Déficit Cognitivo: Estudos de Adaptação, Validação e Normalização do Montreal Cognitive Assessment (MoCA)*. PhD thesis, Universidade de Coimbra.
- [39] Freitas, S., Simões, M. R., Alves, L., and Santana, I. (2011). Montreal Cognitive Assessment (MoCA): Normative study for the Portuguese population. *Journal of Clinical and Experimental Neuropsychology*, 33(9):989–996.
- [40] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 1606–1611.
- [41] Glikmann-Johnston, Y., Oren, N., Hendler, T., and Shapira-Lichter, I. (2015). Distinct functional connectivity of the hippocampus during semantic and phonemic fluency. *Neuropsychologia*, 69:39–49.
- [42] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 1, pages 517–520. IEEE.
- [43] Goldberg, Y. and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method.
- [44] Gonçalves Oliveira, H. and Gomes, P. (2014). ECO and Onto.PT: A flexible approach for creating a Portuguese wordnet automatically. *Language Resources and Evaluation*, 48(2):373–393.

- [45] Google Cloud (2019a). Cloud Speech-to-Text - Speech Recognition | Cloud Speech-to-Text API.
- [46] Google Cloud (2019b). Cloud Storage - Online Data Storage | Cloud Storage | Google Cloud.
- [47] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning Word Vectors for 157 Languages.
- [48] Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2019). Word vectors for 157 languages · fastText.
- [49] Haapala, A., Määttä, E., and Necas, D. (2017). GitHub - ztane/python-Levenshtein: The Levenshtein Python C extension module contains functions for fast computation of Levenshtein distance and string similarity.
- [50] Hendelman, W. (2015). *Atlas of Functional Neuroanatomy*. CRC Press.
- [51] Henry, J. D. and Crawford, J. R. (2004). A Meta-Analytic Review of Verbal Fluency Performance Following Focal Cortical Lesions.
- [52] Henry, J. D., Crawford, J. R., and Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: A meta-analysis. *Neuropsychologia*, 42(9):1212–1222.
- [53] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, (November):82–97.
- [54] Ho, A. K., Sahakian, B. J., Robbins, T. W., Barker, R. A., Rosser, A. E., and Hodges, J. R. (2002). Verbal fluency in Huntington's disease: A longitudinal analysis of phonemic and semantic clustering and switching. *Neuropsychologia*, 40(8):1277–1284.
- [55] Hurks, P. P., Hendriksen, J. G., Vles, J. S., Kalff, A. C., Feron, F. J., Kroes, M., Van Zeben, T. M., Steyaert, J., and Jolles, J. (2004). Verbal fluency over time as a measure of automatic and controlled processing in children with ADHD. *Brain and Cognition*, 55(3):535–544.
- [56] IBM (2019). Watson Speech to Text.
- [57] International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*.
- [58] IPAQ Research (2005). Guidelines for data processing and analysis of the international physical activity questionnaire (IPAQ)—Short and long forms.
- [59] Jurado, M. B. and Rosselli, M. (2007). The elusive nature of executive functions: A review of our current understanding.
- [60] Kasten, M., Bruggemann, N., Schmidt, A., Klein, C., Weintraub, D., Siderowf, A. D., and Xie, S. X. (2010). Validity of the MoCA and MMSE in the Detection of Dementia in Parkinson Disease. *Neurology*, 75(5):478–479.
- [61] König, A., Linz, N., Tröger, J., Wolters, M., Alexandersson, J., and Robert, P. (2018). Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and Geriatric Cognitive Disorders*, 45(3-4):198–209.

- [62] König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., Manera, V., Verhey, F., Aalten, P., Robert, P. H., and David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 1(1):112–124.
- [63] Kral, V. A. (1962). Senescent forgetfulness: benign and malignant. *Canadian Medical Association journal*, 86:257–260.
- [64] Landauer, T. K. and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211–240.
- [65] Lang, A. E. and Lozano, A. M. (1998). Parkinson's Disease. *New England Journal of Medicine*, 339(16):1130–1143.
- [66] Laurin, D., Verreault, R., Lindsay, J., MacPherson, K., and Rockwood, K. (2001). Physical activity and risk of cognitive impairment and dementia in elderly persons. *Archives of neurology*, 58(3):498–504.
- [67] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.
- [68] Levy, R. (1994). Report Aging-Associated Cognitive Decline. *International Psychogeriatrics*, 6(1):63–68.
- [69] Lezak, M. D., Howieson, D. B., Loring, D. W., and Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press, USA.
- [70] Linz, N., Tröger, J., Alexandersson, J., and König, A. (2017). Using Neural Word Embeddings in the Analysis of the Clinical Semantic Verbal Fluency Task. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers.*, volume 1.
- [71] Lopez, O. L., Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., and Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome.
- [72] Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.
- [73] Manning, C. D., Raghavan, P., Schutze, H., Manning, C. D., Raghavan, P., and Schutze, H. (2012). Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, pages 100–123. Cambridge University Press, Cambridge.
- [74] Mayr, U. (2002). On the dissociation between clustering and switching in verbal fluency: Comment on Troyer, Moscovitch, Winocur, Alexander and Stuss.
- [75] michmech (2018). GitHub - michmech/lemmatization-lists: Machine-readable lists of lemma-token pairs in 23 languages.
- [76] Microsoft Azure (2019). Speech to Text API.
- [77] Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). Introduction to WordNet: On-line Database. *International journal of lexicography*, 3(4):235–244.
- [78] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

- [79] Nasreddine, Z. (2019). MoCA Montreal - Cognitive Assessment.
- [80] Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.
- [81] NLTK Project (2019). Natural Language Toolkit. <https://www.nltk.org/>.
- [82] Nogueira, D. S., Reis, E. A., and Vieira, A. (2017). Verbal fluency tasks: Effects of age, gender, and education. *Folia Phoniatrica et Logopaedica*, 68(3):124–133.
- [83] Nutter-Upham, K. E., Saykin, A. J., Rabin, L. A., Roth, R. M., Wishart, H. A., Pare, N., and Flashman, L. A. (2008). Verbal fluency performance in amnesic MCI and older adults with cognitive complaints. *Archives of Clinical Neuropsychology*, 23(3):229–241.
- [84] Orengo, V. M. and Huyck, C. (2001). A stemming algorithm for the Portuguese language. In *Proceedings - 8th Symposium on String Processing and Information Retrieval, SPIRE 2001*, pages 186–193.
- [85] Pakhomov, S. V., Eberly, L., and Knopman, D. (2016). Characterizing cognitive performance in a large longitudinal study of aging with computerized semantic indices of verbal fluency. *Neuropsychologia*, 89:42–56.
- [86] Pakhomov, S. V. and Hemmy, L. S. (2014). A computational linguistic measure of clustering behavior on semantic verbal fluency task predicts risk of future dementia in the Nun Study. *Cortex*, 55(1):97–106.
- [87] Pakhomov, S. V., Hemmy, L. S., and Lim, K. O. (2012). Automated semantic indices related to cognitive function and rate of cognitive decline. *Neuropsychologia*, 50(9):2165–2175.
- [88] Parks, R. W., Loewenstein, D. A., Dodrill, K. L., Barker, W. W., Yoshii, F., Chang, J. Y., Emran, A., Apicella, A., Sheramata, W. A., and Duara, R. (1988). Cerebral metabolic effects of a verbal fluency test: A PET scan study. *Journal of Clinical and Experimental Neuropsychology*, 10(5):565–575.
- [89] Patwardhan, S. and Pedersen, T. (2006). Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense Bringing Computational Linguistics and Psycholinguistics Together*, number Proceedings of the Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together, pages 1–8.
- [90] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.
- [91] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- [92] Peter, J., Kaiser, J., Landerer, V., Köstering, L., Kaller, C. P., Heimbach, B., Hüll, M., Bormann, T., and Klöppel, S. (2016). Category and design fluency in mild cognitive impairment: Performance, strategy use, and neural correlates. *Neuropsychologia*, 93:21–29.

- [93] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations.
- [94] Piatt, A. L., Fields, J. A., Paolo, A. M., Koller, W. C., and Tröster, A. I. (2003). Lexical, Semantic, and Action Verbal Fluency in Parkinson's Disease with and without Dementia. *Journal of Clinical and Experimental Neuropsychology*, 21(4):435–443.
- [95] Porto Editora (2019). Infopédia - Dicionários Porto Editora.
- [96] Povey, D. (2019). Kaldi ASR.
- [97] Raoux, N., Amieva, H., Le Goff, M., Auriacombe, S., Carcaillon, L., Letenneur, L., and Dartigues, J. F. (2008). 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.
- [98] Raskin, S. A., Borod, J. C., and Tweedy, J. (1990). Neuropsychological aspects of Parkinson's disease. *Neuropsychology Review*, 1(3):185–221.
- [99] Regard, M., Strauss, E., and Knapp, P. (1982). Children's Production on Verbal and Non-Verbal Fluency Tasks. *Perceptual and Motor Skills*, 55(3):839–844.
- [100] Reis, A. and Castro-Caldas, A. (1997). Illiteracy: A cause for biased cognitive development. *Journal of the International Neuropsychological Society*, 3:444–450.
- [101] Rodrigues, I. T., Ferreira, J. J., Coelho, M., Rosa, M. M., and Castro-Caldas, A. (2015). Action verbal fluency in Parkinson's patients. *Arquivos de Neuro-Psiquiatria*, 73(6):520–525.
- [102] Rodríguez-Aranda, C., Waterloo, K., Johnsen, S. H., Eldevik, P., Sparr, S., Wikran, G. C., Herder, M., and Vangberg, T. R. (2016). Neuroanatomical correlates of verbal fluency in early Alzheimer's disease and normal aging. *Brain and Language*, 155-156:24–35.
- [103] Romanowski, B. (2011). GitHub - romanows/WordSequenceAligner: Align word sequences and calculate metrics like word error rate (WER).
- [104] Ryan, J. O., Pakhomov, S. V., Marino, S., Bernick, C., and Banks, S. (2013). Computerized Analysis of A Verbal Fluency Test. *Proceedings of ACL*, pages 884–889.
- [105] Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of aging*, 30(4):507–14.
- [106] Santana, I., Farinha, F., Freitas, S., Rodrigues, V., and Carvalho, A. (2015). Estimativa da prevalência da demência e da doença de Alzheimer em Portugal. *Acta Médica Portuguesa*, 7:182–188.
- [107] Senhorini, M. C., Cerqueira, C. T., Schaufelberger, M. S., Almeida, J. C., Amaro, E., Sato, J. R., Barreiros, M. A., Ayres, A. M., Castro, C. C., Scazufca, M., Menezes, P. R., and Busatto, G. F. (2011). Brain activity patterns during phonological verbal fluency performance with varying levels of difficulty: A functional magnetic resonance imaging study in Portuguese-speaking healthy individuals. *Journal of Clinical and Experimental Neuropsychology*, 33(8):864–873.
- [108] Simões, A. and Guinovart, X. G. (2014). Bootstrapping a Portuguese WordNet from Galician, Spanish and English Wordnets. In *Advances in Speech and Language Technologies for Iberian Languages*, Springer, pages 239–248.

- [109] Smith, T., Gildeh, N., and Holmes, C. (2007). The Montreal cognitive assessment: Validity and utility in a memory clinic setting. *Canadian Journal of Psychiatry*, 52(5):329–332.
- [110] Swank, A. M. (1996). Physical Dimensions of Aging. *Medicine & Science in Sports & Exercise*, 28(3):398,399.
- [111] Tombaugh, T. . N. ., Kozak, J. ., and Rees, L. . (1999). Normative data stratified by age and education for two measures of verbal fluency : FAS and Animal Naming. *Archives of Clinical Neuropsychology*, 14(2):13–14.
- [112] Tröger, J., Linz, N., König, A., Robert, P., and Alexandersson, J. (2018). Telephone-based Dementia Screening I: Automated Semantic Verbal Fluency Assessment. *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare - PervasiveHealth '18*, pages 59–66.
- [113] Troyer, A. K. (2000). Normative Data for Clustering and Switching on Verbal Fluency Tasks. *Journal of Clinical and Experimental Neuropsychology*, 22(3):370–378.
- [114] Troyer, A. K., Moscovitch, M., and Winocur, G. (1997). Clustering and switching as two components of verbal fluence: Evidence from younger and older healthy adults. *Neuropsychology*, 11(1):138–146.
- [115] Troyer, A. K., Moscovitch, M., Winocur, G., Alexander, M. P., and Stuss, D. (1998a). Clustering and switching on verbal fluency: the effects of focal frontal- and temporal-lobe lesions. *Neuropsychologia*, 36(6):499–504.
- [116] Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., and Freedman, M. (1998b). Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. *Journal of the International Neuropsychological Society*, 4(2):137–143.
- [117] Tucha, O., Mecklinger, L., Laufkötter, R., Kaunzinger, I., Paul, G. M., Klein, H. E., and Lange, K. W. (2005). Clustering and switching on verbal and figural fluency functions in adults with attention deficit hyperactivity disorder. *Cognitive Neuropsychiatry*, 10(3):231–248.
- [118] United Nations (2019). World Population Prospects - Population Division.
- [119] Weide, R. (2005). The Carnegie mellon pronouncing dictionary [cmudict. 0.6].
- [120] Whiteside, D. M., Kealey, T., Semla, M., Luu, H., Rice, L., Basso, M. R., and Roper, B. (2016). Verbal Fluency: Language or Executive Function Measure? *Applied Neuropsychology:Adult*, 23(1):29–34.
- [121] Woods, D. L., Wyma, J. M., Herron, T. J., and Yund, E. W. (2016). 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.
- [122] World Health Organization (2019). Dementia.
- [123] Xiaowei Xu, Ester, M., Kriegel, H.-P., and Sander, J. (2002). A distribution-based clustering algorithm for mining in large spatial databases. In *Proceedings 14th International Conference on Data Engineering*, pages 324–331. IEEE Comput. Soc.

- [124] Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The Microsoft 2017 Conversational Speech Recognition System. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2018-April, pages 5934–5938.
- [125] Yergeau, F. (1996). UTF-8, a transformation format of Unicode and ISO 10646. Technical report.
- [126] Zhao, Q., Guo, Q., and Hong, Z. (2013). Clustering and switching during a semantic verbal fluency test contribute to differential diagnosis of cognitive impairment. *Neuroscience Bulletin*, 29(1):75–82.