Computer-based Characterization of Language Alterations Throughout the Alzheimer's Disease Continuum

by

Laura Elena HERNÁNDEZ DOMÍNGUEZ

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FORTHE DEGREE OF DOCTOR OF PHILOSOPHY Ph.D.

MONTREAL, JANUARY 24, 2019

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Sylvie Ratté, Thesis Supervisor
Software and IT Engineering Department at École de technologie supérieure

Mr. Gerardo Eugenio Sierra Martínez, Thesis Co-supervisor
Engineering Institute at Universidad Nacional Autónoma de México

Mr. Mohamed Cheriet, President of the Board of Examiners
Automated Production Engineering Department at École de technologie supérieure

Mr. Luc Duong, Member of the jury
Software and IT Engineering Department at École de technologie supérieure

Mrs. Simona Maria Brambati, External Evaluator
Department of Psychology at Université de Montréal

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

ON JANUARY 17, 2019

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

FOREWORD

This Ph.D. dissertation presents my research work carried out between 2013 and 2018 at École de technologie supérieure, under the supervision of Professor Sylvie Ratté. This is a thesis based on articles required for the culmination of the Ph.D. program in engineering (applied research).

The present work is part of the *Cécilia* project, whose main objective is to study the language alterations caused by Alzheimer's disease and other dementias. The studies carried out by the *Cécilia* team aim at characterizing changes in verbal and non-verbal communication through the analysis of transcriptions, speech, facial and corporal expression.

The objective of this research is to propose computer-based methodologies to characterize alterations in verbal communication that occur throughout the Alzheimer's disease continuum. The resulting methodologies will be later incorporated into multi-modal analyses of patients to aid physicians in the detection and monitoring of Alzheimer's patients.

This work resulted in a total of 3 journal and 4 conference papers, published or under peer review, plus 1 book chapter and 1 public outreach article, for which I am the first author. With relation to this project, I am also coauthor of 2 journal and 2 conference papers. This dissertation is centered around the journal papers for which I am the first author, which are presented in Chapters 2, 3 and 4. A concomitant contribution of my Ph.D. project was the creation of the Latin-American cohort of the Carolinas Conversations Collection. This work was presented in one conference paper and is the focus of Chapter 5. Chapter 6 provides a general discussion, and the final conclusions and future work are presented in Chapter 7. All the publications derived from my doctoral work are listed in Appendix II.

ACKNOWLEDGMENT

I would like to profusely thank Prof. Sylvie Ratté for her encouragement and guidance during my Ph.D. studies. I consider her to be an excellent researcher and professor, but it is her humane, empathic and compassionate nature that made working alongside her such a rewarding experience. I will always be grateful to her for proposing me to work in the Cécilia project, which I cherish deeply, and I am proud to have helped grow. Her support during the tribulations that I faced while pursuing my degree are priceless and could never be overstated.

I also wish to thank Prof. Gerardo Sierra Martínez for his counsel as codirector of my Ph.D. studies, and his support during my stay at the Linguistic Engineering Group in Mexico. His help with coordinating the recollection and transcription of conversations provided us with invaluable resources for the Cécilia Project.

Further, I would like to thank the jury members, Prof. Luc Duong, Prof. Mohamed Cheriet and Dr. Simona Brambati, for accepting to review my thesis, and for their meaningful comments and discussions about my project.

A very special thanks also to Prof. Boyd Davis and Charlene Pope for all their guidance and suggestions to help us in the creation of the Latin-American cohort of the Carolinas Conversations Collection. Likewise, to all the participants of the Ecuador and Mexico recollections, to their caregivers for their time and patience, and to the interviewers and transcribers, who have greatly contributed to the creation of this cohort. I would also like to thank the Psychiatric Hospital “Fray Bernardino Álvarez” in Mexico City, especially Dr. Andrés Roche Bergua, Dr. Alexiz Bojorge and Dr. Janet Jiménez Genchi, for their support and clinical insights. My sincere thanks also go to all the staff members of the “Fundación Perpetuo Socorro” in Quito, Ecuador, especially Dr. Edwin Velasco and Mrs. Amparo Sarabia.

I would also like to express my great appreciation to Prof. Annette Gerstenberg, from Potsdam University, who gave us access to her wonderful LangAge corpus, for her insights, comments and fruitful discussions.

For sure, I want to express my gratitude to all LINCS and LIVE members—old and new—, for their help, their insights and, above all, for always making me feel at home in the lab. Also, to the MFO team, particularly Mizar, Luc, Laura and Isabelle, who helped me regain my strength and keep my sanity. Moreover, I want to thank my friends for their conversations, companionship and support: Lorena and Alejandro, in situ; and Paola and Rafael, notwithstanding the distance. Similarly, all my appreciation to Raymond, for his invaluable help and advice with Arrow, and to Marc and Foti for their awesome camaraderie.

I also thank my parents, for being on Skype every Sunday, and for all their help and food through the hardest moments of this stage of my life. To my brother, for suggesting Danby. And lastly, but most importantly, I cannot put into words my appreciation for Edgar, for being there, in health and in sickness, as my accomplice in this adventure.

Finally, I would like to acknowledge the *Fonds de recherche du Québec - Nature et technologies* (FRQNT; 177601) and the *Consejo Nacional de Ciencia y Tecnología* (CONACYT; 231979), the joint project of the *Ministère des Relations internationales et de la Francophonie* Quebec-CONACYT, and the PAPIIT project IA400117, which all made this project possible.

Caractérisation assistée par ordinateur des altérations du langage tout au long du développement de la maladie d'Alzheimer

Laura Elena HERNÁNDEZ DOMÍNGUEZ

RÉSUMÉ

Selon la Société de l'Alzheimer du Canada, et l'Alzheimer Society des États-Unis, il est impératif de rechercher des méthodes de détection précoce de la maladie d'Alzheimer. Beaucoup d'études ont mis en évidence les nombreux avantages de la détection de la maladie au stade préclinique pour les patients, les membres de leurs familles et les gouvernements. Cependant, à ce stade, les changements sont très subtils, ce qui rend difficile leur détection.

Des altérations des fonctions langagières ont été constatées des années avant le stade de démence de la maladie d'Alzheimer. Pour cette raison, de nombreux chercheurs ont concentré leurs efforts sur la recherche de méthodes permettant d'identifier des indices de la présence de la maladie cachés dans le langage.

Les tâches standardisées de description d'images font partie de tests cognitifs couramment utilisés en pratique clinique. Ces tâches ont pour objectif d'encourager les patients à décrire un stimulus visuel. Ces tests présentent l'avantage de limiter le discours des patients à un thématique restreinte; cela a pour effet de circonscrire le vocabulaire et de faciliter les comparaisons inter-patients et inter-langues. Cependant, ils limitent également la diversité des structures syntaxiques en entravant certaines analyses linguistiques. De plus, comme ils font partie des examens cliniques habituels, ils peuvent augmenter la nervosité chez certains patients.

L'étude de conversations spontanées est une alternative à l'utilisation de tâches de description d'image pour l'analyse du langage. Les conversations spontanées ont l'avantage de permettre l'utilisation de structures syntaxiques sans contrainte et d'un vocabulaire idiosyncratique. Elles sont également moins stressantes pour les patients et pourraient être conduites avec une infirmière, un soignant ou une personne familière au patient. Néanmoins, de nombreux facteurs, tels que les différences sociodémographiques et culturelles, peuvent définir les caractéristiques linguistiques des individus. Par conséquent, une caractérisation des changements dans les fonctions du langage qui se produisent pendant le développement de la maladie pourrait être utile dans le monitoring des changements spécifiques du patient.

Cette thèse de doctorat présente une méthodologie assistée par ordinateur qui évalue les performances des patients durant les tâches standardisées de description d'images, et l'évaluation des fonctions linguistiques dans le contexte de ces tâches et dans des conversations spontanées. Nous pensons que les deux évaluations peuvent se complémenter mutuellement et constituer une méthode peu coûteuse et non invasive de monitoring des fonctions langagières.

Dans la pratique, les tâches de description d'images peuvent être tenus de manière routinière chez le médecin, tandis que les conversations spontanées peuvent avoir lieu à intervalles plus réguliers, aux endroits plus pratiques pour le patient.

Dans cette thèse, nous avons comparé les performances langagières et les fonctions linguistiques des patients durant l'exécution des tâches de description d'images à celles d'une population présentant des caractéristiques sociodémographiques similaires. Pour cela, notre méthode a évalué l'informativité et la pertinence des descriptions des patients, ainsi que leur richesse lexicale. Nous avons entraîné des algorithmes d'apprentissage automatique avec nos métriques pour estimer leur capacité à différencier les patients d'Alzheimer des témoins en santé. Nous avons obtenu une surface sous la courbe de 0,83 pour cette tâche. De plus, pour la classification des témoins en santé et des patients présentant une déficit cognitive légère, qui est souvent un précurseur préclinique de la maladie d'Alzheimer, nous avons atteint une surface sous la courbe de 0,79.

En outre, nous avons proposé une méthode automatisée d'évaluation de la richesse lexicale, de la distribution du vocabulaire, de la fluidité de la parole et de l'utilisation de structures syntaxiques spécifiques des personnes âgées francophones durant des conversations spontanées. Nous avons décrit les changements subis par quatre locuteurs lors du passage d'un état de santé à une forme de maladie cognitive, comprenant la maladie d'Alzheimer. Nous avons observé des différences marquées dans les mesures que nous proposons chez les individus susceptibles de développer une maladie cognitive et des témoins apparemment sains, et ce, même lors de l'analyse des transcriptions de conversations qui ont eu lieu dix ans avant le diagnostic.

En tant que contribution concomitante de ce travail de doctorat, nous avons conçu le protocole et créé la cohorte espagnole du *Carolinas' Conversations Collection*. Cette cohorte comprend des enregistrements vidéo longitudinaux et des transcriptions de conversations spontanées avec des personnes âgées hispanophones au Mexique et en Équateur. Ces récoltes sont le résultat des efforts combinés de six institutions de quatre pays différents, et seront disponibles à des fins de recherche sur demande. Cette entreprise vise à réduire la rareté des données de ce type et à encourager la recherche sur la langue et la communication chez les personnes âgées.

Mots-clés: maladie d'Alzheimer; détection précoce; traitement du langage naturel; apprentissage automatique; fonctions langagières; tâches de description d'images; conversations spontanées.

Computer-based characterization of language alterations throughout the Alzheimer's disease continuum

Laura Elena HERNÁNDEZ DOMÍNGUEZ

ABSTRACT

According to the American and Canadian Alzheimer's Associations, research into methods for the early detection of Alzheimer's disease is imperative. Many studies have emphasized the numerous advantages for patients, family members and governments of detecting the disease at the pre-clinical stage of its continuum. However, at this stage, changes are very subtle, making their detection a challenging task.

Alterations in language functions have been found years before the dementia stage of the disease continuum. For this reason, many researchers have focused their efforts on investigating methods for identifying cues of the presence of the disease hidden in language.

One type of cognitive test commonly used in this type of research consists of standardized picture description tasks. These tasks elicit the speech of patients through a visual stimulus, and are usually part of cognitive assessment batteries used in clinical practice. The tasks have the advantage of presenting patients with a single constrained thematic, which limits the vocabulary and facilitates comparisons across patients and languages. However, they also limit the variety of syntactic structures, hindering some linguistic analyses, and being a part of usual clinical examinations, may increase nervousness in some patients.

The study of spontaneous conversations is an alternative to using picture description tasks for language analyses. Spontaneous conversations have the advantage of allowing the use of unconstrained idiosyncratic syntactic structures and vocabulary. They are also less stressful to patients and could be conducted with a nurse, a caregiver or a person familiar to the patient. Nevertheless, many factors, such as socio-demographic and cultural differences, may define the linguistic characteristics of individuals. Consequently, a characterization of the changes in language functions that occur during the continuum of the disease could be helpful in the monitoring of patient-specific changes.

This doctoral thesis presents a computer-based methodology for evaluating patients' performance during standardized picture description tasks, and for assessing language functions in the context of these tasks and in spontaneous conversations. We believe that both evaluations can complement each other and provide an inexpensive and noninvasive method for monitoring language functions. In practice, picture description tasks could be realized routinely at the doctor's office, while spontaneous conversations could be held at more regular intervals and at more convenient locations for the patient.

For our work, we compared the computed performance and language functions of patients during standardized picture description tasks against a population with similar socio-demographic characteristics. For this, our proposed method evaluated the informativeness and pertinence of the descriptions of patients, as well as their lexical richness. Using our metrics, we trained machine learning algorithms to estimate their adeptness at differentiating Alzheimer's patients from healthy controls. We obtained an area under the curve of 0.83 in this task. We also achieved an area under the curve of 0.79 for classifying healthy controls and patients with mild cognitive impairment, which is often a pre-clinal precursor of Alzheimer's disease.

In addition, we proposed an automated method for evaluating lexical richness, vocabulary distribution, speech fluidity and the use of specific syntactic structures among older French speakers during spontaneous conversations. We characterized the changes that four speakers underwent as they transitioned from a healthy state to some form of cognitive disease, including Alzheimer's disease. We observed marked differences in our proposed metrics between those individuals that would develop a cognitive disease and healthy matched controls, even when analyzing transcriptions of conversations from up to ten years before the time of diagnosis.

As a concomitant contribution of this doctoral work, we designed the protocol and created the Spanish cohort of the Carolinas' Conversations Collection. This cohort includes longitudinal video-recordings and transcriptions of spontaneous conversations of older Spanish speakers in Mexico and Ecuador. These recollections are the result of the combined efforts of six institutions from four different countries, and will be available for research purposes upon request. This undertaking is aimed at lessening the scarcity of data of this type, and at encouraging research on language and communication in the older population.

Keywords: Alzheimer's disease; early detection; natural language processing; machine learning; language functions; picture description tasks; spontaneous conversations.

TABLE OF CONTENTS

Page

INTRODUCTION 1
0.1 Problem statement and motivation 1
0.2 Research objectives and contributions 3
0.3 Outline 7

CHAPTER 1 LITERATURE REVIEW 9
1.1 Alzheimer's disease overview 9
1.2 Detection of language alterations in Alzheimer's disease 13
1.2.1 State of the art on the automatic analysis of picture description tasks 15
1.2.2 State of the art on the automatic analysis of spontaneous conversations 22
1.2.3 Summary 32

CHAPTER 2 COMPUTER-BASED EVALUATION OF AD AND MCI PATIENTS DURING A PICTURE DESCRIPTION TASK 35
2.1 Abstract 35
2.2 Introduction and Motivation 36
2.2.1 Information coverage 37
2.2.2 Linguistic characteristics 39
2.2.3 Phonetic analysis 40
2.3 Methods 42
2.3.1 Corpus 42
2.3.2 Extraction of information coverage measures 43
2.3.3 Extraction of linguistic and phonetic characteristics 46
2.3.4 Automatic classification 47
2.4 Results 48
2.4.1 Feature analysis 48
2.4.2 Binary classification 48
2.5 Discussion 53
2.5.1 Comparison to other approaches 54
2.5.2 Study advantages and limitations 55
2.5.3 Future work 56

CHAPTER 3 AUTOMATED DIFFERENTIATION OF ALZHEIMER'S AND MCI PATIENTS FROM HEALTHY CONTROLS USING ENGLISH AND SPANISH TRANSCRIPTIONS OF DESCRIPTION TASKS 57
3.1 Abstract 57
3.2 Introduction 58
3.3 Materials and Methods 60
3.3.1 Corpora 61
3.3.2 Preprocessing 64
3.4 Results 69
3.5 Discussion 71

3.6 Conclusions....76

CHAPTER 4 AGING WITH AND WITHOUT COGNITIVE DISEASES: CHARACTERIZING 10 YEARS OF LANGUAGE DIFFERENCES IN OLDER FRENCH SPEAKERS....79
4.1 Abstract....79
4.2 Introduction....80
4.3 Materials and Methods....82
4.3.1 LangAge Corpus....82
4.3.2 Pre-processing....83
4.3.3 Extraction of characteristics....84
4.4 Results....90
4.5 Discussion....94
4.5.1 Behavior of the composite variables....94
4.5.2 Differentiation of CI and healthy controls....100
4.5.3 Mini case study: Participant 13 and matched control 37....102
4.5.4 Mini case study: Participant 25 and matched control 11....103
4.5.5 Mini case study: Participant 48 and matched control 47....104
4.5.6 Mini case study: Participant 27 and matched control 18....106
4.6 Limitations....107
4.7 Conclusions....108

CHAPTER 5 CONVERSING WITH THE ELDERLY IN LATIN AMERICA: A NEW COHORT FOR MULTIMODAL, MULTILINGUAL LONGITUDINAL STUDIES ON AGING....111
5.1 Abstract....111
5.2 Introduction....112
5.3 Methodology....113
5.4 Description of the samples....115
5.5 Implications, applications and prospects....116
5.5.1 Improving communication....118
5.5.2 Medical applications....118
5.6 Conclusions and future work....119

CHAPTER 6 GENERAL DISCUSSION....121
6.1 Evaluation of performance and language functions in elicited speech in cognitive testing settings....121
6.2 Classification of healthy and cognitively impaired individuals from restricted and semi-restricted discourses....123
6.3 Longitudinal characterization of language alterations in spontaneous speech....126
6.4 The Latin-American cohort of the Carolinas' Conversation Collection....127

CONCLUSION AND FUTURE WORK....129

BIBLIOGRAPHY....137

LIST OF TABLES

Page

Table 2.1 Linguistic characteristics selected to evaluate patients' language functions 41

Table 2.2 Distribution of interviews used for experimentation 42

Table 2.3 Active voice linguistic patterns used for the coverage measure 43

Table 2.4 Correlations* of features with the severity of cognitive impairment and with the MMSE 49

Table 2.5 Correlations* of features with sociodemographic variables 50

Table 2.6 Performance* of classifiers separating HCs from AD patients 51

Table 2.7 Performance* of classifiers separating HCs from cognitively impaired patients (AD or MCI, indistinctly) 52

Table 3.1 Distribution of the cohorts in the constrained-discourse corpora used for experimentation 61

Table 3.2 Lexical richness features 68

Table 3.3 Features significantly ($p < .05$) correlated with severity of cognitive impairment, controlled for age, education, gender and number of words in the description..70

Table 3.4 Performance metrics of the learners with both corpora for classification of healthy controls and individuals with MCI and/or AD. 71

Table 4.1 Sample of cognitive impaired subjects and their closest healthy control 83

Table 4.2 Measures for lexical richness evaluated 85

Table 4.3 Correlation with severity of cognitive impairment. Bold font indicates negative correlation 91

Table 4.4 Table of the rotated loadings matrix for the action POS n-gram ratios group*..92

Table 4.5 Correlation with severity of cognitive impairment (column 1), and factor loadings for composite variables* (loading coefficients less than 0.3 were excluded).....93

Table 5.1 Socio-demographic overview of the participants in the collection.....116

Table 5.2 Prevalence of the main mental health disorders in each cohort.....119

LIST OF FIGURES

Page

Figure 1.1 The Cookie Theft picture from the Boston Diagnostic 14

Figure 2.1 Linguistic variations of ICUs in the Cookie Theft……..….45

Figure 2.2 Data set partitioning during a 10-fold cross-validation process.....47

Figure 4. Typical TFID curve 87

Figure 4.2 Behavior of the composite variable from lexical richness measures over time: **a)** Distribution in box plots of each type of aging. **b)** Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample. 95

Figure 4.3 Behavior of the composite variable from the characteristics based on vocabulary distribution over time: **a**) Distribution in box plots of each type of aging. **b**) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample. 96

Figure 4.4 Behavior of the *skewness of utterances' subjectivity* variable over time: **a**) Distribution in box plots of each type of aging. **b**) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample. 97

Figure 4.5 Behavior of the Passive POS n-gram ratios compound variable over time: a) Distribution in box plots of each type of aging. b) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample. 99

Figure 4.6 Behavior of the use of simple verbs' and the future and past tense verbs' (action POS n-gram) ratios' compound variables over time: a) Distribution in box plots of each type of aging. b) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.....100

Figure 4.7 Scatter plot of the distribution of interviews with Components 1 and 3 from the PCA using composite variables. The label near each point indicates the participant ID-interview number. Interviews 1, 2 and 3 where held in 2005, 2012 and 2015, respectively. The hue difference indicates normal or cognitively declined aging processes. Circle, square and rhomboid markers indicate healthy control (HC), mild cognitive impairment (MCI) and severe CI, respectively, at the time of the interview. 101

Figure 4.8 Comparison of the behavior of all significant composite variables and PCA components over time in participants 13 (CI group; dashed lines) and 37 (healthy matched control; continuous lines). 102

Figure 4.9 Comparison of the behavior of all significant composite variables and PCA components over time in participants 25 (CI group; dashed lines) and 11 (healthy matched control; continuous lines). 104

Figure 4.10 Comparison of the behavior of all significant composite variables and PCA components over time in participants 48 (CI group; dashed lines) and 47 (healthy matched control; continuous lines). 105

Figure 4.11 Comparison of the behavior of all significant composite variables and PCA components over time in participants 27 (CI group; dashed lines) and 18 (healthy matched control; continuous lines). 107

Figure 5.1 Multimodal studies available in the corpus……...117

LIST OF ABREVIATIONS

AD	Alzheimer's disease
AUC	Area under the curve
BBVA	Banco Bilbao Vizcaya Argentaria
CI	Cognitive impairment
CONACYT	Consejo Nacional de Ciencia y Tecnología
CSF	Cerebrospinal fluid
FRQNT	Fonds de recherche du Québec - Nature et technologies
HC	Healthy control
ICU	Information content unit
IDF	Inverse document frequency
IPA	International phonetic alphabet
KMO	Kaiser-Meyer-Olkin
MCI	Mild cognitive impairment
MFCC	Mel-frequency cepstral coefficients
MMSE	Mini-mental state examination
MRI	Magnetic resonance imaging
NLP	Natural language processing
PCA	Principal component analysis
POS	Parts of speech
RF	Random Forests
SVM	Support vector machine
TF	Term frequency
TFIDF	Term frequency × inverse document frequency

INTRODUCTION

Alzheimer's Disease is the most prevalent form of dementia, making for about 65% of cases (Alzheimer Society of Canada, 2014). In 2015, almost 50 million people worldwide were living with dementia, and it is expected that this number doubles every 20 years. If the current trend continues, by 2050, there will be over 130 million cases worldwide, with almost 70% of patients living in low and middle income countries (Alzheimer's Disease International et al., 2015).

Despite its high incidence, AD is often diagnosed at the dementia stage of its continuum, which appears several years after onset (Alzheimer's Association, 2018a). The most anticipated research advancement on the disease is the finding of treatments to stop or delay its progression. Nevertheless, when available, these treatments will require that the disease be detected at its earliest (Knopman, Boeve, & Petersen, 2003). Therefore, considerable effort is being invested in the identification of early AD biomarkers.

Some biomarkers have shown promising results in helping in the early diagnosis of Alzheimer's disease (Alzheimer's Association Research Center, 2016). However, most of these biomarkers, such as the extraction of cerebrospinal fluid (CSF) and magnetic resonance imaging (MRI), are expensive and particularly invasive for the growing elderly population. For this reason, these biomarkers should be regarded as tools to support and confirm the diagnosis, but not as regular monitoring mechanisms. Non-invasive and inexpensive tools that can be regularly used in clinical practice to alert of early signs of Alzheimer's disease must therefore be devised.

0.1 Problem statement and motivation

A wide variety of studies have found that language alterations may manifest many years, even decades (Snowdon et al., 1996), before the dementia stage of AD. These alterations are inconspicuous, and usually go unnoticed by humans, but numerical analyses have shown that

these subtle differences are statistically significant and could be used as an early alert mechanism for physicians.

Multiple studies on computer-based language alterations have been done in the context of standardized cognitive tests. These tests have the advantage of being widely known and used in clinical practice, and their normalization facilitates comparisons between studies and multiple languages. However, these tests also present disadvantages, since they can produce nervousness and stress in the patients being tested; as well, some produce a "practice effect" in the patient when performed frequently (Smith & Bondi, 2013). Furthermore, since they present a limited variety of syntactic structures, some linguistic phenomena are not observable, which represents a limitation in terms of evaluating pragmatic processes and discourse (Boschi et al., 2017).

Many of the alterations that have been detected in AD patients were first identified by performing deep manual quantitative linguistic analyses on spontaneous speech and writing. With the aid of Natural Language Processing tools, some of these analyses are apt to be automated. In recent years, a few studies have focused their attention on the automatic analysis of connected speech in spontaneous conversations in the elderly, especially among English speakers (Boschi et al., 2017).

Automatic analyses of language alterations in spontaneous conversations of Alzheimer's patients are usually carried out by comparing average tendencies between healthy and cognitive impaired individuals. However, multiple factors, such as sociodemographic and personality traits, may affect an individuals' linguistic performance. Therefore, it is desirable to have longitudinal personalized analyses that reflect the specific changes that each patient undergoes through time. However, due to the scarcity of datasets of this nature, there have been no computer-based longitudinal studies that show the evolution of patients' language from healthy to dementia stages in a personalized manner.

In this work, we propose techniques for monitoring the language functions of elderly patients in two settings: cognitive testing and spontaneous conversations. Both techniques could be combined by including the first type of analysis during the usual cognitive testing at the hospital, while analyses of spontaneous conversations undertaken with nurses, family members or caregivers could be held on a more frequent basis to complement evaluations.

0.2 Research objectives and contributions

The objective of this research is to develop methods for analyzing transcriptions of elderly speakers to **1**) evaluate their performance and condition of language functions from elicited speech in cognitive testing settings (Table 0.1; picture description tasks), and **2**) characterize the alterations in verbal communication that occur in individuals during their transition from healthy to dementia stages using spontaneous speech. The second objective is divided into sections that **2a**) contrast the significance of variables used in cognitive testing settings to its use in more unconstrained discourses (Table 0.1; object description), and **2b**) perform longitudinal analyses of free spontaneous conversations with individuals that transitioned from an apparently cognitive intact stage into cognitive impairment or Alzheimer's disease. In an effort to diversify the target population commonly used in these studies, and to test their robustness, our proposed methods were evaluated with native speakers of different languages.

Table 0.1 Discourse characteristics of the tasks employed for the assessment of connected speech in AD in this dissertation.

	Picture description tasks	**Object description**	**Spontaneous conversations**
Type of discourse	semi-spontaneous	semi-spontaneous	spontaneous
Thematic	restricted	semi-restricted	free
Syntactic structures	limited	partially limited	diverse
Stimulus	visual image	mental image	conversational flow
Lexicon	limited	partially limited	diverse

Three main contributions were made in a bid to achieve these goals:

1) **Automatic evaluation of picture description tasks**: Some of the least stressful tests for patients to perform are standardized picture description tasks. In these tests, patients are shown an image, and they are asked to describe the scene with as much detail as possible. We proposed a computer-based methodology to evaluate some language functions of the patients, as well as to evaluate their performance during the task by adapting an information coverage measure. Our evaluation method significantly correlated with the results of the Mini-Mental State Examination and the severity of cognitive impairment. Our method can be adapted to different populations, languages and pictures. Our first contribution resulted in the publication of the following paper:

    - **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra-Martínez, Andrés Roche-Bergua. "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task". *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10, pp. 260–268, 2018.

2) **Classification of healthy and cognitively impaired individuals from restricted and semi-restricted discourses**: We proposed two new metrics to evaluate the coverage of information and pertinence of the discourse based on the use of generic and specific vocabulary in healthy and cognitively impaired individuals. Using these metrics, we evaluated transcriptions of standardized picture descriptions by English speakers. Furthermore, and moving towards our second objective of evaluating language functions in free spontaneous conversations, we evaluated transcriptions of older Spanish speakers describing common objects. In this setting, although patients were all describing the same objects, their descriptions were grounded in their personal experiences and conception of these objects. The type of discourse in this task is of a semi-spontaneous nature, with a more limited lexicon than that of free spontaneous conversations.

We also evaluated other linguistic features, such as lexical richness and the use of specific linguistic patterns that could provide an insight into the types of syntactic structures that are most affected by cognitive impairment. Our experiments were carried out with native speakers of Spanish and English to test the multilingual robustness of our proposed metrics.

We found that our metrics of information coverage and pertinence based on the use of specific vocabulary were the most correlated with the severity of cognitive impairment for both tasks. When using these features, along with lexical richness measures and specific linguistic patterns to train a support vector machine learner, our results compared favorably against those of the state-of-the-art methods that rely on manual annotations or manual extraction of information content units. The results derived from this contribution were presented in two papers:

- **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra-Martínez. "Automated differentiation of Alzheimer's and MCI patients from healthy controls using English and Spanish transcriptions of description tasks". In *Computers in Biology and Medicine*. Under review.
- **Laura Hernández-Domínguez**, Edgar García-Cano, Sylvie Ratté, Gerardo Sierra-Martínez. "Detection of Alzheimer's disease based on automatic analysis of common objects descriptions". In the *Association for Computational Linguistics' 7th Workshop on Cognitive Aspects of Computational Language Learning*, Berlin, Germany, pp. 10–15. 2016.

3) **Longitudinal characterization of language alterations in spontaneous speech**: Spontaneous conversations have the advantage of feeling more natural and being less stressful to patients than description tasks (see Table 0.1). Also, the conversations can be carried out by a caregiver, a family member or a trusted member of the patient's community. To attain the second objective, a dataset with longitudinal transcriptions of spontaneous conversations with elderly French speakers was analyzed. All

participants of the recollections started as apparently healthy individuals, and some went on to develop different forms of cognitive impairment. We extracted several linguistic features, and based on our second contribution, proposed an adaptation of the vocabulary distribution measures to characterize the changes that occur in patients through time. Each patient that developed some form of cognitive impairment was contrasted against an individual that remained apparently healthy throughout the entire recollections. There were significant differences in the estimated measures based on vocabulary distribution, as well as on lexical richness metrics, linguistic patterns, and fluency descriptors between the groups with a healthy and cognitively impaired aging. Some of the differences were apparent up to ten years prior to the diagnosis of a cognitive disease. The results of this contribution were presented in the following paper:

- **Laura Hernández-Domínguez**, Sylvie Ratté, Annette Gerstenberg, Gerardo Sierra-Martínez. "Aging with and without cognitive diseases: characterizing 10 years of language differences in French elderly speakers". In *Computer Speech and Language*. Under review.

One of the biggest challenges when performing longitudinal studies on communication strategies in spontaneous conversations of elderly speakers is a lack of data resources. To the best of our knowledge, only three datasets of this type are available for research: the *Carolinas Conversation Collection* (Pope & Davis, 2011) with English speakers, and the CorpAGEst (Bolly & Boutet, 2018) and the *LangAge* (Gerstenberg, 2011) corpora, with French speakers. In the face of this scarcity, one of the objectives of the *Cécilia* project is to increase the resources available for such studies. A concomitant contribution of my doctoral studies was the creation of the Mexican and Ecuadorian cohorts of the *Carolina Conversation Collection* with Spanish speakers. This included the elaboration of the interview, recollection, transcription and annotation protocol, as well as the coordination and realization of in-site video-recorded interviews with elderly participants of both countries. This ongoing contribution has been presented in two conferences and in one chapter in a published book:

- **Laura Hernández-Domínguez**, Sylvie Ratté, Charlene Pope, Boyd Davis. “Conversing with the elderly in Latin America: a new cohort for multimodal, multilingual longitudinal studies on aging.”. In the *Association for Computational Linguistics’ 7th Workshop on Cognitive Aspects of Computational Language Learning*, Berlin, Germany, pp. 16–21. 2016.
- Sylvie Ratté, **Laura Hernández-Domínguez**, Andrés Roche-Bergua, Gerardo Sierra-Martínez, Boyd Davis. " Cécilia Project: an international multidisciplinary collaboration on the study of language in later life". In the *3rd International Conference CLARE, Encounters in Language and Aging Research*, Berlin, Germany. 2017.
- **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra-Martínez, Andrés Roche-Bergua, Janet Jiménez-Genchi. “El Proyecto Cécilia: estudios del lenguaje en pacientes con demencia tipo Alzheimer”. In *Psicogeriatría: Temas selectos*. Part 2: Geriatric Psychiatry, pp. 353–363. 2017.

0.3 Outline

The organization of this thesis is as follows. In **Chapter 1**, we present an overview of Alzheimer’s disease and present a review of relevant works on the analysis of alterations of verbal communication due to AD. **Chapter 2** introduces our proposed methodology for an automatic evaluation of a picture description task and of language functions, published in the *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* journal. **Chapter 3** presents a comparison between the use of vocabulary among Alzheimer’s patients and healthy elderly individuals in restricted and semi-restricted discourse settings. This chapter was submitted to the journal of *Computers in Biology and Medicine.* **Chapter 4** presents a characterization of verbal communication changes in apparently healthy individuals while transitioning to a cognitive impairment diagnosis over a ten-year time span. This chapter was submitted to the journal of *Computer Speech and Language*. **Chapter 5** presents our efforts to overcome the scarcity of longitudinal data for studies in communication changes in the elderly

population. This chapter presents the creation of and advances made in the Latin-American cohort of the Carolinas Conversations Collection, as presented in the *ACL's 7th Workshop on Cognitive Aspects of Computational Language Learning*. **Chapter 6** summarizes the main contributions of this dissertation and discusses its limitations. **Chapter 7** presents the final conclusions and the possibilities for future works. Finally, **Appendix II** provides a complete list of the publications related to this study.

CHAPTER 1

LITERATURE REVIEW

The objective of this chapter is to familiarize the reader with the fundamentals of Alzheimer's disease, and to portray the state of the art in the computer-based methods for studying language alterations caused by this disease. The chapter starts with an overview of the general aspects of AD, including its incidence, symptoms, methods for diagnosis and a description of its continuum. Then, a critical review of the literature on the best-known computer-based methods that exist on studies of language changes in AD during picture description tasks and in natural language conversations is given. Finally, this chapter concludes with a summary of the advantages, limitations and research possibilities derived from these state-of-the art approaches.

1.1 Alzheimer's disease overview

With the reduction of the birth rate and the rapid growth of the elderly community, we are fast turning into an aged population. As live expectancy increases, age-related disorders increase as well, bringing with them great economic challenges for governments and society in general. Important efforts are being made all around the world in the search for methods of detecting and treating these disorders in effective and inexpensive ways.

From the wide range of age-related disorders, dementing illnesses are highly common, being Alzheimer's disease the most prevalent condition with 65% of all cases (Alzheimer Society of Canada, 2014). The prevalence of AD increases with age; 10% of people age 65 and older are affected by the disease, and this number increases to almost a third of the population over 85 years old (Alzheimer's Association, 2018a). Over 80% of AD cases are from people aged 75 and older.

Alzheimer's disease is a degenerative brain disease whose main causes remain unknown. Although there is a genetic component associated to AD, it is estimated that only 1% of the

cases are caused by a genetic mutation (Bekris, Yu, Bird, & Tsuang, 2010). It is believed that AD is caused by a combination of multiple factors, rather than by a unique cause. The main risk factors of AD are older age, having a first-degree relative diagnosed with AD, and having inherited the APOE-e4 risk gene. However, several modifiable risk factors also play a crucial role in the development of the disease (Baumgart et al., 2015), such as smoking, obesity, hypertension, and high cholesterol levels.

The American Alzheimer's Association (2018, 2011) identifies early detection of Alzheimer's disease as one of the biggest challenges faced by physicians when dealing with dementing illnesses. It is highly common to diagnose a person suffering from Alzheimer's dementia several years after the disease has initiate, which means that by the time a proper diagnosis has been made, the condition has already made severe damage to the patient.

The revised AD diagnostic guidelines (Sperling et al., 2011) establish the progression of signs and symptoms that occur along the disease continuum. This continuum starts with brain changes that may begin 20 years prior to the appearance of any symptoms (Villemagne et al., 2013), yielding a potential for early diagnosis (Alzheimer's Association, 2018a).

The disease progression starts with the slow accumulation of the protein fragment beta-amyloid outside neurons, and the accumulation of a mutation of the protein tau inside neurons. Beta-amyloid interferes with neurons synapses, while tau tangles impede nutrients to reach inside neurons (Alzheimer's Association, 2018a).

At the first stage, the brain is able to compensate for these changes, allowing the individual to function normally for several years. In individuals with high education or that regularly perform cognitively demanding and mentally stimulating activities, this compensation mechanism, also known as the *cognitive reserve*, allows them to function with normality for a longer time (Almeida et al., 2015). However, through time, the amount of damage reaches a point in which a slow decline in cognitive functions is evident in the patient.

The first stage of symptomatic AD is known as Mild Cognitive Impairment (MCI). At this stage, patients often refer to symptoms like loss of memory and concentration. However, their symptoms are not severe enough to interfere in the patient's daily life activities. While MCI is a part of the normal progression of AD, it is important to note that a diagnosis of MCI impairment is not necessary an AD sentence, since not all MCI cases are caused by AD (Smith & Bondi, 2013). Some MCI symptoms may be derived from depression, obstructive sleep apnea, vitamin B_{12} deficiency or even from certain medications. Some MCI individuals will remain in that state indefinitely, while others might even revert to a normal cognition. An estimated 32% of individuals with an MCI diagnosis will progress to AD in the next 5 years (Ward, Tardiff, Dye, & Arrighi, 2013). The timely identification of these patients remains a major goal on AD research (Alzheimer's Association, 2018a).

After MCI, the next stage is *Alzheimer's dementia*, where the brain is no longer able to compensate for the changes. At the moderate stage of the dementia, there are noticeable memory, thinking and behavioral symptoms that are severe enough to interfere with a person's daily life activities. Patients experiment confusion with time and place. They also may present mood changes that are caused by confusion, suspicion, depression and anxiety. Their working memory functions are severely affected, impairing their ability to retain recent information. They have difficulties following plans or instructions, and with concentration. Language functions in patients with Alzheimer's dementia are noticeably altered, making them struggle with vocabulary and joining conversations. This stage is usually the longest (Alzheimer's Association, 2018a).

At the final stage, patients require continuous help with basic activities of daily living, and their ability to communicate is limited. In this phase, the damage in the brain is so extensive that it interferes with the areas in charge of movement. Patients with advanced Alzheimer's dementia eventually become bed-bound and lose control of their ability to swallow, making them vulnerable to aspiration pneumonia, which is the leading cause of death among individuals with AD (Burns, Jacoby, Luthert, & Levy, 1990).

Detecting Alzheimer's diseases at the MCI stage may present multiple benefits for patients and the general society (Alzheimer's Association, 2018a). Control of blood pressure, mental activity stimulation, aerobic exercise, smoking cessation and stroke prevention appear to reduce the risk of progression from MCI to dementia (Langa & Levine, 2014). An early diagnosis also gives time to individuals for planning for the future and to make important choices before their cognitive abilities fade, such as legal directives, including end-of-life care and planning. An opportune diagnosis also provides patients with the opportunity of joining clinical trials available (Dubois et al., 2016), which may allow a better and inexpensive monitoring of the patient, as well as access to new therapies. Finally, patients that have a timely diagnosis and start arrestive treatments, remain longer in their communities, which reduces stress and costs to patients and caregivers (Dubois et al., 2016).

According to (Alzheimer's Association, 2018a), the methods and studies that are being most researched for early diagnosis of Alzheimer's are biomarkers, brain imaging/neuroimaging (magnetic resonance imaging and computed tomography), cerebrospinal fluid proteins, and genetic risk profiling. However, most of these studies are expensive, invasive or may expose the patient to unnecessary pain or risk. It is therefore important that these tests are used as a mechanism of confirmation of diagnoses, rather as in regular medical practice.

The most usually evaluated cognitive abilities when diagnosing dementia are episodic memory, executive function, perceptual speed, verbal ability, visuospatial skill, attention and language (Taler & Phillips, 2008). From the wide variety of cognitive abilities altered in Alzheimer's patients, language alterations are of great interest to researchers since many studies have found them even at the beginning of the disease (Schröder, Wendelstein, & Felder, 2010). There have been numerous recent studies that have focused their attention on the analysis of language production of Alzheimer's patients as a non-invasive and economical means to detect this disease.

1.2 Detection of language alterations in Alzheimer's disease

Several studies have analyzed the changes and variation of language in people suffering from Alzheimer's disease. Most studies have been based on standard tests, such as asking the patient to remember words from a list previously given, or asking the patient to retrieve certain types of words, like names of animals, fruits, vegetables, home duties, words starting with the same letter, etc. (Sabat, 1994; Taler & Phillips, 2008). However, some authors (Bucks, Singh, Cuerden, & Wilcock, 2000; Sabat, 1994) coincide in the view that these tests not necessarily report the real participant performance in normal language interactions, and that these tests have proven insensitive to early communication deficits that are observed in natural conversations. For these reasons, they argue that it is advisable to create methods for analyzing language alterations in spontaneous speech.

Picture description tasks are a type of standardized test often given to elderly patients in clinical practice, where patients are shown an image and are requested to describe it with as much detail as possible. AD patients have been found to provide less informative descriptions than healthy individuals during picture description tasks even at early stages of the disease (Ahmed, de Jager, Haigh, & Garrard, 2013). Since this type of task elicit a semi-spontaneous speech (Prins & Bastiaanse, 2004) in the patient within a restricted context, it is a good instrument for studying early detection of AD signs through language analysis. One of the best-known tests of this type is the Cookie Theft picture (Figure 1.1) description task from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983).

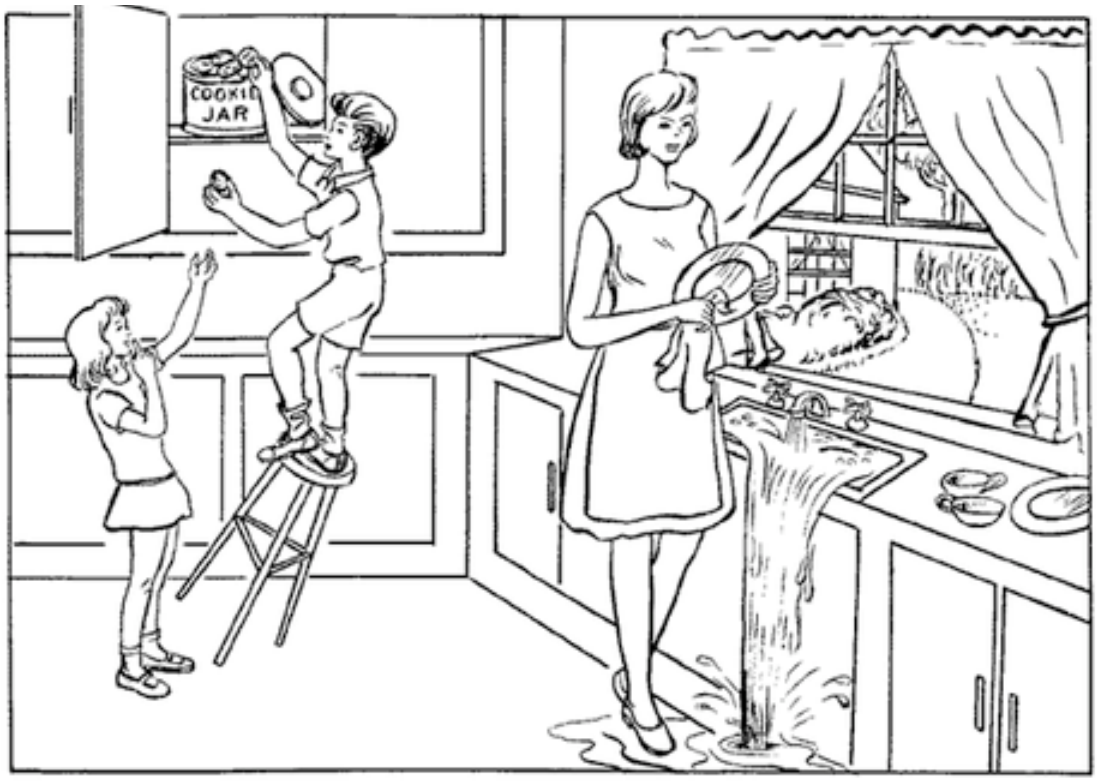

Figure 1.2 The Cookie Theft picture from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983)

Despite being relatively easy tasks for the participants, some authors argue that cognitive testing in a clinical setting tend to produce anxiety in patients, and also some produce a "practice effect" in the patient when performed frequently (Smith & Bondi, 2013). For this work, we propose that a good monitoring mechanism could involve the combination of the application of sporadic picture description task with the analysis of more regular spontaneous conversations.

The following critical literature review focuses on the best-known computer-based approaches for detection of Alzheimer's disease in 1) picture description tasks and 2) spontaneous conversations. The characteristics of the studies, methods, results and best performing features are highlighted in bold letters, while the major points for improvement and observations are underlined.

1.2.1 State of the art on the automatic analysis of picture description tasks

The following studies are the most well-known and cited works on computer-based approaches of language analysis of AD patients during picture description tasks. A more complete and deeper systematic review of studies in connected speech in these tasks has been recently presented by Slegers *et al*. (Slegers, Filiou, Montembeault, & Brambati, 2018). For this review, we did not include commonly cited works that have a main focus on acoustic features and speech processing, since this type of study is out of the scope of this dissertation.

1.2.1.1 Kavé *et al.*

Kavé *et al*. have published several articles (Kavé & Goral, 2016; Kavé, Goral, & Dassa, 2018; Kavé & Levy, 2003) on the analysis of picture description tasks in **elderly Hebrew speakers**. In their first paper (Kavé & Levy, 2003), they evaluated transcriptions of picture description tasks from **14 patients diagnosed with AD** and **48 healthy controls**. AD patients presented memory problems and evidence of gradual deterioration of cognitive functions, but their mini-mental state examination score was above 17/30, which indicated that all **AD participants were at the mild stage of Alzheimer's dementia**.

As part of their first evaluation, the authors analyzed the *conceptual semantic* aspect of descriptions of the **Cookie Theft picture**. They measured 1) the information content of the description by analyzing the number of *information content units* (ICUs) that participants mentioned. For this, they had a pre-defined list of 25 ICUs as proposed by (Croisile et al., 1996). 2) The number of circumlocutionary comments, such as "*you can see*" and "W*hat can I say?*". 3) The pronoun rate.

The authors also described the *syntax*, by estimating the number of words per clause, the proportion of each type of clauses (independent, dependent or complete) and the proportion of each type of sentences (declarative, head-complement, existential, relative, conjoined and impersonal), and the proportion of nouns with respect to verbs+nouns. The *morphology* of the

language was studied by estimating the proportion of inflected words out of the size of the text, the proportion of lemmas of verbs (verb roots) out of the total forms of verbs mentioned.

As part of their results, the authors found a **significant difference in the total number of information units** provided by controls when compared to AD patients. **AD patients provided significantly less ICUs related to objects and actions**. AD patients also produced **more circumlocutionary comments** and a **higher proportion of pronouns relative to nouns+pronouns**. Also, AD patients **produced more words per clause**. There was no significant difference in the rest of the characteristics with AD.

This first work of Kavé *et al*. showed linguistic differences between AD patients and HC, although some characteristics, like the use of circumlocutionary comments and the type of clause and sentence were manually extracted and its extraction by automatic means is not trivial. Also, it is common that authors use lists of pre-defined ICUs to evaluate the completeness of the information provided by patients. However, these lists are not standardized, and may not reflect cultural and socio-economic differences that could be observed and mentioned by patients from different backgrounds in the pictures. For instance, a Latin-American population could notice that all subjects in the Cookie Theft picture (Figure 1.2) are blond, or that the dimensions of the kitchen are bigger than an average home in, say, Mexico City. Another limitation of this work was the reduced size of its sample.

In their 2016 work (Kavé & Goral, 2016), the authors continued studying older **Hebrew speakers** during descriptions of the **Cookie Theft picture**. Their sample was comprised of **20 AD patients** and **20 matched healthy controls**. For this experiment, they estimated the total number of words, ratio of content words, noun ratio, pronoun ratio, Type-token ratio (TTR), and an adaptation of TTR just for nouns, mean frequency of words and nouns, and mean word length. **All linguistic features were extracted automatically** using a tool for scoring essays in Hebrew using lexical and grammatical measures. The **selection of "content" nouns was made manually**. An interesting aspect of this work is that the **authors compared the use of nouns with that of younger participants** performing the same picture description task.

The authors found that **healthy older individuals used nouns that are less frequent**, while their younger counterparts used nouns that were more common. This phenomenon was explained as the effect of **increased and richer vocabulary in a healthy older population** due to their continual language development. However, the authors noted that individuals with AD tended to revert to the use of more common nouns since their semantic network has diminished. This findings correspond to previous literature on semantic deficits in AD patients (Adlam, Patterson, Bozeat, & Hodges, 2010).

For their most recent work (Kavé & Goral, 2018), Kavé *et al*. studied manual transcriptions of descriptions of the **Cookie Theft picture** from **35 AD** patients, and **35 healthy controls**. All participants were **Hebrew speakers**. Similar to their work in 2016, the authors used an **automatic tool created for scoring essays in Hebrew to extract the linguistic features**. The authors studied the same linguistic characteristics as in their 2016 work but incorporated the analysis of information content that was used in their 2003 study with the 25 ICUs proposed by Croisile (Croisile et al., 1996).

As part of their findings, the authors detected a significant **increased use of number of words**, **pronouns** and **mean word frequency by AD patients**, **and** a **decreased** use of **prepositions** and **content words**, and **lower TTR values**. The authors also found that **AD patients tend to inflect verbs in the most common pattern** in Hebrew (PAAL).

As part of their pre-processing, the authors removed the use of incomplete words and interjections. These features were not included in their analysis, despite previous literature (R. Alegria, Gallo, et al., 2013; Habash, 2012) that have found a correlation between these characteristics and AD. Two interesting findings came from the latest studies of this research group, the frequent employment of more commonly used words, and the inflection of verbs in the most common linguistic patterns of its language by AD patients.

1.2.1.2 Orimaye *et al.*

Orimaye *et al*. have presented a series of studies (Sylvester O. Orimaye, Wong, Golden, Wong, & Soyiri, 2017; Sylvester Olubolu Orimaye, Tai, Sze-Meng Wong, & Piau Wong, 2015; Sylvester Olubolu Orimaye, Wong, & Golden, 2014) on **English speakers** performing picture description tasks. For all their works, the authors have used the **Pitt corpus** of the (Becker, Boiler, Lopez, Saxton, & McGonigle, 1994) **DementiaBank dataset**. This corpus contains audio recordings and transcriptions of participants describing the **Cookie Theft picture**. In the dataset, there are samples form healthy controls, AD and MCI patients, and also patients suffering from vascular dementia and other dementias, and memory disorders.

In their first study (Sylvester Olubolu Orimaye et al., 2014), the authors divided the participants into healthy control and dementia patients, which included the participants with dementia **(AD, vascular and unidentified),** the **MCI patients and** the group with **memory complaints**. To have a balanced sample, the authors only used a part of the sample of patients with dementia, leaving **242 healthy controls** and **242 participants with any form of dementia**. To extract their linguistic features, they used the **Stanford parser** (Klein & Manning, 2003).

The linguistic features evaluated were number of coordinated sentences (conjunction tags), subordinate sentences (preposition tags), reduced sentences (verb in gerund or present participle, and verb in past participle tags), number of predicates, dependency distance, number of dependencies, production rules, revisions, repetitions, total number of utterances, use of function words, hapax legomena, total word count, total character count, number of sentences, repetitions, bigrams and morphemes.

The authors found a significant **lower production of sentences and predicates**, as well as **shorter utterances**, and a **higher number of utterances**, **repetitions** and **revisions in the dementia group**. These features were used to train five learners: an **SVM** with radial basis kernel, **Naïve Bayes**, **J48**, a **Neural Network** with back propagation and **Bayesian networks**. With a **10-fold cross validation** experiment, the authors reported a **75% precision** and **69%**

recall (**F-score = 0.72**) **for detecting** patients in the **dementia group** by **using neural networks**, and a **74% precision** and **73% recall** (**F-score = 0.73**) using SVM**.**

This work corroborated several findings with linguistic features that have been tested in spontaneous speech. A downside of this investigation is that the dementia group was comprised of participants with different diagnosis and levels of dementia, not allowing to examine which features are more associated with which types of disease. This is common with small datasets. However, the DementiaBank presents an ample sample of AD and MCI patients that could be studied separately. The authors addressed this issue in their next works.

For their second work (Sylvester Olubolu Orimaye et al., 2015), the authors present a novel approach for distinguishing **19 MCI patients** from **19 healthy controls** from the Pitt corpus. In their work, they used a measured based on skip-grams, which is a technique in which the frequency of appearance of sequences of several consecutive letters is estimated, but also, a number of words are skipped in this sequence. As an example, the 1-skip-3-grams of the sentence "*I am in love*" are: "*I_in_love*" and "*I_am_love*".

The authors performed several experiments using a different number of the top skip-grams as features. For their experiments, they trained four algorithms: **SVM**, **Naïve Bayes**, **Decision Trees** and **logistic regression**. With a **10-fold cross-validation** and using the top **200 skip-n-grams as features**, the authors reported a **98% precision** and **97% recall** (**F-score = 0.97**, **AUC = 0.99**) **for detecting MCI patients** using SVM, Naïve Bayes, and Logistic regression**.**

Although the proposal of the skip-n-grams is new and interesting, there was no statistical evaluation of the correlation of individual skip-n-grams with MCI or with the mini-mental state examination (MMSE) scores of patients. A further point is that the authors observe a very high, almost perfect accuracy in detecting MCI using these 200 features alone. However, the authors do not mention the process followed to tune the hyperparameters, and it is not clear whether they used a separate validation set, or whether their experiment is overfitting the data.

In their latest paper (Sylvester O. Orimaye et al., 2017), Orimaye *et al*. focus their attention in the binary classification of **AD patients** and **healthy controls**. For their study, they again used the **Pitt corpus**. As their sample, they selected **99 AD patients** and **99 healthy controls**. For this study, the authors extracted all the linguistic features that they tested in their first paper (Sylvester Olubolu Orimaye et al., 2014), plus number of incomplete words and fillers (interjections), part-of-speech entropy, content density and pause rate. Finally, they extracted the frequency of bigrams and trigrams mentioned by participants.

The authors found that **AD patients** produced **less reduced sentences** and **number of predicates**, as well as **shorter utterances**. These patients also had a **higher number of repetitions**, **revisions**, **word replacements**, **incomplete words**, **filler words** and **trailing offs**. Using a combination of the linguistic characteristics that were significantly correlated with AD and the bigrams and trigrams ranked highest according their information gain, the authors formed a set of **1,000 features**. With these features, they trained an SVM implementation and reported an **AUC = 0.93**.

There are some questions about the appropriateness of using 1,000 features in a dataset of only 200 samples with SVM. Also, similar to their previous paper, the authors did not enter into detail about the cross-validation portion of the set used for calibrating the parameters of the SVM. Another limitation is that the authors did not include all samples of AD patients and healthy controls and chose to only work with a subset of both. Finally, there was no statistical evaluation of the feature against the MMSE scores of participants, which could have provided with a sense of comparison to formal medical assessment.

1.2.1.3 Rudzicz *et al.*

Prof. Frank Rudzicz is the director of the Signal Processing and Oral Communication Laboratory (SPOClab) of the Department of Computer Science at the University of Toronto and the Toronto Rehabilitation Institute. Part of his research focuses on the study of speech-language pathologies and rehabilitation engineering. This laboratory is responsible for two of

the best-known works on automatic detection of AD based on picture description tasks (Fraser, Meltzer, & Rudzicz, 2016; Yancheva & Rudzicz, 2016).

In their first work (Fraser et al., 2016), the authors took a sample of transcriptions of descriptions of the **Cookie Theft** picture from native **English speakers**. The descriptions were provided by AD patients and healthy controls from the **Pitt Corpus**. Their sample was comprised of **233 transcriptions of AD** participants, and **240 of healthy controls**. Despite using the same corpus as Orimaye *et al*., Fraser *et al*. used a smaller sample without providing a clear explanation as to why some participants were excluded from their study.

In total, Fraser *et al*. extracted **370 features** that included **part-of-speech ratios**, **syntactic complexity**, **grammatical constituents**, **lexical richness**, **information content**, **repetitiveness** and **acoustic features**. To determine the amount of information content provided by a participant, the authors extracted the most *relevant items* from a **manually-made list of information content units** (Croisile et al., 1996). These features were binary and indicated the **presence or absence of these specific items**. As noted by the authors, one disadvantage of this metric is that it does not allow to detect whether the subject is mentioning the item in the appropriate context.

Using **logistic regression**, the authors performed a **classification of AD patients from healthy controls** with a 10-fold cross-validation procedure. Through a Pearson's correlation analysis, the authors ranked their features according to their significance and performed a feature selection process in which they tested the n top ranked features. Their **best average accuracy was of 81.92%** using the **35 top-ranked features**.

As part of their analysis, the authors found that an increase in pronouns and the use of high-frequency words were suggestive of a vaguer discourse. Also, they observed that a decrease in the number of prepositional phrases indicated less-detailed descriptions. Regarding the information content, Fraser *et al*. noted a reduction in the number of key words by the AD cohort, which pointed to more uninformative descriptions.

A different work (Yancheva & Rudzicz, 2016) by Rudzicz's team was based on vector-space topic models to detect signs of Alzheimer's disease. For this work, the authors studied a sample of **255 transcriptions** of descriptions of the **Cookie Theft** picture **by AD patients** and **241 transcriptions of healthy controls** from the **Pitt corpus**.

In this study, Yancheva *et al*. used the **same linguistic features** used in their previous work (Fraser et al., 2016). However, the most innovative aspect of this study is that, instead of using a pre-defined list of information content units made by a specialist, the authors proposed a model to automatically crate this list by using vector-space topic models. The extracted information content units **only considered verbs and nouns**. Using distances from the topic-modeling clusters, the authors defined a measure of **idea density** and **idea efficiency**.

Using a Random Forest classifier trained with a distance-based, the authors performed a binary classification of AD patients and healthy controls with a combination of the extracted features. The authors reported an **accuracy and F-score of 80%**. One of the findings of this work was the discovery of a new information content unit (*apron*) that had not been part of any of the pre-defined standard lists. This finding suggest that an automatic extraction of information content units could provide population-specific referents for the evaluation of picture description tasks. Before this, all lists were dependent on the subjective perception of the creators of the list, and their consideration of what is "important" to mention during the task.

1.2.2 State of the art on the automatic analysis of spontaneous conversations

1.2.2.1 Bucks *et al.*

In 2000, Bucks *et al.* (Bucks et al., 2000) presented a research work in which they analyzed eight linguistic measures to assess their importance in the automatic discrimination between healthy and demented individuals, particularly individuals with mild to moderate dementia of Alzheimer's type. These measures were taken on spontaneous conversational speeches of **8**

individuals with a diagnosis of **Alzheimer's disease, and 16 healthy controls**. The participants were equally distributed between females and males, and had similar age ranges, however, there was a significant higher educational level in the controls.

For their study, Bucks *et al*. collected **24 conversations in English** with a production of approximately **1000 words each**. The conversations were obtained with little intervention from interviewers, who only encouraged the participants to talk about their lives and experiences, providing as little stimuli and interruptions as possible. The conversations were transcribed, only on the side of the participant. The authors ignored multiple attempts of producing the same word or phrase, and stereotypical phrases as "you know", "oh, boy", etc.

The authors studied eight linguistic measures: **pronouns**, **nouns**, **adjectives**, and **verbs rates** per 100 words (part of speech); **Type token ratio**, Brunet's Index (Brunet, 1978), and Honore's Statistics (Honoré, 1979) (vocabulary richness); and semantic cohesion measured in the rate **of *clause-like semantic units* (CSU)** per 100 words (CSU is consider a noun and verb phrase that a speaker can produce; these units were automatically found by using linguistic rules (Singh, 1996)).

For their analysis, the authors made analyses of covariance to control for differences obtained due to the difference of years in education, finding that only the pronoun rate had a correlation of over 5% with the number of years in education. On the one hand, as part of their results, the authors found that factors such as age, years of education, mental evaluation scores, and duration of illness did not correlate significantly with any of the linguist measures. On the other hand, the authors reported **significant differences between control and experimental sets in all linguistic measures, except for the CSU rate**.

Bucks *et al*. performed a **Principal Component Analysis** (PCA) with the eight linguistic features and found two principal components: lexical richness (PC1 contrasted noun, adjective, pronoun and verb rates, as well as type token ratio, Honore's Statistics and Brunet's Index), and phrase making factor (PC2 had a positive load on CSU rate).

Finally, the authors performed **linear discriminant analysis** to establish the performance of each measure in binary classification of participants. The authors found that the **most important measures** were **noun rate**, **pronoun rate** and **Brunet's index**. Also, the authors report a **classification between healthy subjects and Alzheimer's sufferers with 87.5%** of accuracy using cross-validation evaluation.

This work was possibly the firsts that performed automatic classification of patients with Alzheimer's disease and healthy elderly individuals. Some of the metrics used in this paper, such as the CSU rate have now been replaced with more sophisticated NLP measures for cohesion, but the authors findings set a precedent on which features are more likely to be of use for this particular problem.

Apart from its small size, the major issue with this study are the high differences between the number of years of education that have been found in many recent experiments as a crucial factor to consider in cognitive decline. Another characteristics that the authors purposefully left out of the experiment were the incomplete words and discourse markers, such as "*you know*", which other authors have explored in more depth and found relevant (Yi-hsiu Lai & Lin, 2012). Finally, there was no longitudinal component to this study.

1.2.2.2 Alegria *et al.*

Renne Alegria *et al.* have conducted a **series of researches since 2008** (R. Alegria, Bolso, et al., 2013; R. Alegria, Gallo, et al., 2013; R. Alegria, Bottino, & Ines, 2011; Renne P. Alegria, Ferreira, Marques, Bottino, & Nogueira, 2010; Renne P. Alegria, Perroco, Marques, Barbosa, & Bottino, 2008; Renné P. Alegria, Perroco, Marques, Nogueira, & Bottino, 2009) regarding the effects of Alzheimer's disease in patients' **discourses in Brazilian Portuguese**. Their study subjects were part of PORTER (Old Age Research Program of the Institute of Psychiatry, the University of Sao Paulo Medical School). Their work progress was annually presented in the Alzheimer's Association International Conference on Alzheimer's disease.

The authors have used **on average discourses of eleven Alzheimer's patients and eight controls**, although every year they have increased the number of participants.

From 2008 to 2013, Alegria *et al.* focused their attention in the **vocabulary that is retained by Alzheimer's patients**, rather than in the type of words they have forgotten. The authors concluded that Alzheimer's patients are able to communicate effectively in the initial stages of the disease despite having troubles finding some words, because they are able to **remember the words that are related to familiar themes —**family, religion, profession, food, health and education**—**, thus concluding that this vocabulary should be used in order to improve communication with Alzheimer's patients.

According to Alegria *et al.*, Alzheimer's patients' use of thematic words is significantly higher in frequency than in healthy individuals. This might be a clue to follow in the detection of language alterations caused by Alzheimer's disease.

In 2013, Alegria *et al.* added a new approach in their research focusing on the use of certain grammatical categories in patients with Alzheimer's. For this research the authors analyzed discourses in Portuguese of **twenty-three patients** from the PORTER program **and twenty-three healthy controls**. Each participant had a **twenty minutes conversation**, which was transcribed for later analysis. In their work the authors found no difference in the proportion of adjectives and conjunctions used by Alzheimer's patients. However, they found a **significant difference in the use of interjections**, **adverbs**, **pronouns** and **prepositions between the controls and the patient**s. Furthermore, they found that **these differences increase as the disease progress**.

The last approach of this research might give certain clues regarding the nature of alterations in languages with grammatical constructions similar to Portuguese (e.g. Spanish, Italian, and French). Although the number of participants is still low, this was one of the works with the biggest number of patients. It also followed patients longitudinally. However, in the sample

there were no patients that transitioned from a healthy stage to cognitive impairment during the five years of reported advances.

1.2.2.3 Jarrold *et al.*

In their research of Jarrold et al (2010), present evidence that it is possible to asses some mental disorders through the application of **data-mining and text analytics**. In particular, the authors focused in detection of **pre-symptomatic Alzheimer's disease**, cognitive impairment and clinical depression.

As a basis for their work, the authors took the results from the "Nun Study" (Snowdon et al., 1996), an analysis of autobiographical writings of nuns in their twenties, which concluded that *idea density* was a strong predictor of Alzheimer's disease, even if the diseases presented itself 50 years later.

For this work, the authors used the Western Collaborative Group Study as their dataset. They took a sample of **22 interviews in English** with individuals that were **declared cognitively normal at the time of the interviews** (1988**), but their cause of dead was clinically verified as Alzheimer's disease**. Their **controls were semi-structured interviews made to 23 age-matched men** never diagnosed with dementia.

The authors extracted linguistic features from the dataset using three lexical analyzers. The first one, POST, extracted a part of speech frequency vector describing the rate of use of nouns, adjectives, verbs, etc. The second analyzer, LIWC, was used to count the frequency of words from a pre-defined list based on certain categories, like positive emotions, first person words, etc. The third analyzer, CPIDR, was used to extract the density of propositional ideas expressed by the speakers. From these measurements, the authors selected those features in which they found significant variations with the presence or absence of the disorder.

Jarrold *et al.* trained three different machine learning algorithms with the previously selected features: **logistic regression**, **J48** and **multilayered perceptron**. With these algorithms the authors were able to **predict which individuals were going to develop Alzheimer's with an accuracy of 73%** with their best performing learner (the authors didn't go into depth about the individual performance of each learner or into which specific features were preserved). For this evaluation the authors used 5-fold cross-validation repeating the evaluation over 100 times with different combinations of testing samples and presenting the mean accuracy.

From these results, the authors concluded that the **most valuable feature** for detecting Pre-Alzheimer's disease **was *idea density*** measured with CPIDR. They found that this feature was highly significant regardless educational level, age, age squared and cognitive impairment measures of the sample (these findings coincide with those of Bucks *et al.*). The authors determined that it is possible to extrapolate the results of the "Nun Study" to speech, rather than writing productions, to an elder population and to both genders.

Jarrold *et al.* show a very interesting work with a clear and precise methodology description. However, their experiments should be replicated into larger datasets and with both genders, in order to determine whether these findings can be extrapolated to participants with different characteristics. Also, it is interesting that their analysis was performed in patients at a time that they were considered cognitively healthy, and who later were diagnosed with Alzheimer's disease at their time of death. Nevertheless, it would be very interesting to have followed these patients to the mild and moderate stages of Alzheimer's, to observe the full range of changes produced by AD. However, this works provides major indications of features that have good potential for early detection of AD.

The authors also note that their sample labeling was based on clinical diagnosis made by the physicians who signed the death certificate of the patients. However, the goal standard method of Alzheimer's assessment is through the analysis of brain tissue during an autopsy, which leads to a possible error margin caused by possible misdiagnosis. This is a factor that apply for

most datasets available for Alzheimer's studies of cognitive decline and remains as one of the challenges of working with this disease.

Finally, as a future project, the authors proposed applying a multi-agent framework to face the problem by training one different learner for each of the questions in the interviews and combining them into one single meta-learner to test their performance. This is a highly interesting idea, however I consider that different leaners should be trained in less specific dataset partitions, such as partitions based on age frame, gender and education level. With this, instead of developing algorithms for classifying specific structured interviews, it would be possible to develop more general classifiers based on spontaneous conversations.

In their most recent study (Jarrold et al., 2014), the authors used **logistic regression**, **multilayer perceptron** and **decision trees** to differentiate 9 healthy controls from 48 patients with different types of dementia (9 AD patients among them). In this work they studied 10 minutes spontaneous speech samples from each participant by extracting **acoustic** and **lexical features**. These speech samples where composed of the answer to a semi-structured interview and a picture description task.

As part of their features, they extracted 14 parts-of-speech frequency counts, and the distribution of words into 81 categories such as emotional, cognitive, function words, verb tenses and negations.

As part of their findings, they detected that AD patients have an **increased use of pronouns**, **verbs and adjectives**. The authors reported an 88% accuracy in distinguishing between AD patients and controls, with 83% sensitivity and 90% specificity for AD when using layered perceptron learners. AD patients in this study were already at the dementia stage of the continuum, with a mean mini-mental state examination score of 18/30, which usually means that the patient already presents disorientation, mild impairment in household tasks and impaired problem solving. This study was not longitudinal, and there is a question on the

appropriateness of the use of layered perceptron in such a small sample (9 controls and 9 AD patients, with oven 90 features).

1.2.2.4 Habash and Guinn

Habash and Guinn in 2012 (Guinn & Habash, 2012a; Habash, 2012) presented a study of some linguistic metrics in order to detect which features could proof useful in the automatic detection of Alzheimer's disease in spontaneous conversations with **English** speakers. In their study they analyzed grammar and syntax (part-of-speech), lexical richness (Type token ratio, Brunet's Index (Brunet, 1978), and Honore's Statistics (Honoré, 1979)), filler words (rate of short phrase utterance), repetitions, incomplete words, syllables per minute, go-ahead utterances, and paraphrasing (direct, reflexive and indirect).

Eighty conversations from The Carolina's Conversations Collection (Pope & Davis, 2011) were used as part of their dataset. The conversations selected by the authors were **conversations with Alzheimer's patients**. The authors used the dialogs produced by the patients as their sample of Alzheimer's patients, and **the dialogs produced from the interviewers as part of their control group**. Also, the authors used the SWITCHBOARD corpus (Godfrey, Holliman, & McDaniel, 1992) as a control group.

The findings of the authors were that **part-of-speech is not a good metric in the detection of Alzheimer's**. Also, the authors **didn't find a relevant difference** between the **lexical richness** of Alzheimer's patients and their interviewers in the Carolina's Conversations corpus; however, they found differences in this metric when comparing to the SWITCHBOARD corpus. They found that **interviewers in the Carolina's Conversations corpus produce more go-ahead utterances than those in the SWITCHBOARD corpus**. The Guinn and Habash also found that **Alzheimer's patients produce more incomplete words**, **filler words**, **repetitions**, and have a **slower rate of speech**. Also, interviewers of Alzheimer's patients exhibit a **more extended use of paraphrasing**.

The paper includes several interesting computational linguistic metrics that can be used in order to detect Alzheimer's in conversations with patients; however, the findings of Guinn and Habash contradict those of other authors in respect of the usefulness of part-of-speech and lexical richness for detecting Alzheimer's trough language analysis. Nevertheless, the control groups in this study aren't comprised by people with similar characteristics to the ailing sample —such as age and education level. Likewise, using the interviewers in the Carolina's Conversations corpus as a control group is not equitable since the authors of this corpus state that "all interviewers receive training in more effective ways to speak with and listen to older people in natural conversations", which means that their aim is not to conversate freely and naturally, but to promote the flow of conversations.

The authors also tested the performance of three different machine learning algorithms: **k-neighbors** (with k=1, k=3 and k=5), **decision tree** and **support vector machine** (SVM). For their evaluation, they used the **features that they have found being statistically significant**. The machine learning algorithms were tested according to their combined performance in two tasks: identification of non-Alzheimer's patients, and identification of Alzheimer's patients. According to the authors, the **decision tree algorithm was the one with highest accuracy, with 79.5%**.

In their findings, the authors show optimism in the fact that they have a high false negative rate (63.48%) but a low false positive rate (16.14%), meaning that "these classifiers rarely say someone exhibits signs of dementia when, in fact, they do not have dementia". However, as mentioned before, their control groups have very different characteristics from the ailing group, which may imply that the ability of correctly identifying cognitive healthy patients can be related to those differences, and not specifically to the differences in the language caused by Alzheimer's progression.

In 2014 (Guinn, Singer, & Habash, 2014), the authors addressed the issue of comparing Alzheimer's patients with interviewers, admitting to the misleading nature of their past studies.

For this study, the authors compared conversations of 28 elderly **English** speakers with AD, and 28 otherwise healthy individuals from the Carolinas Conversations Collection.

The **features that better** helped in **identifying AD** patients were **Type-Token Ratio** (TTR), **Brunet's index**, percentage of **go-ahead utterances** and **incomplete words**. However, they still did **not find significant differences with noun**, **verb**, **adjective** or **pronoun rate**, nor with **pauses**, **repetitions**, **filler phrases** or **syllables per minute**.

With the features that were statistically significant, the authors applied two learning algorithms: a **Bayesian classifier** and a **decision tree** classifier. The authors had an 80.80% precision and **75% recall** for detecting AD patients.

In this study, the authors still found contradicting results with previous literature with respect of part-of-speech clauses. Also, this is not a longitudinal analysis, and it considers participants that are already at the dementia stage of the AD continuum. An interesting analysis of this study is that it includes the number of go-ahead utterances used by the interviewer, which makes it **an approach that also analyses the exchanges between the participant and its interlocutor**. Nevertheless, for the Carolinas Conversation Collection, the interviewers are trained to exhort AD participants and other participants with difficulties to continue their speech. This makes it difficult to extrapolate their findings to spontaneous conversations with non-trained interviewers and may present important differences even among trained interviewers due differences in their own personal communication strategies.

1.2.2.5 Khodabakhsh *et al.*

Khodabakhsh et al presented a series of studies (Khodabakhsh & Demiroglu, 2015; Khodabakhsh, Kusxuoglu, & Demiroglu, 2014; Khodabakhsh, Yesil, Guner, & Demiroglu, 2015) on Alzheimer's detection from unstructured conversational speech. For their study, they used **10 minutes unstructured conversations** of **28 AD patients** and **51 age and education-matched control subjects**. All participants were **native Turkish speakers**. For their

recollections, there was no specific thematic and the questions were followed according o the flow of the conversations.

As classifiers, the authors used an **SVM with a linear kernel**, a **nearest neighbor with Euclidean distance**, **and classification trees** algorithms. The authors make **special emphasis in acoustic features**, which is out of the scope of this dissertation, but they also extracted several linguistic features, such as TTR, Honoré's statistics, Brunet's index, suffix ratio, word entropy, number and date ratios, question ratio, fillers and incomplete sentence ratio, part-of-speech frequencies, and unintelligible word ratio.

Authors trained different learners with features that were significantly correlated with AD but with different types of controls. Learners used features controlled by education, age and gender. The **best learners** overall for differentiating between AD patients and healthy controls using linguistic features were **SVM** with **73.4% accuracy and 60.7% recall**. The most relevant features were **word entropy**, **question ratio**, **noun and pronoun ratios**, **pronoun-to-noun ratio**, **number and date ratios** and **Brunet's index**. The authors reported **83.5% of accuracy** and **64.3% recall using acoustic features**.

The main problem with this study is that AD participants were already at the late stages of the disease. However, it is important to note that some of the vocabulary richness features, such as Brunet's index and pronoun rate were significantly correlated with Turkish AD patients, a finding that coincides with Brazilian Portuguese and English AD patients. These studies seem to provide important clues for performing cross-linguistic studies.

1.2.3 Summary

Our literature review presented an overview on the computer-based works that have studied methods for detecting signs of Alzheimer's disease in two contexts: picture description tasks and spontaneous conversations. Most of these studies agree on the importance of lexical

richness features and part-of-speech ratio on the detection of the disease. However, **these studies differ on which specific measures and parts-of-speech are the most relevant for the task**. Further analysis on the significance of these types of features in both contexts is still need.

On the context of picture description tasks, most studies have measured the amount of information provided by the participants by using lists of information content units provided by specialists. To create a list, a specialist decide what is "important" from the picture, and therefore, what should be mentioned by the patients. This poses **several disadvantages**: first, the subjectivity of the task has led to a **lack of consensus in the evaluation of patients** undertaking the task, since **different authors include different units on the list** depending on their perception of their importance. Second, **the list is created by a person that differ vastly from the population for which the list is created**, which could lead to a gap between the observations from the specialist and from the patients. Third, the **lists are not generalizable to other types of pictures, populations and cultures**. A study that focused on the automatic detection of information content units was able to detect that a specific older population was including an item that has been ignored by previous lists' authors. This finding suggests that a computer-based creation of referents to evaluate the amount of information provided during a picture description could potentially help in the automatic evaluation of these tasks.

In the context of spontaneous conversations, **previous works have explored the differences on retained lexis between healthy controls and AD patients**. In their work, these authors found that AD patients tend to preserve a set of preferential lexis that they use to maintain communication abilities. A further exploration into the differences of use of specific and generic vocabulary between AD patients and healthy controls, not only in the context of spontaneous conversations but in description tasks is needed.

Due to the higher availability of corpora of English speakers, **most studies have focused their attention to this language**. It is important to include studies with different languages to expand

the knowledge about the forms that the disease manifests in different languages, and to explore the possibility of creating methods that are robust and language independent.

Finally, **few longitudinal studies have been made to observe language changes through the progression of the disease**. The inclusion of MCI patients to the studies, and the analysis of the language alterations that occur through time as patients transition from a healthy condition to cognitive impairment could provide a deeper understanding about the mechanisms that are affected by AD and help in its detection in pre-clinical stages.

CHAPTER 2

COMPUTER-BASED EVALUATION OF AD AND MCI PATIENTS DURING A PICTURE DESCRIPTION TASK

Laura Hernández-Domínguez[1], Sylvie Ratté[1], Gerardo Sierra-Martínez[2] and Andrés Roche-Bergua[3]

[1] Software and IT Engineering Department,
École de technologie supérieure, Montreal, Canada

[2] Engineering Institute, Universidad Nacional Autónoma de México (UNAM)

[3] Psychogeriatric Unit, Hospital Psiquiátrico Fray Bernardino Álvarez, Mexico

Email: laura.hzdz@gmail.com, sylvie.ratte@etsmtl.ca, gsierram@iingen.unam.mx

This article was published in *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, Elsevier, on March 13, 2018. DOI: 10.1016/j.dadm.2018.02.004.

2.1 Abstract

Introduction: We present a methodology to automatically evaluate the performance of patients during picture description tasks.

Methods: Transcriptions and audio recordings of the Cookie Theft Picture description task were used. With 25 HC samples and an information coverage measure, we automatically generated a population-specific referent. We then assessed 517 transcriptions (257 AD, 217 HC, 43 MCI) according to their informativeness and pertinence against this referent. We extracted linguistic and phonetic metrics which previous literature correlated to early-stage AD. We trained two learners to distinguish HC from cognitively impaired individuals.

Results: Our measures significantly ($p < .001$) correlated with the severity of the cognitive impairment and the Mini Mental State Examination score. The classification sensitivity was: between HCs and AD, 81% (AUC=.79); between HCs and AD&MCI, 85% (AUC=.76).

Conclusion: An automatic assessment of a picture description task could assist clinicians in the detection of early signs of cognitive impairment and AD.

Keywords: Alzheimer's disease (AD), Mild cognitive impairment (MCI), Picture description task, Automatic assessment, Information coverage, Linguistic analysis, Phonetic features, Machine learning

2.2 Introduction and Motivation

Multiple studies have assessed language functions as early markers of Alzheimer's disease (AD) (Szatloczki, Hoffmann, Vincze, Kalman, & Pakaski, 2015). Consequently, language is now widely accepted to be one of the first cognitive abilities affected by this dementia. Some of the most commonly used tests in clinical practice are Verbal Fluency by categories, Picture Description, the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983), and the Token Test (De Renzi & Vignolo, 1962), which measure expository speech, oral expression and comprehension of commands, respectively (Spreen & Strauss, 1998).

This exploration of the changes in language functions derived from AD has attracted significant attention among scientists outside the field of medicine (Laske et al., 2015). Researchers, especially those working in Natural Language Processing (NLP), have proposed computer-based approaches for automatic and semi-automatic analysis of language in patients suffering from AD (R. Alegria, Gallo, et al., 2013; Bucks et al., 2000; Fraser & Hirst, 2016; Homan et al., 2014; Khodabakhsh et al., 2014; König et al., 2015; López-de-Ipiña et al., 2015; Zhou, Fraser, & Rudzicz, 2016).

In this work, we propose a methodology to automatically describe patients' performance during a picture description task (Goodglass & Kaplan, 1983). We selected this type of test because it elicits spontaneous speech from patients, allowing to describe not only patients' ability to retrieve information from a visual stimulus, but also some of their linguistic characteristics. Our evaluation describes three aspects: the informativeness and pertinence of

the description provided by the patient, some linguistic characteristics, such as vocabulary richness and general use of part-of-speech categories, and a phonetic overview.

2.2.1 Information coverage

One of the key objectives of a picture description task is to measure the amount and quality of the information that a patient can provide from a visual stimulus. Even early in the course of the disease, AD patients have been shown to provide less informative descriptions than cognitively intact elderly adults (Ahmed, de Jager, et al., 2013). This measure is generally made by comparing the description provided by the patient to a list containing the main information content units (ICUs) of the image, namely, actors, objects, actions and places. Over the years, several authors have come up with pre-defined lists of ICUs for the Cookie Theft picture description task (Croisile et al., 1996; Forbes-McKay & Venneri, 2005; Hier, Hagenlocker, & Shindler, 1985; Yi hsiu Lai, Pai, & Lin, 2009; Nicholas, Obler, Albert, & Helm-Estabrooks, 1985; Yorkston & Beukelman, 1980). However, one of the disadvantages of using pre-defined lists to evaluate elderly patients is that the list author does not necessarily have a similar education level, age, focus, cultural background and interests as the target population. Also, different authors may come up with different lists, depending on their idiosyncrasies, their own observations, and what they may consider "important" from the picture.

2.2.1.1 Related computational works

Hakkani-Tür et al. (Hakkani-Tür, Vergyri, & Tur, 2010) used a manually pre-defined list as referent to automatically compare descriptions of the Western Aphasia Battery's Picnic Picture. The authors found a high correlation between the traditional manual assessment and their automatic approach. However, the automatic evaluation had trouble handling ICUs expressed in multiple ways.

Pakhomov et al. (Pakhomov et al., 2010) used manual transcriptions of descriptions of the Cookie Theft picture to assess the performance of patients with frontotemporal lobar degeneration. They compiled a list of pre-defined ICUs based on (Yorkston & Beukelman, 1980), and manually extended it to include lexical and morphological variants of words and phrases. One drawback of this method is that it entails the manual creation of a list that considers as many variants as possible for each ICU.

Fraser et al. (Fraser et al., 2016) used a semi-automatic approach to automatically classify Alzheimer's patients and healthy elderly controls (HC) by analyzing manual transcriptions of descriptions of the Cookie Theft picture in the Pitt corpus (Becker et al., 1994). As a referent, the authors used the pre-defined list proposed by (Croisile et al., 1996), and evaluated the frequency of key words used to name the ICUs in different ways. As in Pakhomov's work, manually considering all the ICUs and their linguistic variations is a time-consuming task.

Yancheva (Yancheva & Rudzicz, 2016) automatically extracted the main ICUs retrieved by elderly adults in the Pitt corpus. The authors contrasted automatically extracted ICUs to a combination of several pre-defined lists of ICUs. They retrieved most of the human-selected ICUs. Additionally, they found that some participants mentioned the object *apron*, a new ICU that none of the specialists had perceived before. They also observed that HCs were more prone than AD patients to mention this object in their descriptions.

The appreciation of the fact that a woman is wearing an apron while doing housework could be attributed to a generational and cultural perception of what the object *apron* represented to elderly participants taking the test back in the 1980s. Different remarks may be attributable to cultural differences. For example, a non-Caucasian-predominant population may remark on the fact that all the subjects in the Cookie Theft picture are blond. Hence, we consider that a fairer referent for comparison in this task should be constructed by healthy participants of the target population. As such, it would be possible to create referents that are adapted to specific populations from different generations, cultures, and educational and general socio-economic backgrounds.

2.2.1.2 The coverage measure

We identify three important tasks for performing an automatic evaluation of a picture description task:

1) Creating a population-adapted referent.
2) Evaluating the *informativeness* of descriptions: estimate how much of the information in the referent is being covered by the participant.
3) Evaluating the *pertinence* of utterances: determine how much of what the participant is saying is covered by the referent. Some participants, particularly those with AD, can drift off-topic. While this situation is easily detected when performing a manual evaluation, it is a challenging task for an automatic analysis.

With these tasks in mind, we selected the information coverage measure proposed by Velazquez (Velázquez-Godínez, 2017). He originally proposed the method for comparing the coverage of information in news articles, although it could be used in different contexts.
Velazquez proposes a methodology for creating a referent, providing a subject of comparison for evaluating the information coverage. One distinguishing feature of his measure is that it uses linguistic patterns that allow the consideration of the context. Additionally, the measure allows a two-way analysis of the information coverage, from the referent by the subject of comparison and vice versa. These two measures would allow the estimation of informativeness and pertinence, respectively.

2.2.2 Linguistic characteristics

There is extensive literature covering the analysis of the linguistic characteristics of AD patients (Ahmed, Haigh, De Jager, & Garrard, 2013; R. Alegria, Bolso, et al., 2013; R. Alegria, Gallo, et al., 2013; Bucks et al., 2000; Fraser et al., 2016; Guinn & Habash, 2012b; Jarrold et al., 2010; Kemper et al., 1993; Khodabakhsh et al., 2015; Snowdon et al., 1996). As part of our evaluation, we selected those that most authors have found to correlate significantly with the

disease, and that could be used in picture description tasks (Table 2.1). In Section 2.3.2, we provide further information about the methodology and tools used for extracting these characteristics.

2.2.2.1 Part-of-speech distribution

We made an evaluation of the frequency and ratio of adjectives, conjunctions, nouns, prepositions and verbs per 100 words. We also evaluated the frequency of auxiliary verbs, and their ratio to the total number of verbs.

2.2.2.2 Vocabulary richness

Several measures have been explored to evaluate the richness of an author's language. These same measures can be used to evaluate the variability of the vocabulary of patients during a picture description task.

2.2.3 Phonetic analysis

Several authors (Fraser et al., 2016; Khodabakhsh & Demiroglu, 2015; Lopez-de-Ipina et al., 2015; Pakhomov et al., 2010; Rudzicz, Chan Currie, Danks, Mehta, & Zhao, 2014; Satt et al., 2014) have found significant differences in the audio signals produced by AD patients as compared to cognitively intact elderly individuals. Mel Frequency Cepstral Coefficients (MFCCs) are among the most used features for automatic speech analysis. Only the first 12 to 13 MFCCs are usually used, since most of the information about the transfer function of the vocal tract is in the lower range of frequencies.

Table 2.1 Linguistic characteristics selected to evaluate patients' language functions

Measure	**Equation**	**Interpretation**
Text size	N	Number of words used in a text
Vocabulary size	V	Number of different lemmas*
Hapax legomena	V_1	Number of lemmas mentioned only once
Hapax dislegomena	V_2	Number of lemmas mentioned exactly twice
Brunet's W index (Brunet, 1978)	$W = N^{V^{-c}}$	Rationalization of the size of the vocabulary and the length of the text. *W* is stable when *c* has values between 0.165 and 0.172 (Holmes & Forsyth, 1995). We used *c*=0.172, the original value proposed by Brunet.
Honoré's R statistics (Honoré, 1979)	$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}}$	A measure based on the ratio of hapax legomena, vocabulary size, and the length of the text
Type Token Ratio	$TTR = \frac{V_1}{V}$	TTR measures the ratio of hapax legomena and the size of the vocabulary. It can be sensitive to the size of the sample (McEnery & Oakes, 2000).
Sichel's S (Sichel, 1975)	$S = \frac{V_2}{V}$	Similar to TTR, but using hapax dislegomena, being more robust against samples of different sizes (Tweedie & Baayen, 1998)
Yule's characteristic *K* (Miranda-García & Calle-MartÍn, 2005)	$K = 10^4 \frac{[\sum_{i=1}^{N} i^2 V(i,N)]}{N^2} - \frac{1}{N}$	Yule's is a measure of lexical repetition considered to be text length independent. In this measure, the number of lemmas of frequency i ($V(i,N)$) is estimated to measure the frequency distribution of a text.
Entropy	$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$	Entropy measures the uniformity of the vocabulary. In the equation, $p(x)$ is the probability of a word x occurring in the text X. We measured the general entropy of the complete text, and the average entropy of sentences.

*lemmas refer to words without inflections (in their canonical form).

2.3 Methods

2.3.1 Corpus

For this work, we used the Pitt corpus (Becker et al., 1994) of the DementiaBank database. This corpus contains audio recordings and manual transcriptions of participants undertaking the standard Cookie Theft picture description task from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). This password-protected dataset is available upon request for research purposes.

Table 2.2 Distribution of interviews used for experimentation

Variable	**All (n=517)**	**AD (n=257)**	**HC (n=217)**	**MCI (n=43)**
Participants	262	169	74	19
Gender				
Male	189	87	75	27
Female	328	170	142	16
Education (y)				
6-9	55	51	2	2
10-12	200	112	79	9
13-16	209	76	111	22
17+	53	18	25	10
Age (y)				
Under 50	6	0	5	1
50-59	81	21	57	3
60-69	188	81	94	13
70-79	190	111	57	22
80+	52	44	4	4

Abbreviations: n, number of tests; AD, Alzheimer's disease; HC, healthy elderly control; MCI, mild cognitive impairment

The participants of the corpus are mainly HCs, probable and possible AD patients, and Mild Cognitively Impaired (MCI) subjects. We excluded other diagnoses from this study due to their scarce numbers in the corpus. In this work, we did not differentiate between probable and possible diagnoses of AD. The main inclusion criterion for our study was that both the transcripts and audio files of the participant were present for each test. We studied 262 participants, with a total of 517 tests (see Table 2.2). 25 other HC subjects and their tests were set aside for creating the referent. These subjects were not part of the experimentation sample.

2.3.2 Extraction of information coverage measures

2.3.2.1 Adaptation of the coverage measure

Velazquez's (Velázquez-Godínez, 2017) measure uses duplets of linguistic patterns to find the degree to which a referent *R* is covered by a subject of comparison *S*. We selected the active voice patterns proposed by Velazquez, given the expository speech nature used during picture description tasks (see Table 2.3).

Table 2.3 Active voice linguistic patterns used for the coverage measure[1]

p* in *R	***p* in *S***	**Interpretation**	**Example**
N-V	N-V	Subject + action	"boy stealing"
V-N	V-N	Action over an object	"stealing cookies"
P-N	P-N	Locations, Indirect objects	"in kitchen"
N-V-N	N-V-N	Subject + action + object	"woman washing dishes"

Abbreviations: p, pattern; R, referent; S, subject of comparison; N, noun; V, verb; P, preposition.

[1] Taken from (Velázquez-Godínez, 2017) with the author's permission.

Velazquez splits the text into sentences; for our study, we split it into utterances. The comparison of utterance patterns follows the equation:

$$coverage(R,S) = \frac{\sum_{p\in\{R\}} MaxSim(p,S) \times \alpha_p}{\sum_{p\in\{R\}} \alpha_p} \quad (2.1)$$

where *R* is the referent, *S* is the document that is the subject of comparison, and *p* is a linguistic pattern. The parameters α are used to modify each pattern's weight. For this work, all patterns were considered to weigh equally; all parameters α_p were thusly set to 1.

2.3.2.2 Automatic pre-processing

We cleaned the original raw text to apply the information coverage measure as follows:

1) We removed all marks of repetitions, hesitations, incomplete words and pauses, as well as any introductory statements such as "*it looks like*".
2) We standardized the names of the most prominent ICUs. For example, all mentions of the words "*brother*", "*lad*", "*kid*", etc., were automatically replaced by "*boy*" following Figure 2.1.
3) We used FreeLing 4.0 (Padró & Stanilovsky, 2012) for tagging the transcripts with their *lemmas* and their part-of-speech.
4) Two consecutive nouns were considered a single noun divided by a forward slash.
5) Some authors have found differences in the use of adjectives between AD patients and HCs (Homan et al., 2014). During the picture description task, it is common that participants describe objects with adjectives. To take these rich descriptions into account, we joined an adjective preceded by the verb "*to be*" by means of a forward slash.
6) All part-of-speech tags that were not in the linguistic patterns were discarded for the comparison.

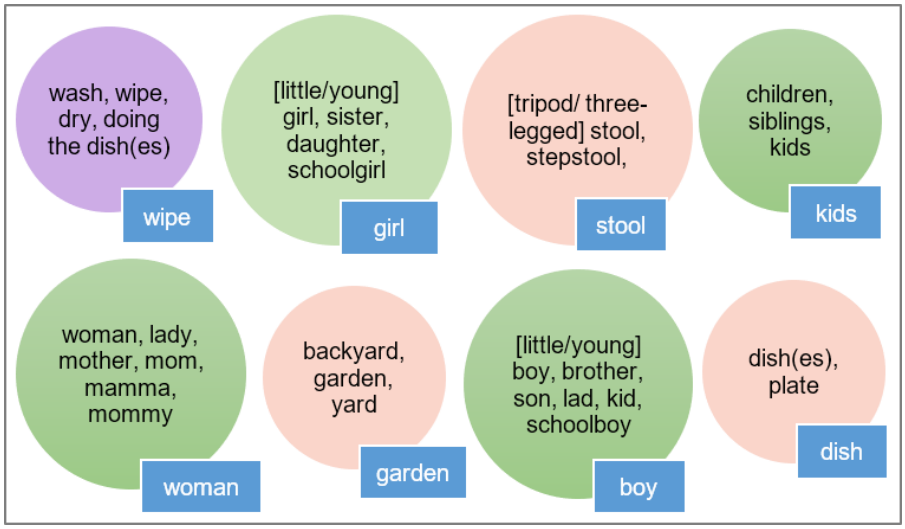

Figure 2.1 Linguistic variations of ICUs in the Cookie Theft picture description task. The standardized name of each group is shown. Abbreviation: ICUs, information content units.

2.3.2.3 Creation of the referent

We created a referent formed from tests taken by HCs from the same corpus. All HCs whose entrances were used to create the referent were excluded from the evaluation sample. To create the referent, we selected all 25 HCs with only one test. We aimed for the referent to be as diverse as possible, while simultaneously significantly avoiding reducing the number of samples left for the evaluation.

Using Velazquez's coverage measure, we created a referent that included the patterns extracted from the 25 HCs. For each utterance, if the utterance was not already at least 80% covered by the referent, the patterns were added to the referent. Thus, we automatically created an incremental referent that considered different manners used by HCs to describe similar actions and situations. The following are real examples of patterns in the referent:

- water(N) run(V)
- water(N) overflow(V)
- water(N) spill(V)

- water(N) flow(V)
- water(N) splash(V)
- spill(V) water(N)
- kitchen/water(N) overflow(V)

2.3.2.4 Scoring participants' performance

Informativeness was estimated by measuring how much of the information in the referent was covered by the participant. The more the referent was covered by a participant, the more informative the associated descriptions were. To measure the *pertinence*, we estimated how much of what the participant said was covered by the referent. A low pertinence coverage may indicate that the participant was drifting off-topic.

Emulating a typical clinical scoring of a picture description test, we counted the number of utterances from the referent that exceeded an *informativeness* and a *pertinence* threshold. To that end, we tested three different thresholds: 60%, 80% and 100%. We also estimated the sum of the *informativeness* and *pertinence*.

2.3.3 Extraction of linguistic and phonetic characteristics

For extracting the linguistic characteristics, we conducted a usual NLP pre-processing by removing all marks of repetitions, hesitations, incomplete words and pauses. We used FreeLing 4.0 (Padró & Stanilovsky, 2012) to automatically tag the transcripts with their lemmas and part-of-speech. We then automatically extracted the linguistic characteristics described in section 1.2.

We used python_speech_features 0.6 (Lyons, 2017) to estimate the first 13 MFCC values of the soundwaves in 25 millisecond segments. As per Fraser et al. (Fraser et al., 2016), our features consisted of the mean, kurtosis, skewness and variance of the values of each MFCC.

2.3.4 Automatic classification

To automatically discriminate between HCs and cognitively impaired individuals, we used two widely recognized machine learning (ML) algorithms, namely, Support Vector Machine (SVM) (Smola & Schölkopf, 2004) and Random Forests (RFC) (Breiman, 2001). In (Asgari, Kaye, & Dodge, 2017) is presented a succinct and elegantly simplified explanation of both algorithms and of their use in a linguistic analysis for detecting MCI.

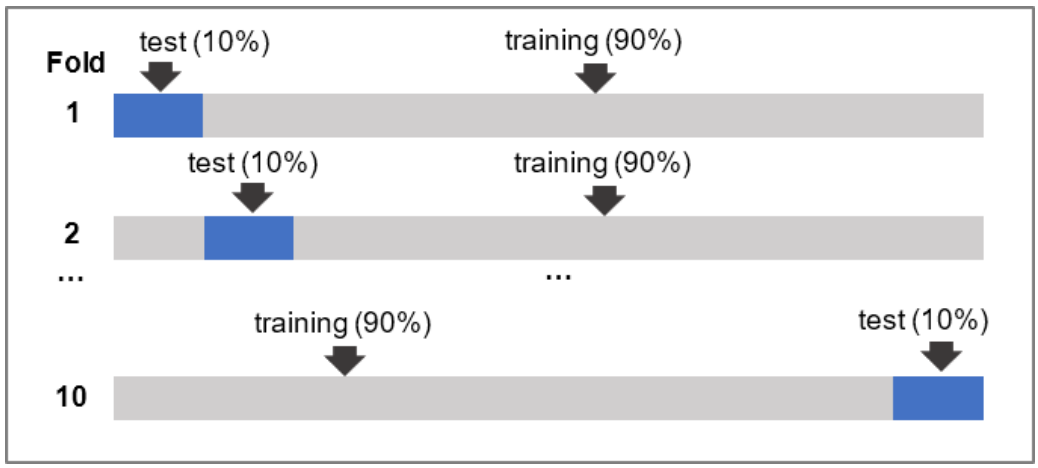

Figure 2.2 Data set partitioning during a 10-fold cross-validation process to evaluate classifiers. The blue section indicates the part of the dataset that is being used as test, while the remaining gray area indicates the part of the dataset being used as training set in each fold.

For our evaluation, we performed two binary classification experiments: the first one consisted of a classification between participants with AD and HCs, while for the second, we added the MCI participants to the sample and classified HCs and cognitively impaired participants. The sample of MCI participants was too small to be used as a learning class.

For this work, we used 90% of the evaluation sample as the training set, and 10% as the test set. We performed a 10-fold cross-validation (see Figure 2.2). We report average of the ten test classifications.

2.4 Results

2.4.1 Feature analysis

Our evaluation of participants' picture descriptions covered a total of 105 features, divided into information coverage measures and linguistic and phonetic characteristics. We estimated their correlation with the severity of the cognitive impairment diagnosis (healthy=0, MCI=1, AD=2) and with the Mini-Mental State Examination (MMSE) results of participants. These correlations are reported in Table 2.4.

We additionally analyzed the correlation of the features with respect to age, gender and education. Our findings are reported in Table 2.5.

2.4.2 Binary classification

We tested the performance of the algorithms with each type of feature independently, and then combinations of them. The results of the first and second experiments are shown in Table 2.6 and Table 2.7, respectively. The best model represents the performance of the algorithms with a higher area under the curve of receiver operating characteristics (AUC) during the 10-fold cross-validation process.

Table 2.4 Correlations* of features with the severity of cognitive impairment and with the MMSE

Correlation to cognitive impairment		**Correlation to MMSE score**	
Variable	*Corr.†*	*Variable*	*Corr.†*
Informativeness t=100%	-0.408	Informativeness t=100%	0.443
Informativeness t=80%	-0.388	Informativeness t=80%	0.437
Informativeness score	-0.334	Informativeness score	0.429
Informativeness t=60%	-0.333	Informativeness t=60%	0.372
Informativeness variance	-0.257	Informativeness variance	0.338
Hapax legomena	-0.254	Auxiliary verb frequency	0.305
Pertinence t=100%	-0.222	Hapax legomena	0.265
Auxiliary verb frequency	-0.216	Auxiliary verb rate	0.241
MFCC-12 kurtosis	0.205	Noun frequency	0.226
Pertinence t=80%	-0.201	Preposition rate	0.194
MFCC-8 kurtosis	0.198	Pertinence t=100%	0.192
MFCC-12 skewness	-0.185	General entropy	0.189
Noun frequency	-0.183	Vocabulary size	0.187
Honoré's R statistics	-0.180	Pertinence t=80%	0.183
MFCC-10 kurtosis	0.163	Honoré's R statistics	0.177
Conjunction rate	0.163	Preposition frequency	0.175
Vocabulary size	-0.156	MFCC-12 skewness	0.173

Abbreviations: MMSE, Mini-Mental State Examination; t, threshold; MFCC, Mel Frequency Cepstral Coefficients.

*All correlations presented with P value < .001. Variables are shown in descending order with respect to the strength of their correlation.

†Controlled for education, age and gender.

Table 2.5 Correlations* of features with sociodemographic variables

Age		**Gender**		**Education**	
Variable	*Corr.*†	*Variable*	*Corr.* ‡	*Variable*	*Corr.*§
MFCC-3 kurtosis	-0.200	MFCC-10 kurtosis	-0.239	Preposition freq.	0.230
Conjunction freq.	0.182	MFCC-12 variance	0.181	Hapax legomena	0.222
Brunet's W index	0.179	MFCC-13 variance	0.179	Vocabulary size	0.219
General entropy	0.177	MFCC-5 skewness	0.175	Text size	0.207
Auxiliary verb freq.	0.175	MFCC-5 variance	0.174	General entropy	0.201
MFCC-1 variance	0.172	MFCC-8 skewness	-0.169	Adjective freq.	0.200
MFCC-6 kurtosis	-0.168	MFCC-10 skewness	0.163	Conjunction freq.	0.191
MFCC-5 kurtosis	-0.164			Noun freq.	0.190
MFCC-9 kurtosis	-0.158			Informativeness t=60%	0.190
Informativeness score	0.156			Auxiliary verb freq.	0.184
				Verb freq.	0.172
				Informativeness score	0.169
				Brunet's W index	0.167

Abbreviations: MFCC, Mel Frequency Cepstral Coefficients; freq., frequency.
*All correlations presented with P value < .001. Variables are shown in descending order with respect to the strength of their correlation.
†Controlled for education, gender, and cognitive status.
‡Controlled for age, education, and cognitive status.
§Controlled for age, gender, and cognitive status.

Table 2.6 Performance* of classifiers separating HCs from AD patients

Learner	Features	Accuracy	Sensitivity	Specificity	Precision	F-score	AUC
Average performance							
RFC	Ling	0.72	0.76	0.67	0.74	0.75	0.72
SVM	Ling	0.75	0.75	0.74	0.77	0.76	0.75
RFC	Cov	0.73	0.78	0.67	0.73	0.75	0.72
SVM	Cov	0.74	0.80	0.67	0.74	0.77	0.74
RFC	Phon	0.59	0.66	0.52	0.62	0.64	0.59
SVM	Phon	0.62	0.70	0.52	0.63	0.66	0.61
RFC	Cov+Ling	0.78	0.84	0.72	0.78	0.81	0.78
SVM	**Cov+Ling**	**0.79**	**0.79**	**0.78**	**0.82**	**0.80**	**0.79**
RFC	Best†	0.75	0.78	0.71	0.76	0.77	0.74
SVM	**Best†**	**0.79**	**0.81**	**0.77**	**0.81**	**0.81**	**0.79**
Best model							
RFC	Ling	0.81	0.77	0.86	0.87	0.82	0.82
SVM	Ling	0.85	0.85	0.86	0.88	0.86	0.85
RFC	Cov	0.85	0.88	0.82	0.85	0.87	0.85
SVM	Cov	0.85	0.88	0.82	0.85	0.87	0.85
RFC	Phon	0.67	0.65	0.68	0.71	0.68	0.67
SVM	Phon	0.72	0.84	0.57	0.70	0.76	0.71
RFC	**Cov+Ling**	**0.94**	**1.00**	**0.86**	**0.90**	**0.95**	**0.93**
SVM	Cov+Ling	0.88	0.81	0.95	0.95	0.88	0.88
RFC	Best†	0.85	0.85	0.86	0.88	0.86	0.85
SVM	Best†	0.87	0.80	0.95	0.95	0.87	0.88

Abbreviations: AUC, area under the curve of receiver operating characteristics; RFC, Random Forests Classifier; SVM, Support Vector Machine classifier; Ling, set of all linguistic features; Cov, set of all information coverage features; Phon, set of all phonetic features; Cov+Ling, a combination of all linguistic and information coverage features.

*The best results are indicated in bold.

†A combination of all features with P value < .001 when correlating with cognitive impairment.

Table 2.7 Performance* of classifiers separating HCs from cognitively impaired patients (AD or MCI, indistinctly)

Learner	Features	Accuracy	Sensitivity	Specificity	Precision	F-score	AUC
Average performance							
RFC	Ling	0.70	0.78	0.59	0.73	0.75	0.69
SVM	Ling	0.72	0.80	0.61	0.74	0.77	0.70
RFC	Cov	0.74	0.83	0.61	0.75	0.79	0.72
SVM	Cov	0.73	0.86	0.56	0.73	0.79	0.71
RFC	Phon	0.59	0.79	0.31	0.61	0.69	0.55
SVM	Phon	0.61	0.81	0.33	0.62	0.70	0.57
RFC	Cov+Ling	0.76	0.84	0.66	0.77	0.81	0.75
SVM	**Cov+Ling**	**0.78**	**0.85**	**0.68**	**0.78**	**0.82**	**0.76**
RFC	Best†	0.77	0.82	0.69	0.78	0.80	0.75
SVM	Best†	0.75	0.82	0.65	0.76	0.79	0.73
Best model							
RFC	Ling	0.78	0.80	0.76	0.83	0.81	0.78
SVM	Ling	0.87	0.87	0.86	0.90	0.88	0.87
RFC	Cov	0.83	0.87	0.77	0.84	0.85	0.82
SVM	Cov	0.83	0.90	0.73	0.82	0.86	0.81
RFC	Phon	0.67	0.90	0.36	0.66	0.76	0.63
SVM	Phon	0.65	0.90	0.29	0.64	0.75	0.59
RFC	Cov+Ling	0.85	0.87	0.82	0.87	0.87	0.84
SVM	Cov+Ling	0.85	0.87	0.82	0.87	0.87	0.84
RFC	**Best†**	**0.87**	**0.87**	**0.86**	**0.90**	**0.88**	**0.87**
SVM	Best†	0.83	0.83	0.82	0.86	0.85	0.83

Abbreviations: AUC, area under the curve of receiver operating characteristics; RFC, Random Forests Classifier; SVM, Support Vector Machine classifier; Ling, set of all linguistic features; Cov, set of all information coverage features; Phon, set of all phonetic features; Cov+Ling, a combination of all linguistic and information coverage features.

*The best results are indicated in bold.

†A combination of all features with P value < .001 when correlating with cognitive impairment.

2.5 Discussion

We presented a methodology for an automatic evaluation of a picture description task. This evaluation aims not only to score participants' performance during the task itself, but also to analyze their language and phonetic productions in a single commonly used non-invasive clinical test. Our objective is to provide clinicians with computational aids for the early detection of signs that might alert of the presence of MCI or AD.

From the features observed in Table 2.4, the strongest correlations with the severity of the cognitive impairment were obtained with the information coverage measures. The less informative or pertinent the picture description, the higher the severity of the impairment. These correlations were consistent with the participants' scores on the MMSE, and were mostly independent of age, gender and education.

Our findings on the correlation of linguistic and phonetic features with cognitive impairment were consistent with previous literature, and provided a broader evaluation of the participants' performance. In general, vocabulary richness and syntactic complexity measures were inversely correlated with the severity of the disease. These variables were also positively correlated with the number of years of education. An increased rate of conjunctions was correlated with cognitive impairment. We hypothesize that a high use of conjunctions in a picture description task could indicate hesitation or confusion.

Phonetic variables were naturally highly correlated with age and gender. Also, we observed an increased use of conjunctions with age, as well as an increase in the entropy of the description. This may indicate more chaotic or disorganized descriptions.

We tested SVM and RFC first with each type of feature independently, and then with combinations of same. When we experimented with all three types of features together, we carried out a pre-selection of the best features. For this selection we chose features with $P < .001$ when correlating to the severity of the cognitive impairment.

2.5.1 Comparison to other approaches

Contrasting our results against previous works on automatic evaluation of picture description tasks can be difficult for multiple reasons. First and foremost, it is not customary in NLP to provide performance metrics such as AUC and specificity. While accuracy, precision, recall and F-score, are usually illustrative in classes with similar sample sizes, these values could become misleading when the classes are skewed.

An additional challenge in contrasting these methods is that not every author works with the same data distribution even when using the same dataset. With ML algorithms, the ways the samples are distributed along the dataset and in the training and test set lead to slightly different results. Authors tend to report the results obtained with a distribution in which their algorithms performed at their best.

Finally, despite using the Pitt corpus, previous works differ in the number of samples used during their evaluation. Fraser et al. (Fraser et al., 2016) used 233 HC and 240 AD samples; Yancheva et al. (Yancheva & Rudzicz, 2016) used 241 HC and 255 AD samples; for this work, we used 242 HC and 257 AD samples (about 10% of the HC sample was used to form the referent, and was not included in the evaluation). There is no clear explanation from previous authors regarding why they did not include all the samples in their experimentation.

Fraser et al. reported an accuracy of 81.92%, while Yancheva et al. reported an accuracy, precision, recall and F-score of 80%. In both works, the authors performed a classification between HCs and AD participants, without including the MCI sample. In our work, two SVM classifiers tied with the highest AUC at 0.79 in this task (Table 2.6). The first learner used a combination of all the information coverage and linguistic features, while the second used a combination of all features with $P < .001$ when correlating with cognitive impairment. The second algorithm presented a higher sensitivity (81%) and a higher F-score (81%), comparable to state of the art work (Fraser et al., 2016) which use a manually-made list of ICUs.

When we incorporated the MCI sample into the experiment (Table 2.7), we observed that the SVM learner, trained with information coverage and linguistic features, performed at the highest AUC (0.76). There was an expected increase in the false-negative rate (specificity = 68%). However, the sensitivity was still high at 85%.

The best model of an experiment represents the highest performance achieved by an algorithm during the cross-validation process. This indicates the highest potential of the algorithms in classifying new data with similar characteristics to the sample. For the first experiment (Table 2.6), the best model had an AUC of 0.93, with excellent sensitivity and a true-negative rate of 86% when classifying HCs and AD patients. When the MCI sample was incorporated (Table 2.7), the best model had an AUC of 0.87, with sensitivity=87% and specificity =86%.

2.5.2 Study advantages and limitations

One of the advantages of our proposed methodology is that the informativeness and pertinence measures are estimated against an automatically created referent. This referent has the particularity of being adaptable to differences in population or even to different pictures for description.

Previous automatic works present difficulties at considering linguistic variabilities for expressing similar notions. With our proposed approach, the referent is created from examples of descriptions from healthy age-related individuals. Hence, it incorporates different ways of expressing similar ideas, and even what could be considered as *normal* deviations from topics. It also allows for the consideration of context through the accounting of linguistic patterns of phrases, rather than just of isolated words. In this regard, the bigger the sample set aside for creating the referent, the richer and more variate the referent.

To our knowledge, this is the first time that an automatic measure of pertinence has been implemented in a picture description task. While most computational approaches focus only

on the information coverage, one advantage of our measure is that it helps to detect when patients drift off topic, a highly challenging task in automatic analysis.

One disadvantage of our approach is that it sacrifices part of the HC group to create the referent, reducing the availability of HC samples for training the algorithms. Our study also presented a limitation when evaluating MCI patients, yielding a high false-negative rate.

2.5.3 Future work

In both experiments, we observed that our phonetic characteristics were not sufficiently discriminative or had little to no effect in the performance of the algorithms. As previous authors have reported, the use of more complex acoustic and rhythm features could significantly increase the automatic classification performance of HCs and AD patients.

In future work, we propose to extend the research scope with the evaluation of the performance of the information coverage metrics in descriptions of different picture description tasks or even in different restricted-discourse tests. Also, there is a potential to perform multilingual studies, since all the features proposed in this work are language-independent or can be adapted for studies in different languages. Finally, we intend to research the effects of different HC sample sizes for creating the referent.

CHAPTER 3

AUTOMATED DIFFERENTIATION OF ALZHEIMER'S AND MCI PATIENTS FROM HEALTHY CONTROLS USING ENGLISH AND SPANISH TRANSCRIPTIONS OF DESCRIPTION TASKS

Laura Hernández-Domínguez[1], Sylvie Ratté[1] and Gerardo Sierra[2]

[1] Software and IT Engineering Department,
École de technologie supérieure, Montreal, Canada
[2] Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico

E-mail: laura.hzdz@gmail.com, sylvie.ratte@etsmtl.ca

This article was submitted to *Computers in Biology and Medicine*, Elsevier, on November 4, 2018.

3.1 Abstract

Background: The diagnosis of Alzheimer's disease (AD) at early stages is a research priority. Description tasks have traditionally been used as part of the battery of tests for the cognitive assessment and early detection of AD. These tasks evaluate a patient's ability to focus and observe, as well as the assessment of some language functions in the patient.

Method: One corpus of standard picture descriptions in English and one of descriptions of common objects in Spanish were used for this study. We proposed automatically extracted information coverage and pertinence measures based on the use of task-specific vocabulary. We used these measures, in combination with lexical richness and other linguistic measures, to train Random Forest (RF) and Support Vector Machine (SVM) classifiers to distinguish AD and mild cognitive impairment (MCI) patients from healthy controls.

Results: With our best model, we obtained an F_1-score of 0.97 and 0.83 for a classification differentiating AD patients from healthy controls in the Spanish and English datasets, respectively. For MCI patients versus healthy controls, we obtained an F_1-score of 0.80.

Conclusions: Our proposed information coverage and pertinence measures proved to be highly relevant for the classification process since they were always part of the selected features after backward elimination, and had a significant correlation with the severity of cognitive impairment in both corpora. These results compared favorably against other computational state-of-the-art methods for AD and MCI detection using these datasets. Our proposed method presents an inexpensive and non-invasive alternative for detection of early signs of AD.

Keywords: Alzheimer's disease; Mild cognitive impairment; automated diagnosis; natural language processing (NLP); picture description; language alterations.

3.2 Introduction

Alzheimer's disease (AD) is the leading cause of dementia, with a high prevalence in the elderly population(Alzheimer's Association, 2018b). In the United States, 3% of people aged 65-74 have dementia of the Alzheimer's type, and the risk increases with age. At age 85 and older, the prevalence of the disease is estimated at 32%.

Currently, there is no biomarker or test that can diagnose AD with certainty, especially in its early stages. The Alzheimer's association identifies mild cognitive impairment (MCI) as a desirable stage for detecting early signs of AD (Alzheimer's Association, 2015, 2018b). Although an MCI diagnosis is not necessarily an AD sentence, it is estimated that 32% of MCI cases will progress to the Alzheimer's dementia stage in the following five years (Ward et al., 2013).

When diagnosed, Alzheimer's disease is usually identified as "probable" or "possible" AD, with post-mortem confirmation by autopsy. Since only 1% or less of all AD cases are considered to be the result of a genetic mutation (Bekris et al., 2010), genetic screening for AD is not appropriate as a diagnostic tool for the majority of the population.

The current procedure for diagnosing AD is based on a combination of approaches, and usually involves different medical specialists, such as neurologists and geriatricians (Alzheimer's Association, 2018b). The family history and reports from family members regarding changes in skills and behavior are taken into account for the diagnosis. Additionally, cognitive and blood tests, as well as brain imaging, are carried out to determine the state of the patient and to discard other potential causes.

Language function assessment has been widely recognized as an early marker for the disease (Szatloczki et al., 2015), even before the manifestation of symptoms. The most commonly used cognitive tests in clinical practice for assessing language functions are picture description tasks, verbal fluency by categories, the Boston naming test (Kaplan et al., 1983), and the Token test (De Renzi & Vignolo, 1962). These tests are used to evaluate patients' oral expression, expository speech and comprehension of commands.

Since the assessment of language functions became an early marker of AD, multiple computer-based analyses of changes in language caused by Alzheimer's disease and Mild Cognitive Impairment have surged. Most of these studies have been done with English speakers, since the majority of clinical datasets available for research are in this language.

Some computer-based works (Drummond et al., 2015; Fraser et al., 2016; Kavé & Goral, 2016; Kavé et al., 2018; Sylvester O. Orimaye et al., 2017; Sylvester Olubolu Orimaye et al., 2014; Yancheva & Rudzicz, 2016) have been carried out through analyses of standardized picture descriptions tasks. These tasks elicit a semi-spontaneous speech (Prins & Bastiaanse, 2004) with a predictable structure in a constrained context, which facilitates comparisons across

studies and languages. However, these tasks tend to limit the variety of syntactic structures (Boschi et al., 2017), not allowing further analysis of certain linguistic phenomena.

Other works (R. Alegria, Gallo, et al., 2013; Renné P. Alegria et al., 2009; Bucks et al., 2000; Guinn et al., 2014; Jarrold et al., 2014; Khodabakhsh & Demiroglu, 2015) have focused on analyses of spontaneous conversations, which feel more natural and are less stressful to patients. These studies allow for deeper linguistic analyses, but are difficult to compare across patients and studies, in addition to being prone to presenting differences depending on the subject and duration of the interview. For these reasons, such studies would seem to be better suited for personalized medicine and longitudinal intra-patient comparisons.

In this work, we evaluate the different manifestations of Alzheimer's disease in the language of patients in two different description tasks. The first task is the standard Cookie Theft picture description from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). The second task consists of a description of six common objects. While the objects to be described are the same for all speakers, the descriptions derive from their personal experiences and views of each object. This type of elicited speech is less restrictive in context and syntactic structures than a standardized picture description task, but still constrains the vocabulary and limits the scope of the discourse.

3.3 Materials and Methods

For this study, we compared the use of the parts of speech (POS) and specific vocabulary, as well as traditional lexical richness measurements of healthy elderly controls (HC) and Alzheimer's (AD) patients. Our main goal was to provide clinicians with a non-invasive and economical computer-based tool to aid in the detection of the disease.

3.3.1 Corpora

Two main clinical corpora, available under request for research purposes, were used for this study: the BBVA linguistic corpus of definitions of sematic categories by healthy and Alzheimer's afflicted elderly people (Peraita & Grasso, 2010), in Spanish, and the Pitt Corpus of the DementiaBank (Becker et al., 1994) dataset, in English. Additionally, to obtain a sample of general vocabulary, we used two free-discourse corpora: the oral corpus of reference of contemporaneous Spanish (Marcos Marín, 1992) (CORLEC) and the Carolinas' Conversations Collection (CCC) (Pope & Davis, 2011), in English.

3.3.1.1 Constrained-discourse corpora

The BBVA and Pitt corpora (see Table 3.1) contain transcripts of the elicited speech of elderly people, with and without dementia. In both corpora, participants were recorded while performing specific description tasks, which limits their vocabulary to specific subjects.

Table 3.1 Distribution of the cohorts in the constrained-discourse corpora used for experimentation

BBVA Corpus				**Pitt Corpus**				
	HC	**AD**	**Total**		**HC**	**MCI**	**AD**	**Total**
Transcriptions	30	39	**69**	**Transcriptions**	242	43	257	**542**
Gender				**Gender**				
Male	14	19	**33**	**Male**	88	27	87	**202**
Female	16	20	**36**	**Female**	154	16	170	**340**
Education (yrs.)				**Education (yrs.)**				
6-11	12	33	**45**	**6-9**	4	2	51	**57**
12-15	13	5	**18**	**10-12**	88	9	112	**209**
16+	5	1	**6**	**13-16**	123	22	76	**221**
				17+	27	10	18	**55**
Age (yrs.)				**Age (yrs.)**				
Under 60	0	1	**1**	**Under 50**	6	1	0	**7**
60-69	11	11	**22**	**50-59**	62	3	21	**86**
70-79	13	19	**32**	**60-69**	101	13	81	**195**
80-89	6	7	**13**	**70-79**	68	22	111	**201**
90+	0	1	**1**	**80+**	5	4	44	**53**

Abbreviations: HC, healthy control; MCI, mild cognitive impairment; AD, Alzheimer's disease.

The BBVA Corpus

The BBVA Corpus consists of manual transcriptions of descriptions of six objects: *dog*, *pine*, *apple*, *chair*, *car* and *pants*. To elicit their speech, all participants were given the instruction "*Tell me everything that you can about* [object]". Recordings were made as a part of the EMSDA evaluation battery for semantic memory (Peraita Adrados, González Labra, Sanchez Bernardos, & Galeote Moreno, 2001), and the cognitive status of participants was assessed by neurologists. All recollections were obtained following written informed consent by the participants (Grasso, Díaz-Mardomingo, & Peraita-Adrados, 2011). We were granted access by the authors to use this this corpus for research purposes.

Participants were divided into healthy controls and AD patients. All AD patients undertook the Mini-mental state examination (MMSE). The mean score of the population over 65 is 27 (±1.7) out of a total of 30 points, with an average decrease of 3-4 points per year after the onset of the Alzheimer's dementia stage (Cockrell & Folstein, 2002). A score of less than 26 points is in the range of mild dementia, while less than 20 points is considered a moderate stage (Perneczky et al., 2006). The average score of the AD cohort of the BBVA corpus was 18/30 points.

The BBVA corpus contained samples of speakers from Spain and Argentina. However, the transcriptions of the Argentinian sample were not readily available for research at the time of our request. For this reason, only the Spanish sample was used in this study.

The Pitt Corpus

The Pitt Corpus from the DementiaBank dataset (Becker et al., 1994) was created by the University of Pittsburg School of Medicine in the '90s. The corpus is accessible upon approval by its authors. Participants in the corpus are elderly HC, MCI and AD patients, as well as participants with memory complaints and vascular dementia. For our study, we only considered the first three groups.

The corpus contains transcriptions of responses to the Cookie Theft picture description task, from the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983), as well as to the word fluency, story recall, and sentence constructions tasks. The last three tasks were undertaken only by the dementia group; hence, for this study, we only used the transcriptions for the Cookie Theft task.

3.3.1.2 Free-discourse corpora

To capture what could be considered "generic" conversational vocabulary and phrases, we used two corpora containing spontaneous conversations with healthy people in a similar age-range and talking in the same language as our clinical population in the constrained-discourse corpora.

The CORLEC corpus

The CORLEC (Marcos Marín, 1992) corpus is a free use dataset of an orthographically transcribed oral corpus containing over one million words. It was created at the Universidad Autónoma de Madrid, in Spain. The recollections took place from January 1991 to February 1992. The corpus was formed with different types of spontaneous speech, from natural conversations to spontaneous academic presentations. Speakers spoke in Spanish and ranged widely in age. There was no reference to the cognitive status of the older speakers, but we assumed them as healthy controls.

From this corpus, we were interested in capturing the most common vocabulary used in "generic" Spanish as spoken language by elderly people. For this reason, we used the transcriptions in the following categories: natural and/or familiar conversations, ludicrous conversations (such as those in TV game contests), humanistic talks, and news interviews. We selected speakers who were over 55. Our sample consisted of 79 speakers (37 female; 42 male) with an average of 62 years of age.

The Carolinas' Conversations Collection (CCC)

The CCC (Pope & Davis, 2011) is a longitudinal collection of conversations with elderly people living in North and South Carolina, USA. It is an ongoing project that began in 2008. It contains over 400 natural conversations with adults over 60, mostly with English speakers, although a Latin American Spanish cohort is currently being added to the collection (Hernandez-Dominguez, Ratte, Pope, & Davis, 2016). Conversations are held with HC and AD participants and revolve around health and other common life issues.

As with the CORLEC corpus, our aim was to obtain a generic vocabulary of spoken English among healthy older speakers. To this end, we used a total of 341 conversations (267 female; 74 male) with healthy controls. The average age of the speakers was 72.

3.3.2 Preprocessing

Markings, such as pauses, noises and interruptions were erased from all four corpora. Then, each cleaned utterance in the corpus was lemmatized and tagged with its part of speech (POS) using FreeLing 4.0 (Padró & Stanilovsky, 2012) for Spanish and English. The lemmatization process consists in transforming all words to their canonical form (*lemmas*). This process allows to reduce variability when considering word inflections (e.g., *swim, swimming, swam*), transforming them all to the same lemma (*swim*).

We developed an automatic tool for natural language processing to handle each corpus. This tool allows the automatic organization of turns of entire dialogs in a corpus of conversations. It preserves the turns of each speaker, their order, and the speaker's demographic information. It also provides methods for automatically extracting vocabulary, n-grams, and some popular lexical richness metrics for individual speakers or for entire cohorts.

From the utterances by the elderly participants of the Pitt and BBVA corpora, the following features were automatically extracted:

3.3.2.1 Information coverage

In a previous work (Hernández-Domínguez, Ratté, Sierra-Martínez, & Roche-Bergua, 2018), we proposed a methodology to evaluate a picture description task based on the information coverage measure proposed by Velazquez et al. (Velázquez-Godínez, 2017). The main disadvantage of this method is that it requires sacrificing a sample of the healthy cohort of the dataset to create a *referent* to estimate the amount of information provided by the participant during a description task. In this study, we propose a new method that measures the coverage of information based on the amount of *specific vocabulary* that the speaker covers during the task.

The words and phrases that are common to general speech, independently of the subject, can be considered as *generic* vocabulary. Prepositions and connectors, and even idiomatic phrases such as *you know*, and *let's see* are part of this type of vocabulary. In order to learn the *specific vocabulary* from a description task, it is first necessary to know what generic or everyday vocabulary is, and then find the words and phrases that are commonly used during the task, but that are not part of the generic vocabulary.

To illustrate this notion, in Figure 3.1, we show a general representation of the subset (**R-S**) that represents the most frequent specific vocabulary used in a constrained-discourse corpus. In the case of this study, the constrained-discourse corpora (**R**) are the Pitt Corpus and the BBVA for English and Spanish, respectively, and the spontaneous free-discourse corpora (**S**) correspond to the English CCC and Spanish CORLEC corpus.

Due to the small size of the datasets, especially in the case of the Spanish corpora, for the extraction of the vocabulary, we only kept the verbs, nouns, adjectives, and prepositions. With this, we capture the main actors, actions and their relationships. For example, in the phrase "The boy steals a cookie from the jar", we would only keep: *boy/N steal/V cookie/N from/P jar/M*. Additionally, two consecutive verbs and two consecutive nouns would be combined into a single verb/noun in order to take composed verbs into consideration (*is/V*

watching/V=be_watch/V; cookie/N jar/N=cookie_jar/N). As such, we reduced the variability of the language in terms of the estimation of the frequency of n-grams.

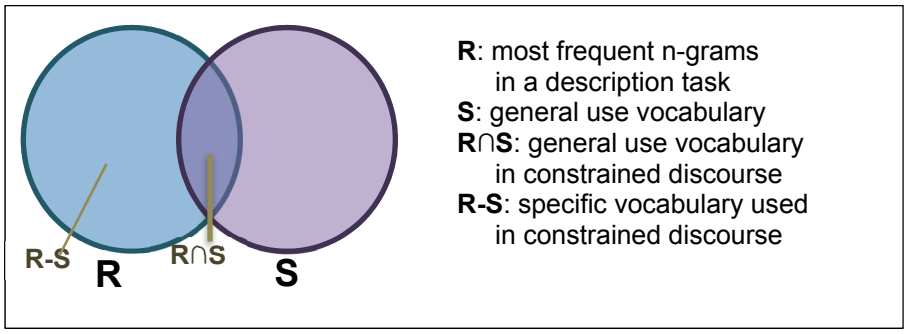

Figure 3.1 Extraction of specific vocabulary commonly used in a constrained-discourse corpus. R is the set of the most frequent n-grams in a constrained-discourse corpus; S is the set of the n-grams with the highest frequency in a spontaneous free-discourse corpus.

An information coverage measure is the estimation of the proportion of words and phrases in the specific vocabulary that the participant provided during the description task. We also divided the specific vocabulary according to parts of speech in order to observe whether there were differences in the coverage of specific parts of speech.

To determine the *pertinence* of the descriptions, we estimated the percentage of n-grams that the participant uttered that were part of the specific vocabulary.

3.3.2.2 Extraction of general vocabulary

From CORLEC and CCC, we obtained the frequency of n-grams (n=1 to 4) uttered by the speakers. An *n-gram* is a contiguous sequence of n tokens. These tokens can be words, lemmas, or POS. For example, in the phrase "*The boy steals cookies"*, the *1-grams* are: *the, boy, steals*, and *cookies*; the *2-grams* are: *the-boy, boy-steals*, and *steals-cookies*; the 3-grams are: *the-boy-steals*, and *boy-steals-cookies;* the only *4-gram* is *the-boy-steals-cookies*.

Since these are the spontaneous free-discourse corpora, we considered the most frequent lemmas in these corpora to be a sample of the generic vocabulary in their respective languages (Figure 3.1, **S**). We selected 1071 (top 2%) and 1185 (top 1%) n-grams from CORLEC and CCC, respectively.

3.3.2.3 Extraction of specific vocabulary

From the BBVA and Pitt corpora, we also extracted the most frequent n-grams (n=1 to 4) uttered by participants (Figure 3.1, **R**). The n-grams that were respectively at the top 5% and 10% of the BBVA and the Pitt corpus were compared against the generic vocabulary extracted from the corpus in their respective languages. The most frequent n-grams of the BBVA or the Pitt Corpora, that were not among the most frequently mentioned in the CORLEC or the CCC, respectively (Figure 3.1, **R-S**), were selected as a sample of specific vocabulary. In total, we extracted 2533 and 2902 specific n-grams from the Pitt and the BBVA corpora, respectively.

3.3.2.4 Lexical richness features

Various lexical richness measures have been used in computer-based analysis for dementia screening (R. Alegria, Gallo, et al., 2013; Bucks et al., 2000; Fraser et al., 2016; Guinn & Habash, 2012a; Hernández-Domínguez, García-Cano, Ratté, & Sierra-Martínez, 2016; Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018; Snowdon et al., 1996). Lexical richness in healthy older speakers has been found to remain stable (Gerstenberg, 2015), and even to increase in healthy individuals as old as 90+ years [42]. However, in the case of individuals with cognitive impairment, these measures tend to present a significant reduction (Fraser et al., 2016; Guinn & Habash, 2012b; Hernández-Domínguez et al., 2018). Although some works still debate the relevance of each measure for identifying signs of cognitive impairment (Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018), most studies agree on their importance. For this work, we extracted the following measures (See Table 3.2): vocabulary size, hapax legomena, hapax dislegomena, Brunet's Index(Brunet, 1978), Honoré's statistics

(Honoré, 1979), Type-token ratio (TTR), Sichel's *S* (Sichel, 1975), Yule's characteristic *K* (Miranda-García & Calle-MartÍn, 2005), and Entropy.

Table 3.2 Lexical richness features

Feature	**Formula/Variable**
Text size	*N* (total number of words)
Vocabulary Size	*V* (number of different lemmas)
Hapax legomena	V_1 (number of lemmas mentioned only once)
Hapax dislegomena	V_2 (number of lemmas mentioned twice)
Brunet's Index	$W = N^{V^{-c}}$ with $c = 0.172$(Tweedie & Baayen, 1998)
Honoré's statistics	$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}}$
Type Token Ratio	$TTR = \frac{V_1}{V}$
Sichel's *S*	$S = \frac{V_2}{V}$
Yule's characteristic *K*	$K = 10^4 \frac{\left[\sum_{i=1}^{N} i^2 V(i,N)\right]}{N^2} - \frac{1}{N}$
Entropy	$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$ * * Where $p(x)$ is the probability of the word x occurring in a text X.

3.3.2.5 Parts of speech and linguistic patterns

Several authors (R. Alegria, Gallo, et al., 2013; Guinn & Habash, 2012b; Jarrold et al., 2010) have found differences in the use of certain parts of speech between healthy and cognitively impaired individuals during spontaneous conversations. From the BBVA and Pitt Corpus, we extracted the rates of parts of speech, and n-grams of parts of speech to observe whether these differences are also observable during description tasks despite the reductions or the syntactic structures in this type of discourse.

3.4 Results

We performed a correlation analysis of the extracted features with the severity of the cognitive impairment. In the case of the BBVA corpus, there were only two cohorts: healthy controls and Alzheimer's patients. For the Pitt corpus, there were three cohorts, namely, healthy controls and patients with MIC or AD. Table 3.3 shows all the features that were significantly correlated ($p < .05$). Positive correlations are highlighted in bold font.

We used a Support Vector Machine (SVM) and Random Forest (RF) to observe the features' adeptness at distinguishing healthy individuals from patients with MCI and AD. For the BBVA corpus, we performed a single classification between healthy controls and AD patients using the entire sample. For the Pitt Corpus, we carried out two classifications: 1) healthy controls and AD patients; and 2) healthy controls and MCI patients.

Since the sizes of the cohorts in the Pitt Corpus vary significantly, we balanced them by selecting the cohort for classification with the fewest samples, and we selected the same number of samples from the other class, making sure that the participants were of similar age, gender and education levels as the smallest cohort.

For each classification experiment, we shuffled the dataset and performed a 10-fold cross-validation process. We divided the dataset into two parts: 90% for training and 10% for testing. The training set was, in turn, divided into three parts to perform cross-validation for parameter tuning. The best model from the cross-validation process was used to classify the testing set (completely unseen data for the learner). We repeated this process 10 times, ensuring that each tenth element of the dataset was used as part of the testing set at least once.

Table 3.3 Features significantly ($p < .05$) correlated with severity of cognitive impairment, controlled for age, education, gender and number of words in the description.

Pitt Corpus		BBVA	
Information coverage			
Information coverage	-0.482**	Information coverage	-0.552**
Verb n-gram coverage	-0.472**	Prep. n-gram coverage	-0.536**
Noun n-gram coverage	-0.447**	Verb n-gram coverage	-0.533**
Prep. n-gram coverage	-0.361**	Noun n-gram coverage	-0.516**
Adj. n-gram coverage	-0.148*	Adj. n-gram coverage	-0.502**
Pertinence			
Pertinence	-0.366**	Pertinence	-0.477**
Noun pertinence	**0.291****		
Verb pertinence	-0.257**		
Lexical richness			
Entropy	-0.285**		
Vocabulary size	-0.255**		
Hapax dislegomena	-0.244**		
Hapax legomena	-0.242**		
Yule's characteristic *K*	**0.202****		
Total different n-grams	-0.191**		
Ratio of original 2-grams	-0.149*		
Ratio of original 1-grams	-0.137*		
Type-token ratio (TTR)	-0.117*		
Sichel's *S*	-0.113*		
Honoré's statistics	-0.099		
POS rate			
Pronoun rate	**0.281****	Pronoun rate	**0.300**
Adverb rate	**0.280****	Interjection rate	**0.285**
Determiner rate	-0.211**		
Verb rate	-0.147*		
Noun rate	-0.141*		
Conjunction Rate	**0.121***		
Linguistic Patterns			
Noun + verb	-0.204**	Noun + adj.	-0.326*
Noun + verb + adj.	-0.122*	Verb + verb + verb	**0.258**
Adj. + noun	**0.121***	Verb + noun + verb + verb	**0.255**
Noun + verb + noun + prep	-0.105		
Adj. + adj.	**0.097**		
Adj. + prep	**0.088**		

To train the algorithms, we selected the features that were significantly correlated with diagnosis for each corpus (see Table 3.3). For reducing the number of features, we performed a feature reduction with backward elimination. For this, we started with the entire set of features shown in Table 3.3 for each corpus. From the set, we tested removing one feature at a time, and trained the algorithm without that feature. We repeated the process by removing a different feature each time. At the end of the iteration, we eliminated the feature that was the least significant according to the classification and repeated the process of eliminating a feature in the next iteration.

In Table 3.4, we report the average performance of the 10-fold cross-validation process with the features that produced the highest values for the area under the curve for each experiment.

Table 3.4 Performance metrics of the learners with both corpora for classification of healthy controls and individuals with MCI and/or AD.

Cohorts (size†)	Learner	Accuracy	Sensitivity	Precision	F_1-score	AUC
BBVA Corpus						
HC & AD (30 & 39)	RF	0.971	0.974	0.974	0.974	0.970
	SVM	0.985	**0.974**	1	**0.98**	**0.987**
Pitt Corpus						
HC & AD (242 & 242)	RF	0.794	0.788	0.798	0.793	0.794
	SVM	0.834	**0.821**	0.842	**0.831**	**0.834**
HC & MCI (42 & 42)	RF	0.726	0.714	0.731	0.722	0.726
	SVM	0.797	**0.809**	0.790	**0.800**	**0.797**

† In number of participants

Abbreviations: HC, healthy control; AD, Alzheimer's disease; MCI, mild cognitive impairment; RF, random forests; SVM, support vector machine; AUC, area under the curve.

3.5 Discussion

The features with the highest correlations with cognitive impairment in both corpora (see Table 3.3) were our proposed measurements of information coverage based on the use of specific vocabulary. Cognitively impaired individuals produced less informative descriptions overall.

This phenomenon was observed for all the types of vocabulary-specific n-grams extracted, which estimated the coverage of actors (noun n-grams), actions (verb n-grams), characteristics (adjective n-grams) and their relationships (preposition n-grams). This finding is in accordance with previous literature on picture description tasks (Hakkani-Tür et al., 2010; Hernández-Domínguez et al., 2018; Kavé et al., 2018; Yancheva & Rudzicz, 2016), and appears to also be applicable to object descriptions.

The overall measure of pertinence also correlated significantly with the severity of the cognitive impairment in both corpora. In the case of the BBVA corpus, no specific pertinence's part of speech was correlated, and this could be caused by the less constrained nature of the discourse, which makes it more difficult to detect when participants drift from the topic at hand. For the Pitt corpus, cognitively impaired individuals used fewer specific verbs in their descriptions, but at the same time, used more specific nouns. This could be due to a higher use of common verbs and less informative phrases, such as "*there is a boy*", which includes the specific noun *boy*, but does not include any task-specific verb.

The lexical richness measures were significantly correlated with the severity of cognitive impairment when analyzing the Pitt corpus. This result is in agreement with several previous works (Fraser et al., 2016; Hernández-Domínguez et al., 2018; Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018). However, in the case of the BBVA corpus, we did not find this type of correlation. This phenomenon could be explained by the high variability of the number of words during the descriptions across participants, since most of these measurements are highly sensitive to text size.

The cognitively impaired cohorts of both datasets produced a significantly higher rate of pronouns. In spontaneous speech, this effect has been explained by suggesting that the use of pronouns is a strategy used to compensate for difficulties in recalling proper names (Singh & Bookless, 1997). AD patients in the BBVA corpus also produced a higher number of interjections, which may indicate a reduced fluency. In line with previous studies (R. Alegria, Gallo, et al., 2013; Fraser et al., 2016; Guinn & Habash, 2012b; Jarrold et al., 2010), other rates

of parts of speech, such as determiner, noun, verb and conjunction rates, were significantly lower in the cognitively impaired cohort of the Pitt corpus.

When analyzing specific linguistic patterns in the BBVA corpus, AD patients produced a significant lower number of *noun + adjective* patterns. This finding corresponds with a previous manual evaluation of this corpus (Grasso et al., 2011), in which the authors found a significantly lower number of *evaluative* attributes mentioned by AD participants. These attributes are mainly related to the physical evaluation of the objects described, such as shape, color, texture, etc., as well as to abstract concepts associated with the object, such as kindness, sympathy, appeal, etc. (Grasso et al., 2011). In turn, AD patients produced a higher number of *verb + verb + verb* and *verb + noun + verb + verb* patterns. When analyzing these patterns, we observed that they are present when the participant is stuck in their speech and is trying to find additional characteristics to mention. The most repeated form of this pattern was "*tiene tiene tiene*" ("has has has").

For the Pitt corpus, the cognitively impaired cohort produced a significantly lower number of phrases with the *noun + verb* form (e.g., "mother dries", "boy reaches", "girl laughs", *noun + verb + adjective* (e.g., "window is open", "cookie-jar is full", "stool is crooked"), and *noun + verb+ noun + preposition* (e.g., "boy gets cookie of", "mother dries dishes while", "girl has finger on"). This indicates that a smaller number of actions, details and complex syntactic contractions are formulated by this cohort. In addition, cognitively impaired individuals produced a higher number of phrases containing adjectives, a finding consistent with previous studies (Forbes, Venneri, & Shanks, 2002; Jarrold et al., 2014; Vincze, 2016).

When performing the classification for both corpora, SVM classifiers with linear kernels outperformed RF in all experiments (see Table 3.4). For the BBVA corpus, we obtained a very high area under the curve (AUC), and an equally very high sensitivity for classifying AD patients from controls (0.98 and 0.97, respectively) as compared to those of the Pitt corpus (0.83 and 0.82). To verify that there was no overfitting in the classification of the BBVA corpus, we ran the SVM implementation of the sci-kit learn Python library (Pedregosa et al.,

2011) with feature selection, but using the default values without parameter tuning. The AUC of this experiment was 0.95. However, there are several aspects of the BBVA that may have favored the classification of this corpus.

First and foremost, the number of words in the samples of healthy controls in the BBVA corpus was significantly higher than those of AD patients. On average, healthy controls produced 600 words per task, while AD patients only produced 205. This significant difference may be attributable to the length and complexity of the task, which requires that participants describe, with as much detail as possible, six objects without any grounded reference. Compared to the Pitt corpus, which requires a single description task, assisted by a visual stimulus, the BBVA's is arguably a very onerous task for cognitively impaired individuals. In fact, over 10% of AD patients in the BBVA corpus were not able to provide descriptions for all six objects.

Another aspect that may facilitate the classification in the BBVA corpus is the significant difference in the number of years of education between healthy controls and AD participants: 85% of AD patients had only primary education, 13% finished their high school education, and just 2% had a college degree. In contrast, 16% of healthy controls had a college degree, 43% finished high school, and 41% had only primary education.

A further issue with the BBVA corpus resides in the apparent ambiguity of the question asked to participants. When carefully observing their responses, it is apparent that participants tend to have different interpretations of the instruction "*Tell me everything that you can about* [object]". Some participants just list parts and functions, while others give full recounts of their previous experiences with said objects. AD patients seem to have received or interpreted the instruction in vaguer terms, while healthy controls have more uniform answers. Part of this difference could be due to the fact that AD patients were questioned in the context of a mental examination (EMSDA (Peraita Adrados et al., 2001)) in a hospital setting, while healthy controls were examined in a natural context, and seem to have more interactions and clarifications during the performance of the task.

Compared to previous works on the BBVA corpus, a classification of AD patients and healthy controls using Bayesian networks (J. M. Guerrero, Martínez-Tomás, Rincón, & Peraita, 2015) reported an AUC of 0.9621. This work used manually-extracted information component units as features, as well as the socio-demographic information of participants as *a priori* deterministic inputs. In contrast to this work, our features were fully automatically-extracted. Furthermore, we did not rely on socio-demographic information as a feature for our classification, since the sample selected for the corpus is not representative of the general population, and such reliance could cause the learner to overfit this particular sample. Our results also outperformed those of our previous study (Hernández-Domínguez et al., 2016) on this same corpus, in which we reported an F_1-score of .880 using an SVM trained only with automatically-extracted linguistic features (part-of-speech rates and lexical richness). The inclusion of our information coverage and pertinence features based on the use of specific vocabulary improved these results. The selected features used by our best SVM model in the BBVA corpus were: information coverage; preposition, verb and noun n-gram coverage; pertinence; pronoun and interjection rates; and *noun + adjective* and *verb + verb + verb* linguistic patterns.

In the case of the classification in the Pitt Corpus, several previous studies have been found. Orimaye et al. presented a first work (Sylvester Olubolu Orimaye et al., 2014) classifying cognitively impaired (AD, MCI and other dementia) patients versus healthy controls using syntactic and lexical features using SVM. In the work, they reported an F_1-score of 0.73. In comparison, we obtained an F_1-score of 0.80 when classifying AD and MCI patients as a single group, from healthy controls. In a more recent work (Sylvester O. Orimaye et al., 2017), Orimaye et al. used a variant (Platt & Others, 1998) of SVM trained with 1000 features, which included syntactic and lexical features, as well as the use of bigrams and trigrams as features to discriminate between AD patients and healthy controls. The authors reported an AUC of 0.93, although they performed the parameter tuning using samples of descriptions of the same participants they were classifying, but taken on a different date, which could arguably lead to a form of overfitting. This is a plausible scenario, especially considering that the authors are using a set of features ten times larger than their number of examples.

Another previous work (Fraser et al., 2016) that classified AD patients and healthy controls on the Pitt corpus using logistic regression reported an average accuracy of .81 using 35 features, which included part-of-speech ratios, acoustic features (Mel-frequency cepstral coefficients), and a list of manually extracted information content units proposed by Croisile et al. (Croisile et al., 1996). Compared to this work, we obtained an accuracy of 0.83 for this task, and the evaluation of all features was fully automatic.

Yancheva et al. (Yancheva & Rudzicz, 2016) developed a method for automatic extraction of information content units based on vector-space topic models. In their work, the authors trained a Random Forest learner with these vectors, idea density measures, and the same acoustic and lexical and syntactic features used in (Fraser et al., 2016). With this combination of features, they reported an F_1-score of 0.80 for differentiating AD patients from healthy controls.

There is one previous work (Santos et al., 2017) on the classification of MCI patients versus healthy controls in the Pitt corpus. For this study, the authors used word embeddings to enrich complex networks, and bags of words and lexical diversity measures to train an SVM. Their reported accuracy for this task was of 0.65. Our classification of MCI patients had a 0.79 accuracy.

The features selected for the classification of AD patients and healthy controls were: verb and noun n-gram coverage; noun pertinence; entropy; pronoun, determiner, verb and conjunction rates; vocabulary size; number of different n-grams; TTR; and *noun* + *verb*, *noun* + *verb* + *noun* + *preposition*, *adjective* + *noun* and *adjective* + *preposition* linguistic patterns.

3.6 Conclusions

In this work, we presented a method for the automated classification of AD and MCI patients versus healthy controls using transcriptions of description tasks. Our work was tested in two different description task settings: a standardized picture description task in English and a task

involving the description of six common objects in Spanish. The first setting had the advantage of presenting patients with a visual stimulus that constrained their vocabulary to a single scene. The second setting allowed participants a higher variety of syntactic structures since their descriptions were grounded on their mental images of, and personal experiences with, the objects being described.

Unlike most of the previous literature on the evaluation of description tasks, our proposed method for assessing information coverage does not rely on the manual selection of information content units. Instead, our method automatically extracts a generic vocabulary from a corpus of spontaneous speech of older speakers, and using this generic vocabulary, it captures the specific vocabulary used for a description task. We then evaluate the use of specific vocabulary in terms of coverage (how much of the specific vocabulary is used by the speaker) and pertinence (how much of the speaker's vocabulary corresponds to task-specific vocabulary).

We trained SVM and RF algorithms to differentiate between AD and MCI patients and healthy controls. We used our proposed coverage and pertinence measurements, coupled with lexical richness features and parts of speech and use of specific linguistic patterns. In all our experiments, SVM with linear kernels outmatched RF classifiers.

For the Spanish BBVA corpus, our results outperformed previous computer-based feature extraction methods and compared favorably with the state of the art that relies on manually extracted information content units and socio-demographic information for their classification. For the English Pitt corpus, we significantly outperformed previous classifications of MCI patients and healthy controls, which represent a challenging and crucial step in the early detection of AD signs (Alzheimer's Association, 2018b). We also achieved a modest improvement in the classification of AD patients versus healthy controls when compared to previous computer-based studies.

Since the description of common objects favors a wider variety of syntactic structures, we hypothesized that the BBVA corpus could be appropriate for studying linguistic changes caused by Alzheimer's disease in greater depth. However, the small size of the sample, the heterogeneity of the samples, the complexity of the task for the AD cohort, and the significant differences in education levels between AD patients and healthy controls make it difficult to study this phenomenon. As future work, it would be interesting to observe linguistic changes in spontaneous conversations, especially in the context of longitudinal analyses for personalized medicine applications.

CHAPTER 4

AGING WITH AND WITHOUT COGNITIVE DISEASES: CHARACTERIZING 10 YEARS OF LANGUAGE DIFFERENCES IN OLDER FRENCH SPEAKERS

Laura Hernández-Domínguez[1], Sylvie Ratté[1], Annette Gerstenberg[2] and Gerardo Sierra[3]

[1] Software and IT Engineering Department,
École de technologie supérieure, Montreal, Canada

[2] Potsdam University, Potsdam, Germany

[3] Universidad Nacional Autónoma de México (UNAM), Mexico City, Mexico

Email: laura.hzdz@gmail.com, sylvie.ratte@etsmtl.ca

This article was submitted to *Computers Speech and Language*, Elsevier, on October 29, 2018.

4.1 Abstract

Background: Language functions have been identified as early markers of several forms of cognitive impairment. Multiple authors have found that it is possible to asses some language functions through the analysis of spontaneous conversations. However, most studies to date have examined short-term language changes in English speakers. In this work, we present a study on semi-structured interviews conducted over a ten-year span with 4 older French speakers. The speakers started out as apparently cognitively healthy individuals, and years later, were diagnosed with some form of cognitive impairment. The language changes in these individuals are compared against those of cognitively healthy matched pairs.

Method: We automatically estimated different measures that have been associated with cognitive impairment, like lexical richness, part-of-speech ratios, hesitations, and unclear words. We also propose a metric based on vocabulary distribution to determine the use of generic and specific vocabulary.

Results: Using a two-step principal component analysis, we obtained two components that were able to clearly discriminate the interviews of the individuals that had a diagnosis of cognitive impairment at the time of the interview. Furthermore, these components also separated participants that were healthy at the time of the interview, but that would end up developing some form of cognitive impairment up to ten years later.

Conclusions: Our proposed method describes the behavior of some language functions throughout the development of cognitive impairment in French speakers. This non-invasive and inexpensive method could potentially be used for follow-up of patients and for early detection of signs of cognitive impairment.

Keywords: Cognitive impairment; language functions; spontaneous speech; French; natural language processing; vocabulary distribution; lexical richness; early detection.

4.2 Introduction

New guidelines for the diagnosis of Alzheimer's disease (AD) put a particular emphasis on the understanding of the disease as a continuum (Alzheimer's Association, 2018a). It starts with the preclinical stage, where the first changes in the brain begin, years before any symptom is evident; this is followed by mild cognitive impairment (MCI) due to AD, where the symptoms are mild and do not affect everyday living; finally, there is the *Alzheimer's dementia* phase, where symptoms such as memory, thinking and behavioral changes affect the individual's daily life, reducing their cognitive and, ultimately, their physical functions.

Currently, the disease is usually diagnosed at the MCI, or even at the dementia phase, when the symptoms are already affecting the patient's life. The 2018 Alzheimer's association's facts and figures report (Alzheimer's Association, 2018a) estimates that between 15 and 20 percent of the older population (65 years and older) suffers from mild cognitive impairment (MCI). Among this group, individuals that also present memory problems are at a higher risk of developing Alzheimer's disease (AD). Around 32% of MCI patients will develop Alzheimer's

dementia within 5 years of their initial diagnosis (Ward et al., 2013). However, MCI does not represent an automatic sentence for dementia, and many patients remain stable or even return to a normal condition.

Currently, the disease is usually diagnosed at the MCI, or even at the dementia phase, when the symptoms are already affecting the patient's life. The 2018 Alzheimer's association's facts and figures report (Alzheimer's Association, 2018a) estimates that between 15 and 20 percent of the older population (65 years and older) suffers from mild cognitive impairment (MCI). Among this group, individuals that also present memory problems are at a higher risk of developing Alzheimer's disease (AD). Around 32% of MCI patients will develop Alzheimer's dementia within 5 years of their initial diagnosis (Ward et al., 2013). However, MCI does not represent an automatic sentence for dementia, and many patients remain stable or even return to a normal condition.

Although research in Alzheimer's detection is focused and rapidly advancing on biomarkers for early diagnosis, positron emission tomography (PET), magnetic resonance imaging (MRI) and cerebrospinal fluid (CSF) testing are costly, and could be uncomfortable and invasive for older patients. It is therefore preferable to screen patients such that only those that are suspected to be at a high risk of developing MCI undergo such testing.

Language function assessment represents one of the earliest (Szatloczki et al., 2015), non-invasive and inexpensive markers of AD. There have been multiple studies (Ahmed, Haigh, et al., 2013; R. Alegria, Gallo, et al., 2013; Asgari et al., 2017; Asgari, Kaye, Mattek, & Dodge, 2015; Bucks et al., 2000; Fraser, 2016; Guinn & Habash, 2012b; Hakkani-Tür et al., 2010; Hernández-Domínguez et al., 2016; Jarrold et al., 2010, 2014; Kavé & Goral, 2018; Kemper et al., 1993; Khodabakhsh et al., 2014; Lehr, Prud, Shafran, & Roark, 2012; Schröder et al., 2010; Snowdon et al., 1996; Thomas, Keselj, Cercone, Rockwood, & Asp, 2005; Wankerl, Nöth, & Evert, 2016; Wendelstein, Felder, & Schröder, 2011; Zhou et al., 2016) on the linguistic changes that occur as a result of MCI and Alzheimer's disease (AD). Most studies have been made with English speakers, although some work has been done with Spanish

(Hernández-Domínguez et al., 2016), Portuguese (R. Alegria, Gallo, et al., 2013) and Turkish (Khodabakhsh et al., 2014) speakers. Most of the latter were conducted in cognitive test settings, with a few, nevertheless in spontaneous conversation contexts.

In this work, we present a computer-based analysis and characterization of the evolution of the linguistic features of four senior French speakers that developed cognitive impairment (CI), and four age-, gender-, education-, profession- and multilingual-matched cognitively unimpaired controls. The four patients in the CI group started interviews as apparently healthy individuals, with no symptoms of memory or thinking problems, and over the course of ten years, developed CI.

4.3 Materials and Methods

4.3.1 LangAge Corpus

For our study, we used the LangAge Corpus (Gerstenberg, 2011), which is comprised of biographical interviews of older French speakers. The first set of interviews in the corpus was conducted in 2005, and the same participants were interviewed again twice, once in 2012, and then in 2015. This corpus was originally intended to characterize the evolution of language in a healthy aged population. However, over the years, family members and caregivers reported that some participants developed some form of cognitive impairment, with one case diagnosed as Alzheimer's disease. Given the original intent of the study, no additional medical information was available for the participants in the corpus.

The LangAge corpus currently has more than 150 interviews (Gerstenberg, n.d.). For this work, we selected all four participants that were reported as having cognitive issues by a trusted family member, and four healthy controls (HC) that were their closest match by age, gender, multilingualism, education level and type of profession. The distribution of our sample is shown in Table 4.1.

Table 4.1 Sample of cognitive impaired subjects and their closest healthy control matches from the LangAge corpus

ID	Age	Gender	Educ.*	Profession†	Bilingual	Cognitive status	Diagnosis
13	**83**	**Female**	**1**	**2**	**No**	**Alzheimer's disease**	**2013**
37	79	Female	1	1	No	Matched healthy control	-
25	**84**	**Female**	**2**	**2**	**No**	**Memory issues, confusion**	**2012**
11	86	Female	2	2	No	Matched healthy control	-
27	**76**	**Male**	**3**	**4**	**No**	**Progressive cognitive disease**	**2015**
18	76	Male	3	4	No	Matched healthy control	-
48	**70**	**Male**	**2**	**3**	**Yes**	**Cognitive impairment**	**2015**
47	75	Male	3	4	Yes	Matched healthy control	-

*Education: 1-CEP (*Certificat d'études primaires*; from 11 to 13 years old); 2-BEP (*Brevet d'études professionnelles*; from 15 to 16 years old); 3-BAC (*Baccalauréat*; from 17 to 18 years old)

†Profession: 1-worker; 2-qualified employee; 3-highly qualified employee; 4-high academic/management

4.3.2 Pre-processing

The interviews in the LangAge corpus were transcribed and time-aligned following a set of rules (Gerstenberg, Annette Hekkel & Kairet, 2018) to reduce variation in the forms of marking pauses, interjections, incomplete words and other phenomena. We extracted all the utterances from the interviewees and removed repeated n-grams (e.g., *sur la ~~sur la~~ ~~sur la~~*), incomplete words and interjections. This was done to minimize the error rate that these markings tend to produce in automatic part of speech (POS) taggers. We used FreeLing 4.0 (Padró & Stanilovsky, 2012) to lemmatize (obtain the canonical form of each word) and tag the participants' utterances with their POS.

After the cleaning and tagging process, we standardized the size of the samples for all participants. This was done to reduce variations in lexical characteristics that arise from comparing texts of different sizes. For this process, we found the participant's transcription with the smallest number of words (N=1,547), and then we cut all other participants'

transcriptions to this size while respecting the integrity of utterances. This was done by incorporating the first utterances of the participants into the sample until each sample reached 1,547 words. We then cut the last word of the utterance. The average number of words per sample per participant was 1,552.

4.3.3 Extraction of characteristics

4.3.3.1 Lexical richness measures

Several lexical richness measures have been proposed for computer-based analysis for dementia screening (Fraser et al., 2016; Guinn & Habash, 2012b; Hernández-Domínguez et al., 2018; Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018) and in normal aging, where lexical richness measures have been shown to remain stable (Gerstenberg, 2015), with a continuous increase in vocabulary in healthy individuals as old as 90+ years (Goral et al., 2007). Although there have been some mixed results (Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018) with regards to their relevance for the task, most studies agree on their importance. In this study, we evaluate these characteristics in older French speakers and analyze their behavior in spontaneous conversations through the years. The lexical richness measures (see Table 4.2) that we selected were **Brunet's Index** (Brunet, 1978) (with the most commonly used value (Tweedie & Baayen, 1998) of $c = 0.172$), **Honoré's statistics** (Honoré, 1979), **Yule's characteristic *K*** (Miranda-García & Calle-MartÍn, 2005), **Sichel's *S*** (Sichel, 1975), **Type-token ratio** (TTR), and **Entropy**.

Table 4.2 Measures for lexical richness evaluated

Measure	**Equation**
Brunet's index	$W = N^{V^{-c}}$ *with c=0.172*
Honoré's statistics	$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}}$
Type Token Ratio	$TTR = \frac{V_1}{V}$
Sichel's *S*	$S = \frac{V_2}{V}$
Yule's characteristic *K*	$K = 10^4 \frac{[\sum_{i=1}^{N} i^2 V(i,N)]}{N^2} - \frac{1}{N}$
Entropy	$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$

NOTE. *N* refers to the size of the text (number of words); *V* is the size of the vocabulary (number of different lemmas); V_1 corresponds to the number of lemmas that occur only once; V_2 is the number of lemmas that occur twice, and *p(x)* is the estimated probability of a lemma x occurring in a text X.

4.3.3.2 Characteristics based on vocabulary distribution

For each interview, we estimated the participants' **vocabulary size** *V* and the number of **hapax legomena** (lemmas that are mentioned only once) and **hapax dislegomena** (lemmas that are mentioned twice). We also estimated the **percentage of distinct n-grams** (n = 1-4) over the total number of n-grams.

Additionally, we propose the use of TFIDF to extract lexical richness measures that are based on the distribution of n-gram frequencies in the vocabulary. These measures incorporate the use of term frequency (TF) and the inverse document frequency (IDF) or term specificity (Sparck Jones, 1972). The TFIDF estimation was performed over n-grams (*n* = 1-4) following equation 4.1.

$$TFIDF(n\_gram, d_i) = TF(n\_gram, d_i) \times IDF(n\_gram, D) \quad (4.1)$$

where TFIDF is the value of an n-gram in document d_i in the sample of documents D. TF corresponds to the number of times said n-gram occurred in d_i, and IDF is defined as the inverse document frequency of the n-gram in the sample of N documents D:

$$IDF(n\_gram, D) = \log\left(\frac{N}{|\{d \in D: n\_gram \in d\}|}\right) \quad (4.2)$$

The TFIDF statistic is a common measure to indicate the importance of a word in a document with respect to the rest of the documents in a corpus. The intuition behind this measure is that, if a word is very common in a specific document, but its use is rare in the rest of the documents, the TFIDF value of the word in that document will be high, since it means that is highly specific to the document. On the contrary, if a word is highly used in a document, but it is also prevalent in the rest of the documents, it is not considered to be a specific word, but rather, a generic one in that corpus.

The typical behavior of the TFIDF curve in a document follows a logarithmic curve pattern as shown in Figure 4.1, where the words in the horizontal axis are ranked from the highest TFIDF value to the lowest. The top section of the curve (left) corresponds to the TFIDF values of the most specific words in the document, while the bottom corresponds to the most generic words (right).

In a previous study (**Chapter 3**), we proposed the use of similar measures to evaluate the use of specific and generic vocabularies in the older population in a restricted discourse context, where all participants were given the task of describing the same image. For this work, we aim to adapt these measures to determine the differences between the use of what could be considered an idiosyncratic vocabulary and a generic one.

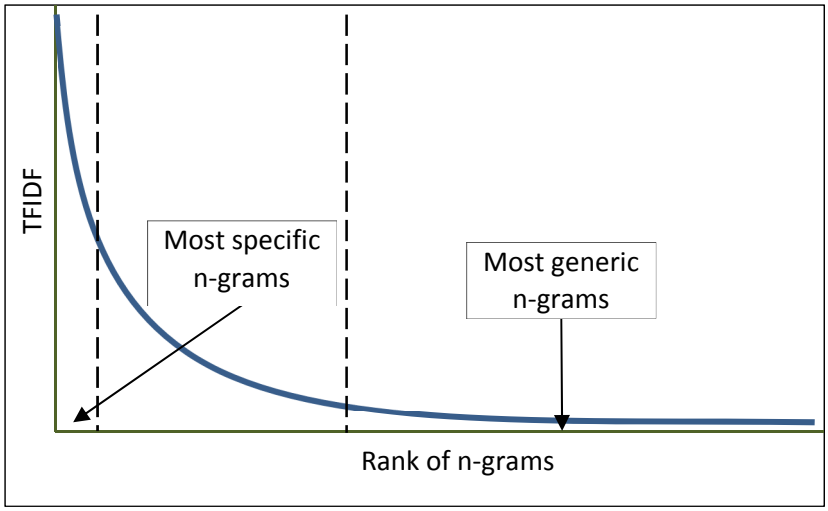

Figure 4.1 Typical TFID curve

We estimated the TFIDF statistics to evaluate two different phenomena: 1) the **general TFIDF** of the n-grams used by a participant during each interview, calculated against all the interviews from all other participants; and 2) the **individual TFIDF**, which was estimated by contrasting a participant's interview against all other interviews of that same participant.

The upper section of the individual TFIDF will correspond to those n-grams that the participants used the most during an interview, but that they did not use in their other two interviews. It measures the variation between the specific vocabulary that the participant used for each interview. This same section on the general TFIDF curve corresponds to those words that that the participants used the most during an interview and that were not used by other interviewees. It is a specialized or thematic vocabulary with respect to the other speakers.

In the case of the general TFIDF curve, the lower part contains the n-grams that the participant had in common in this interview with the rest of the participants (generic vocabulary and phrases). For the individual TFIDF, the lower section of the curve contains the n-grams that the participants mentioned the most in all their interviews. The n-grams that are present in the lower section of the individual TFIDF curve, but not in the lower section of the general TFIDF curve, correspond to the ***idiosyncratic* n-grams**, that is, those particular n-grams that the participant used in all his/her interviews, but that were not common among other participants.

To determine the threshold for cutting the upper and lower sections of the TFIDF curves, we approximated the area under the curve as the summation of the TFIDF values for all n-grams in each interview. The top sections of the TFIDF curves were cut at the n-gram in the highest-ranking n-grams, where the summation reached 10% of the area under the curve. The bottom sections were cut at 5% of the summation of the lowest ranking n-grams in the curve. These thresholds were selected after observation of the n-grams from both sections of the curves. All n-grams with a TFIDF value of zero were always included, since they do not contribute to the summation.

4.3.3.3 Sentiment polarity and subjectivity

Some works have found evidence that patients with AD are able to experience emotions with an intensity similar to that of cognitively intact individuals (Henry, Rendell, Scicluna, Jackson, & Phillips, 2009). The assessment of emotional experience has generally been based on self-reported emotion experience, and the expression of emotions has been corroborated by analyses of facial expressions and using electromyographic recordings of muscle activity (Burton & Kaszniak, 2006). These works have found that the expression of positive emotions is particularly affected by AD and some other forms of cognitive impairment.

To determine whether there was a difference in the verbal expression of emotions in our sample, we used the Textblob-fr 0.2.0 (Loria, 2013) python library, the French language support for the TextBlob 0.15.1 (Loria et al., 2018) sentiment analysis module, to determine the **sentiment polarity** and **subjectivity** scores (between -1 and 1) for each utterance. To determine the polarity and subjectivity of words, Textblob-fr uses a dictionary that assigns a sentiment and subjectivity score to each word. The reported utterance score corresponds to the average score of all its words.

4.3.3.4 Ratio of use of Part-of-speech (POS) patterns

Following previous works (R. Alegria, Gallo, et al., 2013; Guinn & Habash, 2012b; Jarrold et al., 2010), we estimated the **ratio of use of parts-of-speech (POS)**. We studied the proportion of linguistic patterns formed by **n-grams of POS** (where n=1-4). Previous studies (Beber, da Cruz, & Chaves, 2015) have shown marked difficulties in verb production and processing in AD patients. For that reason, the part-of-speech n-grams ratios were subdivided into two groups: **action patterns** (patterns including a verb) and **passive patterns** (the rest).

4.3.3.5 Utterance-level characteristics

We estimated several characteristics for each utterance of a participant during an interview. For each participant, we calculated the characteristic's average, kurtosis and skewness across each interview:

- Number of interjections**,** which were counted separately depending on their function in "**back channel interjections**", such as "*m-hm*", "*hein*", "*hm*", and "*ah*"; and "**hesitation interjections**", "*euh*".
- **Number of unclear words**.
- **Number of effective words**, which were comprised of all tokens that did not correspond to an interjection or that were deemed unclear or incomplete by the transcriber.
- **Number of syllables**. To approximate the number of syllables in French, we used the Epitran 0.56 (Mortensen, Dalmia, & Littell, 2018) python library. With this library, we did a transliteration of the orthographic text of the transcriptions into the IPA (International Phonetic Alphabet). Since the IPA considers a vowel as a syllable center, we counted the number of phonetic vowels ("a", "ɑ", "ɛ", "ə", "e", "œ", "ø", "i", "o",

"ɔ", "u", and "y") as the number of syllables, except for those words that ended with the phoneme "ə", which is usually silent in French. For example, the phrase « *L'homme et la femme sont là* » ("*the man and the woman are there*"), is phonetically transliterated as: "*l'ɔmə ɛt la fɛmə sɔnt la*". The syllables were separated as: *l'ɔmə* **-** *ɛt* **-** *la* **-** *fɛmə* **-** *sɔnt* **-** *la* (6 syllables).

- **Duration** (in milliseconds). For this, we used the time markings in the aligned transcriptions.
- Speech rate characteristics: **number of syllables per word**, **words per second**, and **syllables per second**.

4.4 Results

We performed a correlation analysis of the extracted characteristics with the severity of the cognitive impairment, controlling for age, gender, education, profession and bilingualism. Table 4.3 shows the characteristics that had a significant correlation[2] ($p < .05$).

We used principal component analysis (PCA) to create composite variables from the groups of characteristics shown in Table 4.3. The suitability of PCA was assessed prior to the analysis. Inspections of the correlation matrices for all groups of characteristics showed that all variables had at least one correlation coefficient greater than 0.3. Bartlett's test of sphericity was statistically significant ($p < .0005$) for all groups of characteristics, indicating that the data was likely factorizable.

[2] Only characteristics that had a significant correlation with the severity of cognitive impairment appear in the table. Other correlations, such as the sentiment polarity and idiosyncratic n-grams, are discussed in section 4.

Table 4.3 Correlation with severity of cognitive impairment. Bold font indicates negative correlation

Characteristic	*r*	*p*	Characteristic	*r*	*p*
Lexical richness measures			**Action POS n-gram ratios***		
Brunet's Index	0.737	< .001	pron + verb	0.677	< .001
Honoré's Statistics	**-0.703**	< .001	pron + verb + advb	0.685	< .001
TTR	**-0.721**	< .001	verb + advb	0.643	< .001
Sichel's S	**-0.651**	< .001	verb + conj + pron+ verb	0.519	< .05
Entropy	**-0.637**	< .001	pron+ verb + verb	**-0.471**	< .05
Vocabulary distribution measures			verb + verb	**-0.462**	< .05
Percentage of different 2-grams	**-0.822**	< .001	**Passive POS n-gram ratios***		
Percentage of different 3-grams	**-0.812**	< .001	prep	**-0.625**	< .001
Ratio of upper individual TF-IDF n-grams	**-0.731**	< .001	noun + prep	**-0.652**	< .001
Hapax legomena	**-0.717**	< .001	noun + prep + det	**-0.596**	< .001
Vocabulary size	**-0.690**	< .001	det + noun + prep	**-0.580**	< .001
Percentage of different 1-grams	**-0.694**	< .001	pron	0.577	< .01
Percentage of different 4-grams	**-0.695**	< .001	prep + det	**-0.570**	< .05
Hapax dislegomena	**-0.672**	< .01	det + noun + prep + det	**-0.549**	< .05
Number of different n-grams	**-0.631**	< .01	noun + prep + noun	**-0.509**	< .05
Ratio of upper general TF-IDF n-grams	**-0.570**	< .05	prep + det + noun	**-0.502**	< .05
Ratio of lower general TF-IDF n-grams	0.458	< .05	noun + prep + det + noun	**-0.492**	< .05
Subjectivity			prep + det + noun + prep	**-0.491**	< .05
Skewness of utterances' subjectivity	0.507	< .01	**Hesitation interjections**		
Kurtosis of utterances' subjectivity	0.495	< .01	Skewness of hesitation interjections	0.572	< .05
Average utterances' subjectivity	0.480	< .05	Kurtosis of hesitation interjections	0.507	< .05
Unclear words					
Average unclear words per utterance	0.617	< .05			

*Abbreviations: *pron*, pronoun; *advb*, adverb; *conj*, conjunction; *prep*, preposition; *det*, determiner.

To select the number of composite variables from each group, we chose the components that were able to account for at least 75% of the cumulative variance (cv). We extracted one composite variable each from the *lexical richness measures* (cv=84.5%), the *vocabulary distribution measures* (cv=75.4%), the *passive POS n-gram ratios* (cv=80.8%), and the *hesitation interjections* (cv=98.3%) groups. Two composite variables (see Table 4.4) were extracted from the *action POS n-gram ratios* group (cv=80.2%). The first variable seems to be related to the use of simple verbs, while the second one appears to be related to the use of two consecutive verbs, such as in the case of using future and past tenses.

Since there is only one significant variable in the *unclear words* group, there was no point in performing PCA on it. Finally, the Kaiser-Meyer-Olkin (KMO) for the *subjectivity* group was 0.465, which made it 'unacceptable' according to Kaiser's classification of measure values (Kaiser, 1974). For that reason, we did not use PCA for this group. Instead, we choose the characteristics with the highest correlation with cognitive impairment (s*kewness of utterances' subjectivity*) to represent this group.

Table 4.4 Table of the rotated loadings matrix for the action POS n-gram ratios group*.

Characteristic	**Composite Variable**	
	1ǂ	**2**†
pron + verb + advb	.924	
verb + advb	.892	
pron + verb	.890	
verb + conj + pron+ verb	.535	
pron+ verb + verb		.991
verb + verb		.941

***** Loading coefficients less than 0.3 were excluded. The varimax with Kaiser normalization was used for the rotation.

ǂ Use of simple verbs

†Use of future and past tenses

We ran PCA to better observe the relationship between the composite variables and the severity of cognitive impairment. The suitability of PCA was assessed prior to the analysis. An inspection of the correlation matrix showed that all composite variables had at least one correlation coefficient greater than 0.3. The KMO measure was 0.73 and the Bartlett's test of sphericity was statistically significant ($p < .0005$).

PCA revealed three components (see Table 4.5) that had eigenvalues greater than one, and which explained 56.2%, 16.6% and 12.7% of the total variance, respectively. A visual inspection of the scree plot indicated that the three components were relevant (Cattell, 1966).

Table 4.5 Correlation with severity of cognitive impairment (column 1), and factor loadings for composite variables* (loading coefficients less than 0.3 were excluded).

Composite Variable	***r*ǂ**	**Components**		
		1	**2**	**3**
Passive POS n-gram ratios	-.540	0.914		
Lexical richness measures	-.689	0.847	-0.302	
Action POS n-gram ratios (simple verbs)	.619	-0.845	0.34	
Vocabulary distribution-based measures	.669	0.834	-0.47	
Average unclear words per utterance	.617	-0.673		-0.484
Skewness of utterances' subjectivity	.428		0.941	
Hesitation interjections	.552		0.905	
Action POS n-gram ratios (use of past/future)	-.471			0.942

* The varimax with Kaiser normalization was used for the rotation.

ǂ All correlations were significant at $p < .05$ level. Negative correlations are highlighted in bold font.

Another correlation analysis over the most commonly used n-grams in the corpus showed a significant positive correlation of negation words and phrases with the severity of cognitive impairment. These included a higher use of *pas* (auxiliary for negation), *non* (no), *rien* (nothing), *même* (even), and *être_pas* (not + to be). There was also a significant higher use of

common verbs, such as *être* (to be), *savoir* (to know), *faire* (to do), *penser* (to think), *avoir* (to have). The use of the n-gram *je_aller* (I + go to), which indicates the use of future tense, was negatively correlated with the severity of CI. All n-grams that were significantly correlated (p <.05) with the severity of CI can be seen in Appendix I, Table-A I-1.

4.5 Discussion

4.5.1 Behavior of the composite variables

In keeping with the findings of multiple authors (Fraser et al., 2016; Hernández-Domínguez et al., 2018; Khodabakhsh et al., 2015; Shinkawa & Yamada, 2018), the correlation analysis in Table 4.3 showed that the **lexical richness measures** were significantly correlated with the severity of CI. From the six measures, only Yule's characteristic *K* was not present in this list. The resulting composite variable created with PCA using these measures also showed a significant (p <.001) inverse correlation with CI.

Figure 4.2 shows that there was a significant difference between participants in the healthy aging and the CI groups for the lexical measures' composite variable from the first interviews. In b), it can be observed that the mean tendency of the HC remained relatively stable over time, while the CI group presented a steep descent in the last interview. During all interviews, the HC group attained higher values of this composite variable than the general median. Our findings support the importance of observing vocabulary diversity when studying early signs of CI.

In the case of the **characteristics based on vocabulary distribution**, commonly tested metrics, such as vocabulary size, hapax legomena and hapax dislegomena, were inversely correlated with the severity of cognitive impairment (p <.001, p <.001, p <.01 respectively). Similarly, all the percentages of different n-grams over the total number of n-grams (n=1 to 4) had a significant inverse correlation (p <.001) with CI.

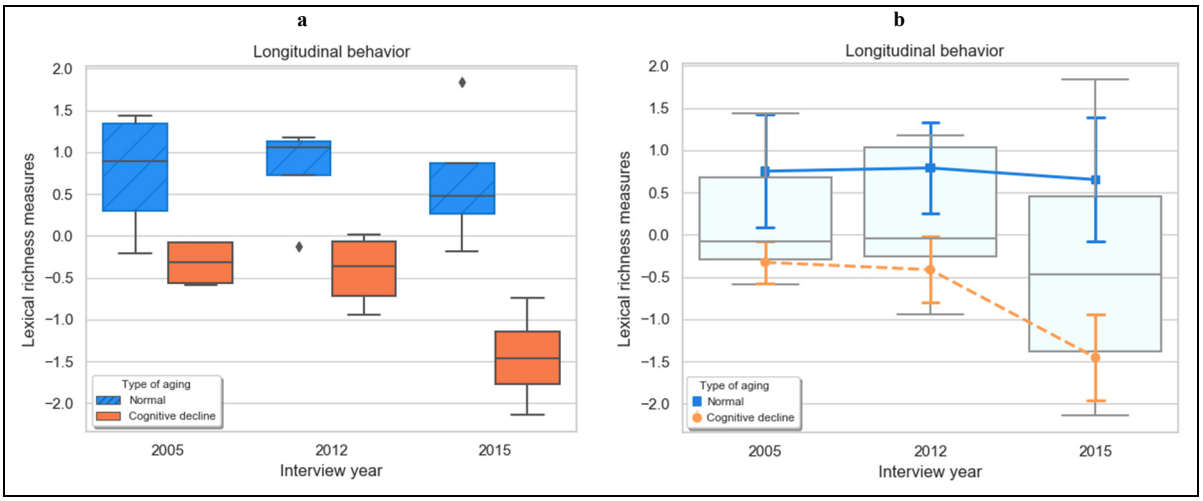

Figure 4.2 Behavior of the composite variable from lexical richness measures over time: **a)** Distribution in box plots of each type of aging. **b)** Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.

Our proposed metric, the r*atio of upper individual TFIDF n-grams,* was significantly (p <.001) inversely correlated with CI. Having a higher value in this metric implies that the participant used more variant specific vocabulary among his/her interviews. In our findings, participants with cognitive impairment had a lower tendency to use original vocabulary between interviews, which may reflect a disposition to stay in comfortable topics.

Another two of our proposed metrics, *Ratio of upper general TFIDF n-grams* and *Ratio of lower general TFIDF n-grams* were significantly correlated (p <.05) with CI. The first, which presented a negative correlation, showed that participants with CI were prone to using less "original" vocabulary when compared against the rest of the participants in the sample. The second metric indicated a propensity to use more generic vocabulary by the CI population. We were, however, unable to find any correlation with the use of *idiosyncratic vocabulary* and CI.

The composite variable created from the characteristics based on vocabulary distribution presented a significant (p <.001) inverse correlation with CI. The inter- and intra-speaker comparison of the use of vocabulary provided valuable information regarding the tendency of CI participants to shift their vocabularies towards more general terms. These findings are

consistent with similar results found in patients with semantic dementia (Garrard, Rentoumi, Gesierich, Miller, & Gorno-Tempini, 2014; Hoffman, Meteyard, & Patterson, 2014).

Similar to the behavior in the lexical richness measures, Figure 4.3 shows that there was a sustained significant difference between the vocabulary-based distribution composite variable of HC and CI groups from interview 1. This difference was more evident for the third interview, where the central tendency of this composite variable for the CI group presents a steep decline. The mean values of the vocabulary-based distribution composite variable for the CI group were maintained below the general median across all interviews.

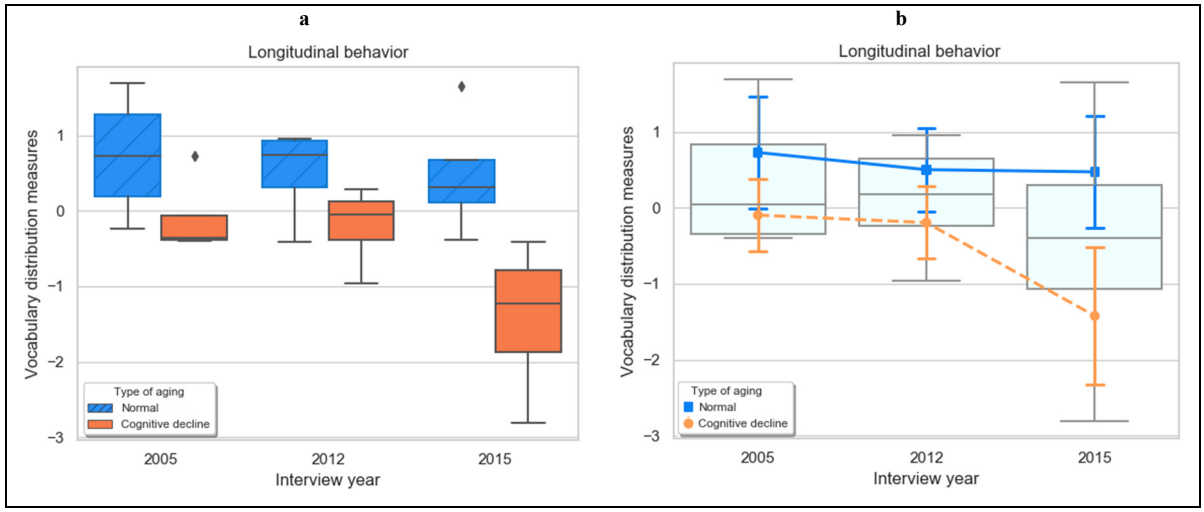

Figure 4.3 Behavior of the composite variable from the characteristics based on vocabulary distribution over time: **a**) Distribution in box plots of each type of aging. **b**) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.

The markedly higher use of negative phrases and constructs (*pas*, *non*, *rien*, *même*, *être_pas*) by CI participants led us to hypothesize that this phenomenon could be caused by a more emotionally negative discourse. We performed a **sentiment polarity and subjectivity** analysis to evaluate this phenomenon and to explore evidence of differences in the verbal expression of emotions among the CI group. However, we found no correlation with the use of negative sentiment terms. On examining the n-grams, we observed that these constructs were most

frequently used to compose phrases that indicated doubt, such as « *je* [*ne*] *sais pas* » (I don't know), « *je* [*ne*] *sais rien* » (I know nothing), « *je* [*ne*] *suis pas sûr* » (I'm not sure). This finding is in accordance with previous findings in patients with semantic dementia (Garrard & Forsyth, 2010).

We also found that CI participants presented a more subjective discourse. This seemed to show, on average, a higher propensity to speak in vaguer and less-defined terms than their healthier counterparts, using more words that related to feelings than to facts. This tendency to use subjective terms appeared to be widespread during the whole interview, and not just in a few utterances (as indicated by a higher skewness and kurtosis levels). However, an analysis of the graphics in Figure 4.4 shows that these results may also have been slanted by the presence of outliers.

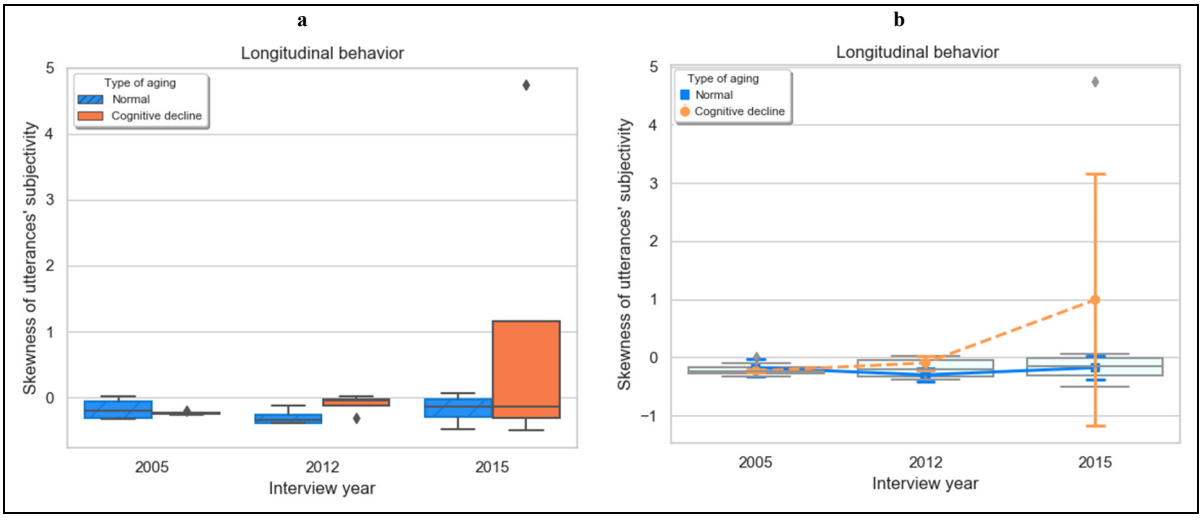

Figure 4.4 Behavior of the *skewness of utterances' subjectivity* variable over time: **a**) Distribution in box plots of each type of aging. **b**) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.

PCA could not be applied to create a single composite variable from the subjectivity measures since it did not pass the KMO measure of sampling adequacy test. For this reason, the highest correlated (p <.01) measure of subjectivity (*skewness of utterances' subjectivity*) was selected to represent this group of variables in the general PCA analysis. Figure 4.4 shows a relatively

similar behavior for both groups for this characteristic. A closer inspection of this variable shows that the mean tendency of the CI group is being raised significantly thanks to an outlier. The confidence intervals for the CI group extend well above and below the intervals of the HC group. A study with a bigger sample is needed to enable a better interpretation of the tendencies shown by this variable.

When studying the **ratio of use of part-of-speech (POS) patterns**, we saw that prepositions and noun phrases were used at a significantly lower rate by the CI group. This finding has been observed in previous studies (Ash et al., 2013; Fraser et al., 2016; Jarrold et al., 2014). At the same time, we observed a significantly increased use of pronouns by the CI sample. This phenomenon has been observed in patients with the semantic variant of primary progressive aphasia (Wilson et al., 2010), Alzheimer's patients (Fraser et al., 2016) and patients with semantic dementia (Garrard & Forsyth, 2010). This increase in the use of pronouns has been explained as a compensation mechanism resulting from the inability of aphasic patients to remember proper names, and its use as "a crutch" for sentence structuring (Singh & Bookless, 1997).

In Figure 4.5, we can observe that the medians of the composite variable made from the ratio of use of passive POS patterns have relatively close values for both groups of participants in the first two interviews. Notwithstanding a marked decline for this value in the third interview for both groups, the CI group presented a steeper slope. The central tendency of the CI group consistently performed below, but inside the confidence interval of the mean tendency of the normal aging group. In the first two interviews, the mean of the passive POS patterns composite variable of the CI group was almost at the median level of the general population. At the third interview, the mean value of this group fell to the limit of the first quartile of the general population. This seems to indicate that the difference in the use of prepositions, nouns, and pronouns is a sign that appears later for the types of cognitive impairment observed in our sample.

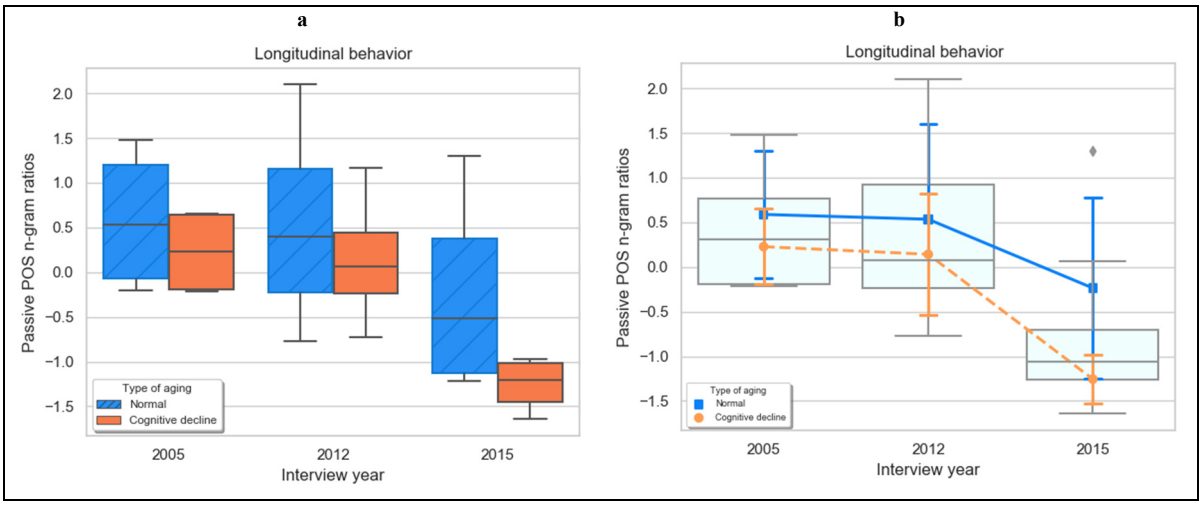

Figure 4.5 Behavior of the Passive POS n-gram ratios compound variable over time: a) Distribution in box plots of each type of aging. b) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.

With respect to the **action POS n-gram ratio** compound variable, we observed a decrease in the use of past and future verb tenses, as well as a significant increase in the use of simple verbs during the third interview for the CI impaired group. We also observed that there was a high correlation with an increase in the number of common verbs, such as *être* (to be), *savoir* (to know), *faire* (to do), *penser* (to think), *avoir* (to have), and CI. These findings correspond to those of previous authors (Hoffman et al., 2014), who deduced that semantic dementia patients have a tendency to produce more highly frequent verbs. Similarly to the passive POS patterns, the difference becomes more evident at 10 years after the first interview (see Figure 4.6), which may indicate that this is also a late marker for CI.

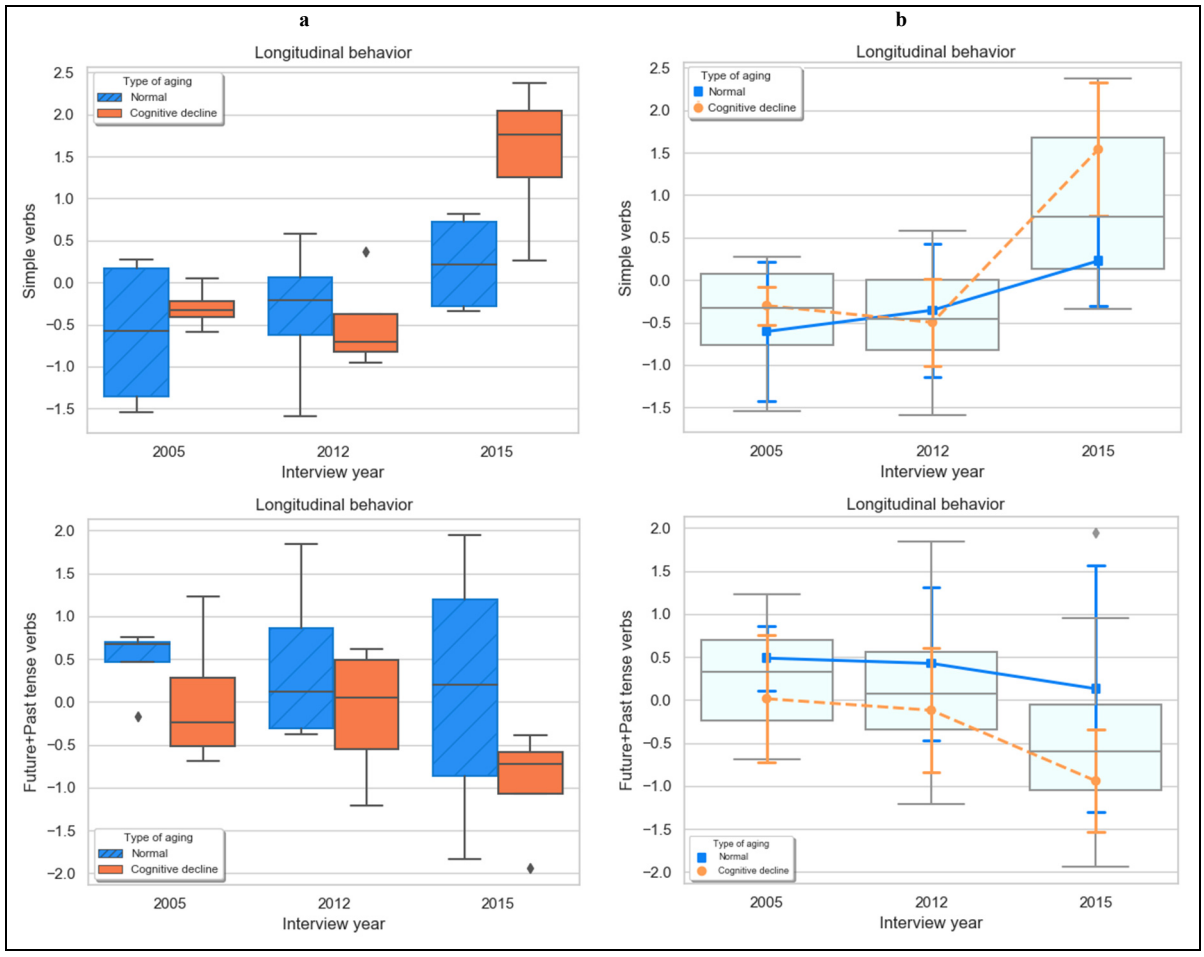

Figure 4.6 Behavior of the use of simple verbs' and the future and past tense verbs' (action POS n-gram) ratios' compound variables over time: a) Distribution in box plots of each type of aging. b) Central (mean) tendencies of normal and cognitively declined aging. Confidence intervals correspond to the standard deviation. Boxplots in the background indicate distributions of the entire sample.

4.5.2 Differentiation of CI and healthy controls

We performed a correlation analysis of the severity of CI with the components of the PCA analysis that incorporated all the composite variables (see Table 4.5). Components 1 and 3 were significantly negatively correlated (p <.001 and p <.05, respectively) with the severity of CI. A scatter plot of the distribution of the participants with these two components is shown in Figure 4.7.

The scatter plot in Figure 4.7 shows, in the lower left corner, a clear separation (dotted line) of the interviews of severely cognitively impaired participants. Furthermore, there is another threshold (dashed line) that separates the interviews of participants with healthy aging processes from those of the CI group (except for Participant 18, whose specific metrics we will discuss later). This finding suggests that the PCA Components 1 and 3 from the composite variables could differentiate individuals that may develop some form of CI from those with normal aging processes, even ten years before any symptoms are present.

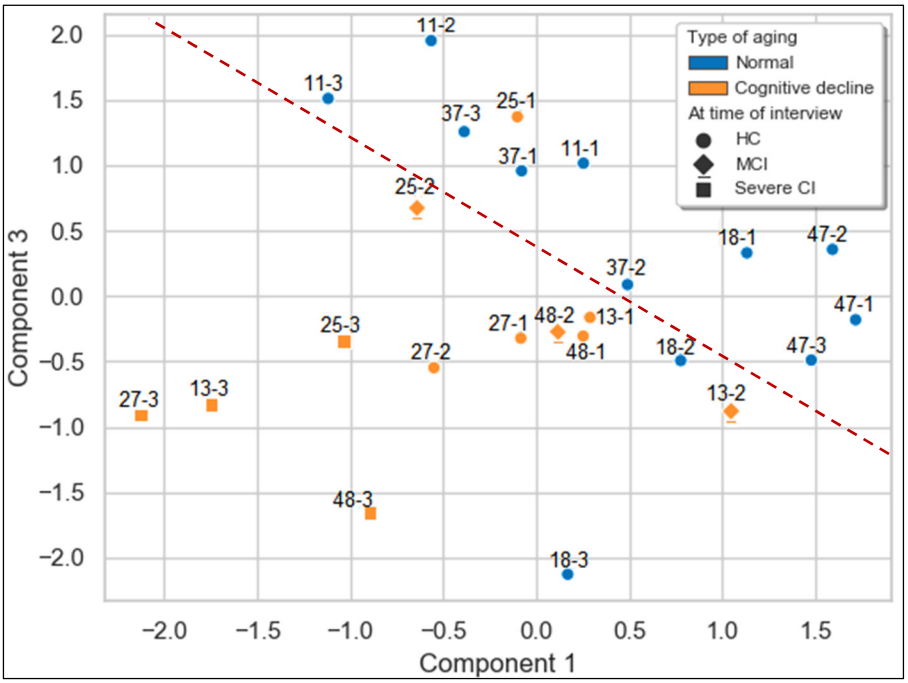

Figure 4.7 Scatter plot of the distribution of interviews with Components 1 and 3 from the PCA using composite variables. The label near each point indicates the participant ID-interview number. Interviews 1, 2 and 3 where held in 2005, 2012 and 2015, respectively. The hue difference indicates normal or cognitively declined aging processes. Circle, square and rhomboid markers indicate healthy control (HC), mild cognitive impairment (MCI) and severe CI, respectively, at the time of the interview.

4.5.3 Mini case study: Participant 13 and matched control 37

Figure 4.8 shows the behavior over time of all the composite variables and PCA components that were significantly correlated with CI for participants 13 and 37. Participant 13 is a female who was diagnosed with Alzheimer's disease in 2013. Both females had low education levels (primary studies) and were 93 and 89 years old each by the time of the last interview.

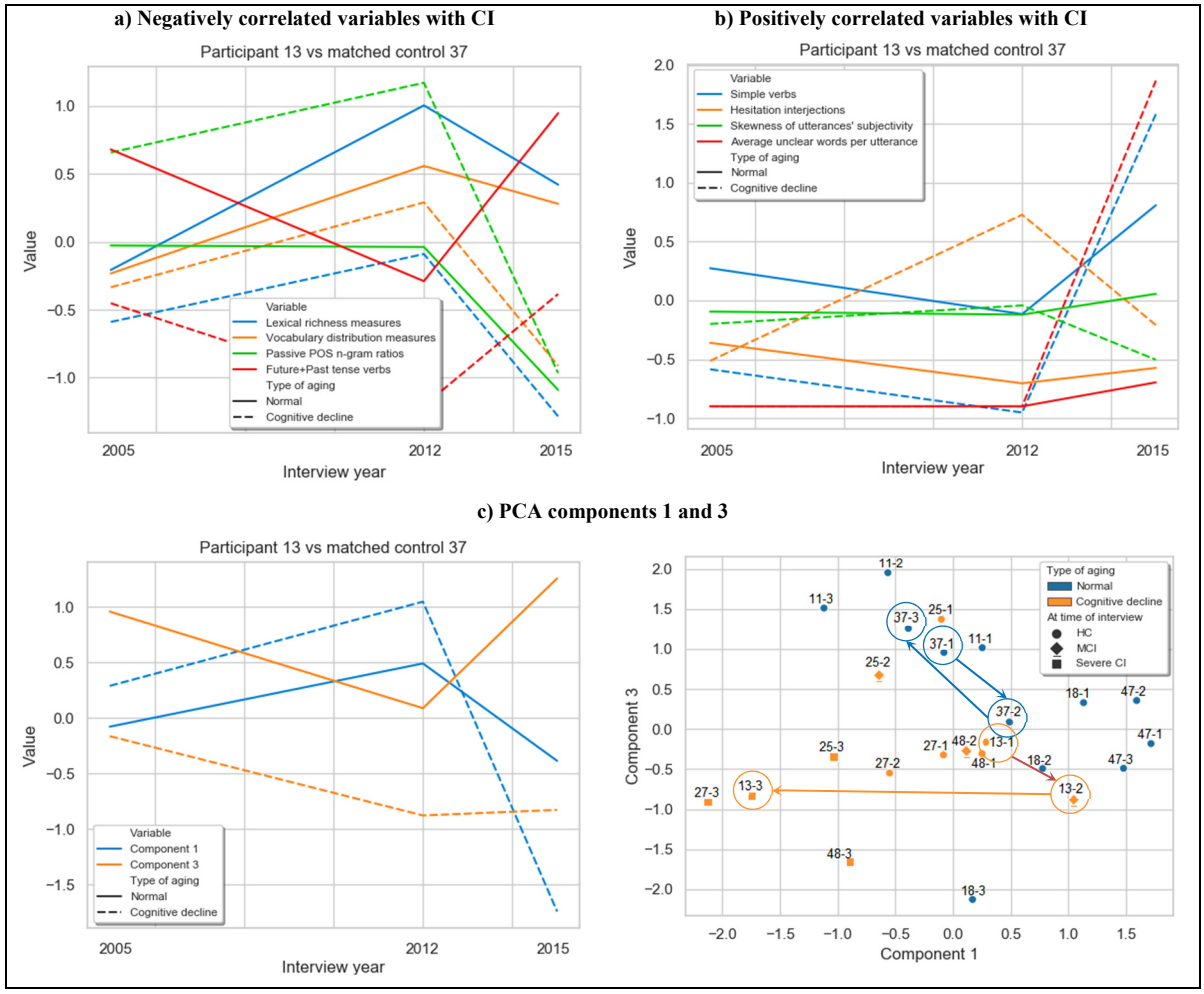

Figure 4.8 Comparison of the behavior of all significant composite variables and PCA components over time in participants 13 (CI group; dashed lines) and 37 (healthy matched control; continuous lines).

In Figure 4.8-**a)**, the lexical richness measures, vocabulary distribution and passive POS patterns ratios had a similar behavior to that of her healthy counterpart until 2012. By the time of the third interview, Participant 13 had a much steeper drop in these three variables. In Figure 4.8-**b)**, Participant 13 had a significant increase in the average number of unclear words and simple verbs, compared to Participant 37 at their third interview. However, Participant 13 had a similar use of hesitation interjections 10 years after her first interview. With respect to the PCA components, in Figure 4.8-**c)**, Participant 13 had a higher value for the first interviews, with a steep decrease at the last one. The pattern of the line for component 3 was not very different between the two women, but Participant 3 consistently performed worse in this component. In the scatter plot, both participants are relatively close until the third interview.

4.5.4 Mini case study: Participant 25 and matched control 11

Participant 25 is a female who started presenting memory issues in 2012 and was showing signs of confusion in 2015. Participant 11 is a matched healthy control for Participant 25. Both women had completed their professional studies and held jobs as qualified employees. By the time of the last interview, they were 94 and 96 years old, respectively. Figure 4.9 shows the behavior over time of all the composite variables and PCA components that were significantly correlated with CI for these participants.

In Figure 4.9-**a)**, Participant 25 had a constant increase in the use of past and future tense verbs. However, her lexical richness and vocabulary distribution metrics present a steep decline in the last three years. The behavior of her use of passive POS patterns is similar to that of her healthy counterpart, and only presents a slight decline over time. In Figure 4.9-**b)**, Participant 25 presents an increase in all four composite variables, with a notable slope at the third interview. For both PCA components in Figure 4.9-**c)**, Participant 25 presents a consistent decline over time from the second interview. However, Participant 11 has a similar behavior for component 1. In the scatter plot, both participants start very close together, but Participant 25 slowly starts drifting towards the lower left corner.

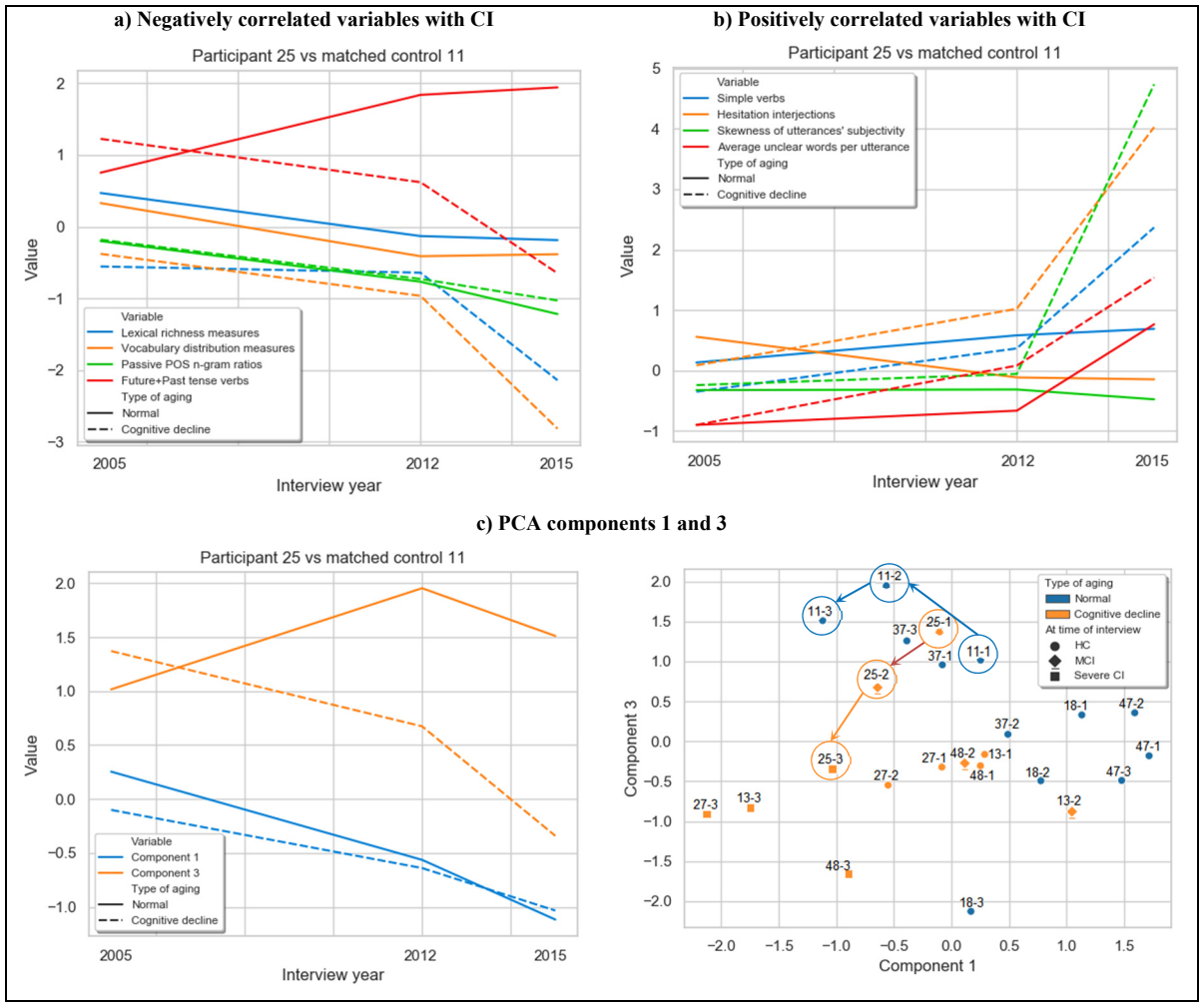

Figure 4.9 Comparison of the behavior of all significant composite variables and PCA components over time in participants 25 (CI group; dashed lines) and 11 (healthy matched control; continuous lines).

4.5.5 Mini case study: Participant 48 and matched control 47

Participant 48 is a male reported by his spouse, in 2015, as having cognitive impairment. Participant 47 is his matching healthy counterpart. Although Participant 48 had a slightly lower education level, Participant 47 was five years older. Both participants were bilingual speakers (French and German). In Figure 4.10, the line plot **a)** shows that Participant 47 presented a continuous decline in the vocabulary distribution composite variable. In the rest of the variables, he presented a slight increase at interview two, with a steep decline at interview 3.

Despite some slight changes at interview 2, Participant 47 seemed to maintain the values of four metrics after ten years. In **b)**, Participant 48 presents a continuous increase in the number of unclear utterances over time. At the time of the third interview, however, he seems to maintain similar levels of the other 3 composite variables to those at the first interview. Participant 47 presents an increase in the use of hesitation interjection and simple verbs. For the PCA components in **c)**, Participant 47 maintains similar levels between the first and last interviews, while Participant 48 presents a steep decline in both metrics after 2012. In the scatter plot, Participant 47 maintained very similar patterns throughout all his interviews, while Participant 48 presents a clear drift towards the bottom left of the graphic at his third interview.

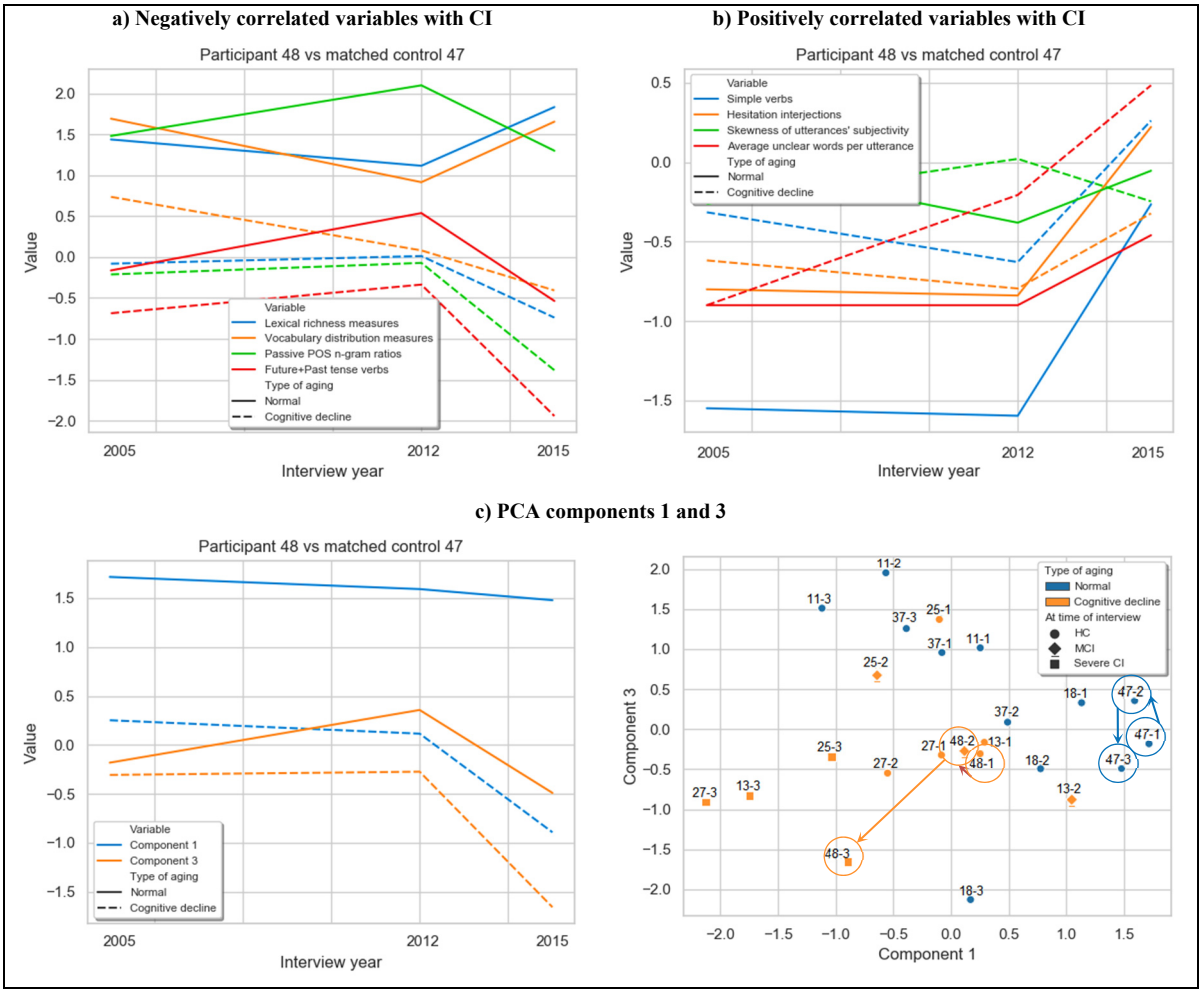

Figure 4.10 Comparison of the behavior of all significant composite variables and PCA components over time in participants 48 (CI group; dashed lines) and 47 (healthy matched control; continuous lines).

4.5.6 Mini case study: Participant 27 and matched control 18

Participant 27 is a male that was diagnosed with a progressive cognitive disease in 2015. Participant 18 is his matched healthy control. Both participants have high education levels and have worked in high academic/management professions.

In Figure 4.11-**a)**, both participants have very similar patterns in their composite variables. For some variables, such as the use of past and future tense, the healthy control presents a steeper decline than the CI counterpart. This is the only case in the sample in which this behavior occurs with a healthy participant. However, although the tendency with all his measures seems to be on steep decline, Participant 18 started with higher values than Participant 27 in all these metrics. In **b)**, Participant 18 presents an increase in the use of hesitation interjections, average unclear words per utterance and unclear words, and a slight increase in the number of simple verbs. In the case of Participant 27, he presented an increase in the number of unclear words after the second interview — which he maintained until the third — and a significant increase in the use of simple verbs. For both PCA components in **c)**, both participants have a decline, although for Participant 27, this decline is much steeper in component 3, and for Participant 18, it is in component 1.

In the scatter plot, we can see the path both participants follow towards the lower and left sides of the graphic. Following the path that most CI participants present in the scatter plot, it might seem that a decline in component 3 is not as meaningful as it is in component 1. Also, we believe that the consistent decline of Participant 18 in most metrics might be an indication of a possible underlying condition. However, as of the time of writing this paper, Participant 18 has not received any unfavorable diagnosis. Continuous monitoring of this participant would be very valuable for the future of this research.

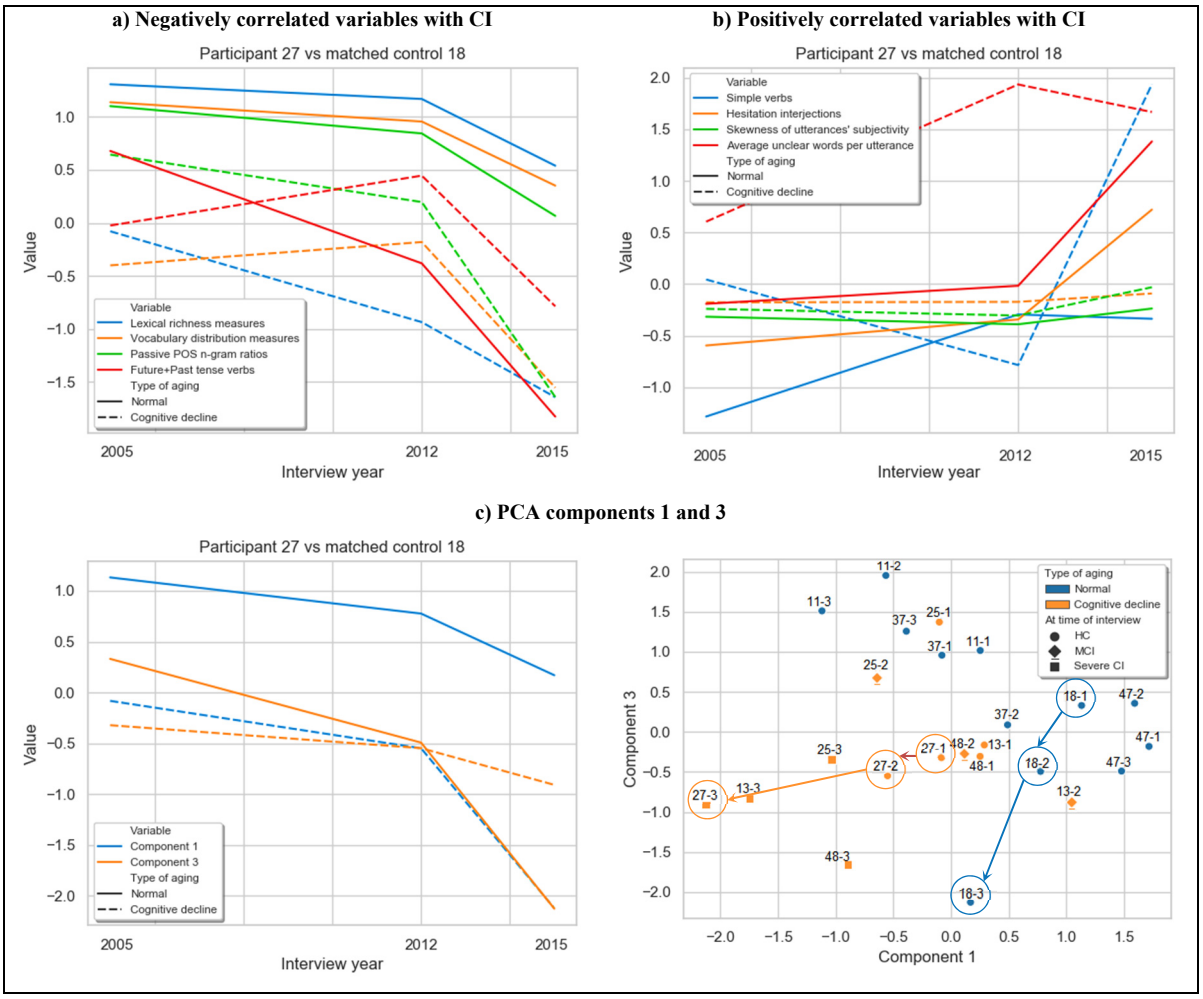

Figure 4.11 Comparison of the behavior of all significant composite variables and PCA components over time in participants 27 (CI group; dashed lines) and 18 (healthy matched control; continuous lines).

4.6 Limitations

Due to the nature of this study, one of its limitations is the small size of its sample. In the future, this limitation may be overcome with the incorporation of similar datasets that are currently under construction, such as the Carolinas' Conversations Collection in English (Pope & Davis, 2011) and Spanish (Hernández Domínguez, Ratté, Pope, & Davis, 2016), and the CorpAGEst (Bolly & Boutet, 2018). This inclusion would not only augment the number of participants, but also allow for deeper inter-language comparisons.

Another limitation of this work is that since it was originally intended to describe the language evolution in healthy older individuals, there was no medical history or formal cognitive assessments of the participants available. The diagnoses of the cognitive impairment group were annotated as described by the spouses, family members and caregivers of the participants. Still, the nature of the speech data seems to balance these limitations, as the individuals did not perceive themselves as patients; the stereotypes underlying cognitive testing and the supposed decline in old age could lead to a lower language production performance (Hess, Hinson, & Hodges, 2009).

4.7 Conclusions

In this work, we presented a computer-based longitudinal analysis of linguistic changes in older French speakers with cognitive impairment. This study was done using a sample from the LangAge corpus (Gerstenberg, 2011), in which three sets of interviews were conducted with the participants over a span of ten years. All participants where apparently healthy individuals at the time of the first interview, and four developed cognitive impairment by the end of the process. Through a computer-based quantitative analysis, we compared the evolution of language alterations in these four participants against four healthy control participants matched by age, gender, educational level and bilingualism. To the best of our knowledge, this is the first work using spontaneous conversations to study the language changes occurring in participants as they transition from a healthy, pre-clinical stage, to cognitive impairment.

Few previous works have been conducted on spontaneous conversations, with those done mostly being for English speakers. In this study, we extracted the main characteristics that have consistently been found to be altered by the presence of CI and studied their evolution over time. We also proposed the use of four linguistic measures based on vocabulary distribution to observe the use of general and specific vocabulary by the speakers.

We found that most participants in the CI impaired group presented a continuous drop in the lexical richness measures throughout the ten years of the study. Also, for most metrics, HC individuals started with better scores beginning with the first interview. Our proposed metrics based on vocabulary distribution showed a tendency of CI participant to reduce the number of original words and phrases over time, and a predisposition to talk about similar topics through most interviews. It also showed that CI participants will tend to use more generic vocabulary and use more pronouns and fewer nouns and verbs, except for common use verbs.

After a PCA, we found that two main components were highly correlated with the severity of cognitive impairment. In a scatter plot we observed a clear separation of the severely impaired individuals from the rest. Furthermore, with the exception of one participant, there also seems to be a clear separation between apparently healthy individuals that later developed CI from those who remained healthy throughout the ten-year span.

One apparently healthy participant had a pattern of decline in most metrics similar to that in the CI group. Despite this tendency, this participant consistently had higher scores in the metrics than his CI counterpart. This participant is a highly educated individual who held a high academic or management profession, factors that have been linked to a higher cognitive reserve (Almeida et al., 2015). In 2015, he reportedly underwent a medical exam, performed by a neurologist, confirming the absence of CI indicators. At the time of writing of this paper, there was no evidence that the participant had developed any form of CI; however, it would be beneficial to monitor this individual's progression.

It is our belief that our proposed method and analysis could help in the following up of patients trough time and develop more personalized analyses of cognitive status. The study of spontaneous speech represents an inexpensive and non-invasive process for detecting early signs of cognitive impairment. Our metrics evinced a significant difference between individuals that would age as cognitively intact individuals and those who would develop a form of cognitive impairment even up to ten years before the time of diagnosis.

CHAPTER 5

CONVERSING WITH THE ELDERLY IN LATIN AMERICA: A NEW COHORT FOR MULTIMODAL, MULTILINGUAL LONGITUDINAL STUDIES ON AGING

Laura Hernández-Domínguez[1], Sylvie Ratté[1], Charlene Pope[2] and Boyd Davis[3]

[1] Software and IT Engineering Department,
École de technologie supérieure, Montreal, Canada

[2] Medical University of South Carolina (MUSC), Charleston, SC, United States of America

[3] University of North Carolina at Charlotte, Charlotte, NC, United States of America

Email: laura.hzdz@gmail.com, sylvie.ratte@etsmtl.ca

This article was published at the *ACL's 7th Workshop on Cognitive Aspects of Computational Language Learning*, in Berlin, Germany, on August 11, 2016.

Foreword: As observed I the literature review and in the limitations presented in Chapter 4, one of the biggest obstacles for performing studies on longitudinal analyses of language alterations in spontaneous conversations is the scarcity of data available for research. Besides the direct study of these alterations, a concomitant contribution of my doctoral studies was oriented to increase the available data of this nature by collaborating in the creation of the Latin-American cohort of the Carolinas' Conversation Collection. The recollections for this cohort started in 2015, and although this data was not used for analysis, since the recollections and transcriptions are still in process, it is expected that it will be used for the continuation of the multi-modal analyses for the Cécilia Project.

5.1 Abstract

Many studies have found that language alterations can aid in the detection of certain medical afflictions. In this work, we present an ongoing project for recollecting multilingual conversations with the elderly in Latin America. This project, so far, involves the combined efforts of psychogeriatricians, linguists, computer scientists, research nurses and geriatric

caregivers from six institutions across USA, Canada, Mexico and Ecuador. The recollections are being made available to the international research community. They consist of conversations with adults aged sixty and over, with different nationalities and socio-economic backgrounds. Conversations are recorded on video, transcribed and time-aligned. Additionally, we are in the process of receiving written texts---recent or old---authored by the participants, provided voluntarily. Each participant is recorded at least twice a year to allow longitudinal studies. Furthermore, information such as medical history, educational background, economic level, occupation, medications and treatments is being registered to aid conducting research on treatment progress and pharmacological effects. Potential studies derived from this work include speech, voice, writing, discourse, and facial and corporal expression analysis. We believe that our recollections incorporate complementary data that can aid researchers in further understanding the progression of cognitive degenerative diseases of the elderly.

5.2 Introduction

The *Carolinas Conversations Collection* (Pope & Davis, 2011), a project for recollecting conversations with elderly people that live in North and South Carolina, started in 2008. This project was initially supported by the USA National Library of Medicine. For the collection, the conversations were transcribed, marked, time-aligned and made available to the international research community by means of a secured website[3]. The collection has grown steadily since then, having, at present, over 460 conversations with adults over sixty years old, either healthy or suffering from any medical condition. A fourth of these conversations were made with participants afflicted with Alzheimer's disease.

In 2015, we started to increase the coverage of this collection to incorporate different languages. The first additional language to be incorporated is Latin-American Spanish. We are currently adding conversations with new participants; elderly Spanish speakers from Ecuador and Mexico. Additionally, we are incorporating new information and language modalities to

[3] http://carolinaconversations.musc.edu/

increase the robustness of possible studies that may use this corpus. So far, this project has engaged involvement through combined efforts of six institutions across four different countries.

5.3 Methodology

The recollections are being made at least twice a year with each participant. In Ecuador, we are working in collaboration with *Universidad Técnica Particular de Loja* (UTPL), and with the *Perpetuo Socorro* Foundation, a home for elderly people. In Mexico, the psychogeriatricians from the Psychiatric Hospital *Fray Bernardino Álvarez* have agreed to work as our medical experts and advisors for this project. Furthermore, the Foundation and the Psychiatric Hospital have made arrangements to allow us to communicate with their residents, patients and their guardians, and invite them to participate in our Latin American recollections.

In the case of Ecuador, none of the involved institutions has an Institutional Review Board (IRB) for protection of human subjects, or any formal ethics guidelines. For this reason, our institutional IRB took over that role. Consequently, a person authorized via the protocol and having a Canadian or American certification of training in ethics for research with human subjects, must be present, in person, during all recollections. In the case of Mexico, the hospital has its own IRB, and their staff are trained in ethics. This allows them to recollect the conversations without any member of the team from Canada or the USA needing to be present.

Before the recordings, the participants and their caregivers are given a short explanation of the project and its aims. Provided they agree to participate in the project, they sign an informed consent form, and with the help of their primary psychiatric care providers or their primary caregiver, we fill a questionnaire with the medical information of the participant. In this questionnaire we request all the medications that the participants are actively taking, as well as their medical conditions. With first-time participants, we also record their demographic data, such as birth date, gender, educational level, occupation (prior to retirement), first language, and ethnic affiliation. To protect the privacy of the participants, all names are replaced by

aliases. In the case of Ecuador, aliases are randomly chosen from a pool of names of characters or writers of classic Latin American novels; in the case of Mexico, they are chosen from names of congresspeople. We select aliases that correspond with the gender of the participants.

The interviewers are the caregivers at the Foundation (Ecuador), and the primary psychiatric care providers (Mexico). All interviews take place in the Foundation's and the psychiatric hospital's facilities. We believe that having free topics, and a familiar interviewer and environment, helps provide a more comfortable atmosphere for the participants.

All our interviewers have been trained with techniques to motivate the participants to talk, even if they are afflicted by some type of cognitive impairment. We've created animated videos and other training materials to instruct interviewers on how to incite free conversations with patients. The strategies that we provide, come from practices that have been developed during the years of experience interviewing elderly participants in North and South Carolina for this collection. These materials are available online[4] to facilitate the long-distance knowledge exchange.

While training the interviewers, we usually start by explaining the context of the project. We then emphasize the importance of letting the participants talk and express themselves as much as possible. We ask the interviewers to be patient and allow the participants some time to process their questions and then answer. We also give them cues such as repeating the last utterance of the participants when they are stuck; giving encouraging feedback and signs of interest, such as making eye contact, responding with interjections, corporal and facial expressions according to the mood of the conversation; and keeping the flow of the conversation by mentioning any information that they have gathered about the participants during the time of knowing them.

[4] https://goo.gl/E7xeOO (English and Spanish subtitles are available)

The conversations are free in the sense that there is no specific theme to talk about, although the most common topics are the early lives of the participants, their hobbies, their health and their views on life in general. There is no time limit to these conversations. Some of the common questions to start the flow of the conversation are: "*Tell us about your life*", "*What do you like to do?*", "*How was your childhood?*", "*Do you have any hobbies?*", "*Who is accompanying you today?*", "*Do you have any pet?*", "*What did you use to do for a living?*".

The conversations from Mexico and Ecuador are being manually transcribed and time-aligned by our collaborators of the Linguistic Engineering Group (LEG) at the National University of Mexico. We selected the LEG group due to their vast experience in the creation of corpora[5] in Spanish. The transcriptions are labelled with markings that indicate pauses, interruptions, external noises, participant's noises (e.g., laughter, crying, coughing), intonation and emphasis (e.g., whispering, yelling), actions (e.g., winking, hand gesturing, finger snapping, clapping), and unconventional pronunciations.

In addition to the recordings of the conversations, at Mexico we are also asking the participants and/or their guardians for copies (digital or physical) of written texts, such as old letters, messages, etc., authored by the participants, recently or in years prior to this study, including letters from their youth or middle age. This is to encourage research in written analysis, such as the famous Nun Study (Snowdon et al., 1996).

5.4 Description of the samples

Recollections in Ecuador started in May, 2015. For the first series we interviewed 12 participants, and recorded a total of 15 conversations. The second recollection was made on January, 2016, and it incorporated 4 new participants and a total of 10 interviews. So far, the cumulative recorded time of conversations in Ecuador is over six hours and 45 minutes, and

[5] http://www.corpus.unam.mx/

the average length of the conversations is 16 minutes. The participants' ages range from 70 to 91 years old, with an average age of 83 years old (see Table 5.1).

Table 5.1 Socio-demographic overview of the participants in the collection

		Women	**Men**	**Global**
México	Participants	9	0	9
	Conversations	9	0	9
	Avg. age	69	-	69
	Avg. education (years)	5.5	-	5.5
Ecuador	Participants	14	6	20
	Conversations	25	11	36
	Avg. age	83	82	82.7
	Avg. education (years)	7	13.2	9.8
USA	Participants	71	16	87
	Conversations	368	94	462
	Avg. age	79.3	79.1	79.3
	Avg. education (years)	13.1	14.1	13.3

As shown in Table 5.1, the majority of our participants in all countries are female. We attribute this phenomenon to two main factors: first and foremost, women have shown a significantly higher willingness, in comparison to men, to participate in this project, especially in Mexico. Secondly, the age expectancy of women is higher than men, for which the elderly male population is smaller. We are currently making efforts to increase the number of male participants to balance the sample.

5.5 Implications, applications and prospects

The longitudinal, multilingual and multimodal attributes of our collection, as well as the registration and follow up of the medical treatments taken by the participants and their

demographic information, will allow researchers to perform a wide variety of studies. Some of these studies have already been tackled before. However, in most cases authors have used small monolingual and homogeneous samples that do not allow the possibility of generalizing. Furthermore, many of the datasets used for these studies are not shared to the research community, limiting the advancement of research.

Our collection has the advantage of containing a multiethnic sample, not to mention the heterogeneity gained by including participants from three different countries. These attributes will make for robust research that will support the study of intra-language and inter-language variations, as well as intermodal linguistic analyses (see Figure 5.1). Additionally, it will allow control for alterations attributable to race, demographic factors, specific diseases, medications and treatments. Longitudinal studies will allow following the course of aging in the elderly, and the differences between a healthy versus a pathological decline. This collection also provides data to improve automatic transcription and face recognition for this particular cohort, which tends to present particular challenges. Some of the clearest research possibilities to be performed with this collection are those focused on the improvement of communication with the elderly, and medical applications.

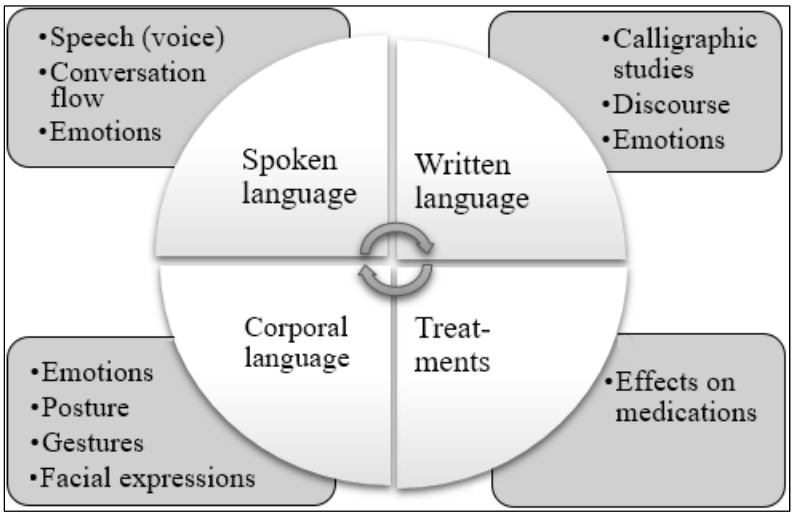

Figure 5.1 Multimodal studies available in the corpus.

5.5.1 Improving communication

It is important to maintain and preserve communication with the elderly, especially since it has been suggested (Arkin, 2007) that maintaining language-enriched conversations along with exercise can delay the effects of dementia. Our collection not only contains the utterances and transcriptions of the elderly participants, but it also includes the entire transcription of the exchanges with the interviewers. This allows performing studies to improve communication by analyzing which strategies prove more successful in promoting conversations with the elderly. Other authors (Davis, 2005; Davis, Maclagan, Karakostas, Liang, & Shenk, 2011) have made a strong emphasis on the importance of preserving communication with elderly people, and have worked in the development of specific communication strategies, particularly with those suffering from dementia.

In addition to explicit linguistic barriers, there are other factors that limit our ability to communicate with elderly people. For example, Freudenberg et al. (Freudenberg, Adams, Kleck, & Hess, 2015) found out that young people have trouble correctly interpreting facial expressions in the elderly, often perceiving neutral expressions as negative emotions. This in part makes studying emotions in this population a challenge, but in doing so, could provide insights on how to preserve an effective communication with them. However, analysis of emotions have other purposes, since alterations in the expression of emotions can show signs of certain disorders (Adams & Oliver, 2011; Hamm, Pinkham, Gur, Verma, & Kohler, 2014).

5.5.2 Medical applications

Automatic language analysis for studying neurodegenerative diseases in elderly people has been gaining momentum in recent years. Authors like Jarrold (Jarrold et al., 2010, 2014), Schröder (Schröder et al., 2010), Prud'hommeaux and Roark (Prud'hommeaux & Roark, 2011), Lehr (Lehr, Shafran, Prud'hommeaux, & Roark, 2013), Gonzalez-Moreira (Gonzalez-Moreira, Torres-Boza, Garcia-Zamora, Ferrer-Riesgo, & Hernandez-Gomez, 2014), Khodabakhsh (Khodabakhsh et al., 2014, 2015), Guerrero (José María Guerrero, Martínez-

Tomás, Rincón, & Peraita-Adrados, 2015), López-de-Ipiña (López-de-Ipiña et al., 2015), and König (König et al., 2015), have studied language alterations that may aid in the automatic detection, or even prediction, of Mild Cognitive Impairment and Alzheimer's disease in its mild and moderate stages, with promising results. Additionally, Goberman (Goberman, Blomgren, & Metzger, 2010), Holtgraves (Holtgraves, Fogle, & Marsh, 2013) and Cardona (Cardona et al., 2013), have studied the linguistic features associated with Parkinson's disease. To support the furthering of these types of research, we prioritize the inclusion of participants suffering from different cognitive and mental afflictions (see Table 5.2).

Table 5.2 Prevalence of the main mental health disorders in each cohort

	Mexico	**Ecuador**	**USA**
Participants	**9**	**20**	**87**
Alzheimer's disease	2	11	47
Parkinson's disease	0	1	0
Depression	1	3	9
Schizophrenia	1	0	1
Bipolar disorder	0	1	1
Healthy control	2	8	23

5.6 Conclusions and future work

In this paper we presented a report of our first recollections of conversations with elderly people in Latin America, as well as the characteristics of this ongoing multidisciplinary multicenter research project. We envisage to continue these recollections for the following two to five years. Additionally, we are initiating the necessary collaboration agreements with Canadian institutions to incorporate a cohort with Canadian French-speakers and English-speakers to our collection. With this cohort we will add a new language and an English variation. Furthermore, in Ecuador we are making arrangements to incorporate some elderly

Quechua-speakers to our sample. To our knowledge, there is no available research on linguistic analysis of this indigenous population. Finally, we are currently working on our first research using this corpus. We believe that our recollections can be of use for performing speech, voice, writing, discourse, and facial and corporal expression-based analysis to further our understanding about the progression of cognitive degenerative diseases, and ultimately to help improving our communication strategies with the elderly, thus ameliorating their quality of life.

CHAPTER 6

GENERAL DISCUSSION

This thesis has addressed the general problem of characterizing language alterations caused by Alzheimer's disease. The main objective was to propose methods for monitoring language functions of elderly patients in two settings: cognitive testing and spontaneous conversations. The combination of both methods could provide clinicians with low-cost non-invasive techniques that could alert of changes that might be related to a form of cognitive impairment, particularly with Alzheimer's disease.

Chapter 1 presented a literature review of the best-known works on computer-based evaluation of picture description tasks. It also introduced the state-of-the art methods for automated analysis of language functions in older speakers in the context of spontaneous conversations and picture description tasks. At the end of this chapter, the limitations and research possibilities of these works were presented.

6.1 Evaluation of performance and language functions in elicited speech in cognitive testing settings

One of the best-known tests used in clinical practice for evaluating cognitive skills and language functions is comprised of picture description tasks. These tasks consist in showing an image to patients and asking them to describe the depicted scene with as much detail as possible. This task is usually evaluated by comparing the description of the patient against a manually pre-defined list of information content units that serves as a referent as to what the specialist considers to be "important" and should be part of a description of the picture.

The majority of computer-based works that have studied methods for the automatic evaluation of a patient's performance during this task have used these same lists of information content units, and have observed whether an item in the list was present during the description, without

considering the context in which it had been mentioned. The use of these lists poses some limitations, since depending on the author, patients would get different results in their evaluations. Also, the lists are not generalizable to different types of pictures and populations, and the referent for the evaluation is constructed by a person that usually does not share the same socio-demographic characteristics of the participants, which creates a gap between what the population observes and mentions and what the evaluator assesses. In fact, the only work that had proposed a method for automatically creating a list of information content units discovered items that the healthy cohort of the population was mentioning, and that all previous authors of reference lists had overlooked.

In **Chapter 2**, we presented a computer-based method for evaluating patients' performances during picture description tasks. Our method consisted of an adaptation of an information coverage metric to create a referent that considered not only the mentioning of information units, but also their context. This referent was created from a sample of healthy older individuals that matched the sociodemographic characteristics of the evaluated cohorts. Furthermore, by using this adaptation of the metric, we proposed a method for estimating the pertinence of patients' utterances during their descriptions. To the best of our knowledge, this was the first work that had presented a method for measuring not only the completeness, but also the efficiency of the discourse during these tasks.

We deepened the evaluation of the task by also extracting lexical richness measures and acoustic features. By combing these features with the information coverage and pertinence metrics, we trained a support vector machine learner and performed a 10-fold cross-validation classification of AD patients and healthy controls with an average F-score of 0.82 and an area under the curve of 0.76. These results compared favorably against state-of the art methods that relied on manually-created information content units.

Some of the previously proposed works presented difficulties when dealing with linguistic variabilities for expressing similar notions. For example, "the boy is stealing a cookie" and "the lad is taking a cookie" are describing the same action on the Cookie theft picture, while

using different words. An advantage of our method is that since it creates a referent from examples, it benefits from larger and diverse sources to include a wide variety of expressions. It also could be used to create demographic-specific referents for different types of populations and pictures.

A clear disadvantage of our proposed method is that a part of the healthy control sample must be excluded from the experiment in order to create the referent. This is an important limitation since most available datasets of clinical studies are already modest in size.

6.2 Classification of healthy and cognitively impaired individuals from restricted and semi-restricted discourses

In **Chapter 3**, we performed a study to contrast the differences in various linguistic variables—that have been found in correlation with cognitive decline—in the constrained discourse of picture description tasks against the more spontaneous type of discourse of describing objects. In this context, although patients are all describing the same objects, no visual reference is provided, so they describe them using their own mental image, perception and personal experiences with said objects. This increases the variety of syntactical structures and the diversity of vocabulary used when compared to standardized picture description tasks and moves us a step closer to the type of analysis and challenges to be expected when studying completely spontaneous conversations.

Based on the previous literature on analysis of spontaneous speech, we were also interested in observing the use of specific and general vocabulary, and its significance when trying to differentiate healthy controls from AD patients during both description tasks. We proposed a new metric for evaluating coverage of information and pertinence of the discourse based on the use of generic and specific vocabulary in healthy and cognitively impaired individuals. We also evaluated other linguistic features, such as lexical richness and the use of specific linguistic patterns that could provide an insight into the types of syntactic structures that are

most affected by cognitive impairment. Our experiments were carried out with native speakers of Spanish and English to test the multilingual robustness of our proposed metrics.

Our new proposed metric for coverage of information solved the biggest limitation of the metric proposed in **Chapter 2**. Instead of using a part of the healthy cohort to create the referent, we first extracted the general vocabulary of a healthy older population from two free-discourse corpora, one in Spanish and one in English. By contrasting the most-used vocabulary of the free-discourse corpora against the description tasks, we were able to extract the specific vocabulary for each task in its respective language. We then measured how many instances of the specific vocabulary participants were covering during their descriptions (information coverage), and how much of their vocabulary corresponded to task-specific vocabulary (pertinence).

We used these features, along with lexical richness measures and specific linguistic patterns to train a support vector machine and random forests learners. We found that our newly proposed metrics of information coverage and pertinence based on the use of specific vocabulary were the highest correlated with the severity of cognitive impairment for both types of tasks. For both corpora, the best results were obtained with the support vector machine learner with linear kernels. On a 10-fold cross-validation experiment for classifying AD patients from healthy controls, we obtained an average F-score and area under the curve of 0.98 for the object description tasks, and an average F-score and area under the curve of 0.83 for the standard picture-description task. Our results compared favorably to those of the state-of-the-art methods for both tasks, and to those of our pervious works.

We corroborated that our high results for the object description task were not caused by an overfitting problem by testing the classification without performing hyper-parameter tuning. We obtained an AUC of 0.95 using the default parameters of the SVM implementation. Also, previous literature has reported AUC results of 0.96 and 0.97 on the same corpus. The apparently vast difference in the performance of our classifiers between standardized picture

description and object description tasks had more to do with the tasks and the characteristics of the cohorts in the corpus than with the features themselves.

From our experiments, we observed that the task of describing six common objects without any visual stimulus was highly taxing for patients with cognitive impairment, with over a tenth of the participants with Alzheimer's disease being unable to complete the task. Furthermore, the education levels and conditions of the settings in which the task was performed favored the healthy cohort considerably. There was a significantly higher level of interaction between participants and examiners for the healthy cohort, which produced, on average, descriptions three times longer than those of the Alzheimer's group. Unfortunately, all these factors made it difficult to compare the influence of the features in distinguishing AD patients from healthy controls in the contexts of standardized picture description tasks and of object descriptions.

We were able, however, to observe some phenomena related to the use of parts of speech that were consistent both in English and Spanish speakers. A significantly increased use of pronouns and nouns without verbs by the AD cohorts is a finding that has been previously detected for AD English and Portuguese speakers, and that we also found now in AD Spanish speakers. However, we also found that the variety of syntactic structures in the discourse for the object description task was still partially restricted due to the nature of the discourse and did not allow us to make deeper observations in various linguistic patterns. This motivates our research progress into the study of spontaneous conversations to evaluate differences in these types of structures.

To evaluate the adeptness of our metrics in detecting signs of AD at one of its earlier stages, we performed a classification of MCI patients and healthy controls using the standardized picture description corpus. We obtained an average AUC of 0.79, an F-score of 0.80, and 0.79 accuracy for this task. This is an important improvement from the previous literature, which reported an accuracy of 0.65, since detecting MCI is challenging, even for specialists.

6.3 Longitudinal characterization of language alterations in spontaneous speech

In **Chapter 4**, we presented a work assessing changes in language functions from spontaneous conversations. In that work, we used a corpus containing conversations of older French speakers that were followed up for ten years. Since the original intent of this corpus was to observe the changes in the language of healthy older speakers, all participants in the corpus started as cognitively healthy individuals. However, over the years, four of them developed some form of cognitive impairment (CI), including Alzheimer's disease. Despite their diagnoses, these participants were part of the sample during all recollections.

For our work on characterization of language changes, we extracted lexical richness measures, performed sentiment and subjectivity analyses, computed part-of-speech ratios and measured elements of speech fluency. Inspired by our findings in **Chapter 3**, we also performed an analysis of distribution of vocabulary, especially of differences between general, specific and idiosyncratic vocabularies. We estimated these features for all the transcriptions of conversations of the four participants of the CI group, and we compared their behavior against four age-, gender-, education-, profession- and bilingualism-matched controls that remained apparently healthy during the 10-year span of the recollections.

We observed the individual behaviors of these measures through time and were able to observe that most presented low values for the CI group, even at the time of the first interviews, when all participants were apparently healthy. The variables that were the most different between both, even from the first interviews, were lexical richness and our proposed metrics based on vocabulary distribution.

After performing a principal component analysis, we obtained two components that were able to clearly differentiate the interviews of the four participants after the time of diagnosis. Moreover, these components separated transcriptions of participants that were healthy at the time of the interview, but that were part of the group that developed some form of CI up to ten years later.

To the best of our knowledge, this is the first work that has studied the changes in language functions of older speakers as they transition from a healthy to a cognitively impaired diagnosis, using spontaneous conversations.

6.4 The Latin-American cohort of the Carolinas' Conversation Collection

Perhaps the biggest limitation in performing longitudinal analysis on changes in language functions from the spontaneous speech of older speakers is the scarcity of available datasets, especially in languages other than English. A parallel contribution of this doctoral research was oriented to aiding solve this problematic with the protocol design and creation of the Latin-American cohort of the Carolinas' Conversation Collection (CCC). This contribution was presented in Chapter 5, and consisted of the inclusion of recollections of spontaneous conversations with older Spanish speakers from Ecuador and Mexico to the existing CCC.

These recollections started in 2015 for the Ecuadorian cohort, and in 2016 for the Mexican one. They are the product of the combined efforts of École de technologie supérieure, the Medical University of South Carolina, the University of North Carolina at Charlotte, Universidad Nacional Autónoma de México, the Psychogeriatric unit of the Psychiatric Hospital "Fray Bernardino Álvarez" in Mexico, and the "Perpetuo Socorro" home for the elderly in Quito, Ecuador. So far, the recollections include recordings of spontaneous conversations with older speakers, but there is a plan to also include a standardized picture description task.

For this cohort, we incorporated information about the medical history of the participants, as well as their current diagnoses and medications. As a part of the CCC, this dataset will be available for research purposes upon request and after a formal approval by the corresponding Ethics committees. This dataset could serve for future longitudinal multimodal studies on language and communication with older people.

CONCLUSION AND FUTURE WORK

Alzheimer's disease (AD) is often diagnosed at the dementia stage of its continuum, which appears several years after onset. The most anticipated research advancement on the disease is the finding of treatments to stop or delay its progression. However, when available, these treatments will require that the disease be detected at its earliest. Hence, considerable effort is being invested in the identification of early AD biomarkers.

Some biomarkers have shown promising results in helping in the early diagnosis of Alzheimer's disease. However, most of these biomarkers, such as the extraction of cerebrospinal fluid and magnetic resonance imaging (MRI), are invasive and expensive, particularly for the elderly population. For this reason, these biomarkers should be regarded as tools to support and confirm the diagnosis, but not as regular monitoring mechanisms.

The study of language alterations can be an inexpensive and noninvasive method for detecting early signs of Alzheimer's disease, since changes in language functions are detected years before the dementia stage. However, there are many factors, including socio-economic and cultural circumstances, that can determine the linguistic characteristics of each individual. Consequently, it is important to characterize the changes in language functions that patients undergo through time, as they transition from a healthy cognition to dementia. This could help in the design of monitoring tools that could detect patient-specific changes that could alert of the presence of Alzheimer's disease.

In this doctoral work, we presented a computer-based approach for assessing language functions in two contexts: description tasks and spontaneous conversations. Moreover, we presented a method for an automated evaluation of patients' performance at picture description tasks by analyzing the informativeness and pertinence of their descriptions.

In the context of picture description tasks, we obtained an average F-score and AUC of 0.83 on a 10-fold cross validation SVM classification of AD patients and healthy controls. This was

done by combining lexical richness features and our proposed metrics for information coverage and pertinence. Our results compared favorably to state-of-the-art methods that relied on manually-extracted lists of information content units to assess the informativeness of descriptions. When we performed a similar experiment, but classifying MCI patients and healthy controls, we obtained an F-score of 0.80 and an AUC of 0.79. These are very encouraging results since detecting MCI is a challenging task, even for clinicians, and MCI is often considered a potential precursor of Alzheimer's disease. The state-of-the-art method that has worked with the MCI cohort of the dataset reported an accuracy of 0.65.

We observed that our informativeness and pertinence metrics were the features that correlated the most with the severity of cognitive impairment, and were always selected during the feature selection processes as relevant. Since the nature of the description tasks limits the variety of syntactic structures and vocabulary in the discourse, the features based on these factors were not always consistently correlated with CI in this context.

In order to characterize the changes in language functions that occur as patients transition from a cognitively healthy status to a form of cognitive impairment, we proposed an automated method for analyzing transcriptions of spontaneous conversations. We proposed the evaluation of the distribution of participants' vocabulary with respect to their use of specific and generic words and phrases. We also estimated features such as lexical richness, part-of-speech ratios, specific syntactic structures and speech fluency.

We computed our proposed metrics in the transcriptions of four older French speakers that started as cognitively healthy, but over a ten-year span went on to develop some form of cognitive impairment, including AD. We observed the longitudinal behavior of our metrics, and contrasted it with those of four age-, gender-, education- and profession-matched participants that remained apparently healthy throughout the same period. We found that there were significant differences in our metrics and in their longitudinal behavior between transcriptions of healthy and CI individuals. Moreover, when applying a principal component analysis, combining our metrics into two components enable the differentiation between the

transcriptions belonging to healthy individuals from those of the CI group up to ten years before the time of diagnosis.

We believe that the combination of analyses of standardized picture description tasks and spontaneous conversations could provide a thorough, inexpensive and noninvasive mechanism to help in the monitoring of patients. Our methods could be applied in a context of comparison of patients' metrics against similar cohorts, but they also could be used in personalized longitudinal analyses to observe changes within patients through time.

The research in this area is still at its early stages, and there are multiple research avenues that are still unexplored. For instance, all the analyses in this work relied on manual transcriptions of speech. The use of dentures and changes in vocalization and pitch with age make the use of generic tools for automatic speech recognition difficult, especially in languages other than English. The creation of tools tailored to this specific population could greatly facilitate the inclusion of our analyses in clinical practice.

Multi-modal studies that evaluate not only verbal expression, but also paralinguistic cues, corporal and facial expression and eye movement, could allow a more complete assessment of communication alterations that could alert of a pathological aging process early on. Furthermore, these studies could also be used to improve communication strategies with older people, with and without cognitive disease: research that could culminate in an improvement in their quality of life.

APPENDIX I

Table-A I-1 Full list of n-grams that were significantly correlated ($p < .05$) with the severity of CI.

Specific n-grams	***r***	***p***
pas/advb (negation aux.)	0.751	<0.001
non/advb (no)	0.742	<0.001
être/verb (to be)	0.678	<0.001
rien/pron (nothing)	0.696	<0.001
même/advb (even)	0.683	<0.001
être_pas/verb+advb (not + to be)	0.681	<0.001
ce_que/conj (that)	0.639	<0.01
je_savoir/pron+verb (I + to know)	0.64	<0.01
loin/advb (far away)	0.609	<0.01
á_un/prep+det (at one)	**-0.580**	<0.01
je_être/pron+verb (I + to be)	0.556	<0.05
que_être/pron+verb (that + to be)	0.552	<0.05
faire_de_le/verb+prep+det (to do/make)	0.553	<0.05
gens/noun (people)	0.520	<0.05
ce/det (that)	**-0.520**	<0.05
le_gens/det+noun (the + people)	**-0.516**	<0.05
je_aller/pron+verb (I + to go)	**-0.516**	<0.05
je_vouloir/pron+verb (I + to want)	**-0.509**	<0.05
je_penser/pron+verb (I + to think)	0.51	<0.05
avoir_de/verb+prep (to have)	0.506	<0.05
qui/pron (who/which)	0.499	<0.05
á/prep (at)	**-0.497**	<0.05
et/conj (and)	**-0.499**	<0.05
faire_de/verb+prep (to do/make)	0.494	<0.05
enfin/advb (finally)	0.488	<0.05
mais/conj (but)	0.485	<0.05
mais_ce/conj+pron (but that)	0.485	<0.05
ce_être/pron+verb (that + to be)	0.481	<0.05
nous_avoir/pron+verb (we + to have)	0.469	<0.05
dans/prep (in)	**-0.467**	<0.05
se/pron (reflexive pronoun)	**-0.465**	<0.05
mais_ce_être/conj+pron+verb (but + that + to be)	**-0.457**	<0.05

APPENDIX II

Publications during PhD studies

- **Laura Hernández-Domínguez**, Sylvie Ratté, Annette Gerstenberg and Gerardo Sierra-Martínez. "Aging with and without cognitive diseases: characterizing 10 years of language differences in French elderly speakers". In *Computer Speech and Language in Biology and Medicine*. Under review.

- **Laura Hernández-Domínguez**, Sylvie Ratté and Gerardo Sierra-Martínez. "Automated differentiation of Alzheimer's and MCI patients from healthy controls using English and Spanish transcriptions of description tasks". *Computers in Biology and Medicine*. Under review.

- **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra-Martínez and Andrés Roche-Bergua. "Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task". In *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 10, pp. 260–268, 2018.

- **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra-Martínez, Andrés Roche-Bergua and Janet Jiménez-Genchi. "El Proyecto Cécilia: estudios del lenguaje en pacientes con demencia tipo Alzheimer". In *Psicogeriatría: Temas selectos*. Part 2, Geriatric Psychiatry, pp. 353–363. 2017.

- Sylvie Ratté, **Laura Hernández-Domínguez**, Andrés Roche-Bergua, Gerardo Sierra-Martínez and Boyd Davis. "Cécilia Project: an international multidisciplinary collaboration on the study of language in later life". In the *3rd International Conference CLARE, Encounters in Language and Aging Research*, Berlin, Germany. 2017.

- **Laura Hernández-Domínguez**, Edgar García-Cano, Sylvie Ratté and Gerardo Sierra-Martínez. "Detection of Alzheimer's disease based on automatic analysis of common objects' descriptions". In the *Association for Computational Linguistics' 7th Workshop on Cognitive Aspects of Computational Language Learning*, Berlin, Germany, pp. 10–15. 2016.

- **Laura Hernández-Domínguez**, Sylvie Ratté, Charlene Pope and Boyd Davis. “Conversing with the elderly in Latin America: a new cohort for multimodal, multilingual longitudinal studies on aging.”. In the *ACL’s 7th Workshop on Cognitive Aspects of Computational Language Learning*, Berlin, Germany, pp. 16–21. 2016.

- **Laura Hernández-Domínguez**, Sylvie Ratté, Boyd Davis and Charlene Pope. "New contributions to the Carolinas Conversations Collection: A comprehensive dataset for research on language alterations in the elderly". In *Proceedings of the Mid-Atlantic Student Colloquium on Speech, Language and Learning*, Philadelphia. 2016.

- **Laura Hernández-Domínguez** and Sylvie Ratté. "Automatic analysis of language alterations for early detection of Alzheimer's disease". In *2nd International Forum of Mexican Talent, Innovation Match*, Mexico City. 2017.

- **Laura Hernández-Domínguez**, Sylvie Ratté, Gerardo Sierra and Andrés Roche-Bergua. "Automatic detection of Alzheimer’s from picture descriptions". In *Substance ÉTS*. 2018. In Press.

- Anayeli Paulino, Gerardo Sierra, Iria da Cunha, **Laura Hernandez-Dominguez** and Gemma Bel. "Detection of rhetorical relations in Alzheimer's patients' speech and healthy elderly subjects: an approach from the RST". In the *19th International Conference on Computational Linguistics and Intelligent Text Processing*. 2018.

- Anayeli Paulino, Gerardo Sierra, **Laura Hernandez-Domínguez**, Iria da Cunha and Gemma Bel-Enguix. "Rhetorical Relations in the Speech of Alzheimer’s Patients and Healthy Elderly Subjects: An Approach from the RST". In *Computación y Sistemas*, 22(3), pp. 895–905. 2018.

- Anayeli Paulino, Gerardo Sierra, **Laura Hernandez-Domínguez** and Iria da Cunha. “Hacia la aplicación de las relaciones retóricas en la identificación de la Demencia tipo Alzheimer en adultos mayores”. In *Signos*, 52, Alteraciones y deterioro de la competencia lingüística en la enfermedad de Alzheimer, 2019. Under review.

BIBLIOGRAPHY

Adams, D., & Oliver, C. (2011). The expression and assessment of emotions and internal states in individuals with severe or profound intellectual disabilities. *Clinical Psychology Review*, *31*(3), 293–306. http://doi.org/10.1016/j.cpr.2011.01.003

Adlam, A.-L. R., Patterson, K., Bozeat, S., & Hodges, J. R. (2010). The Cambridge Semantic Memory Test Battery: Detection of semantic deficits in semantic dementia and Alzheimer's disease. *Neurocase*, *16*(3), 193–207. http://doi.org/10.1080/13554790903405693

Ahmed, S., de Jager, C. A., Haigh, A.-M., & Garrard, P. (2013). Semantic processing in connected speech at a uniformly early stage of autopsy-confirmed Alzheimer's disease. *Neuropsychology*, *27*(1), 79–85. http://doi.org/10.1037/a0031288

Ahmed, S., Haigh, A.-M. F., De Jager, C. A., & Garrard, P. (2013). Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease.pdf. *Brain*, (136), 3727–3737.

Alegria, R., Bolso, M., Gallo, C., Prisco, C. R., Bottino, C., & Ines, N. M. (2013). Retained lexis in people with Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *9*(4), P486–P487. http://doi.org/10.1016/j.jalz.2013.05.993

Alegria, R., Bottino, C., & Ines, N. M. (2011). Why do Alzheimer's disease patients have more ideological and sociolinguistic lexical items preserved?: Improving communication between patients and caregivers. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *7*(4), S622. http://doi.org/10.1016/j.jalz.2011.05.1778

Alegria, R., Gallo, C., Bolso, M., dos Santos, B., Prisco, C. R., Bottino, C., & Ines, N. M. (2013). Comparative study of the uses of grammatical categories: Adjectives, adverbs, pronouns, interjections, conjunctions and prepositions in patients with Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *9*(4), P882. http://doi.org/10.1016/j.jalz.2013.08.233

Alegria, R. P., Ferreira, R. B., Marques, R. C. G., Bottino, C. M. C., & Nogueira, M. I. (2010). Discourse analysis of Alzheimer's disease patients: From the lexicon to discourse (Vol. 6, p. S337). The Alzheimer's Association. http://doi.org/10.1016/j.jalz.2010.05.1128

Alegria, R. P., Perroco, T. R., Marques, R. C. G., Barbosa, M. A., & Bottino, C. M. C. (2008). A pilot study of the lexical aspects of the oral discourse in patients with Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *4*(4), T651–T652.

Alegria, R. P., Perroco, T. R., Marques, R. de C. G., Nogueira, M. I., & Bottino, C. M. C. (2009). Comparison of brazilian portuguese lexical production of Alzheimer's disease patients and healthy elderly subjects. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, *5*(4), P232.

Almeida, R. P., Schultz, S. A., Austin, B. P., Boots, E. A., Dowling, N. M., Gleason, C. E., … Okonkwo, O. C. (2015). Effect of cognitive reserve on age-related changes in

cerebrospinal fluid biomarkers of Alzheimer disease. *JAMA Neurology*, *72*(6), 699–706. http://doi.org/10.1001/jamaneurol.2015.0098

Alzheimer's Association. (2015). *2015 Alzheimer's Disease Facts and Figures*. Retrieved from https://www.alz.org/facts/downloads/facts_figures_2015.pdf

Alzheimer's Association. (2018a). 2018 Alzheimer's Disease Facts and Figures. *Alzheimer's & Dementia*, *14*(3), 367–429. http://doi.org/10.1016/j.jalz.2018.02.001

Alzheimer's Association. (2018b). *2018 Alzheimer's disease facts and figures*. *Alzheimer's & Dementia* (Vol. 14).

Alzheimer's Association Research Center. (2016). Alzheimer's and Dementia Testing for Earlier Diagnosis: biomarkers for earlier detection.

Alzheimer's Disease International, Prince, M., Wimo, A., Guerchet, M., Ali, G.-C., Wu, Y., & Prina, M. (2015). *World Alzheimer Report 2015. The Global Impact of Dementia.* Retrieved from http://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf

Alzheimer Society of Canada. (2014). What is Alzheimer's disease. Retrieved from http://www.alzheimer.ca

Arkin, S. (2007). Language-enriched exercise plus socialization slows cognitive decline in Alzheimer's disease. *American Journal of Alzheimer's Disease and Other Dementias*, *22*(1), 62–77.

Asgari, M., Kaye, J., & Dodge, H. (2017). Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *3*(2), 219–228. http://doi.org/10.1016/j.trci.2017.01.006

Asgari, M., Kaye, J., Mattek, N., & Dodge, H. H. (2015). Detecting mild cognitive impairment (MCI) in older adults from content of spoken utterances. *Alzheimer's & Dementia*. http://doi.org/10.1016/j.jalz.2015.06.981

Ash, S., Evans, E., O'Shea, J., Powers, J., Boller, A., Weinberg, D., … Grossman, M. (2013). Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology*, *81*(4), 329–336. http://doi.org/10.1212/WNL.0b013e31829c5d0e

Baumgart, M., Snyder, H. M., Carrillo, M. C., Fazio, S., Kim, H., & Johns, H. (2015). Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective. *Alzheimer's & Dementia*, *11*(6), 718–726. http://doi.org/10.1016/j.jalz.2015.05.016

Beber, B. C., da Cruz, A. N., & Chaves, M. L. (2015). A behavioral study of the nature of verb production deficits in Alzheimer's disease. *Brain and Language*, *149*, 128–134. http://doi.org/10.1016/j.bandl.2015.07.010

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

Bekris, L. M., Yu, C. E., Bird, T. D., & Tsuang, D. W. (2010). Genetics of Alzheimer disease. *Journal of Geriatric Psychiatry and Neurology*, *23*(4), 213–227.

http://doi.org/10.1177/0891988710383571

Bolly, C., & Boutet, D. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life. *Corpora*, *13*(2).

Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. *Frontiers in Psychology*, *8*(MAR). http://doi.org/10.3389/fpsyg.2017.00269

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. http://doi.org/10.1023/A:1010933404324

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux: structure et évolution: statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue*. Slatkine.

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. *Aphasiology*, *14*(1), 71–91. http://doi.org/10.1080/026870300401603

Burns, A., Jacoby, R., Luthert, P., & Levy, R. (1990). Cause of death in alzheimer's disease. *Age and Ageing*, *19*(5), 341–344. http://doi.org/10.1093/ageing/19.5.341

Burton, K. W., & Kaszniak, A. W. (2006). Emotional experience and facial expression in Alzheimer's disease. *Aging, Neuropsychology and Cognition*, *13*(3–4), 636–651. http://doi.org/10.1080/13825580600735085

Cardona, J. F., Gershanik, O., Gelormini-Lezama, C., Houck, A. L., Cardona, S., Kargieman, L., … Ibáñez, A. (2013). Action-verb processing in Parkinson's disease: new pathways for motor-language coupling. *Brain Structure & Function*, *218*(6), 1355–73. http://doi.org/10.1007/s00429-013-0510-1

Cattell, R. B. (1966). *The scree test for the number of factors*. *Multivariate Behavioral Research* (Vol. 1). Taylor & Francis. http://doi.org/10.1207/s15327906mbr0102_10

Cockrell, J. R., & Folstein, M. F. (2002). *Mini-mental state examination*. (D. G. Copeland, John R. M.; Abou-Saleh, Mohammed T.; Blazer, Ed.)*Principles and Practice of Geriatric Psychiatry* (2nd ed.). John Wiley & Sons.

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative Study of Oral and Written Picture Description in Patients with Alzheimer's Disease. *Brain and Language*, *53*(1), 1–19. http://doi.org/10.1006/brln.1996.0033

Davis, B. (2005). *Alzheimer talk, text, and context: Enhancing communication*.

Davis, B., Maclagan, M., Karakostas, T., Liang, S., & Shenk, D. (2011). Watching what you say: walking and talking in dementia. *Topics in Geriatric Rehabilitation*, *27*(4), 268–277.

De Renzi, E., & Vignolo, L. A. (1962). The token test: A sensitive test to detect receptive disturbances in aphasics. *Brain*, *85*(4), 665–678.

Drummond, C., Coutinho, G., Fonseca, R. P., Assunção, N., Teldeschi, A., de Oliveira-Souza,

R., … Mattos, P. (2015). Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Frontiers in Aging Neuroscience*, *7*(96). http://doi.org/10.3389/fnagi.2015.00096

Dubois, B., Padovani, A., Scheltens, P., Rossi, A., Dell'Agnello, G., & Dell'Agnello, G. (2016). Timely diagnosis for Alzheimer's disease: a literature review on benefits and challenges. *Journal of Alzheimer's Disease*, *49*(3), 617–631. http://doi.org/10.3233/JAD-150692

Forbes-McKay, K. E., & Venneri, A. (2005). Detecting subtle spontaneous language decline in early Alzheimer's disease with a picture description task. *Neurological Sciences*, *26*(4), 243–254.

Forbes, K. E., Venneri, A., & Shanks, M. F. (2002). Distinct patterns of spontaneous speech deterioration: an early predictor of Alzheimer's disease. *Brain and Cognition*, *48*, 356–361. http://doi.org/10.1006/brcg.2001.1377

Fraser, K. (2016). *Automatic text and speech processing for the detection of dementia*. University of Toronto.

Fraser, K., & Hirst, G. (2016). Detecting semantic changes in Alzheimer's disease with vector space models. *Proceedings of LREC 2016 Workshop: Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)*, (May), 1–8.

Fraser, K., Meltzer, J., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. http://doi.org/10.3233/JAD-150520

Freudenberg, M., Adams, R. B., Kleck, R. E., & Hess, U. (2015). Through a glass darkly: facial wrinkles affect our processing of emotion in the elderly. *Frontiers in Psychology*, *6*.

Garrard, P., & Forsyth, R. (2010). Abnormal discourse in semantic dementia: A data-driven approach. *Neurocase*, *16*(6), 520–528. http://doi.org/10.1080/13554791003785901

Garrard, P., Rentoumi, V., Gesierich, B., Miller, B., & Gorno-Tempini, M. L. (2014). Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, *55*(1), 122–129. http://doi.org/10.1016/j.cortex.2013.05.008

Gerstenberg, Annette Hekkel, V., & Kairet, J. (2018). *Corpus LangAge: Transcription Guide*. Retrieved from https://www.uni-potsdam.de/langage/guide/guide_full.pdf

Gerstenberg, A. (n.d.). *LangAge corpora*. Retrieved from www.langage-corpora.org

Gerstenberg, A. (2011). Generation und Sprachprofile im höheren Lebensalter: Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews. *Analecta Romanica*, *76*.

Gerstenberg, A. (2015). A Sociolinguistic Perspective on Vocabulary Richness in a Seven-Year Comparison of Older Adults. *Language Development: The Lifespan Perspective*, *37*, 109–127.

Goberman, A. M., Blomgren, M., & Metzger, E. (2010). Characteristics of speech disfluency

in Parkinson disease. *Journal of Neurolinguistics*, *23*(5), 470–478.

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 517–520 vol.1). Ieee. http://doi.org/10.1109/ICASSP.1992.225858

Gonzalez-Moreira, E., Torres-Boza, D., Garcia-Zamora, M. A., Ferrer-Riesgo, C. A., & Hernandez-Gomez, L. A. (2014). Prosodic speech analysis to identify mild cognitive impairment. *VI Latin American Congress on Biomedical Engineering (CLAIB)*, 580–583.

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger.

Goral, M., Spiro, A., Albert, M. L., Obler, L. K., & Tabor Connor, L. (2007). Change in lexical retrieval skills in adulthood. *The Mental Lexicon*, *2*(2), 215–238. http://doi.org/10.1075/ml.2.2.05gor

Grasso, L., Díaz-Mardomingo, M. C., & Peraita-Adrados, H. (2011). Deterioro de la memoria semántico-conceptual en pacientes con enfermedad de Alzheimer. Análisis cualitativo y cuantitativo de los rasgos semánticos. *Psicogeriatría*, *3*(4), 159–165. Retrieved from http://www.viguera.com/sepg/pdf/revista/0304/304_0159_0165.pdf

Guerrero, J. M., Martínez-Tomás, R., Rincón, M., & Peraita-Adrados, H. (2015). Bayesian Network Model to Support Diagnosis of Cognitive Impair-ment Compatible with an Early Diagnosis of Alzheimers Disease. *Methods of Information in Medicine*.

Guerrero, J. M., Martínez-Tomás, R., Rincón, M., & Peraita, H. (2015). Diagnosis of Cognitive Impairment Compatible with Early Diagnosis of Alzheimer's Disease. *Methods of Information in Medicine*, *55*(1), 42–49. http://doi.org/10.3414/ME14-01-0071

Guinn, C., & Habash, A. (2012a). Language analysis of speakers with dementia of the Alzheimer's type. In *Association for the Advancement of Artificial Intelligence Fall Symposium* (pp. 8–13). AAAI.

Guinn, C., & Habash, A. (2012b). Language Analysis of Speakers with Dementia of the Alzheimer's Type. In *Association for the Advancement of Artificial Intelligence Fall Symposia: Artificial Intelligence for Gerontechnology* (pp. 8–13). AAAI.

Guinn, C., Singer, B., & Habash, A. (2014). A comparison of syntax, semantics, and pragmatics in spoken language among residents with Alzheimer's disease in managed-care facilities. In *2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)* (pp. 98–103). IEEE. http://doi.org/10.1109/CICARE.2014.7007840

Habash, A. (2012). *Language analysis of speakers with dementia of the Alzheimer's type*. University of North Carolina Wilmington.

Hakkani-Tür, D., Vergyri, D., & Tur, G. (2010). Speech-based automated cognitive status assessment. In *INTERSPEECH-2010* (pp. 258–261).

Hamm, J., Pinkham, A., Gur, R. C., Verma, R., & Kohler, C. G. (2014). Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia.

Schizophrenia Research and Treatment, *2014*.

Henry, J. D., Rendell, P. G., Scicluna, A., Jackson, M., & Phillips, L. H. (2009). Emotion Experience, Expression, and Regulation in Alzheimer's Disease. *Psychology and Aging*, *24*(1), 252–257. http://doi.org/10.1037/a0014001

Hernández-Domínguez, L., García-Cano, E., Ratté, S., & Sierra-Martínez, G. (2016). Detection of Alzheimer's disease based on automatic analysis of common objects descriptions. In ACL (Ed.), *Association for Computational Linguistics' 7th Workshop on Cognitive Aspects of Computational Language Learning* (pp. 10–15). Berlin, Germany.

Hernandez-Dominguez, L., Ratte, S., Pope, C., & Davis, B. (2016). Conversing with the elderly in Latin America: a new cohort for multimodal, multilingual longitudinal studies on aging. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning* (pp. 16–21).

Hernández-Domínguez, L., Ratté, S., Sierra-Martínez, G., & Roche-Bergua, A. (2018). Computer-based evaluation of Alzheimer's disease and mild cognitive impairment patients during a picture description task. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *10*, 260–268. http://doi.org/10.1016/J.DADM.2018.02.004

Hernández Domínguez, L., Ratté, S., Pope, C., & Davis, B. (2016). New contributions to the Carolinas Conversations Collection: A comprehensive dataset for research on language alterations in the elderly. In *Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL)*. Philadelphia.

Hess, T. M., Hinson, J. T., & Hodges, E. A. (2009). Moderators of and mechanisms underlying stereotype threat effects on older adults' memory performance. *Experimental Aging Research*, *35*(2), 153–177. http://doi.org/10.1080/03610730802716413

Hier, D. B., Hagenlocker, K., & Shindler, A. G. (1985). Language disintegration in dementia: Effects of etiology and severity. *Brain and Language*, *25*, 117–133. http://doi.org/10.1016/0093-934X(85)90124-5

Hoffman, P., Meteyard, L., & Patterson, K. (2014). Broadly speaking: Vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*, *55*, 30–42. http://doi.org/10.1016/j.cortex.2012.11.004

Holmes, D. I., & Forsyth, R. S. (1995). The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, *10*(2), 111–127.

Holtgraves, T., Fogle, K., & Marsh, L. (2013). Pragmatic language production deficits in Parkinson's disease. *Advances in Parkinson's Disease*, *2*(1), 31–36.

Homan, C. M., Johar, R., Liu, T., Lytle, M., Silenzio, V., & Alm, C. O. (2014). Toward Macro-Insights for Suicide Prevention: Analyzing Fine-Grained Distress at Scale. In *Acl 2014* (p. 107). http://doi.org/10.3115/v1/W14-3207

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, *7*(2), 172–177.

Jarrold, W. L., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., &

Ogar, J. (2014). Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 27–37). Baltimore, Maryland. http://doi.org/10.3115/v1/W14-3204

Jarrold, W. L., Peintner, B., Yeh, E., Krasnow, R., Javitz, H. S., & Swan, G. E. (2010). Language Analytics for Assessing Brain Health: Cognitive Impairment, Depression and Pre-symptomatic Alzheimer's Disease. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 299–307.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). Boston naming test. Philadelphia, Pa: Lea & Febiger.

Kavé, G., & Goral, M. (2016). Word retrieval in picture descriptions produced by individuals with Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *38*(9), 958–966. http://doi.org/10.1080/13803395.2016.1179266

Kavé, G., & Goral, M. (2018). Word retrieval in connected speech in Alzheimer's disease: a review with meta-analyses. *Aphasiology*, *32*(1), 4–26. http://doi.org/10.1080/02687038.2017.1338663

Kavé, G., Goral, M., & Dassa, A. (2018). Severity of Alzheimer's disease and language features in picture descriptions. *Aphasiology*, *32*(1), 4–26. http://doi.org/10.1080/02687038.2017.1303441

Kavé, G., & Levy, Y. (2003). Morphology in Picture Descriptions Provided by Persons With Alzheimer's Disease. *Journal of Speech Language and Hearing Research*, *46*(2), 341–352. http://doi.org/10.1044/1092-4388(2003/027)

Kemper, S., LaBarge, E., Ferraro, R., Cheung, H., Cheung, H., & Storandt, M. (1993). On the preservation of syntax in Alzheimer's disease: Evidence from written sentences. *Archives of Neurology*, *50*(1), 81–86.

Khodabakhsh, A., & Demiroglu, C. (2015). Analysis of speech-based measures for detecting and monitoring Alzheimer's disease. *Methods in Molecular Biology (Clifton, N.J.)*, *1246*, 159–73. http://doi.org/10.1007/978-1-4939-1985-7_11

Khodabakhsh, A., Kusxuoglu, S., & Demiroglu, C. (2014). Natural language features for detection of Alzheimer's disease in conversational speech. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 581–584). IEEE. http://doi.org/10.1109/BHI.2014.6864431

Khodabakhsh, A., Yesil, F., Guner, E., & Demiroglu, C. (2015). Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 1–15. http://doi.org/10.1186/s13636-015-0052-y

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - ACL'03* (pp. 423–

430). Association for Computational Linguistics. http://doi.org/10.3115/1075096.1075150

Knopman, D. S., Boeve, B. F., & Petersen, R. C. (2003). Essentials of the proper diagnoses of mild cognitive impairment, dementia, and major subtypes of dementia. *Mayo Clinic Proceedings*, *78*(10), 1290–308. http://doi.org/10.4065/78.10.1290

König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., … David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *1*(1), 112–124. http://doi.org/10.1016/j.dadm.2014.11.012

Lai, Y. hsiu, Pai, H. hua, & Lin, Y. te. (2009). To be semantically-impaired or to be syntactically-impaired: Linguistic patterns in Chinese-speaking persons with or without dementia. *Journal of Neurolinguistics*, *22*(5), 465–475. http://doi.org/10.1016/j.jneuroling.2009.03.004

Lai, Y., & Lin, Y. (2012). Discourse markers produced by Chinese-speaking seniors with and without Alzheimer's disease. *Journal of Pragmatics*, *44*(14), 1982–2003. http://doi.org/10.1016/j.pragma.2012.09.002

Langa, K. M., & Levine, D. A. (2014). The Diagnosis and Management of Mild Cognitive Impairment. *JAMA*, *312*(23), 2551–2561. http://doi.org/10.1001/jama.2014.13806

Laske, C., Sohrabi, H. R., Frost, S. M., López-de-Ipiña, K., Garrard, P., Buscema, M., … O'Bryant, S. E. (2015). Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimer's & Dementia : The Journal of the Alzheimer's Association*, *11*(5), 561–78. http://doi.org/10.1016/j.jalz.2014.06.004

Lehr, M., Prud, E., Shafran, I., & Roark, B. (2012). Fully Automated Neuropsychological Assessment for Detecting Mild Cognitive Impairment. In *INTERSPEECH 2012 – 13th Annual Conference of the International Speech Communication Association* (pp. 1039–1042). Portland, Oregon.

Lehr, M., Shafran, I., Prud'hommeaux, E. T., & Roark, B. (2013). Discriminative Joint Modeling of Lexical Variation and Acoustic Confusion for Automated Narrative Retelling Assessment. In *HLT-NAACL* (pp. 211–220).

Lopez-de-Ipina, K., Martinez-de-Lizarduy, U., Barroso, N., Ecay-Torres, M., Martinez-Lage, P., Torres, F., & Faundez-Zanuy, M. (2015). Automatic analysis of Categorical Verbal Fluency for Mild Cognitive impartment detection: A non-linear language independent approach. In *2015 4th International Work Conference on Bioinspired Intelligence (IWOBI)* (pp. 101–104). IEEE. http://doi.org/10.1109/IWOBI.2015.7160151

López-de-Ipiña, K., Solé-Casals, J., Eguiraun, H., Alonso, J. B., Travieso, C. M., Ezeiza, A., … Beitia, B. (2015). Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, *30*(1), 43–60. http://doi.org/10.1016/j.csl.2014.08.002

Loria, S. (2013). textblob-fr 0.2.0. Retrieved from https://github.com/sloria/textblob-fr

Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., … Kolb, J. (2018).

TextBlob 0.15.1. Retrieved from https://github.com/sloria/TextBlob

Lyons, J. (2017). python_speech_features 0.6. Retrieved from https://pypi.python.org/pypi/python_speech_features

Marcos Marín, F. (1992). *Corpus Oral de Referencia de la Lengua Española Contemporánea CORLEC*. Madrid, Spain. Retrieved from http://www.lllf.uam.es/ESP/Corlec.html

McEnery, T., & Oakes, M. (2000). Authorship Identification and Computational Stylometry. In R. Dale, H. Somers, & H. Moisl (Eds.), *Handbook of Natural Language Processing*.

Miranda-García, A., & Calle-MartÍn, J. (2005). Yule's Characteristic K Revisited. *Language Resources and Evaluation*, *39*(4), 287–294. http://doi.org/10.1007/s10579-005-8622-8

Mortensen, D. R., Dalmia, S., & Littell, P. (2018). Epitran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2710–2714). Miyazaki, Japan: European Language Resources Association (ELRA).

Nicholas, M., Obler, L. K., Albert, M. L., & Helm-Estabrooks, N. (1985). EMPTY SPEECH IN ALZHEIMER'S DISEASE AND FLUENT APHASIA. *Journal of Speech and Hearing Research*, *28*, 405–410. http://doi.org/10.1044/jshr.2803.405

Orimaye, S. O., Tai, K. Y., Sze-Meng Wong, J., & Piau Wong, C. (2015). Learning Linguistic Biomarkers for Predicting Mild Cognitive Impairment using Compound Skip-grams. In *NIPS Workshop on Machine Learning in Healthcare*. Montreal, Canada.

Orimaye, S. O., Wong, J. S.-M., & Golden, K. J. (2014). Learning Predictive Linguistic Features for Alzheimer's Disease and related Dementias using Verbal Utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 78–87). http://doi.org/10.3115/v1/W14-3210

Orimaye, S. O., Wong, J. S. M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, *18*(1). http://doi.org/10.1186/s12859-016-1456-0

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In E. L. R. Association (Ed.), *Proceedings of the Language Resources and Evaluation Conference* (pp. 2473–2479). Istanbul, Turkey.

Pakhomov, S. V. S., Smith, G. E., Chacon, D., Feliciano, Y., Graff-Radford, N., Caselli, R., & Knopman, D. S. (2010). Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology : Official Journal of the Society for Behavioral and Cognitive Neurology*, *23*(3), 165–177. http://doi.org/10.1097/WNN.0b013e3181c5dde3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=2078195

Peraita Adrados, H., González Labra, M. J., Sanchez Bernardos, M. L., & Galeote Moreno, M. A. (2001). Evaluation battery for semantic memory deterioration in Alzheimer.

Psychology in Spain, *5*(1), 98–109.

Peraita, H., & Grasso, L. (2010). *Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad del alzheimer*. Retrieved from https://www2.uned.es/investigacion-corpuslinguistico/

Perneczky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J., & Kurz, A. (2006). Mapping scores onto stages: Mini-mental state examination and clinical dementia rating. *American Journal of Geriatric Psychiatry*, *14*(2), 139–44. http://doi.org/10.1097/01.JGP.0000192478.82189.a8

Platt, J. C., & Others. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. *Technical Report MSR-TR-98-14*.

Pope, C., & Davis, B. H. (2011). Finding a balance: The Carolinas Conversation Collection. *Corpus Linguistics and Linguistic Theory*, *7*(1), 143–161.

Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology*, *18*, 1075–1091. http://doi.org/10.1080/02687030444000534

Prud'hommeaux, E. T., & Roark, B. (2011). Extraction of Narrative Recall Patterns for Neuropsychological Assessment. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association* (pp. 3021–3024).

Rudzicz, F., Chan Currie, L., Danks, A., Mehta, T., & Zhao, S. (2014). Automatically Identifying Trouble-indicating Speech Behaviors in Alzheimer's Disease. *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, 241–242. http://doi.org/10.1145/2661334.2661382

Sabat, S. R. (1994). Language function in Alzheimer's disease: a critical review of selected literature. *Language and Communication*, *14*, 331–351.

Santos, L. B. dos, Corrêa, E. A., Oliveira, O. N., Amancio, D. R., Mansur, L. L., & Aluísio, S. M. (2017). Enriching Complex Networks with Word Embeddings for Detecting Mild Cognitive Impairment from Speech Transcripts. http://doi.org/10.18653/v1/P17-1118

Satt, A., Hoory, R., König, A., Aalten, P., Robert, P. H., Sophia, N., … Nice, C. H. U. De. (2014). Speech - Based Automatic and Robust Detection of Very Early Dementia. *Interspeech*, (September), 2538–2542. http://doi.org/10.13140/2.1.1258.8805

Schröder, J., Wendelstein, B., & Felder, E. (2010). Language in the Preclinical Stage of Alzheimer's Disease. Content and Complexity in Biographic Interviews of the ILSE Study. In *Klinische Neurophysiologie* (Vol. 41, p. S360).

Shinkawa, K., & Yamada, Y. (2018). Word Repetition in Separate Conversations for Detecting Dementia:A Preliminary Evaluation on Data of Regular Monitoring Service. *AMIA Joint Summits on Translational Science Proceedings*, *2017*, 206–215.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, *70*(351a), 542–547. http://doi.org/10.1080/01621459.1975.10482469

Singh, S. (1996). *Computational analysis of conversational speech of dysphasic patients.*

University of the West of England at Bristol.

Singh, S., & Bookless, T. (1997). Analysing spontaneous speech in dysphasic adults. *International Journal of Applied Linguistics*, *7*(2), 165–181. http://doi.org/10.1111/j.1473-4192.1997.tb00113.x

Slegers, A., Filiou, R.-P., Montembeault, M., & Brambati, S. M. (2018). Connected Speech Features from Picture Description in Alzheimer's Disease: A Systematic Review. *Journal of Alzheimer's Disease*, *65*(2), 519–542. http://doi.org/10.3233/JAD-170881

Smith, G. E., & Bondi, M. W. (2013). *Mild Cognitive Impairment and Dementia: Definitions, Diagnosis, and Treatment* (illustrate). OUP USA.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199–222. http://doi.org/10.1023/B:STCO.0000035301.49549.88

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Finds from the Nun Study. *Journal of the American Medical Association*, *275*(7), 528–532.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., … Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, *7*(3), 280–292. http://doi.org/10.1016/j.jalz.2011.03.003

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary*. (Department of Psychology. University of Victoria, Ed.) (Third Edit). Victoria, B.C., Canada: Oxford University Press.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in Aging Neuroscience*. Frontiers Research Foundation.

Taler, V., & Phillips, N. a. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of Clinical and Experimental Neuropsychology*, *30*(5), 501–56. http://doi.org/10.1080/13803390701550128

Thomas, C., Keselj, V., Cercone, N., Rockwood, K., & Asp, E. (2005). Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. *IEEE International Conference Mechatronics and Automation, 2005*, *3*(February 2005), 1569–1574. http://doi.org/10.1109/ICMA.2005.1626789

Tweedie, F. J., & Baayen, R. H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, *32*(5), 323–352. http://doi.org/10.1023/A:1001749303137

Velázquez-Godínez, E. (2017). *Caractérisation de la couverture d'information : Une approche computationnelle fondée sur les asymétries*. École de technologie supérieure,

Quebec University.

Villemagne, V. L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K. A., Salvado, O., … Masters, C. L. (2013). Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: A prospective cohort study. *The Lancet Neurology*, *12*(3), 357–367. http://doi.org/10.1016/S1474-4422(13)70044-9

Vincze, V. (2016). Detecting Mild Cognitive Impairment by Exploiting Linguistic Information from Transcripts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)* (Vol. 2, pp. 181–187).

Wankerl, S., Nöth, E., & Evert, S. (2016). An Analysis of Perplexity to Reveal the Effects of Alzheimer's Disease on Language. In *Speech Communication; 12. ITG Symposium* (pp. 1–5). Paderborn, Germany: VDE.

Ward, A., Tardiff, S., Dye, C., & Arrighi, H. M. (2013). Rate of Conversion from Prodromal Alzheimer's Disease to Alzheimer's Dementia: A Systematic Review of the Literature. *Dementia and Geriatric Cognitive Disorders Extra*, *3*(1), 320–332. http://doi.org/10.1159/000354370

Wendelstein, B., Felder, E., & Schröder, J. (2011). Language as an indicator of risk for Alzheimer's disease: Analysis of biographical interviews in preclinical stages. In *Alzheimer's & dementia: the journal of the Alzheimer's Association* (Vol. 7, p. S602).

Wilson, S. M., Henry, M. L., Besbris, M., Ogar, J. M., Dronkers, N. F., Jarrold, W., … Gorno-Tempini, M. L. (2010). Connected speech production in three variants of primary progressive aphasia. *Brain*, *133*(7), 2069–2088. http://doi.org/10.1093/brain/awq129

Yancheva, M., & Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer ' s disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 2337–2346). Berlin, Germany: ACL.

Yorkston, K. M., & Beukelman, D. R. (1980). An analysis of connected speech samples of aphasic and normal speakers. *Journal of Speech and Hearing Disorders*, *45*(1), 27–36. http://doi.org/10.1044/jshd.4501.27

Zhou, L., Fraser, K., & Rudzicz, F. (2016). Speech recognition in Alzheimer's disease and in its assessment. *Proceedings of the 17th Annual Meeting of the International Speech Communication Association (Interspeech)*, 1948–1952. http://doi.org/10.21437/Interspeech.2016-1228