TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AUC	Area Under the Curve
DCL	Déficience Cognitive Légère
DT	Decision Tree
LOO	Leave-One-Out
MA	Maladie d'Alzheimer
PCA	Principal Component Analysis
POS	Part-of-Speech
PS	Patient Sain
RFC	Random Forest Classifier
SOV	Sujet-Objet-Verbe
SVM	Support Vector Machine
SVO	Sujet-Verbe-Objet
TALN	Traitement Automatique de la Langue Naturelle
TF	Term Frequency
TFIDF	Term Frequency Inverse Document Frequency
UCI	Unité de contenu d'information

INTRODUCTION

La maladie d'Alzheimer (MA) affecte de plus en plus de personnes, puisque le pourcentage de la population âgé ne cesse d'augmenter d'année en année (Alzheimer's Association, 2018). Les patients atteints de cette maladie ont de la difficulté à effectuer leurs activités quotidiennes sans une assistance directe et donc il devient difficile et coûteux de s'occuper d'eux. Il est donc crucial de diagnostiquer cette maladie dégénérative de plus tôt possible afin de leur offrir les meilleurs outils pour ralentir sa progression. La littérature propose de plus en plus des solutions assistées par ordinateur afin d'offrir des méthodes de diagnostic moins invasives pour les patients et moins coûteuses pour l'État. Par exemple, certains chercheurs ont tenté d'analyser des transcriptions d'entrevue avec des patients dans le cadre du test de description d'image du Cookie Theft (Hernández-Dominguez et al., 2018; Fraser et al., 2016; Kavé et al., 2003) (Figure 1.1). Durant ce test, il est demandé au patient de décrire ce qu'il voit sur l'image présentée à la Figure 0.1. L'analyse des transcriptions permet d'évaluer les différentes fonctions linguistiques des patients. Plusieurs études ont prouvé que ces fonctions sont parmi les premières à être affecté par la maladie (Szatloczki, Hoffmann, Vincze, Kalman et Pakaski, 2015).

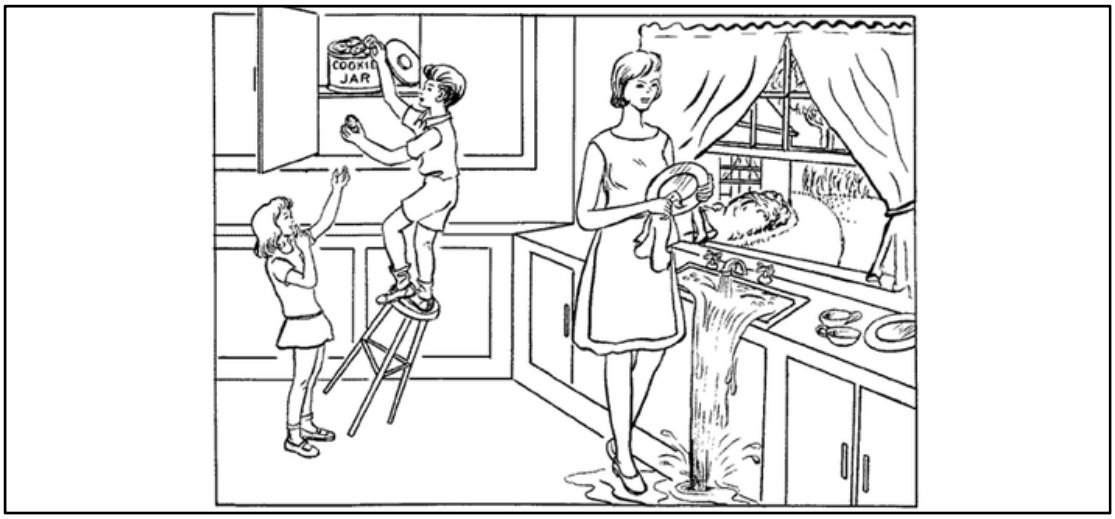

Figure 0.1 Description d'image du Cookie Theft provenant du Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983).

0.1 Problématique

Pour réussir à faire un passage au milieu clinique, il faudra automatiser le processus complet, en particulier, l'analyse comme telle des transcriptions. Lorsque l'on travaille avec des transcriptions, le prétraitement consiste en une étape cruciale. En effet, il est important d'assurer la cohérence et la qualité de nos données afin d'offrir une analyse éclairée. Actuellement, les chercheurs étudient des patients de différentes communautés et donc ils traitent des transcriptions de différentes langues. De plus, les chercheurs effectuent le prétraitement des données manuellement et utilisent parfois des standards d'annotations différents. Il est donc difficile de comparer les résultats lorsque le prétraitement des données diffère d'une étude à l'autre.

0.2 Objectif de recherche

Donc, l'objectif de cette recherche est de présenter un outil de prétraitement multilingue et adaptable à plusieurs contextes, afin de pallier la disparité des tâches de prétraitement et diminuer le temps accordé par les chercheurs au développement de ce type de tâche. Ce prétraitement consiste à nettoyer et normaliser les transcriptions, pour ensuite extraire des mesures permettant de caractériser les capacités linguistiques du patient. Un tel outil pourra servir de référence à la communauté et pourra être amélioré au fil du temps, afin de supporter les différences linguistiques et culturelles; il permettra ainsi d'améliorer la reproductibilité des expériences. Ensuite, à l'aide des mesures extraites durant la tâche de prétraitement, telles que la richesse lexicale ou la distribution des catégories lexicales (POS), il sera possible d'analyser et d'évaluer le niveau linguistique des patients. Ceci permettra aux médecins d'effectuer une analyse longitudinale des altérations linguistiques et phonétiques des patients et d'ainsi être en mesure de surveiller et diagnostiquer plus rapidement la maladie d'Alzheimer, tel qu'illustré à la Figure 0.2.

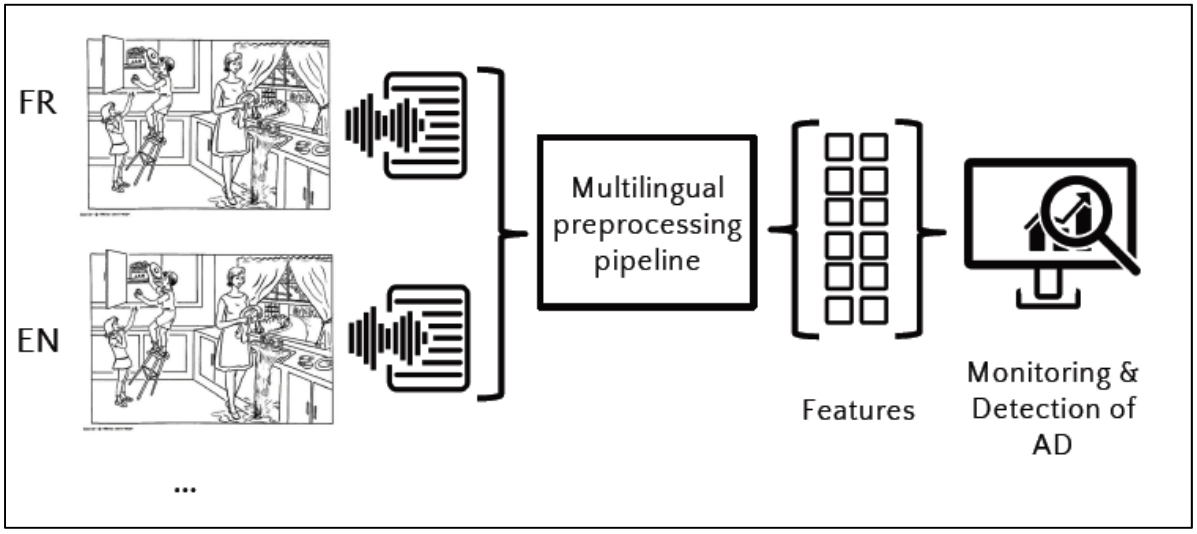

Figure 0.2 Processus multilingue d'extraction de mesures caractérisant le niveau linguistique de patients dans le cadre d'un test de description d'image dans la surveillance et la détection de la maladie d'Alzheimer.

0.3 Données

Afin d'évaluer l'approche proposée, deux corpus de données seront utilisés. Tout d'abord, le Pitt Corpus (anglais), contenant des patients sains (242) et des patients avec la MA ou une déficience cognitive légère (300), permettra de corroborer les résultats obtenus avec différentes études de la littérature. Ensuite, le corpus du CRIUGM (français), contenant de jeunes patients en santé (26) ainsi que des patients âgés (29), permettra de valider l'aspect multilingue du traitement et son efficacité à extraire le niveau linguistique des patients.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

L'objectif principal de cette revue littéraire est de permettre aux lecteurs de mieux comprendre certains concepts et hypothèses développés dans le cadre d'études précédentes. Nous allons donc commencer par décrire ce qu'est la maladie d'Alzheimer et ce qu'elle implique dans notre société d'aujourd'hui. Ensuite, nous ferons un survol des différentes approches récentes pour détecter la maladie. Nous terminons cette revue par une synthèse des différentes études.

1.2 Mise en contexte

1.2.1 Maladie d'Alzheimer

La maladie d'Alzheimer (MA) est une forme de démence qui affecte principalement les fonctions cognitives et qui se caractérise par une accumulation de plaques de la protéine **β**-Amyloïde dans certaines parties du cerveau (Smith et Bondi, 2013). Trois stades généraux ont été déterminés afin de distinguer l'évolution de la maladie. Le premier est la maladie d'Alzheimer préclinique, où des transformations au cerveau commencent à être perçues, mais les symptômes ne sont pas encore apparents. Ensuite vient le stade de la déficience cognitive légère (DCL) dû à la maladie d'Alzheimer, caractérisée par la transformation du cerveau et les symptômes de dégénérescences cérébrales visibles, mais qui n'affecte pas l'exécution des tâches quotidiennes des patients. Finalement, le stade de la démence, dû à la maladie d'Alzheimer, est le dernier stade, où les patients ne sont plus en mesure d'accomplir la majorité de leurs activités quotidiennes de manière autonome. De plus, la perte de mémoire et la confusion deviennent de plus en plus fréquentes et apparentes. Éventuellement, leur fonction motrice qui leur permet d'avaler devient défectueuse, qui en soi, est la cause principale de décès dû à la maladie (Alzheimer's Association, 2018). Les différents stades de la maladie sont caractérisés par l'augmentation des zones affectées dans le cerveau, tel qu'illustré à la

Figure 1.1 (le stade de la DCL se situe entre le stade asymptomatique et le stade léger à modéré de la MA).

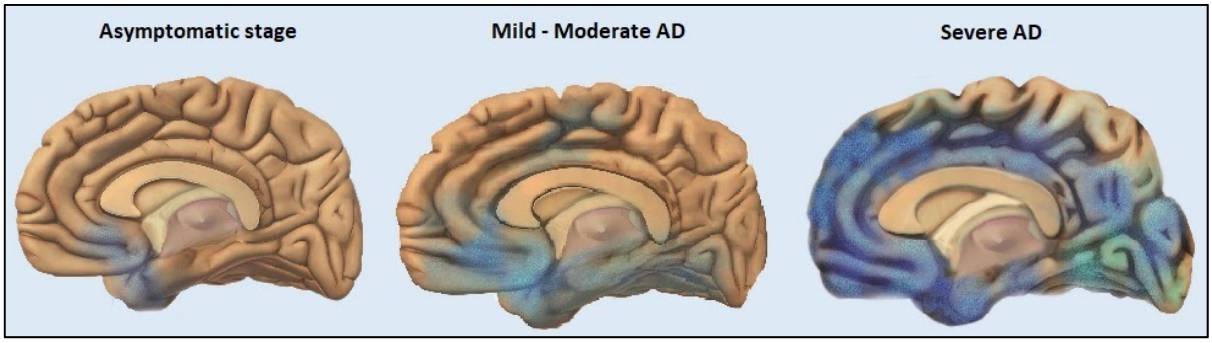

Figure 1.1 Évolution des changements histologiques causée par la maladie d'Alzheimer (Moustris, 2020).

Actuellement, différentes méthodes permettent de diagnostiquer la maladie. Par exemple, les médecins tentent d'analyser l'historique médical du patient et de sa famille (Alzheimer's Association, 2018). Pour l'instant, aucune preuve n'a pu démontrer qu'il y a une cause génétique absolue de la maladie. Toutefois, certains gènes seraient susceptibles ou auraient un lien de causalité avec le développement de la maladie d'Alzheimer. Entre autres, la protéine précurseur de l'amyloïde (APP) sur le chromosome 21 pourrait être en cause malgré que ces gènes ne représentent que 5% des cas répertoriés (Smith et Bondi, 2013). Ensuite, les médecins s'informent auprès des proches des patients pour déterminer quels changements sont perceptibles dans leurs quotidiens. En parallèle, des tests cognitifs et physiques, tel qu'un test de description d'image, permettent de mieux comprendre comment la maladie se développe chez le patient. Finalement, une série de tests sanguins et d'imageries médicales permettent d'approfondir l'analyse scientifique à la source (Alzheimer's Association, 2018). Malheureusement, ces méthodes sont actuellement coûteuses et invasives, ce qui fait en sorte qu'elles ne sont pas préconisées dans le domaine de la santé. De plus, puisqu'elles sont utilisées de manière exceptionnelle et conséquemment, ne permettent pas de suivre l'évolution de la maladie dans le temps.

1.2.2 Impact sur la société

En 2019, Alzheimer's Disease International a recensé près de 50 millions de personnes vivant avec la maladie d'Alzheimer dans le monde. Selon eux, d'ici 2050, près de 152 millions de personnes seront affectées par la maladie (Alzheimer's Association, 2019). Ceci est principalement causé par l'augmentation de l'espérance de vie, ce qui augmente le pourcentage de la population âgée. En effet, au Canada, l'espérance de vie a augmenté, en moyenne, d'environ 11 ans, de 1950 à 2002 (Beaudet et al., 2005). De plus, la proportion des personnes âgées de plus de 65 ans est passé de 8% à 14%, de 1960 à 2009. La Figure 1.2 illustre la tendance de la population vieillissante au Canada (Statistics Canada, 2010).[1]

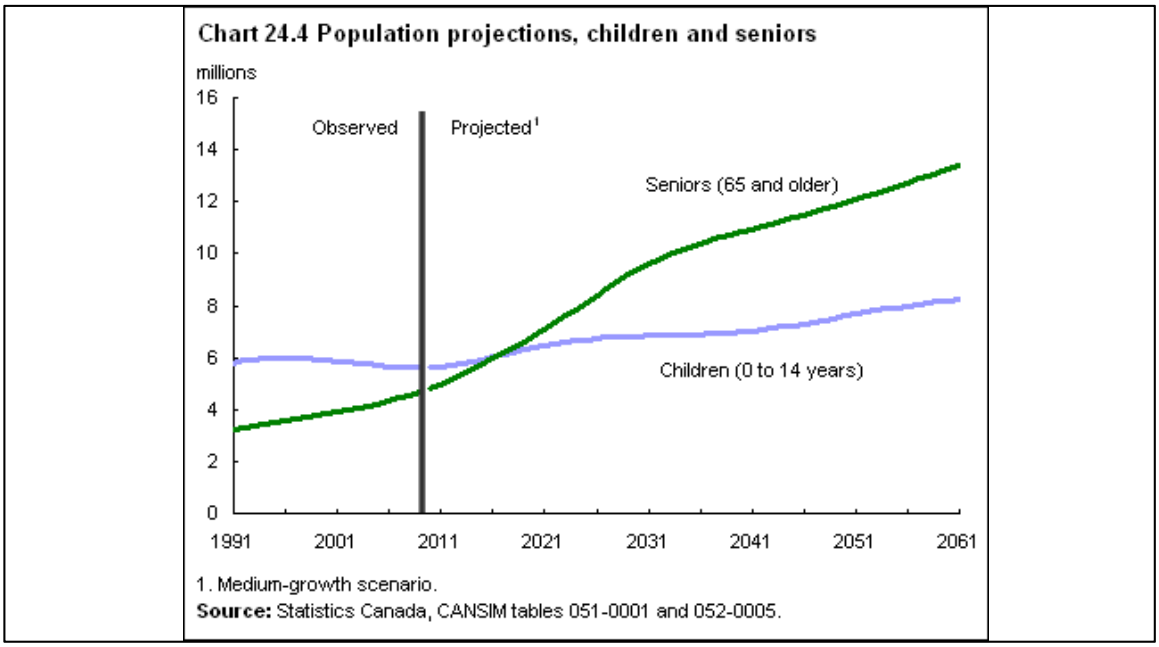

Figure 1.2 Projections de la population, enfants et seniors (Statistics Canada, 2010).

Les coûts associés aux soins accordés aux patients sont très élevés. Lorsqu'une personne atteint le stade de la démence d'Alzheimer, elle n'est plus en mesure de vivre de manière autonome. Des soins personnels sont donc nécessaires. En 2018, il a été estimé que près de 18.4 milliards

[1] Date du dernier grand recensement au Canada.

d'heures ont été nécessaires en assistance non payée pour aider les patients, ce qui représente 232.1 milliards de dollars (Alzheimer's Association 2018). De plus, les techniques actuelles pour diagnostiquer les patients sont très coûteuses. Il faut donc trouver une solution à ce problème, car ces chiffres risquent d'augmenter dans les années à venir.

1.3 Études actuelles

Afin de remédier aux limitations actuelles des méthodes médicales utilisées pour détecter la maladie d'Alzheimer, plusieurs recherches récentes ont proposé des approches basées sur des méthodes d'apprentissage machine. Parmi celles-ci, le développement d'un modèle prédictif à l'aide de mesures évaluant les fonctions linguistiques des patients a su démontrer une avenue intéressante. En effet, les fonctions linguistiques sont généralement altérées par le développement de la maladie, car l'accumulation de la protéine **β**-Amyloïde entraîne des dégradations dans certaines parties du cerveau. Il a été prouvé que les signes d'altération du langage commencent à être visibles dès les premiers stades de la maladie (Szatloczki, Hoffmann, Vincze, Kalman et Pakaski, 2015). Des techniques assistées par ordinateur peuvent donc accélérer le processus de diagnostic. En effet, grâce à ce type de méthode, nous serons éventuellement en mesure d'aider les médecins à détecter les symptômes de la maladie d'Alzheimer plus tôt et ainsi ils pourront diagnostiquer les patients à risque et adopter des mesures préventives.

1.3.1 Hernández-Domínguez et al.

Dans le cadre de sa thèse de doctorat, Hernández-Domínguez a présenté plusieurs études sur la caractérisation de l'altération du langage à travers l'évolution de la maladie d'Alzheimer. Ses travaux ont permis de démontrer l'intérêt de l'analyse des fonctions linguistiques basé sur des techniques informatiques. Afin d'effectuer ces analyses, l'auteure a évalué les performances de différents groupes de patients à l'aide de transcriptions obtenues dans le cadre d'entrevues médicales. Elle a donc extrait des caractéristiques de ces transcriptions à l'aide d'outils de traitements automatiques de la langue naturelle (TALN) pour ensuite développer

des modèles prédictifs pour supporter les médecins dans la détection et la prévention de la maladie.

Dans sa première étude, l'auteure évalue la performance de patients ayant des déficiences cognitives légères. Ensuite, elle évalue celle de patients étant atteints par la maladie d'Alzheimer durant une tâche de description d'image. Pour ce faire, elle a utilisé le Pitt corpus (Becker et al., 1994) provenant de la base de données DementiaBank. Ce corpus contient des enregistrements audio et des transcriptions de patients effectuant le test de description d'image Cookie Theft (Goodglass et Kaplan, 1983). Elle a étudié 74 patients en santé, 43 patients ayant une déficience cognitive légère (DCL) et 169 patients ayant possiblement ou probablement la maladie d'Alzheimer, pour un total de 517 transcriptions et audio. Afin de focaliser ses efforts sur la classification de ces trois catégories, l'auteure a mis de côté les autres diagnostics du corpus.

Afin d'évaluer la performance des participants, l'auteure s'est basée sur certaines caractéristiques linguistiques et phonétiques et a proposé une mesure de couverture d'information. Pour la couverture d'information, elle utilise des unités de contenu d'information pour évaluer l'informativité et la pertinence des transcriptions. Les études actuelles utilisent une liste d'unité de contenu d'information (UCI) (Kavé et al., 2003; Fraser et al., 2016) proposé par (Croisile et al., 1996). Toutefois, Hernández-Domínguez a remarqué qu'il y avait un manque d'objectivité dans la tâche de création de cette liste, et que souvent, cette liste était trop spécifique à un type d'échantillon. Elle a donc proposé la création automatique d'une liste de référents en utilisant un bassin de 25 répondants sains parmi le corpus qu'elle a sélectionné. Ensuite, elle a extrait une série de caractéristiques linguistiques des transcriptions, qui ont été fortement corrélées avec la maladie dans des études précédentes (pour la liste complète des références, voir Hernández-Domínguez, 2018, page 39). Finalement, l'auteure a extrait certaines caractéristiques des segments audio du corpus, telles que la moyenne et la variance des 13 premiers Mel Frequency Cepstral Coefficients (MFCCs).

Afin de créer son modèle prédictif, Dr Hernández-Domínguez a utilisé deux algorithmes d'apprentissage machine et a comparé leurs résultats. L'algorithme Random Forest a offert le meilleur résultat globalement, tandis que l'algorithme Support Vector Machine a obtenu la meilleure performance moyenne. Finalement, une analyse de corrélations a pu démontrer que la mesure de couverture d'information proposée par l'auteure représente une caractéristique prédominante dans la détection de la maladie d'Alzheimer.

Pour sa deuxième étude, l'auteure a tenté de distinguer, de manière automatique, les patients atteints de la maladie d'Alzheimer ou ayant une déficience cognitive légère des patients étant en bonne santé. Elle a recueilli des transcriptions de participants anglophones et espagnols construites dans le cadre d'une tâche de description et de conversations spontanées afin d'effectuer sa recherche. L'objectif était de mettre en valeur les caractéristiques linguistiques ayant une corrélation significative avant que la maladie s'aggrave et aussi de développer un modèle prédictif pour détecter la maladie le plus rapidement possible.

L'auteure s'est basée sur deux types de discours pour cette étude. Tout d'abord, elle a recueilli des discours restreints du Pitt Corpus (Becker et al., 1994), et d'un corpus espagnol, le BBVA Corpus (Peraita et Grasso, 2010). Ce type de discours permet principalement de focaliser les participants sur une tâche précise et donc de s'assurer qu'ils ne s'éloignent pas du sujet principal. Le BBVA Corpus, lui, rassemble un lot de transcriptions de patients espagnols décrivant six objets prédéfinis par les experts.

Ensuite, l'auteure a recueilli des transcriptions de deux corpus comportant des conversations spontanées. Ceci permet d'évaluer leurs fonctions linguistiques durant des discours plus naturels et moins anxiogènes. Le premier corpus (CORLEC) comprend une série de conversations spontanées en espagnol (Madrid, Espagne) tandis que le deuxième (Carolinas' Conversations Collection) contient celles de participants anglophones (Caroline du Nord et du Sud, États-Unis). Ces corpus ont permis à l'auteure de cibler le vocabulaire le plus commun en anglais et en espagnol.

Afin d'extraire des caractéristiques linguistiques des transcriptions, Hernández-Domínguez et son équipe ont développé un outil de prétraitement. Des transcriptions manuellement annotées contiennent souvent des marqueurs décrivant un élément discursif, tel qu'une pause ou une erreur. De ce fait, cet outil de prétraitement a permis d'identifier les pauses, les répétitions, les erreurs et d'autres éléments pouvant rendre la tâche d'extraction des mesures linguistiques plus complexe. Une fois le prétraitement effectué, l'auteure était en mesure d'extraire des caractéristiques, telles que la couverture d'information et la richesse lexicale.

L'analyse statistique effectuée par l'auteure démontre que la couverture d'information a un grand niveau de corrélation avec la sévérité de la maladie, autant pour le Pitt Corpus que le BBVA Corpus. Elle explique donc que ceci pourrait être dû au fait que les participants ayant une déficience cognitive seraient contraints à donner moins d'information durant une tâche de description. Pour ce qui est de la richesse lexicale, une corrélation élevée a été observée pour le Pitt Corpus et non pour le BBVA Corpus. Comme l'auteure le mentionne, cela serait probablement causé par la différence de taille du texte entre ces deux corpus, qui affecte grandement cette variable.

Pour le développement d'un modèle prédictif, l'auteure a effectué une sélection de caractéristiques en éliminant, de manière itérative, les caractéristiques les moins significatives. Les caractéristiques ont ensuite permis de développer plusieurs modèles prédictifs. Parmi ceux-ci, l'algorithme Support Vector Machine a obtenu une performance moyenne (AUC) de 98% pour le BBVA Corpus, et de 83% pour le Pitt Corpus en effectuant une classification binaire (Sains vs Maladie d'Alzheimer) sur une validation croisée. Toutefois, l'auteure mentionne que ces résultats élevés sont probablement dus au fait que le nombre de mots, entre autres, était beaucoup plus élevé dans les discours des patients en santé que ceux atteints de la maladie d'Alzheimer, affectant ainsi les caractéristiques fournies à l'algorithme prédictif. De plus, elle émet comme hypothèse que la différence entre le nombre d'années d'éducation entre ces deux catégories, étant significative, pourrait avoir facilité la classification. En effet, 85% des participants avec la maladie d'Alzheimer n'avaient seulement que l'école primaire comme

éducation (Hernández-Domínguez, 2018). En effet, ceci pourrait affecter la qualité des discours chez les patients ayant un nombre d'années d'éducation moins élevé.

Dans sa troisième étude, l'auteure analyse l'altération des fonctions linguistiques grâce à une expérience longitudinale. Plus précisément, elle a tenté de comprendre les différences entre le vieillissement avec et sans la maladie d'Alzheimer sur une période de 10 ans avec des participants francophones. Quatre patients ont débuté l'étude en ayant, à première vue, une bonne santé mentale. Quelques années après le début de l'expérience, un déclin cognitif leur a été diagnostiqué. Pour chacun de ces patients, leurs résultats ont été comparés à un autre patient, ayant le même âge, sexe, éducation, profession et langues parlées. Les transcriptions de ces patients proviennent du corpus LangAge (Gerstenberg, 2011). Les tests ont été conduits à trois reprises durant la période de 10 ans; la première année, la septième année et à la dernière année de l'étude.

Plusieurs caractéristiques ont été extraites des transcriptions afin d'analyser les corrélations avec la sévérité de la maladie. Tout d'abord, pour ce qui est de la richesse lexicale, l'auteure a utilisé plusieurs mesures reconnues telles que l'index Brunet (Brunet, 1978) et la statistique d'Honoré (Honoré, 1979). Ensuite, elle s'est basée sur la distribution du vocabulaire pour évaluer l'importance de certains mots dans chaque document en utilisant la mesure TFIDF (Sparck Jones, 1972). TF représente la fréquence des termes, tandis que IDF représente la fréquence inverse du document ou, plus précisément, la spécificité des termes. Ceci lui permet de différencier un discours générique d'un discours spécifique dans le cadre d'une description d'image. Aussi, l'auteure a extrait la polarité des sentiments et la subjectivité chez les patients à l'aide d'une librairie Python, TextBlob-fr 0.2.0[2]. Cet outil se sert d'un dictionnaire contenant un pointage de sentiment et de subjectivité pour chaque mot en donnant une moyenne pour un document. Plusieurs études ont prouvé que la positivité perçue chez les patients est grandement affectée par la maladie d'Alzheimer. Finalement, l'auteure a extrait plusieurs caractéristiques au niveau de l'énonciation perçue dans les transcriptions. Par exemple, le nombre

[2] TextBlob-fr 0.2.0, https://github.com/sloria/textblob-fr, 2013

d'interjections, de syllabes ou de mots incohérents lui on permit d'obtenir une moyenne du niveau d'énonciation par patient.

Afin d'analyser l'importance de chacune de ces caractéristiques, l'auteure a effectué une analyse de corrélations avec la sévérité de la maladie. Elle a d'abord réalisé une analyse des composants principaux (PCA) pour ensuite déterminer le niveau de corrélation. Les résultats ont démontré que la richesse lexicale et la distribution du vocabulaire représentent les taux de corrélation les plus significatifs avec la sévérité de la maladie. Aussi, l'auteure a été en mesure de détecter un niveau de corrélation important quant à l'utilisation des négations plus élevée chez les patients avec une déficience cognitive. De plus, un niveau important d'utilisation de verbes communs, tels qu'être, savoir ou faire a été détecté chez ces mêmes patients. Finalement, elle a effectué une séparation linéaire entre les deux composantes les plus significatives selon l'analyse PCA. Ceci lui a permis de démontrer que ces deux composantes pourraient éventuellement être utilisées pour diagnostiquer plus facilement un patient atteint d'une déficience cognitive, même jusqu'à dix ans avant le début de symptômes apparents. De plus, cette analyse permet de détecter des cas particuliers chez certains patients et d'effectuer une analyse approfondie au cas par cas.

1.3.2 Szatloczki, Hoffmann, Vincze, Kalman et Pakaski

L'étude présentée ici tente de mettre le point les changements des habiletés à communiquer décelés à travers l'évolution de la maladie d'Alzheimer. Szatloczki et al. analysent les différentes caractéristiques qui pourraient signaler le début de la maladie. En effet, la phonétique, la phonologie, la morphologie, le lexique, la sémantique, la syntaxe et la pragmatique font partie des branches linguistiques affectées par l'évolution de la maladie d'Alzheimer (Bayles et Boone, 1982). Seules l'articulation et la capacité à répéter ne sont affectées que vers le stade final de la maladie (Appell et al., 1982; Bayles et al., 1992; Croot et al., 2000). Szatloczki et al. ont donc tenté d'analyser ces différentes fonctions linguistiques pour comprendre comment elles sont affectées tout au long du continuum de la maladie.

Au niveau de la phonétique et de la phonologie, des études ont démontré que de plus longues hésitations et une baisse du débit de parole sont perceptibles durant la phase de déficience cognitive légère (Hoffmann et al., 2010; Roark et al., 2011; Jarrold et al., 2014; Satt et al., 2014). Une étude a aussi démontré qu'au stade modéré et sévère de la maladie d'Alzheimer, le débit d'articulation des patients atteints est facilement distinguable de celui des patients en santé (Hoffmann et al., 2010). Une autre étude a observé une tendance au niveau du temps d'entrevue des patients atteints à un niveau modéré. En effet, il a été suggéré qu'ils ont plus de difficulté à entreprendre une longue conversation puisqu'ils se fatiguent rapidement du à l'effort mental requis durant la tâche (López-de-Ipiña et al., 2013). En somme, la majorité des études s'entendent sur le fait qu'une étude temporelle sur des conversations spontanées permettrait de détecter plus facilement la maladie d'Alzheimer à un niveau modéré. L'analyse des différents niveaux de sévérité au stade de la maladie d'Alzheimer est intéressante, mais comporte sont lot complexité. En effet, cette étude distingue la sévérité de la maladie en regroupant les patients avec un niveau léger, modéré et sévère. Ceci rend la tâche d'analyse des caractéristiques linguistiques plus complexe, car cette distinction plus granulaire est plus difficile à détecter.

Szatloczki et al. étudient ensuite le domaine lexical, sémantique et pragmatique de la langue. Les erreurs sémantiques représentent l'une des caractéristiques les plus remarquables durant le stade de la maladie d'Alzheimer (Croot et al., 2000). L'association du test sémantique (SAT) est un outil utilisé pour évaluer la sémantique verbale et visuelle. Des tests effectués avec cet outil ont démontré qu'en général, les patients avec la maladie d'Alzheimer obtenaient des scores significativement moins élevés que ceux qui ne sont pas atteints (Visch-Brink et Denes, 1993). De plus, le milieu de la santé reconnaît que les tests couvrant la sémantique et la phonétique verbale sont de très bons indicateurs de la maladie d'Alzheimer et sont déjà utilisés pour détecter la maladie (Laws et al., 2010). Une autre étude proposée par Duong et al. a permis de comprendre les différences entre l'accès intentionnel et automatique à la mémoire sémantique. Il a été suggéré que l'accès intentionnel serait affecté durant le stade de la déficience cognitive légère et l'accès automatique des années plus tard, durant le stade de la

maladie d'Alzheimer (Duong et al., 2006). Ces études démontrent donc qu'il y a beaucoup de potentiel dans l'analyse sémantique pour la détection de la maladie d'Alzheimer.

1.3.3 Taler et Phillips

L'étude de Taler et Phillips tente d'analyser les différences qui caractérisent les patients en santé, ceux ayant une déficience cognitive légère (DCL) et ceux avec la maladie d'Alzheimer. Leur premier objectif est de comprendre comment différencier ces trois catégories. Ensuite, puisqu'un patient au stade de la déficience cognitive légère ne développera pas nécessairement la maladie d'Alzheimer (Smith et Bondi, 2013), les auteurs tentent de comprendre comment distinguer ceux qui développeront la maladie de ceux qui resteront au stade de la DCL.

Comme les auteurs le mentionnent, l'analyse sémantique, qui cherche à analyser le sens d'une phrase, est une méthode couramment utilisée dans le milieu de la santé pour détecter les déficiences cognitives légères. Ils proposent donc deux types de tâches, exigeant au patient de faire appel à leurs aptitudes à retrouver et produire des mots conformes à un sujet demandé. Tout d'abord, il y a l'analyse phonémique, qui consiste à demander aux participants d'énumérer le plus de mots possible débutant par une lettre demandée (normalement F, A ou S). Ensuite, il y a l'analyse de la fluidité sémantique, qui consiste à demander aux participants de nommer le plus de mots en lien avec une catégorie. Les résultats de ces analyses leur ont permis de détecter une influence plus élevée de la maladie d'Alzheimer au niveau de la fluidité sémantique. Selon les auteurs, la maladie pourrait affecter les fonctions sémantiques des patients, ce qui rendrait ce test plus difficile pour eux. Le test phonémique, lui, requiert moins souvent les fonctions sémantiques. Les auteurs proposent que ceci pourrait expliquer la corrélation moins élevée des mesures de ce test avec la sévérité de la maladie (Henry et al., 2004).

Les auteurs ont ensuite approfondi leur étude sur la fluidité sémantique en analysant le test des catégories, mentionné précédemment. En fait, ils se sont questionnés sur la comparabilité des résultats obtenus à travers différentes études utilisant une variété de catégories. Selon eux, cette variété pourrait éliminer les biais d'apprentissage, car elle empêcherait les patients de se

pratiquer entre les sessions. En effet, la fluidité sémantique pourrait être biaisée, car le patient se sera déjà entraîné à énumérer des mots en fonction d'une catégorie qu'il aurait déjà vue. Toutefois, certaines études contredisent l'hypothèse que la fluidité des catégories est affectée par l'évolution de la dégradation des fonctions cognitives causée par la maladie d'Alzheimer. D'ailleurs, Lambon et al. (2003) ont observé que les sujets étudiés ayant une déficience cognitive légère (DCL) montraient seulement de l'amnésie et que la dégénération des fonctions sémantiques était seulement visible au stade de la maladie d'Alzheimer. Il est donc important, selon eux, de bien distinguer le DCL du DCL amnésique[3], qui sont deux types de déficiences cognitives légères, mais qui ne se caractérisent pas de la même manière. L'étude des mesures de fluidité est, selon les auteurs, une bonne méthode pour détecter l'évolution des déficiences cognitives. Toutefois, elles ne permettent pas nécessairement de prédire la maladie d'Alzheimer. Elles permettent de caractériser l'évolution de la perte de certaines fonctions linguistiques, mais ne peuvent offrir un diagnostic précis quant au type de démence.

Ensuite, Taler et Phillips ont mené une étude sur les difficultés reliées à la recherche de mots justes lorsqu'un patient est appelé à décrire des items. Par exemple, le Boston Naming Test est largement utilisé dans le milieu de la santé (BNT; Kaplan, Goodglass et Weintraub, 1983). Selon les auteurs, les différentes études proposent des conclusions inégales en raison de la grande variété de tests, de nombre d'items à décrire et la différence de taille d'échantillons. Malgré les limitations occasionnées par ces tests, ils ont remarqué une baisse globale du niveau de performance sur les tests de dénomination durant le stade de la maladie d'Alzheimer préclinique, ce qui pourrait porter à croire que la maladie a un impact sur la fonction de la mémoire qui permet de se rappeler certains mots.

Finalement, les auteurs ont effectué une étude sur différentes expériences non standardisées sur les fonctions linguistiques. Tout d'abord, une étude sur l'identification d'un mot a permis de démontrer qu'il y aurait une corrélation entre la rapidité à identifier le mot et la sévérité de la maladie, ce qui porte à croire que les patients avec une bonne santé mentale sont plus rapides

[3] Un DCL amnésique n'affecte que la mémoire.

au niveau associatif (Massoud et al., 2002). Ensuite, une étude du traitement lexical sémantique a su démontrer que de telles approches pourraient grandement aider dans la caractérisation du profil neuropsychologique des patients avec une déficience cognitive légère. Selon les auteurs, les altérations lexicales sémantiques perçues pourraient être dues à une interaction entre les fonctions exécutives et linguistiques puisque les patients avec un DCL auraient de la difficulté à inhiber de l'information lorsqu'ils effectuent une recherche sémantique (Duong et al., 2006). Une autre étude, où les patients résument un texte en une seule phrase, a démontré que les patients avec une déficience cognitive légère ou avec la maladie d'Alzheimer avaient plus de difficulté à généraliser l'idée du texte que ceux en santé (Chapman et al., 2002). Les auteurs ont aussi évalué différentes études analysant les niveaux d'aptitude à écrire. Il est reconnu que les difficultés d'écriture sont couramment observées avec la maladie d'Alzheimer. Une multitude d'études a observé un nombre élevé d'erreurs dans les textes écrits, tels que des phrases incohérentes, des erreurs sémantiques ou graphémiques (Forbes, Shanks et Venneri, 2004).

1.4 Conclusion

Les études présentées ci-dessus permettent de mieux comprendre les différentes altérations des fonctions linguistiques perceptibles tout au long du développement de la maladie d'Alzheimer. Puisqu'il y a plusieurs stades et différents types de déficiences cognitives. Il est important de les distinguer afin de diagnostiquer de manière précise chaque patient. Il y a un grand nombre de caractéristiques linguistiques affectées par le développement de la maladie au niveau cognitif. C'est pourquoi les différentes études servent de bons indicateurs en ce qui a trait à la sélection des caractéristiques pour la création d'un modèle de prédiction de la maladie d'Alzheimer.

Les études de Hernández-Domínguez proposent différentes expériences dans l'objectif d'évaluer l'altération de la langue chez des patients et donc nous permet de mieux comprendre les différentes techniques d'apprentissages machines permettant de créer des modèles prédictifs. De plus, son analyse corrélative sur les différentes mesures en lien avec la sévérité de la maladie met en valeur les variables les plus affectées par le développement de la maladie.

Toutefois, ses conclusions varient d'une étude à l'autre, dû à la variété des données utilisées. Puisque certaines caractéristiques, telles que l'âge, la langue ou le niveau d'éducation, varient d'une étude à l'autre, cela porte à croire qu'il est assez complexe de développer un modèle général dans la détection de la maladie d'Alzheimer. De plus, l'auteure a entrepris une étude longitudinale sur une période de dix ans. Une telle étude permet de mieux percevoir comment les fonctions linguistiques altèrent sur une longue période de temps.

Ensuite, Szatloczki et al. analysent les habiletés à communiquer afin de comprendre comment la maladie les affecte. Plusieurs domaines de la langue sont analysés dont la phonétique, la phonologie et la morphologie. Comme les auteurs le mentionnent, la majorité des études s'entendent sur l'importance d'étudier les caractéristiques temporelles des conversations spontanées pour détecter la maladie. Il a été observé que les hésitations et les pauses, par exemple, deviennent de plus en plus courantes lorsque la maladie se développe. Finalement, l'étude sémantique de la langue chez les patients a démontré un potentiel intéressant qui pourrait permettre de distinguer plus adéquatement la maladie d'Alzheimer, car le taux d'erreurs sémantiques est remarquablement plus élevé lorsqu'il y a une déficience cognitive présente.

Finalement, Taler et Phillips étudient les performances linguistiques à travers les différents niveaux de déficiences cognitives menant à la maladie d'Alzheimer. Leurs objectifs consistent principalement à analyser la détérioration des fonctions linguistiques tout au long du continuum de la maladie. Afin d'y arriver, ils étudient les tests standardisés et non standardisés dans le cadre d'une variété d'études. Ceci leur a permis de mettre le point sur les techniques ayant le plus de succès. Comme la majorité des études, l'une de leurs plus grandes difficultés était de distinguer les différents types de déficiences cognitives, car il y en a beaucoup et chaque caractéristique linguistique n'est pas affectée de la même manière d'un type à l'autre.

Une synthèse de la recherche littéraire mettant de l'avant les informations pertinentes pour notre recherche est présentée au Tableau 1.1.

Tableau 1.1 Synthèse de la revue de littérature

Auteurs	Données (Transcriptions)			Découvertes
Hernández-Domínguez et al. (2018)	*Pitt Corpus (Cookie Theft)*			• Automatisation de la création des ICUs avec un échantillon de PS • Mesure de couverture d'information fortement corrélée avec la sévérité de la maladie
	257 MA	217 PS	43 DCL	
	BBVA Corpus (Discours contraint)			• Extraction de marqueurs discursifs des transcriptions • Corrélation de la couverture d'information élevée pourrait être dû à la difficulté de relever beaucoup d'information par les MA durant la tâche de description • Résultats élevés (98% AUC) peut être causé par la disparité entre la quantité de mots et l'éducation MA vs PS
	39 MA	30 PS		
	Pitt Corpus (Discours contraint)			
	257 MA	242 PS	43 DCL	
	LangAge Corpus (Entrevues biographiques)			• Richesse lexicale et distribution fortement corrélées avec la MA • Corrélation importante au niveau de l'utilisation de négation • Utilisation courante de verbes communs chez les MA
	4 MA	4 PS	--	
Szatloczki et al. (2015	*Comparaison MA, DCL et PS*			• Plus longues hésitations et baisse du débit de parole perçues chez les DCL • Difficulté à entreprendre de longues conversations • Difficile de distinguer MA légère, modéré et sévère • Les patients avec la MA obtiennent un score significativement moins élevé au test sémantique (SAT) que les PS
Taler et al. (2008)	*Comparaison MA, DCL et PS*			• Le test de fluidité sémantique démontre que la maladie affecte les fonctions sémantiques • Fluidité sont de bons détecteurs de l'évolution de la maladie, mais ne prédit pas nécessairement la MA • Corrélation entre la rapidité à identifier un mot et la sévérité de la MA • Plus de difficulté à généraliser une idée chez les DLC et MA

Abréviations: MA, Maladie d'Alzheimer; DCL, Déficience Cognitive Légère; PS, Patient Sain

CHAPITRE 2

MULTILINGUAL DATA PREPROCESSING FOR COMPUTER-BASED DETECTION OF ALZHEIMER'S DISEASE

Frédéric Abiven [a], Sylvie Ratté [b]

[a, b] Département de Génie logiciel, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada, H3C 1K3

Une version écourtée de cet article a été soumise pour publication au journal *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, Wiley.

2.1 Introduction

Monitoring and detecting Alzheimer's disease at an early stage is becoming more crucial as the number of people affected by the disease increases rapidly every year. Currently, nearly 50 million people are living with AD globally, and that number is expected to reach 152 million by 2050 (Alzheimer's Association, 2019). Many studies have covered computer-based approach to evaluate and monitor cognitive functions and detect Alzheimer's Disease (AD) at an early stage (Kong et al., 2019; Weiner et al., 2019; You et al., 2019). The Cookie Theft picture description task has been widely used to monitor and detect the disease. In this study, we analyze transcripts and audio clips from the Pitt Corpus (Becker et al., 1994) (English) and the CRIUGM Corpus (Quebec French). Extracting valuable measures can be strenuous when working with multilingual datasets. Therefore, this work presents a multilingual pipeline approach that preprocesses and extracts multiple linguistic and phonetic characteristics from data. In order to evaluate our approach, we compared the results of both datasets.

2.2 Methods

In this work we present a methodology based on a pipeline architecture for processing transcripts. This allows dividing the work into subprocesses, which makes it easier to approach the multilingual factor. Each subprocess are seen as single entity that can be adapted to different languages and contexts. The pipeline is divided in the following 6 main modules:

typographic normalization and cleaning, part-of-speech (POS) tagging, POS adjustment, POS distribution measurement, linguistic measurement and phonetic measurement. Multilingual modules are identified in blue, as illustrated in Figure 2.1. We will go through each module to explain how they contribute to transcript preprocessing.

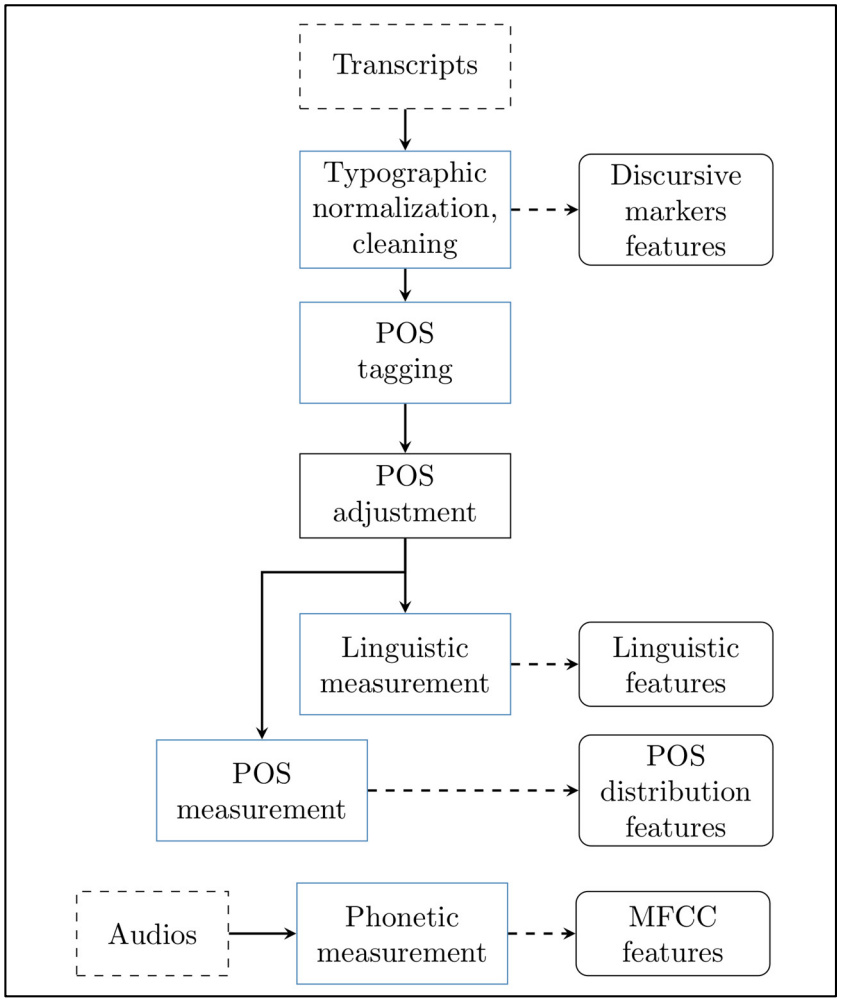

Figure 2.1 Transcript preprocessing pipeline architecture.

2.2.1 Typographic normalization and cleaning

Working with transcripts carries multiple challenges due to the sparsity of related norms. Transcripts may appear in different format, such as plain text files or transcription files (.cha). Also, different discursive marker norms can be used in annotating transcripts since most are produced by hand. Typographic errors could also be injected into transcripts as they are

manually done at the moment. To tackle this problem, the cleaning and normalization task is easily adjustable with configuration files, using a rule-based approach. This allows adapting the process to match different languages and different interview context, as we can specify new rules. The process thus cleans transcripts and extracts discursive markers, which have shown to highly correlate with the disease. In fact, they are widely used in the best performing predictive models to detect AD in English and in French, as shown in Table 1 (respectively 6/10 and 10/10).

In order to make this task multilingual, we incorporated configurable subtasks. First, since interjections (e.g. uhm, mhm) are specific to a language, it is possible to create a list that specifies every interjection that could possibly be found in transcripts. This configuration applies also to expressions (e.g. sigh, laugh) as they vary from one language to another. Finally, since we analyse linguistic patterns in a following task, we want to reduce the complexity of the given patterns. In order to this, we included a synonym reducing task. As illustrated in Figure 2.2, this task reduces synonyms to a single given form. Those synonyms mapping have to be specified in a configuration file. This allows a wider use of this work, since it can be adapted to different languages and contexts. Although we reduce the use of synonyms, we keep track of the count of synonyms that were standardized. Also, we keep copies of transcriptions without this synonym reduction task in order to maintain the true lexical richness of transcriptions, which is evaluated in further tasks.

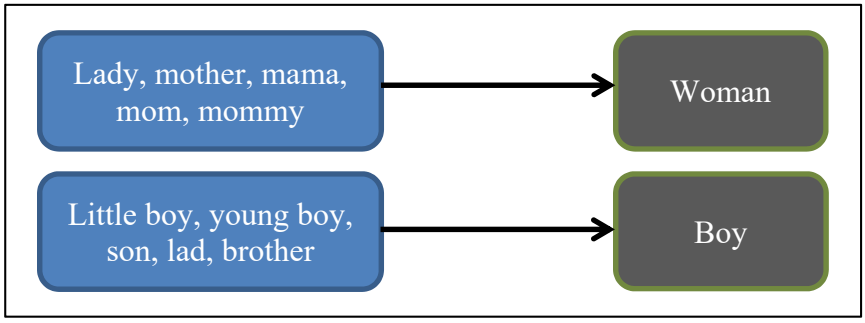

Figure 2.2 Example of the name standardization task.

2.2.2 POS tagging

POS tagging tools have proven their effectiveness in recent years, and are now widely used in NLP tasks. In our work, we used FreeLing 4.0 to analyze and tag transcripts, since it supports many different languages (Padró et Stranilovsky, 2012), although its flexibility in tagging words, tagging norms may vary from one language to another. In fact, some languages can carry a very complex tagging structure and thus, tags may vary in many ways, as illustrated in Figure 2.3. As an addition to this module, we therefore converted tags to a universal form, allowing the following modules to analyze and manipulate transcripts from various corpora. This task was tested on both English and French transcripts, but may be used for numerous languages (Padró et Stranilovsky, 2012).

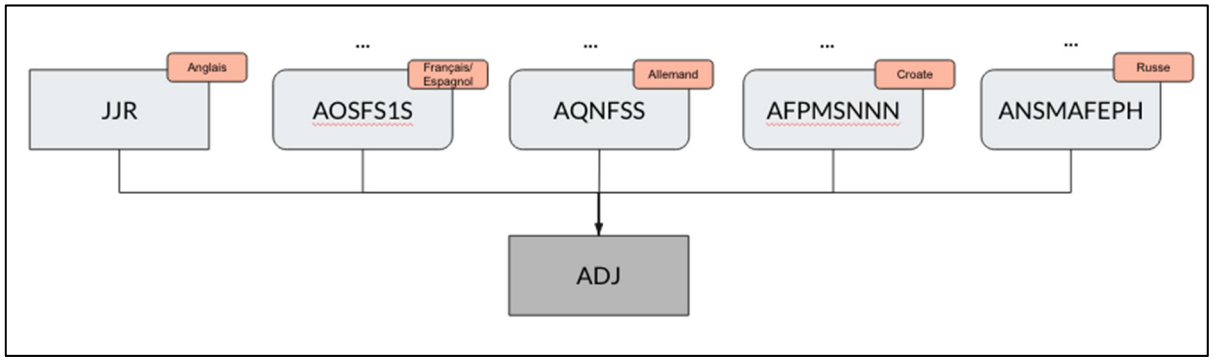

Figure 2.3 Example of the name standardization task.

2.2.3 POS adjustment

Since POS tags are statistically determined, some annotation errors might be introduced into transcripts. This module consists mainly in fixing such mistakes by updating them to their correct form. It evaluates and analyzes the tags, thus allowing improvements in the quality of the results in the following modules. However, this task must be adapted for each language since it depends on a language's structure and rules. In our work, we adapted if for English and French tags.

2.2.4 POS distribution measurement

To measure the distribution of POS tags, the frequency and ratio of the following tags were evaluated: adjectives, conjunctions, nouns, prepositions, verbs and auxiliary verbs. Since the POS tagging module universalizes tags, this process can be applied to different languages.

2.2.5 Linguistic measurement

Linguistic characteristics were automatically extracted within this module. We used the most common linguistic measures utilized in previous works, and which have shown a significant correlation with the disease (Hernández-Dominguez et al., 2018). Since these characteristics are based on the distribution of words and lemmas, this module is language independent. Table 2.1 presents the list of linguistic measures used in this work.

Table 2.1 Linguistic characteristics used to measure patients' language functions

Measure	Equation
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 *W* Index (Brunet, 1978)	$W = N^{V^{-c}}$ with c=0.172 (Tweedie & Baayen, 1998)
Honoré's *R* statistics	$R = \frac{100 \cdot \log N}{1 - \frac{V_1}{V}}$
Type Token Ratio (TTR)	$TTR = \frac{V_1}{V}$
Sichel's *S*	$S = \frac{V_2}{V \neq}$
Yule's characteristic *K* (Miranda-Garcia & Calle-MartÍn, 2005)	$10^4 \frac{[\sum_{i=1}^{N} i^2 V(i,N)]}{N^2} - \frac{1}{N}$

Entropy	$H(X) = -\sum_{x \in X} p(x) \cdot \log_2 p(x)$ Where *p(x)* is the probability of the word *x* being in the text *X*

2.2.6 Phonetic measurement

For phonetic characteristics, we used python speech features 0.6[4] tool to estimate the first 13 MFCCs. We then estimated the mean, kurtosis, skewness and variance of those values. Audios of interviews normally consist in a patient and an interviewer speaking, and so we segmented the audio in order to keep only the patient. In future work, a speaker diarization could be done to extract the patient's voice and thus increase the accuracy of the phonetic measurements.

2.2.7 Modeling

With linguistic and phonetic measures as inputs and the binary target (HC vs. AD), we want to train predictive models that will be able to classify new patients with their cognitive status, with a certain level of confidence. In order to do that, we created an architecture that preprocesses extracted features, selects features and train a model thru a 10-fold cross-validation, as show in Figure 2.4. We started by importing our extracted measures from the preprocessing pipeline. We then iterated thru all possible combination of feature types, which helped us identify the most prominent types of features in AD detection. For each combination, we did a 10-fold cross validation for the English dataset and a Leave-one-out (LOO) cross validation for the French dataset. This decision was based on the fact that the French dataset has limited data.

In this cross validation, we separated our dataset into a training and testing set. We then executed a preprocessing pipeline for the training set and then for the testing set, using the same data scaler. Then, we applied a sequence of feature selection to keep the most valuable

[4] https://pypi.python.org/pypi/python_speech_features

features based on a list of criteria presented in Figure 2.4. Given the selected features, we trained the following algorithms: Support Vector Machine (SVM), Decision Tree (DT) and Random Forest Classifier (RFC). Those algorithms were chosen based on their popularity throughout the literature of AD detection. Finally, we tested each model and measured the mean value of the AUC and other metrics from each iteration of the cross validation.

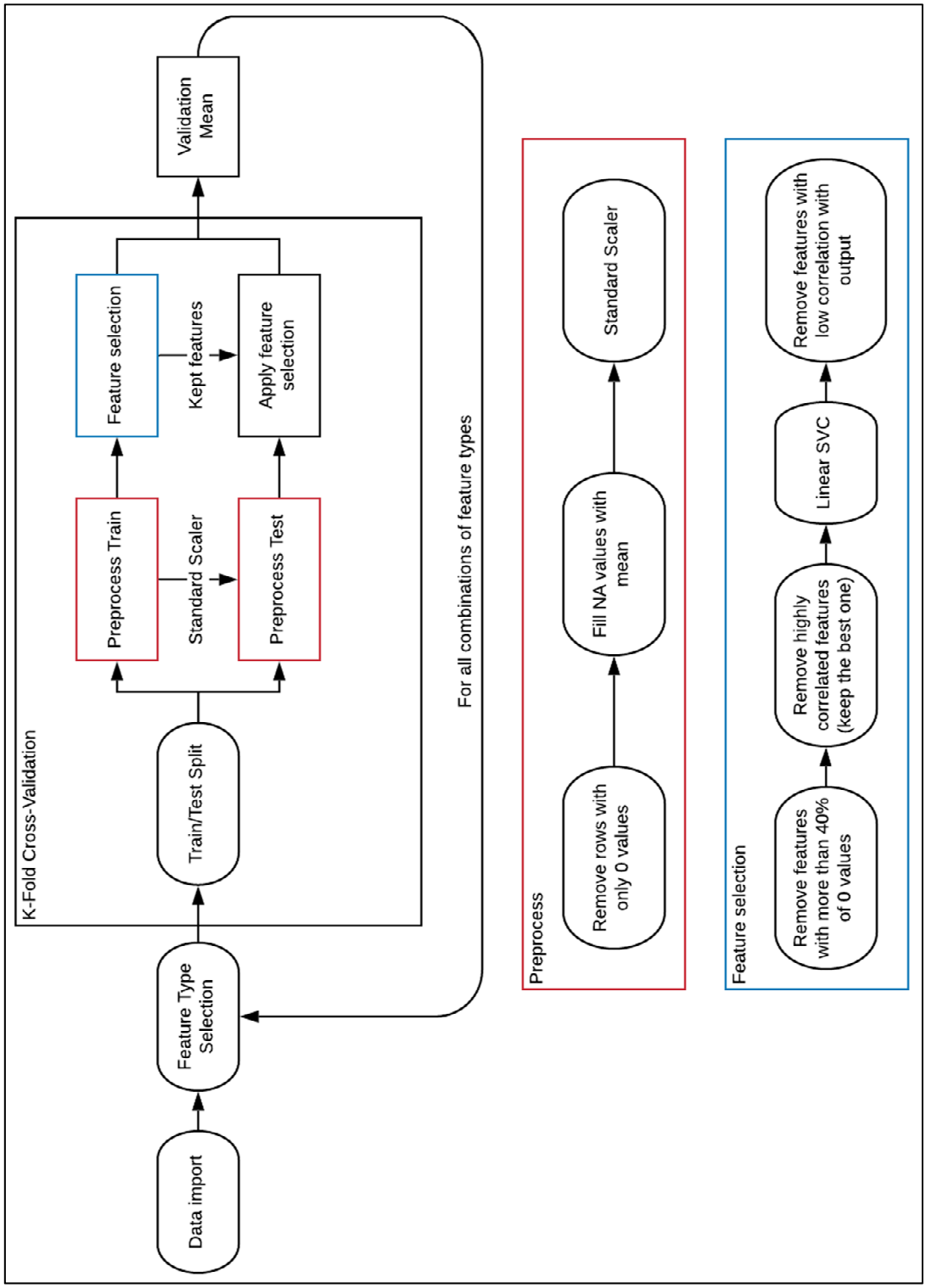

Figure 2.4 Machine learning architecture for AD detection

2.3 Results

In order to understand all our linguistic and phonetic measures and how they interact with AD, we performed correlation analysis. All in all, we extracted 100 features, separated in 4 different categories: discursive markers, POS distribution, linguistic characteristics and phonetic characteristics. We also included information coverage measures that were presented in the work of Hernandez-Dominguez (Hernández-Dominguez et al., 2018). We then ran a feature selection process to extract the most valuable features, as described in Figure 2.4. With the selected features, we trained different predictive models and evaluated their performance with a 10-fold cross-validation, as presented in Table 2.2.

Table 2.2 Average AUC on 10-fold cross validation models with different feature types combination

Pitt Corpus (English)			**CRIUGM Corpus (French)**		
Features	**Model**	**AUC**	**Features**	**Model**	**AUC**
markers-cov-phon-pos	svm	0.76	cov-ling-phon	svm	0.92
markers-cov	svm	0.74	cov-phon-pos	svm	0.92
markers-cov-ling-pos	svm	0.74	cov-ling-phon-pos	svm	0.90
markers-cov-ling	svm	0.73	markers-ling-phon-pos	svm	0.89
markers-phon-pos	svm	0.73	cov-phon	svm	0.89
markers-cov-pos	svm	0.73	markers-ling	svm	0.88
markers-cov-phon	svm	0.73	markers-cov-ling	svm	0.88
markers-ling-phon-pos	svm	0.73	markers-cov-ling	rfc	0.86
markers-ling-pos	svm	0.72	markers-ling-phon	rfc	0.86
markers-cov-ling-phon	svm	0.72	markers-cov-phon-pos	svm	0.86

Abbreviations: cov, Information coverage features; ling, Linguistic features; markers, Discursive markers features; phon, Phonetic features; pos, POS distribution features.

Moreover, we analyzed extracted measures correlation with the disease. By doing this, we were able to analyze linguistic and phonetic features independently. Since we are trying to

implement a multilingual tool, we compared English and French features' correlation to determine correspondence between them, as presented in the Table 2.3.

Table 2.3 Features' correlation with the severity of cognitive impairment

Pitt Corpus (English)		**CRIUGM Corpus (French)**	
Feature	**Correlation**	**Feature**	**Correlation**
Discursive markers			
Repetition rate	0.35	Retracing rate	0.62
Error rate	0.31	Repetition rate	0.37
Incomplete words rate	0.29	Short pause rate	0.31
Retracing rate	0.26	Synonym rate	-0.31
POS distribution			
Conjunction rate	0.18	Auxiliary verb frequency	0.28
Noun frequency	- 0.17	Conjunction rate	- 0.24
Auxiliary verb frequency	- 0.16	Noun rate	- 0.22
Preposition frequency	- 0.13	Verb frequency	0.19
Linguistic characteristics			
Hapax legomena	- 0.24	Yule's K	0.44
Honoré's *R* statistics	- 0.19	Brunet's Index	0.22
Entropy	- 0.17	TTR	- 0.17
Vocabulary size	- 0.16	Hapax legomena	0.16
Phonetic characteristics			
MFCC-9 skewness	0.18	MFCC-9 mean	0.58
MFCC-10 kurtosis	0.17	MFCC-3 skewness	- 0.45
MFCC-8 kurtosis	0.16	MFCC-4 variance	0.44
MFCC-7 kurtosis	0.14	MFCC-11 variance	0.35

2.3.1 Discursive markers

Discursive markers have demonstrated their ability to distinguish healthy controls from AD patients quite remarkably. One of the most correlated features with these markers is the number of retracings in both English and French corpus (respectively 0.26 and 0.62). We hypothesize that patients with AD tend to forget how to describe an object or a person, which forces them to retrace their sentences. Also, we found an inverse correlation with the number of synonyms extracted from transcripts in both languages (respectively -0.13 and -0.31). This could be explained by the fact that AD patients have a smaller vocabulary variety when describing an image. Finally, the number of repetitions detected in both corpora correlates highly with the disease (respectively 0.35 and 0.37), which is consistent with previous studies (Guinn et Habash, 2012; Pompili et al., 2020).

2.3.2 POS distribution

For the POS tags distribution, auxiliary verb frequencies were not correlated in the same way in English and in French. We found that in French, the correlation was positive (0.28) while in English it is was negative (-0.16). This could be due to the fact that auxiliary verbs cannot necessarily be translated the same way between those languages (e.g: Je suis allé à l'école; I went to school) and therefore, measures may vary. Similarly, conjunction and adjectives did not have the same type of correlation between English and French. On the other hand, we found that AD patients tend to use fewer nouns in both languages, which correlates with previous findings (Jarrold et al., 2014). That being said, a POS distribution should be considered and analysed in each language separately, since it does not necessarily have the same representation in each case.

2.3.3 Linguistic characteristics

For the Pitt Corpus, lexical richness correlations were mostly consistent with previous studies (Hernández-Dominguez et al., 2018). With the CRIUGM dataset, most measures were inconsistent with the results obtained with the Pitt Corpus, and indeed, were sometimes highly

correlated with the disease (e.g., Yule's characteristic K (0.44)). We believe that this could be due to the size of the dataset, which is very small, as compared to the English dataset. Nonetheless, this module may be considered as a benchmark, since the results match those of the same experiment conducted on the Pitt Corpus (Hernández-Dominguez et al., 2018).

2.3.4 Phonetic characteristics

Considering phonetic characteristics, results with the Pitt Corpus are relatively consistent with previous studies (Hernández-Dominguez et al., 2018). There may have been some differences in correlation values due to the fact that we segmented the audio to remove the interviewer's voice. For the CRIUGM dataset, some of the MFCCs mean, skewness and variance values were highly correlated with the disease (> 0.4). Again, those high correlations might be explained by the size of the dataset and the manual audio segmentation task, which could bias the results.

2.3.5 Modeling

For both corpora, we tested different combination of feature types, which showed discursive markers to be the most common feature type found in the best predictive models overall. With the Pitt Corpus, our best model had an average AUC of 76%, which is relatively consistent with previous studies (Hernández-Dominguez et al., 2018; Fraser et al., 2016). Looking at the CRIUGM Corpus, our best model had an average AUC of 92%. This result, which is significantly high, may be explained by the very small dataset size and the high correlation found in multiple features.

2.4 Discussion

This work contributes in many ways to improve quality and efficiency of transcript and audio preprocessing to extract measures that characterize linguistic and phonetic functions. Furthermore, we expand its use by making the processing adaptable to many different languages. Results have demonstrated its consistency with previous studies, as well as with a new cohort of French participants. Further research could focus on including languages with

different structure and rules, as it could expand its usage. We would also like to include the information coverage measure extraction as part of a new module in our pipeline, as it has proved its capacity to significantly distinguish AD patients from healthy controls (Hernández-Dominguez et al., 2018). Finally, we believe it would be interesting to compare results between proportionate datasets of different languages to evaluate how the disease may affect cognitive functions in patients differently.

2.5 Acknowledgment

The research presented in this paper was financially supported by NSERC (Natural Sciences and Engineering Research Council of Canada) RGPIN-2018-05714.

CHAPITRE 3

DISCUSSION

Le but de cette étude était de présenter une approche qui permettrait de faciliter la réutilisation d'un algorithme de prétraitement pour l'analyse de transcription, puisque c'est une tâche qui demande beaucoup de temps et d'effort. Il est primordial de normaliser les données afin d'en assurer la cohérence et la fiabilité des résultats. En effet, plusieurs facteurs font en sorte que les données avec lesquels les chercheurs travaillent ne se conforment pas à une seule norme. Par exemple, la langue est un facteur important, car sa structure varie énormément entre les différents dialectes à travers le monde. Aussi, le format des données peut varier puisque la création des transcriptions peut être faite de différentes manières. Certaines transcriptions sont rédigées manuellement, tandis que d'autres sont rédigées avec des outils automatisés. Finalement, le format des fichiers de transcriptions peut être sous un format de conversation (.cha) ou dans un simple fichier de texte. Il est donc important d'avoir une approche qui peut prendre en compte ces différences et les traiter de manière uniforme.

3.1 Prétraitement des données

Pour nous y prendre, nous avons créé un outil d'automatisation basé sur une architecture en pipeline. Cette approche permet de rendre chaque tâche configurable afin de les adapter à la langue du corpus. Notre travail consiste en une série de traitements automatiques de la langue naturelle (TALN). Tout d'abord, la normalisation et le nettoyage des transcriptions permettent d'extraire des mesures de distribution des marqueurs discursifs et d'extraire les anomalies qui pourraient apparaître dans le texte. Ensuite, la tâche d'étiquetage des marqueurs syntaxiques (POS) permet d'apporter une précision sur la nature de chaque mot contenu dans les transcriptions. Nous avons utilisé un outil externe (FreeLing 4.0) pour effectuer cette tâche. Puisque la structure de marqueurs peut varier entre les langues, nous avons effectué une universalisation des marqueurs. Ceci facilite grandement la suite du traitement, puisqu'elle réduit la complexité des marqueurs. Par la suite, la tâche d'ajustement des marqueurs permet

de compenser certaines erreurs qui auraient pu être induites par l'outil utilisé à l'étape d'étiquetage, tel qu'un verbe mal étiqueté. Cependant, cette tâche ne peut être multilingue puisqu'il existe plusieurs structures linguistiques, ce qui complexifie la tâche. En effet, il existe deux structures de langues prédominantes: SVO (Sujet–Verbe–Objet) et SOV (Langus et Nespor, 2010). Dans le cas où il y aurait une erreur d'étiquetage d'un verbe, par exemple, la structure linguistique permettrait de la détecter facilement et de la corriger. Une fois les transcriptions normalisées et adaptées, nous avons été en mesure d'extraire une série de mesures linguistiques et phonétiques. Au total, nous avons extrait plus de 100 mesures caractérisant les fonctions linguistiques et phonétiques des patients. Ces mesures permettent d'analyser et de mieux comprendre comment la MA affecte les fonctions linguistiques et phonétiques des patients. De plus, puisque ces fonctions sont affectées durant les premiers stades de la maladie, ces caractéristiques peuvent être analysées temporellement, offrant une compréhension plus approfondie sur l'évolution de la maladie.

3.2 Entraînement de modèles prédictifs

Une fois les données normalisées, nous avons entraîné plusieurs modèles prédictifs à l'aide des algorithmes d'apprentissage machine Support Vector Machine (SVM), Random Forest Classifier (RFC) et Decision Tree (DT). Pour entraîner ces modèles, nous avons effectué une validation croisée, ce qui permet d'estimer la fiabilité du modèle en profondeur. Pour le Pitt Corpus, nous avons effectué une validation croisée « 10-fold », qui consiste à inter changer 10 fois la portion d'entraînement et de test. Ensuite, puisque les données du corpus du CRIUGM étaient limitées, nous avons effectué une validation croisée Leave-one-out (LOO), qui consiste à inter changer N-1 fois la portion d'entraînement et de test pour un échantillon de grandeur N. C'est une pratique courante lorsque l'on doit travailler avec un échantillon de données limité. Dans cette validation croisée, nous avons effectué une série de sélection de caractéristiques pour extraire les plus prédominantes en termes de corrélation avec la sévérité de la maladie. Cette sélection est réexécutée à chaque itération de la validation croisée afin d'éviter un biais d'apprentissage.

Malgré cette validation croisée, il est difficile d'estimer si nos modèles seront fiables sur de nouvelles données, car le manque de données peut occasionner un surapprentissage. Dans le cas où nous aurions accès à un plus grand nombre de données, les modèles prédictifs seront plus fiables et reflèteront mieux la réalité. Finalement, il serait plus intéressant d'utiliser un corpus en français incluant des patients avec la MA. Ceci permettrait d'évaluer si les mesures extraites en français sont en mesure de bien distinguer les patients atteints de la maladie de ceux en santé.

CONCLUSION ET RECOMMANDATIONS

L'objectif principal de cette recherche était d'offrir une approche universelle pour le prétraitement de transcriptions dans le cadre d'entrevues avec des patients effectuant des tests cognitifs. Cette approche pourra contribuer grandement à plusieurs domaines de recherches pour lesquelles les transcriptions sont requises puisqu'elle permet d'augmenter la capacité de reproduire des expériences scientifiques rapidement et donc concentrer les efforts sur la tâche d'analyse et de compréhension des données. De plus, l'outil de prétraitement est facilement configurable, s'adaptant ainsi à différentes langues et différents contextes de transcriptions, élargissant ainsi son éventail d'applications.

En somme, l'outil de prétraitement normalise les transcriptions et extrait des mesures caractérisant les fonctions linguistiques et phonétiques des patients. Ces mesures sont souvent utilisées pour détecter et surveiller certaines maladies cognitives. Dans notre cas, nous avons développé des modèles prédictifs, à l'aide de ces données, pour détecter la maladie d'Alzheimer. Ces modèles pourront éventuellement être utilisés comme outil d'aide à la décision pour les médecins qui tentent de diagnostiquer des patients ayant potentiellement une maladie cognitive. En effet, ce type d'approche non invasive est intéressante pour les patients qui, très souvent, vivent beaucoup de stress dû à la maladie et aux différents diagnostics actuels. De plus, elle permettrait de réduire significativement les coûts liés au diagnostic.

En appliquant notre outil de prétraitement sur des transcriptions du Pitt Corpus (Anglais) et en développement des modèles prédictifs avec les mesures extraites, nous avons été en mesure de reproduire des résultats similaires avec des études précédentes. Ensuite, nous avons tenté la même expérience sur le corpus du CRIUGM (Français). Les résultats élevés ont soulevé plusieurs questions quant à la performance globale pour ce corpus. En effet, les résultats élevés peuvent être expliqués par le type de patients comparés (jeunes vs. âgés) et par la limitation du nombre de données. Ceci causerait du surapprentissage, car le modèle prédictif s'entraîne sur un échantillon de données trop petit. Toutefois, les résultats démontrent clairement que les mesures distinguent bien le niveau linguistique des jeunes comparés aux personnes âgées. Par

exemple, le nombre de retraçage, étant élevé chez les patients âgés, pourrait être expliqué par la dégradation des fonctions linguistiques dû à la vieillesse naturelle, causant un plus grand nombre d'hésitations durant une tâche de description d'image.

4.1 Perspectives d'avenir

Les techniques assistées par ordinateur pour la détection de maladie cognitive deviennent de plus en plus fréquentes et suscitent de plus en plus d'intérêt dans le domaine médical. Des approches non invasives et automatisées deviennent plus souvent préconisées dues à l'augmentation du nombre de patients atteints de maladies cognitives. Un tel travail permettrait, à long terme, de rendre les tests cliniques plus robustes en offrant des analyses plus complètes des transcriptions. Toutefois, il est difficile d'appliquer de telles mesures sans démontrer la robustesse des systèmes. Donc, il est important d'être en mesure de faire plusieurs expériences et de les reproduire avec un niveau de confiance élevé. Notre outil de prétraitement universel de transcriptions facilitera assurément le travail des chercheurs qui tentent d'analyser comment les maladies cognitives affectent les fonctions linguistiques et phonétiques chez l'humain.

Ensuite, puisque notre travail a été expérimenté avec un échantillon de données en Anglais et en Français, il serait intéressant de tenter l'expérience sur d'autres langues, telle que l'Espagnol, qui fait partie de la même branche linguistique que le Français. Ceci permettrait de valider l'efficacité du pipeline avec une nouvelle langue et d'adapter les modules nécessaires. Ensuite, dans l'éventualité où nous aurions accès à un échantillon de données plus élevé pour le corpus du CRIUGM, incluant des patients avec la MA, il serait intéressant de reproduire l'expérience pour développer un modèle prédictif plus réaliste.

En somme, ceci est un travail incrémental qui devra être adapté au fil du temps afin d'accueillir de nouvelles langues et donc d'augmenter sa capacité à être utilisé pour différentes communautés à travers le monde. Nous sommes certains que ce travail permettra d'accélérer la recherche dans le domaine de la santé mentale et beaucoup de gens pourront en bénéficier.

BIBLIOGRAPHIE

Alzheimer's Association. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *14*(3), 367-429.

Alzheimer's Association. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, *15*(3), 321-387.

Appell, J., Kertesz, A., & Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain and language*, *17*(1), 73-91.

Bayles, K. A., & Boone, D. R. (1982). The potential of language tasks for identifying senile dementia. *Journal of Speech and Hearing Disorders*, *47*(2), 210-217.

Bayles, K. A., Tomoeda, C. K., & Trosset, M. W. (1992). Relation of linguistic communication abilities of Alzheimer's patients to stage of disease. *Brain and language*, *42*(4), 454-472.

Beaudet, M. P., Tully, P., & St-Arnaud, J. U. L. I. E. (2005). Life expectancy. *Health Reports*, *17*(1), 43.

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.

Brunet, E. (1978). *Le 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 française* (Vol. 1). Slatkine.

Chapman, S. B., Zientz, J., Weiner, M., Rosenberg, R., Frawley, W., & Burns, M. H. (2002). Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer Disease & Associated Disorders*, *16*(3), 177-186.

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.

Croot, K., Hodges, J. R., Xuereb, J., & Patterson, K. (2000). Phonological and articulatory impairment in Alzheimer's disease: a case series. *Brain and language*, *75*(2), 277-309.

Duong, A., Whitehead, V., Hanratty, K., & Chertkow, H. (2006). The nature of lexico-semantic processing deficits in mild cognitive impairment. *Neuropsychologia*, *44*(10), 1928-1935.

Forbes, K. E., Shanks, M. F., & Venneri, A. (2004). The evolution of dysgraphia in Alzheimer's disease. *Brain research bulletin*, *63*(1), 19-24.

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*(2), 407-422.

Gerstenberg, A. (n.d.). LangAge corpora. Retrieved from www.langage-corpora.org
Henry, J. D., Crawford, J. R., & Phillips, L. H. (2004). Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. *Neuropsychologia*, *42*(9), 1212-1222.

Guinn, C. I., & Habash, A. (2012, October). Language analysis of speakers with dementia of the Alzheimer's type. In *2012 AAAI Fall Symposium Series*.

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.

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., & Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, *12*(1), 29-34.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, *7*(2), 172-177.

Jarrold, W., Peintner, B., Wilkins, D., Vergryi, D., Richey, C., Gorno-Tempini, M. L., & Ogar, J. (2014, June). 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).

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Kaplan, E.F., Goodglass, H. and Weintraub, S. (1983) The Boston Naming Test. 2nd Edition, Lea & Febiger, Philadelphia.

Kavé, G., & Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *Journal of speech, language, and hearing research*.

Kong, W., Jang, H., Carenini, G., & Field, T. (2019, October). A Neural Model for Predicting Dementia from Language. In *Machine Learning for Healthcare Conference* (pp. 270-286).

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive psychology*, *60*(4), 291-318.

Lambon Ralph, M. A., Patterson, K., Graham, N., Dawson, K., & Hodges, J. R. (2003). Homogeneity and heterogeneity in mild cognitive impairment and Alzheimer's disease: a cross-sectional and longitudinal study of 55 cases. *Brain*, *126*(11), 2350-2362.

Laws, K. R., Duncan, A., & Gale, T. M. (2010). 'Normal'semantic–phonemic fluency discrepancy in Alzheimer's disease? A meta-analytic study. *Cortex*, *46*(5), 595-601.

López-de-Ipiña, K., Alonso, J. B., Travieso, C. M., Solé-Casals, J., Egiraun, H., Faundez-Zanuy, M., ... & Lizardui, U. M. D. (2013). On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis. *Sensors*, *13*(5), 6730-6745.

Massoud, F., Chertkow, H., Whitehead, V., Overbury, O., & Bergman, H. (2002). Word-reading thresholds in Alzheimer disease and mild memory loss: a pilot study. *Alzheimer Disease & Associated Disorders*, *16*(1), 31-39.

Padró, L., & Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Peraita, A. H., & Grasso, L. (2010). *Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad de Alzheimer: una investigación transcultural hispano-argentina* (No. 20107).

Pompili, A., Abad, A., de Matos, D. M., & Martins, I. P. (2020). Pragmatic Aspects of Discourse Production for the Automatic Identification of Alzheimer's Disease. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 261-271.

Roark, B., Mitchell, M., Hosom, J. P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, *19*(7), 2081-2090.

Satt, A., Hoory, R., König, A., Aalten, P., & Robert, P. H. (2014). Speech-based automatic and robust detection of very early dementia. In *Fifteenth Annual Conference of the International Speech Communication Association*.

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

Statistics Canada. 2010. *An aging population*. Statistics Canada Catalogue no. 11-402-X. Ottawa. Version updated October 2016. Ottawa.

Szatloczki, G., Hoffmann, I., Vincze, V., Kalman, J., & Pakaski, M. (2015). Speaking in Alzheimer's disease, is that an early sign? Importance of changes in language abilities in Alzheimer's disease. *Frontiers in aging neuroscience*, *7*, 195.

Taler, V., & Phillips, N. A. (2008). Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5), 501–556.

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.

Visch-Brink, E. G., & Denes, G. (1993). A European base-line test for word-picture processing. *Developments in the assessment and rehabilitation of brain-damaged patients*, 211-216.

Weiner, J., Frankenberg, C., Schröder, J., & Schultz, T. (2019, December). Speech Reveals Future Risk of Developing Dementia: Predictive Dementia Screening from Biographic Interviews. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 674-681). IEEE.

You, Y., Ahmed, B., Barr, P., Ballard, K., & Valenzuela, M. (2019, November). Predicting Dementia Risk Using Paralinguistic and Memory Test Features with Machine Learning Models. In *2019 IEEE Healthcare Innovations and Point of Care Technologies,(HI-POCT)* (pp. 56-59). IEEE.